Legal Knowledge and Information Systems

JURIX 2021: ***The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021***

Editor:
Erich Schweighofer

JURIX 2021

Legal Knowledge and Information Systems

JURIX 2021:
The Thirty-fourth Annual Conference

Editor:
Erich Schweighofer

Traditionally concerned with computational models of legal reasoning and the analysis of legal data, the field of legal knowledge and information systems has seen increasing interest in the application of data analytics and machine learning tools to legal tasks in recent years.

This book presents the proceedings of the 34th annual JURIX conference, which, due to pandemic restrictions, was hosted online in a virtual format from 8 – 10 December 2021 in Vilnius, Lithuania. Since its inception as a mainly Dutch event, the JURIX conference has become truly international and now, as a platform for the exchange of knowledge between theoretical research and applications, attracts academics, legal practitioners, software companies, governmental agencies and judiciary from around the world. A total of 65 submissions were received for this edition, and after rigorous review, 30 of these were selected for publication as long papers or short papers, representing an overall acceptance rate of 46%. The papers are divided into 6 sections: Visualization and Legal Informatics; Knowledge Representation and Data Analytics; Logical and Conceptual Representations; Predictive Models; Explainable Artificial Intelligence; and Legal Ethics, and cover a wide range of topics, from computational models of legal argumentation, case-based reasoning, legal ontologies, smart contracts, privacy management and evidential reasoning, through information extraction from different types of text in legal documents, to ethical dilemmas.

Providing an overview of recent advances and the cross-fertilization between law and computing technologies, this book will be of interest to all those working at the interface between technology and law.

ISBN 978-1-64368-252-5

9 781643 682525

JURIX 2021

ISBN 978-1-64368-252-5 (print)
ISBN 978-1-64368-253-2 (online)
ISSN 0922-6389 (print)
ISSN 1879-8314 (online)

LEGAL KNOWLEDGE AND INFORMATION SYSTEMS

Frontiers in Artificial Intelligence and Applications

The book series Frontiers in Artificial Intelligence and Applications (FAIA) covers all aspects of theoretical and applied Artificial Intelligence research in the form of monographs, selected doctoral dissertations, handbooks and proceedings volumes. The FAIA series contains several sub-series, including 'Information Modelling and Knowledge Bases' and 'Knowledge-Based Intelligent Engineering Systems'. It also includes the biennial European Conference on Artificial Intelligence (ECAI) proceedings volumes, and other EurAI (European Association for Artificial Intelligence, formerly ECCAI) sponsored publications. The series has become a highly visible platform for the publication and dissemination of original research in this field. Volumes are selected for inclusion by an international editorial board of well-known scholars in the field of AI. All contributions to the volumes in the series have been peer reviewed.
The FAIA series is indexed in ACM Digital Library; DBLP; EI Compendex; Google Scholar; Scopus; Web of Science: Conference Proceedings Citation Index – Science (CPCI-S) and Book Citation Index – Science (BKCI-S); Zentralblatt MATH.

Volume 346

Recently published in this series

ISSN 0922-6389 (print)
ISSN 1879-8314 (online)

Legal Knowledge and Information Systems

JURIX 2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8–10 December 2021

Edited by

Erich Schweighofer

University of Vienna

IOS Press

Amsterdam • Berlin • Washington, DC

ISBN 978-1-64368-252-5 (print)
ISBN 978-1-64368-253-2 (online)
Library of Congress Control Number: 2021951513
doi: 10.3233/FAIA346

Publisher
IOS Press BV
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

For book sales in the USA and Canada:
IOS Press, Inc.
6751 Tepper Drive
Clifton, VA 20124
USA
Tel.: +1 703 830 6300
Fax: +1 703 830 2300
sales@iospress.com

Preface

For more than 30 years, the Dutch Foundation for Knowledge Based Systems JURIX (https://jurix.nl) has organised annual conferences on artificial intelligence & law. Starting as a mostly Dutch event, is has spread out to Europe, having taken place in many countries (*inter alia* in Malta, Austria, Belgium, France, Poland, and Czech Republic). This year, the already 34th International Conference on Legal Knowledge and Information Systems (JURIX 2021) takes place in Vilnius, Lithuania. From the point of geography, Lithuania is the heart of Europe; it is not yet but may become so also in the mind of people, reminding us about the richness diversity of the "old continent". Considering participants and speakers, JURIX2021 is now truly a European conference on artificial intelligence & law, with strong outreach to the Americas and Australasia.

This annual international conference has been open for all, in particular academics, legal practitioners, software companies, administrations, parliaments and the judiciary. It is now a place of virtuous exchange of knowledge between theoretical research and applications on artificial intelligence & law. Traditionally, this field has been concerned with legal knowledge representation and engineering, computational models of legal reasoning, and analyses of legal data. However, recent years have witnessed an increasing interest in the application of data analytics and machine learning tools to relevant tasks.

The 2021 edition of JURIX, which runs from December 8 to 10, is hosted by the Mykolas Romeris University in Vilnius. Due to the Covid-19 health crisis, the conference is organised in a virtual format. For this edition, we have received 65 submissions. 13 of these submissions were selected for publication as long papers (10 pages each), 17 as short papers (6–8 pages each) for a total of 30 presentations. We were inclusive in making our selection, but the competition was stiff and the submissions were put through a rigorous review process with a total acceptance rate (long and short papers) of 46%, and a competitive 20% acceptance rate for long papers.

The accepted papers cover a broad array of topics, from computational models of legal argumentation, case-based reasoning, legal ontologies, smart contracts, privacy management and evidential reasoning, through information extraction from different types of text in legal documents, to ethical dilemmas.

Invited speakers have honored JURIX 2021 by kindly agreeing to deliver a keynote lecture: Friedrich Lachmayer and Vytautas Čyras. Friedrich Lachmayer is a retired high-level lawyer of the Austrian administration – the legal service of the Federal Chancellery, a glorified docent (Professor at the University of Innsbruck) and a well-known expert on legal theory and legal visualization. Vytautas Čyras is a professor at the University of Vilnius and has worked for more than 15 years on these topics.

We are very grateful to them for having accepted our invitation and for their interesting and inspiring talks.

Traditionally, the main JURIX conference is accompanied by co-located events comprising workshops and tutorials. This year's edition welcomes six workshops and one tutorial:

- 1st Workshop in Agent-based Modeling & Policy-Making (AMPM 2021)
- AI Approaches to the Complexity of Legal Systems (AICOL 2021)

- CEILI Workshop on Legal Data Analysis (LDA21)
- EXplainable & Responsible AI in Law (XAILA 2021)
- The First International Workshop on Intelligent Regulatory Systems (IRS 2021)
- Use of Information Technology in Judicial Processes (MRU 2021)
- Tutorial on Legal Informatics Topics: Legal Tech & Privacy Impact Assessment (TLIT2021)

We would like to thank the workshops' and tutorials' organizers for their excellent proposals and for the effort involved in organizing the events.

The continuation of well-established events and the organization of entirely new ones provide a great added value to the JURIX conference, enhancing its thematic and methodological diversity and attracting members of the broader community.

Since 2013, JURIX has also hosted the Doctoral Consortium, now in its ninth edition. This initiative aims to attract and promote Ph.D. researchers in the area of AI & Law so as to enrich the community with original and fresh contributions. We owe our gratitude to Monica Palmirani who started the Doctoral Consortium.

Organizing this conference would not have been possible without the support of many people and institutions. Special thanks are due to the local organizing team chaired by Lyra Jakulevičienė and Paulius Pakutinskas of the Legal Tech Centre and Law School, Mykolas Romeris University (Lithuania).

Thanks are also due to the University of Vienna, Arbeitsgruppe Rechtsinformatik, Juridicum and its related organisations, in particular the Wiener Zentrum für Rechtsinformatik (WZRI) and IRI§-Conferences. These efforts were sponsored also by Cybly, Wien/Salzburg and Weblaw, Bern.

This year, we are particularly grateful to the members of the Program Committee for their excellent work in the rigorous review process and for their participation in the discussions concerning borderline papers. Senior Members have provided additional support. Sub-reviewers have done a rigorous check on some papers. Their work has been even more appreciated provided the complex situation we are experiencing due to the pandemic.

Last but not least, this year's conference was organized in partnership with GO Vilnius, Lithuanian Bar Association and Amberlo.

Finally, we would like to thank the former and current JURIX executive committee and steering committee members.

Erich Schweighofer
JURIX 2021 Programme Chair

Partners


amberlo
Go Vilnius


LIETUVOS
AD VOCO
ADVOKATŪRA


IRIS


IZRI


CYBLY
Lawthek • Legalnetics


universität
wien


Arbeitsgruppe Rechtsinformatik
universität
wien
ARI
Centre for Computers and Law

This page intentionally left blank

Programme Committee

Tommaso Agnoloni, Italy
Thomas Ågotnes, Norway
Francisco Andrade, Portugal
Michał Araszkiewicz, Poland
Kevin Ashley, United States
Katie Atkinson, United Kingdom
Trevor Bench-Capon, United Kingdom
Floris Bex, Netherlands
Georg Borges, Germany
Danièle Bourcier, France
Karl Branting, United States
Pompeu Casanovas, Spain
Federico Costantini, Italy
Vytautas Čyras, Lithuania
Luigi Di Caro, Italy
Massimo Durante, Italy
Stefan Eder, Austria
Enrico Francesconi, Italy
Fernando Galindo, Spain
Aldo Gangemi, Italy
Randy Goebel, Canada
Guido Governatori, Australia
Matthias Grabmair, Germany
Rinke Hoekstra, Netherlands
Jeff Horty, United States
Lyra Jakuleviciene, Lithuania
John Joergensen, United States
Jeroen Keppens, United Kingdom
Franz Kummer, Switzerland
Friedrich Lachmayer, Austria
Réka Markovich, Hungary
Thorne McCarty, United States
Ugo Pagallo, Italy
Paulius Pakutinskas, Lithuania
Monica Palmirani, Italy
Ginevra Peruginelli, Italy
Wim Peters, Netherlands
Marta Poblet, Australia
Radim Polčák, Czechia
Henry Prakken, Netherlands
Paulo Quaresma, Portugal
Víctor Rodríguez Doncel, Spain
Antoni Roig, Spain
Antonino Rotolo, Italy
Giovanni Sartor, Italy
Ken Satoh, Japan
Burkhard Schafer, United Kingdom
Erich Schweighofer, Austria
Giovanni Sileno, Netherlands
Clara Smith, Argentina
Christoph Sorge, Germany
Sarah Sutherland, Canada
Leon van der Torre, Luxembourg
Tom van Engers, Netherlands
Marc van Opijnen, Netherlands
Bart Verheij, Netherlands
Fabio Vitali, Italy
Vern Walker, United States
Bernhard Waltl, Germany
Radboud Winkels, Netherlands
Adam Wyner, United Kingdom
Hajime Yoshino, Japan
John Zeleznikow, Australia
Tomasz Zurek, Poland

Subreviewers

Lorenzo Bacci
Matteo Baldoni
Ilaria Canavotto
Roger Ferrod
Mustafa Hashmi
Kaspar Lebloch
Tung Le
Andreas Rauber
Felix Schmautzer
Giovanni Siragusa
Vu Tran
Masaharu Yoshioka

Senior Members

Additional work was provided by the Senior Members advising the chair on meta questions: Kevin Ashley, Trevor Bench-Capon, Danièle Bourcier (honorary), Friedrich Lachmayer (honorary), Thorne McCarty, Henry Prakken, Ken Satoh, Giovanni Sartor, Tom van Engers, Hajime Yoshino (honorary) and John Zeleznikow.

Contents

This page intentionally left blank

1. Visualisation and Legal Informatics

This page intentionally left blank

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210310

Visualization of Legal Informatics

Friedrich LACHMAYER [a] and Vytautas ČYRAS [b,1]
[a] *Vienna, Austria*
[b] *Vilnius University, Lithuania*

Abstract. This paper explores the subject matter of legal informatics. The life-long work of the first author concerning the visualization and coding of statutes is generalized. Besides positive law and customary law, the emergence of machine law is a current topic of focus in the literature. In machine law, legal acts are posited by machines and not by humans (primarily in a situational context). The transformation of a legal act to a legal document can happen in two ways. First, it is a transformation of the legal act into explicit punctuation, for example, for announcement in the case of laws or for written execution in the case of judgments, and, second, as a trend towards electronic documents. Legal theory forms a meta-level to the law and similarly legal informatics forms a meta-level to legal information. Legal informatics in Austria is based on the work of Ota Weinberger, Ilmar Tammelo and Leo Reisinger and has been developed by Erich Schweighofer in the framework of the IRIS conferences. Legal informatics is distinguished from legal information, whereas legal logic and meta-theories appear on top of legal informatics. In terms of syntax, machine culture is characterized by formal notations. Notations of legal logic are just the beginning; the target is a technical notation, a basis for programming. Visualizations are in the middle. On the one hand, visualizations serve to understand people by breaking away from the textual; on the other hand, by emphasizing the formal they form a bridge to machines. Legal text can be translated directly into formal languages, but visualizations can facilitate this task as an intermediate methodological step. Hans-Georg Fill's metamodeling can be seen as a metameta-level.

Keywords. Machine law, legal act, legal document, legal logic, formalization, visualization

1. Transition from Legal Act to Legal Document

To date, the law has known two stages of development: customary law and positive law. A third stage of development is now emerging, namely machine law.

In law, a distinction must be made between the legal act (speech act) and a document (see Figure 1). The legal act of a law consists of the speech act of parliament. Usually, the announcement of the law will be added in a publication gazette, but the resolution of the law takes place in parliament and not in the publication organ. Similarly to a judgment, the announcement of the judgment will be constitutive and written copy can be added. Indeed, the judgment could also only be given in writing, i.e., without prior verbal announcement.

This is similar to legal documentation where the content of the legal act is shown in the document. In legal information, however, there is now a tendency for the speech act

[1] Corresponding author, Institute of Computer Science, Faculty of Mathematics and Informatics, Vilnius University, Didlaukio 47, Vilnius, Lithuania; E-mail: vytautas.cyras@mif.vu.lt.

and document to merge: there is only one integrative act that consists of a legal act and an electronic document at the same time, if, for example, the legal act is already being set electronically by a machine.

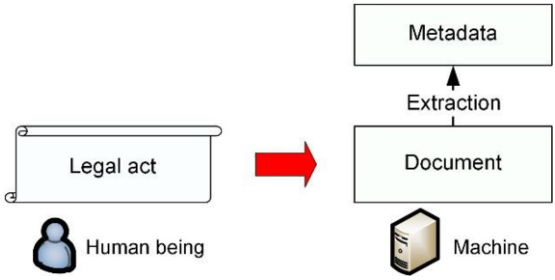


Figure 1. Transition from legal act to legal document.

So far, the law has been extensively posited. Situational norms also exist, such as traffic lights, but these have not been interpreted as their own norms, but rather as elements of the facts to which the norms were linked. Machine law will be posited by machines, especially in situational contexts. It is a question of legal or scientific interpretation whether these machines are interpreted as "persons" and the norm positing is interpreted as a "legal act". The 2001 IRIS conference was dedicated to this topic ("On the way to ePerson"). Nowadays the IRIS (International Legal Informatics Symposium; *Internationales Rechtsinformatik Symposion*) is held annually at the University of Salzburg; see https://iris-conferences.eu/.

The arrow in Figure 1 symbolizes the transition in two ways. On the one hand, it concerns the transformation of the legal act into explicit punctuation, for example, for announcement in the case of laws or for written execution in the case of judgments. On the other hand, there is now the trend towards electronic documents, for example, in RIS (*das Rechtsinformationssystem des Bundes*; the Legal Information System of the Republic of Austria, see https://www.ris.bka.gv.at).

Metadata can be extracted from these documents, providing the advantage of easier access to documents when searching. This means the full text search is no longer required. In addition, words not contained in the full text can be added. Additionally, the metadata can be extracted directly from the text. This is a topic in legal informatics.

The law itself can contain the type of metadata, for example, the legal principles in court decisions. These are generated by the court itself (see, for example, Felix Gantner's manuscript entitled *Digital Transformation of the Law* and also [1]).

2. Legal Theory and Legal Informatics

Legal theory is a meta-level to law, just as legal informatics is a meta-level to legal information (see the middle section of Figure 2, labeled *Meta-level*).

When legal informatics emerged, there were several variants of legal theory, such as traditional legal dogmatics, discourse theory, and, as before, theories of natural law. The scientific discourse at that time (at least in Austria) was also shaped by Hans Kelsen's *Pure Theory of Law* [2], the second edition of which was published in 1960. A peculiarity of the Pure Theory of Law lies in the clear line of thought and language, which is dedicated to the structural knowledge of the law and thus formed an analytical

starting point for the subsequent legal informatics. Pure jurisprudence speaks about logic in law, but contains no formal expressions.

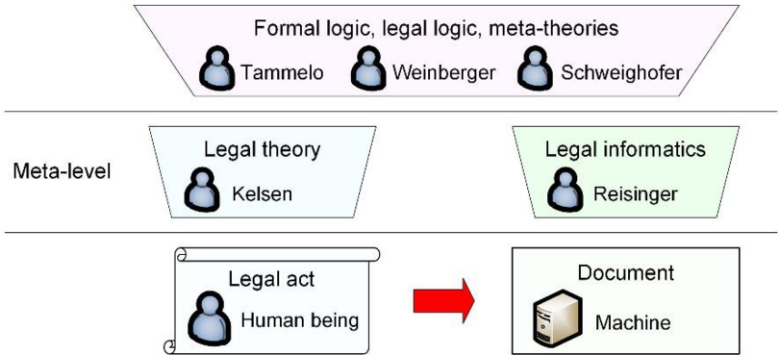


Figure 2. Legal theory and legal informatics on a meta-level.

Nevertheless, legal informatics must be distinguished from legal information. While legal information is usually implemented on a project basis, legal informatics is part of science and belongs to the meta-level of legal information.

Some researchers have viewed legal informatics as a "hyphenated science" because it has two subject areas, namely, law and information. This view of hyphenated science affected the selection of personnel because scientists with a double degree (such as Ota Weinberger, Herbert Fiedler, Leo Reisinger and Erich Schweighofer) gave qualified access.

At the beginning of legal information in the 1970s, there were two concepts for the projects: there was a demand market in which the IT producers had oriented themselves towards the peculiarities of the law and thus incorporated the results of legal theory into the documentation software. Over the course of time, however, this changed in the direction of a supply market: the general documentation software offered is so powerful that (almost) all problems of legal documentation can be solved with it and so it is no longer necessary to take into account (supposed) peculiarities of the law. Here, too, the truth will lie somewhere in the middle, as the vast majority of problems can be solved by general structures and the peculiarities of the law only make up a small but ultimately relevant area of software construction.

Leo Reisinger presented the state of development at that time in his book *Rechtsinformatik*, published in 1977 [3].

3. Legal Logic

For the development of legal informatics, legal logic, which was motivated in the early 1950s by Georg Henrik von Wright [4], constituted an important theoretical basis. Legal logic is clearly illustrated in Figure 2 and acts as a meta-theory.

The topic of legal informatics in German-speaking regions was initially treated theoretically, in particular by Herbert Fiedler [5], Fritjof Haft [6], Lothar Philipps [7], Jürgen Rödig [8] and Spiros Simitis [9].

The situation in Austria was as follows. From the point of view of the first author, the Czech legal philosopher and logician Ota Weinberger was the first to point out the avant-garde position of legal logic in Austria in 1968. The first edition of his book on

legal logic (*Rechtslogik*) was published in 1970 [10]. Consequently, Weinberger became a professor in Graz. His student Alfred Schramm largely devoted himself to legal expert systems.

In 1973 the Estonian legal philosopher Ilmar Tammelo came from Australia to Austria and accepted a position as a professor in Salzburg after Rene Marcic (a representative of natural law). Tammelo was highly innovative and eager to experiment, as well as being in contact with many foreign scholars. The further development of formal notations was an interesting topic for him [11].

Leo Reisinger habilitated as a computer scientist (in Vienna) and a lawyer (in Graz). In the 1970s, he wrote several books on legal informatics. Concerning the logic of law, he adopted the model produced by Carlos E. Alchourrón and Eugenio Bulygin [12].

The first author of this paper has repeatedly taken part in Ilmar Tammelo's seminar in Salzburg. With this tradition in mind, the IRIS was founded in Salzburg in the 1990s together with Erich Schweighofer. The annual IRIS congresses have endeavored to offer a forum for both theory and practice in legal informatics, especially in the form of project culture. Because of Schweighofer's special merits, an extensive conference volume is published and given to participants at the beginning of each congress. With these volumes he creates a knowledge base for legal informatics that can be used in the following years (see e.g. the recent proceedings, IRIS 2021 [13]). In this way, Schweighofer has re-established the Austrian legal informatics community and provided further thematic impulses. Schweighofer has also written about the prospects for legal informatics and legal data science [14].

4. Visualization

Traditionally, law is textual. Jurists transform texts into texts. There are various kinds of texts: laws, contracts, claims, judgements, etc. Text transformations require abstracting, reasoning and other legal methods. Judgements, guidelines and their head notes are formulated in abstract legal terms. Abstracting and extracting are therefore needed and are performed by jurists and secretaries.

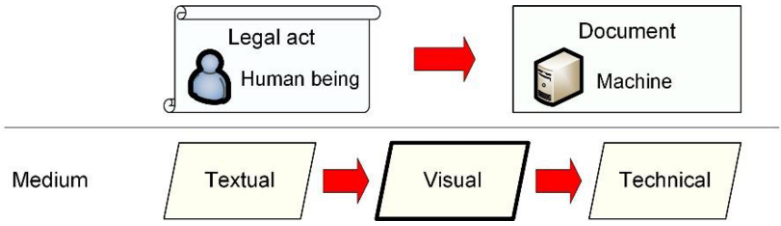


Figure 3. Legal visualization appears in the middle of the multi-arch bridge which leads from textual law to its enforcement by computer.

Hence, positive law, like traditional jurisprudence, is textual. In terms of syntax, machine culture is characterized by formal notations. The logical notations of legal logic are just the beginning. The target area is technical notations as the basis of programming.

Visualization can occupy a middle position (see Figure 3). On the one hand, visualizations can serve to better understand people by breaking away from the textual; on the other hand, by emphasizing the formal, they can represent a bridge to machines.

It is possible that the texts are translated directly into formal language, but it can also be that methodological intermediate steps in the sense of visualizations facilitate this task (see Figure 4 and [15]). The authors have attempted to exhibit such intermediary possibilities in a series of articles (see [15, 16]).

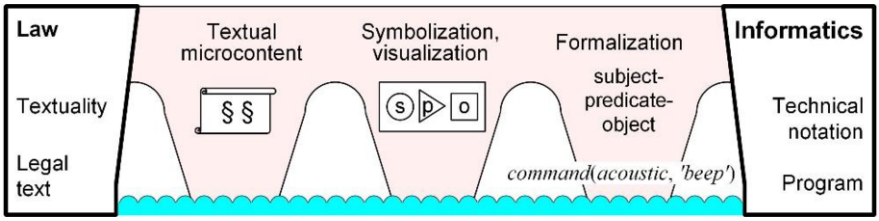


Figure 4. The multibridge metaphor: transformations lead from norm to its machine implementation [15].

The question of whether there is an independent legal logic or if this is simply an application of formal logic to the law is negated when the notation as a syntactic structure is in the foreground. It is entirely possible to develop a special normative notation, just as there is a specific chemical notation, for example, H_2O.

4.1. Transformation from Legal Text to Computer Program

The premise of this paper is that it seems unrealistic to proceed directly in one step from legal texts to their formalization (in the form of logic programming, e.g., Prolog). Intermediate steps are needed. In other words, we hold that a one-arch bridge is unrealistic and advocate a multi-arch bridge of some kind. Hence, an approach in legal informatics is proposed which is called Multi-phase Transformation.

There are many approaches to formalizations in the legal domain. Here, various formalisms, notations, logics and modelling techniques are used. As a one-bridge approach, Tammelo [11] addressed logic-based representation. He was successful in representing short legal texts in the prefix notation of binary operators. However, such formal notation was not easy to read. Sergot et al. [17] employed logic programming while representing the British Nationality Act as a logic program. Grabmair and Ashley [18] examined two transformations: First, the statute text is transformed into an Intermediate Norm Representation, and then to a rulebase.

Whilst the transformation is feasible in the case of a clear statement, difficulties arise with complex texts and a scalability problem is faced. Hence, the quality of transformation is acceptable for small texts only. However, the quantity (scalability) is not acceptable. Here the early attempts of artificial intelligence research on understanding natural language can be recalled. You can succeed in a world of toy blocks, but it would scarcely be possible to represent the meaning in the general case.

4.2. Multi-arch Bridge Implies Multiphase Transformation

The building of a bridge is continued with the observation that legal knowledge representation is needed as an intermediate step. The input/output chain is *Legal text* → *Legal knowledge representation* → *Program*. Next, *Legal knowledge representation* is decomposed into three intermediate stages: *textual microcontent, symbolization/ visualization, and formalization* (see Figure 5).

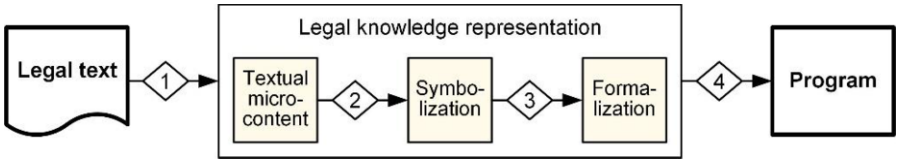


Figure 5. The Multiphase Transformation approach – a multibridge [16].

The four bridging steps in Figures 4 and 5 are represented by 'input → output' pairs:

Step 1. *Microcontenting:* legal text → textual microcontent

Step 2. *Visualizing:* textual microcontent → symbolization/visualization

Step 3. *Formalizing:* symbolization/visualization → formalization

Step 4. *Implementing:* formalization → program.

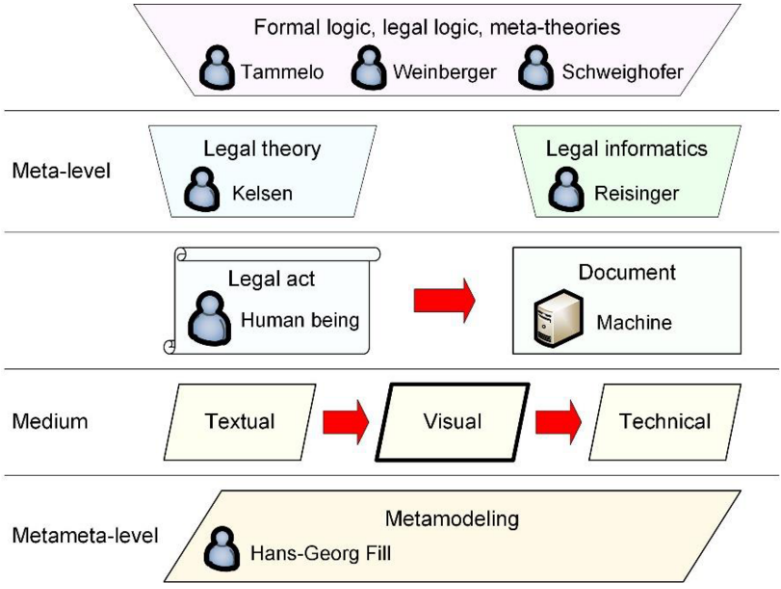


Figure 6. Summary of the discussed topics of legal informatics.

5. Metamodeling

The business informatics specialist Hans-Georg Fill has worked for years on conceptual modeling and visualization in the field of business informatics [19, 20, 21], and also on metamodeling for enterprise systems [22, 23]. We depict Fill's work on a metameta-level in our summarizing Figure 6, in which the relevant section is labeled *Metamodeling*. The semantics conveyed by a visual (i.e. the meaning of the representation) is addressed in [19]. Fill [19, p. 172] holds that "the goal of knowledge explication […] is to explicate knowledge that resides in the heads and minds of people and express it by a visualisation". He lists four basic aims of visualizations: knowledge explication, knowledge transfer, knowledge creation, and knowledge application. Knowledge explication is a primary aim of legal visualization in our approach.

6. Evolution: Animals–Human Beings–Machines

Consider the line of evolution from plants to animals to human beings to machines, as shown in Figure 6 [24]. In the proposed model, biological evolution leads to the development of human beings. The last step, the evolution from humans to machines, however, is a process of technological evolution in which humans produce machines. Moreover, humans strive to give human capabilities to their creatures, thus making machines artificially intelligent, a situation that is reminiscent of the ancient myth of Pygmalion and its modern variations.

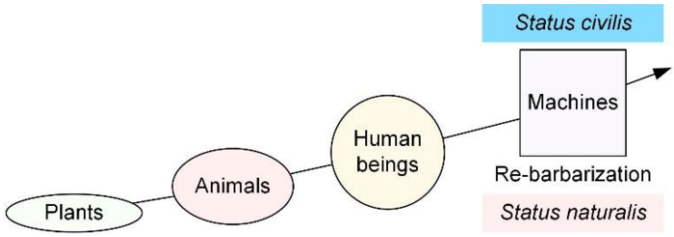


Figure 7. The line of evolution from plants to animals to human beings to machines [24].

One question associated with the evolutionary step from humans to machines is whether machines reside in status civilis or status naturalis. A relapse to status naturalis is a permanent temptation of modern culture, although re-barbarization is a kind of political atavism. Weapons are substitutes for the former raptors. We, however, maintain that machines have to be not monsters.

The theological problem of theodicy,[2] which Leibniz addressed, namely, the place of evil in the Creation, arises again in the case of machines as human creations.

Legal informatics is not just a science that synthesizes between jurisprudence and technology, but it also gives the area of machines a human-like normativity, and it does the same in their role as actors on the human stage, which is transformed into a common stage.

We see the digital ubiquity of an organization, which is examined by Fill [23], as an issue in the evolutionary step to machines.

7. Conclusion

The topics explored within legal informatics are summarized in Figure 6. We hold the belief that the work in progress applies in particular to legal visualization, which acts as a bridge between people's textual understanding of the law and the formal, abstract notations of the machine world.

[2] Theodicy means vindication of God. It is to answer the question of why a good God permits the manifestation of evil, thus resolving the issue of the problem of evil (see Wikipedia, https://en.wikipedia.org/wiki/Theodicy).

References

[1] Gantner F.Theorie der juristischen Formulare. Schriften zur Rechtstheorie, Heft 252. Berlin: Duncker & Humblot; 2010. 178 p.

[2] Kelsen H. Pure theory of law. 2nd ed. (trans: Knight M) (Reine Rechtslehre, 2. Auflage. Deuticke, Wien 1960). Berkeley: University of California Press; 1967. 368 p.

[3] Reisinger L. Rechtsinformatik. Berlin, New York: de Gruyter; 1977. 378 p.

[4] von Wright GH. Deontic logic. Mind 1951 Jan; 60(237):1-15. https://www.jstor.org/stable/2251395.

[5] Fiedler H, Haft F, Traunmüller R, editors. Expert systems in law – impacts on legal theory and computer law. Neue Methoden im Recht vol 4. Tübingen: Attempto; 1988.

[6] Haft F (2010) Das Normfall-Buch: IT-gestütztes Strukturdenken und Informationsmanagement. 4th ed. München: Normfall-GmbH; 2010.

[7] Philipps L. Endliche Rechtsbegriffe mit unendlichen Grenzen: Anthologia. Bern: Editions Weblaw; 2012. 232 p.

[8] Rödig J. Schriften zur juristischen Logik. Editors: Bund E, Schmiedel B, Thieler-Mevissen G. Heidelberg: Springer; 1980. 366 p.

[9] Simitis S, Hornung G, Indra Spiecker Döhmann I, editors. Datenschutzrecht: DSGVO mit BDSG. Baden-Baden: Nomos; 2019. 1474 p.

[10] Weinberger O. Rechtslogik. 2nd ed. Berlin: Duncker & Humblot; 1989. 432 p.

[11] Tammelo I. Modern logic in the service of law. Vienna: Springer; 1978. 196 p.

[12] Alchourrón CE, Bulygin E. Normative systems. LEP Library of Exact Philosophy 5. New York and Vienna: Springer-Verlag; 1971. 232 p.

[13] Schweighofer E, Kummer F, Saarenpää A, Eder S, Hanke P, editors. Cybergovernance: Proceedings of the 24th International Legal Informatics Symposium IRIS 2021, Feb 24-27. Bern: Editions Weblaw; 2021. 411 p. Jusletter IT, 25 February 2021. https://jusletter-it.weblaw.ch/issues/2021/25-Februar-2021.html.

[14] Schweighofer E. From information retrieval and artificial intelligence to legal data science. In: Schweighofer E, Galindo F, Cerbena C, editors. In: Proceedings MWAIL2015, ICAIL multilingual workshop on AI & Law Research, 15th International Conference on Artificial Intelligence and Law (ICAIL 2015). books@ocg.at, vol 313. Vienna: OCG; 2015. p. 13–23. http://fedora.phaidra.univie.ac.at/fedora/get/o:399570/bdef:Content/get.

[15] Čyras V, Lachmayer F. Multisensory legal machines and legal act production. In: 25th IVR World Congress: Law Science and Technology; 2011 Aug 15-20. Series A: Methodology, logics, hermeneutics, linguistics, law and finance, paper No. 026/2012. Goethe University Frankfurt am Main; 2012. http://publikationen.ub.uni-frankfurt.de/files/24884/IVR_World_Congress_2011_No_026.pdf.

[16] Čyras V, Lachmayer F. Multiphase transformation in the legal text-to-program approach. In: Sturm F, Thomas P, Otto J, Mori H, editors. Liber amicorum Guido Tsuno. Vico Verlag, Frankfurt am Main; 2013. p. 57-70. Available at SSRN: https://ssrn.com/abstract=2632045.

[17] Sergot MJ, Sadri F, Kowalski RA, Kriwaczek F, Hammond P, Cory HT (1986) The British Nationality Act as a logic program. Communications of the ACM 1986; 29(5):370-386.

[18] Grabmair M, Ashley K. Towards modeling systematic interpretation of codified law. In: Moens M, Spyns P, editors. Legal knowledge and information systems. JURIX 2005: the eighteenth annual conference. Frontiers in artificial intelligence and applications 134. Amsterdam: IOS Press; 2005. p. 107-108.

[19] Fill HG. Visualisation for semantic information systems. Wiesbaden: Springer Gabler; 2009. 347 p.

[20] Fill HG. Transitions between syntax and semantics through visualization. In: Schweighofer E et al., editors. Zeichen und Zauber des Rechts. Bern: Editions Weblaw, p. 935-44; 2014.

[21] Pittl B, Fill HG. A visual modeling approach for the semantic web rule language. Semantic Web 2020; 11(2):361-89.

[22] Fill HG. Abstraction and transparency in meta modeling. In: Schweighofer E, Kummer F, Hötzendorfer W, editors. Transparency: Proceedings of the 17th International Legal Informatics Symposium IRIS 2014, books@ocg.at, vol 302. Vienna: OCG, p. 435-42; 2014. Jusletter IT 20 February 2014. doi: 10.38023/88d2121f-5631-433e-ab17-bace2ec98211.

[23] Fill HG. Enterprise modeling: from digital transformation to digital ubiquity. In: Ganzha M, Maciaszek L, Paprzycki M, editors. Proceedings of the 2020 Federated Conference on Computer Science and Information Systems. ACSIS vol 21, p. 1-4; 2020. doi: 10.15439/2020F001.

[24] Čyras V, Lachmayer F. Legal visualisation in the digital age: from textual law towards human digitalities. In: Hötzendorfer W, Tschohl C, Kummer F, editors. International trends in legal informatics: festschrift for Erich Schweighofer. Bern: Editions Weblaw; 2020, p. 61-76.

2. Knowledge Representation and Data Analytics

This page intentionally left blank

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210312

Automatically Identifying Eviction Cases and Outcomes Within Case Law of Dutch Courts of First Instance

Masha MEDVEDEVA [a,b,1], Thijmen DAM [a,b], Martijn WIELING [b] and Michel VOLS [a]
[a] *Department of Legal Methods, University of Groningen (UG), The Netherlands*
[b] *Center for Language and Cognition Groningen, UG, The Netherlands*

Abstract. In this paper we attempt to identify eviction judgements within all case law published by Dutch courts in order to automate data collection, previously conducted manually. To do so we performed two experiments. The first focused on identifying judgements related to eviction, while the second focused on identifying the outcome of the cases in the judgements (eviction vs. dismissal of the landlord's claim). In the process of conducting the experiments for this study, we have created a manually annotated dataset with eviction-related judgements and their outcomes.

Keywords. outcome identification, case law, machine learning, judicial decision

1. Introduction

Legal scholars and practitioners are confronted with an enormous and expanding body of case law. For example, in the Netherlands the judiciary dealt with over 1.3 million cases in 2020 alone.[2] Many of these cases involve bulk cases on, for example, family law or labour law. Another area that results in a considerable number of bulk cases is landlord-tenant law. It is estimated that courts have to decide whether or not a tenant needs to be evicted in nearly 20.000 cases every year (1). The Dutch judiciary does not publish all judgements online, but a significant number of cases can be found online in the *Open Data van de Rechtspraak* dataset.[3] Traditionally, legal scholars and practitioners collect and analyse these cases manually (2). Of course, this is time-consuming and will become impossible due to the increasing amount of published case law online.[4]

In this paper we are trying to solve this legal research problem. Specifically, we want to identify judgements concerning eviction within all judgements published by the Dutch judiciary and extract their outcome from the text (i.e. eviction/non-eviction). This work builds upon existing research that until now has been done manually (3), and our goal is to test how much of the data collection and outcome extraction can be automated. Some

[1] Corresponding Author. E-mail: m.medvedeva@rug.nl
[2] https://www.rechtspraak.nl/Organisatie-en-contact/Rechtspraak-in-Nederland/Rechtspraak-in-cijfers (Dutch)
[3] https://www.rechtspraak.nl/Uitspraken/Paginas/Selectiecriteria.aspx (Dutch)
[4] https://www.volkskrant.nl/nieuws-achtergrond/raad-voor-de-rechtspraak-meer-vonnissen-online-publiceren~bf045df7 (Dutch)

of the case law under review has already been annotated by hand and can be used for training machine learning systems.

In this paper, we use 'judgement' to denote the text of a published judgement. The word 'outcome', and its synonyms 'verdict' and 'decision', are used to define a specific closed class (i.e. a limited number) of labels for verdicts. An example of an outcomes is *eviction* or *non-eviction* in the landlord-tenant law context.

2. Related work

This paper deals with the identification of the topic of a judgement (i.e. an eviction case or a non-eviction case). To our knowledge, the number of publications dealing with automatically identifying the topic of a case for dataset creation is limited. Some similar work involves topic modelling techniques that allow one to identify and cluster multiple topics at once (4; 5; 6), and using document similarity to find the documents that deal with similar issues (7; 8), both of which can be particularly hard to evaluate.

Besides the identification of the topic of a judgement, this paper concerns outcome identification (i.e. the extraction of the outcome from the full text). This identification task can be useful in itself, for instance if one wants to know the statistics of cases concerning eviction actually ending in a tenant being evicted.

Depending on the court, identifying the outcome can be more or less complicated. Some courts publish their judgements with meta-data stating the outcome (e.g., the European Court of Human Rights). As a result, one just needs to extract this information in order to get the outcomes. While in other judgements, the wording of the outcome may be standardised and therefore easy to extract (e.g., "The Court of Appeal therefore affirms the decision of the Court of First Instance."), the majority of courts seem to formulate their decisions in free-form natural language, making the task of extracting a specific outcome a more complex task.

There is a small number of studies focusing specifically on identifying the outcome within the judgements. Recent papers extracted outcomes from Appellate Decisions in US State courts (9; F1-score: 0.82), US federal court dockets (10; recall up to 0.96) as well as French courts (11; F1-scores: 0.8-1.0) using various machine learning methods. In this paper we compare the performance of a simpler keyword-search approach (not requiring annotated data) to a simple machine learning system.

3. Data

For our dataset we rely upon the *Open Data van de Rechtspraak*,[5] an official, publicly available, database of the judiciary of the Netherlands. Not all Dutch case law is published online, but merely a subset of judgements that *De Rechtspraak* allows for publication. Unfortunately, exact criteria are not available to the public, though some guidance is provided by a dedicated page on their website.[6] The *Open Data van de Rechtspraak* dataset can be downloaded as one large file archive (>4GB) of XML files containing the texts of the judgements as well as some basic meta-data (e.g., court, date).

[5] https://www.rechtspraak.nl (Dutch)

[6] https://www.rechtspraak.nl/Uitspraken/Paginas/Selectiecriteria.aspx (Dutch)

For this paper, we are specifically interested in the cases of the courts of first instance (*rechtbanken*). A collection of 591 eviction cases between 2000 and 2020 (manually collected and annotated, including the verdicts: eviction or non-eviction) from the courts of first instance was already available (based on existing research from our lab). This dataset was compiled in such a way that it should include the large majority of all published eviction cases between 2000 and 2020. As this dataset only contains a relatively limited number of eviction cases, and no non-eviction cases, we aimed to supplement it by including both cases about eviction, and cases on other topics, but still somewhat related to the subject matter. This was to ensure the task was useful and not trivial, and that we had a larger dataset to train the system.

To increase the likelihood of identifying eviction cases, we used the following procedure. We extracted all (2,641,946) judgements from the *Open Data van de Rechtspraak* dataset between the years 2000 and 2020. From this set, we only included judgements from the courts of first instance, and furthermore selected the judgements that contain at least one of the following words *huurovereenkomst* (rental agreement), *ontruiming* (eviction) or *woning* (home). Subsequently, we narrowed down the selection by only retaining judgements with the label "private law", which is the appropriate label for eviction cases. These relatively simple filters allowed us to reduce the amount of judgements to 24,268 cases. Unfortunately, this number was still rather large. Consequently, we made a further reduction by only including cases from 2016-2018, and excluding cases already included in the original set of 591 cases of the original set, yielding a set of 4,795 judgements. From this set, we randomly sampled 69 judgements (1 hour of manual annotation) to assess the proportion of cases related to eviction. A manual inspection showed that more than half of the judgements (37) were eviction cases. This suggests that our manually curated dataset of 591 eviction cases was missing a substantial amount of eviction-related cases.

To increase the amount of data, we took all 591 manually annotated eviction judgements, and we again randomly sampled from the 2016-2018 judgements, extracting twice the amount (1182) of manually annotated eviction judgements. We then built a simple three-fold cross-validation support vector machine (SVM) only using 1-3 n-grams (i.e. sequences of one to three words from the text of the judgement) as features. When training the model, we treated the 591 judgements as eviction cases and the 1182 judgements as non-eviction cases.[7] Of course this is a sub-optimal class distinction, as potentially many of the 1182 judgements may, in fact, be eviction cases. Consequently, from all cases that were classified as non-eviction cases, we only retained those which were (when included in the test set during the three-fold cross-validation procedure) assigned the non-eviction label with over 99% confidence (using Platt Scaling; 13). This reduced the number of non-eviction judgements in our training set to 809. We then trained the system again (using 809 non-eviction cases and 591 eviction cases) and evaluated it by using the rest of the judgements from between 2000 and 2020. Out of 22,868 judgements, 3,277 (14%) were predicted as eviction-related.

Of course, not all predictions will be correct. To supplement our final correct training dataset, however, we did not use these predictions, but instead used these simply to guide two subsequent manual annotation rounds (the first annotation round included the 69 aforementioned cases). Specifically, we asked two legal experts to spend eight hours

[7] For a more detailed explanation of machine learning classification and its evaluation (i.e. precision, recall, f1-score, accuracy) applied to legal texts, see (12) and (4).

in total on annotating judgements that our model predicted as eviction-related in the second annotation round (under the assumption that many would *not* be eviction related), and an additional four hours in total in a third annotation round focusing on the judgements that our predicted as *non-eviction*. The annotators were provided the full text of a randomly selected judgement and they had to simply identify whether the judgement was concerning an eviction or not. In the allocated time, 716 judgements were annotated. Out of predicted eviction judgements 298 (55%) turned out to be eviction related, while 243 judgements (45%) were not. In addition, and the vast majority of non-eviction cases 161 out 175 (92%) turned indeed out to be non-eviction related. This left us with a dataset with 940 eviction judgements, and 436 non-eviction judgements. Table 1 provides an overview of our final dataset.

Table 1. Number of available data in the initial dataset and after three rounds of annotation.

	Eviction	Non-eviction
Initial dataset	591	0
First annotation round	37	32
Second annotation round (predicted as eviction)	298	243
Third annotation round (predicted as non-eviction)	14	161
Total	940	436

Once we identify the eviction-related judgements, we are also interested in their outcome. In the judgements concerning evictions, the courts of first instance can decide to evict the resident and/or cancel the lease (labelled as *eviction*) or reject the property owner's claim (labelled as *non-eviction*). The cases are decided on by a single judge. All of the eviction cases in the court of first instance are property owner vs. resident, with the latter being the defendant.

4. Experiment I: Identifying Eviction-Related Judgements

4.1. Methodology

From the final dataset (see Table 1), we used 200 judgements (100 eviction-related and 100 non-related) to test and evaluate the model, which left us with 840 eviction and 336 non-eviction judgements to train and fine-tune the system. We then balanced this dataset for training, leaving us with 336 eviction-related judgements and the same number of non-related judgements. We used three-fold cross-validation to fine-tune the parameters and ended up using a linear support vector machine, using as features the frequencies of 1-6 character n-grams (i.e. sequences of one to six characters).[8] The results of the best model can be found in Tables 2 and 3.

[8]The following command, showing all used parameters, was used to fit our final model: `CountVectorizer(analyzer = 'char', ngram_range = (1,6), max_features=None, max_df = 0.7, lowercase=False, binary=True); LinearSVC(C=0.01)`. For more details on each parameter see the `sklearn` documentation available at https://scikit-learn.org/. The full set of parameters we experimented with can be found in our code and data available at https://github.com/masha-medvedeva/EVICT.

Table 2. Results (precision, recall, f1-score and accuracy) for identifying eviction-related judgements using three-fold cross-validation.

	Precision	Recall	F1-score	Support
Non-eviction	0.90	0.88	0.89	336
Eviction	0.88	0.90	0.89	336
Accuracy			0.89	672
Macro avg.	0.89	0.89	0.89	672
Weighted avg.	0.89	0.89	0.89	672

Table 3. Results (confusion matrix) for identifying eviction-related judgements using three-fold cross-validation.

		Actual topic	
		Non-eviction	Eviction
Predicted topic	Non-eviction	294	42
	Eviction	32	304

4.2. Results

Our final results, when evaluating our model on the held-out test set, are shown in Tables 4 and 5.

Table 4. Results (precision, recall, f1-score and accuracy) on the test set for identifying eviction-related judgements.

	Precision	Recall	F1-score	Support
Non-eviction	0.92	0.81	0.87	100
Eviction	0.83	0.95	0.89	100
Accuracy			0.88	200
Macro avg.	0.89	0.88	0.88	200
Weighted avg.	0.89	0.88	0.88	200

Table 5. Results (confusion matrix) on the test set for identifying eviction-related judgements.

		Actual topic	
		Non-eviction	Eviction
Predicted topic	Non-eviction	81	19
	Eviction	5	95

The results suggest that when having a reasonable amount of annotated data, it is possible to identify eviction-related cases with a relatively high accuracy of about 88%. Consequently, this automatic procedure can be suitably used to speed up the process of finding relevant (eviction-related) case law.

When we evaluated the model on all (filtered) judgements published between 2000 and 2020 not included in our dataset, a total of 3,248 out 22,872 cases (all original judgements between 2000 and 2020, excluding all annotated judgements) were classified as eviction-related judgements. With an estimated precision of 83%, we expect about 2,695 cases to be actual eviction-related judgements. Similarly, with an estimated precision of 92% in identifying non-eviction related judgements, we expect an additional 8% of these (i.e. 1569 judgements) to be eviction-related.

5. Experiment II: Identifying the Outcome

5.1. Methodology

Once we identified the eviction-related judgements, we were interested in identifying how many of them actually resulted in the eviction of a resident. Identifying the verdict should not necessarily always be a machine learning task. A simple keyword search could potentially be sufficient. Therefore, we first tried determining words that may be characteristic of a specific outcome. While judgements of the courts of first instance do not have a clear structure, they could potentially use the same wording for the verdict itself. We then compare these results to a more sophisticated machine learning system which is able to take more advanced features into account.

For this experiment we used the full set of 940 eviction-related cases shown in Table 1. Except for the cases included in the initial dataset which already included an annotated outcome, we asked two legal experts to annotate the outcome of each case: *eviction* or *non-eviction*. We excluded 28 cases that had other types of verdicts, such as only cancellation of the lease, but no eviction, etc. The final dataset for this task contained 912 judgements (620 having an eviction outcome, whereas 292 had a non-eviction outcome).

5.1.1. Keyword-Based System

For the keyword-based system, we determined (via manual inspection of several cases) a number of terms that relate to each specific outcome. We then automatically searched for these terms in the *decision* section of the published judgement, and in cases where the decision section was not specified, in the bottom part (2500 characters) of the text. The keywords that we chose for being representative of an *eviction* outcome were (including different forms of the same words): *ontbinding* (cancellation), *ontruiming* (eviction), and *verlaten* (leave). If none of these words were found, our keyword-based system determined that no eviction had been ordered by the court.

We tested the method on all 912 judgements that we had labels for. The results of this system can be found in Tables 6 and 7.

Table 6. Results (precision, recall, f1-score and accuracy) for identifying the outcome of eviction cases using keyword extraction.

	Precision	Recall	F1-score	Support
Non-eviction	0.88	0.65	0.75	292
Eviction	0.85	0.96	0.90	620
Accuracy			0.86	912
Macro avg.	0.87	0.80	0.82	912
Weighted avg.	0.86	0.86	0.85	912

Table 7. Results (confusion matrix) on the test set of identifying the outcome of eviction cases using keyword extraction.

		Actual outcome	
		Non-eviction	Eviction
Predicted outcome	Non-eviction	189	103
	Eviction	25	595

This simple system achieved reasonably good results, although we can see that non-eviction is not categorised very well, 103 (35%) out 299 non-eviction cases were misclassified. However, the issue with a keyword-based system, is that it is very hard to improve upon, unless one can come up with more specific keywords. Moreover, if the keywords from one type of outcome are found in the judgement with a different outcome, this is hard to correct. For instance, a judgement can contain the phrase "at this point, eviction is not necessary". While the word 'eviction' is present in this judgement, the case clearly resulted in no eviction. However since we are just dealing with individual words, it is hard to incorporate all possible nuances.

Nonetheless, as opposed to a system using machine learning, which we will discuss in the next subsection, this system does not require any prior annotation, other than determining the keywords.

5.1.2. Machine Learning System

During the keyword-based experiment, we determined that the outcome usually appears within the last 2500 characters of the judgement. While we experimented with shorter and longer fragments, this subset seemed to work best for both of the experiments in identifying the verdict. We used the initial dataset for training and cases from the first, second and third rounds of annotation for testing. We have built a Linear SVC that uses character (1-7) n-grams, and optimised it for a number of other parameters.[9] The results of the model during the cross-validation stage can be found in Tables 8 and 9.

Table 8. Results (precision, recall, f1-score and accuracy) for identifying the outcome of eviction cases using three-fold cross-validation.

	Precision	Recall	F1-score	Support
Non-eviction	0.97	0.96	0.96	183
Eviction	0.98	0.99	0.98	379
Accuracy			0.98	562
Macro avg.	0.98	0.97	0.97	562
Weighted avg.	0.98	0.98	0.98	562

Table 9. Results (confusion matrix) for identifying the outcome of eviction cases using three-fold cross-validation.

		Actual outcome	
		Non-eviction	Eviction
Predicted outcome	Non-eviction	175	8
	Eviction	5	374

5.2. Results

We then tested the model on the cases that we were able to extract in the previous experiment. The performance on the test set can be found in Tables 10 and 11.

[9]The following command, showing all used parameters, was used to fit our final model: `CountVectorizer(analyzer = 'char', ngram_range = (1,7), max_features=2000, max_df = 0.9, lowercase=True, binary=True); LinearSVC(C=0.001)` The full set of parameters we experimented with can be found in our code and data available at https://github.com/masha-medvedeva/EVICT.

Table 10. Results (precision, recall, f1-score and accuracy) on a test set for identifying the verdict of eviction cases.

	Precision	Recall	F1-score	Support
Non-eviction	0.82	0.94	0.88	109
Eviction	0.97	0.91	0.94	241
Accuracy			0.92	350
Macro avg.	0.90	0.92	0.91	350
Weighted avg.	0.92	0.92	0.92	350

Table 11. Results (confusion matrix) on a test set for identifying the verdict of eviction cases.

		Actual outcome	
		Non-eviction	Eviction
Predicted outcome	Non-eviction	102	7
	Eviction	22	219

As we can see from the results, we were able to achieve a very high performance, especially for the eviction class. When inspecting the cases manually, it is clear that the phrasing of the judgement outcomes varies to a large extent from case to case. Similar as in many other natural language processing tasks, the best-performing model included not word n-grams, but character n-grams (14; 15). While we did try using word n-grams for this experiment, in the hope of identifying additional keywords for the keyword-based approach, we did not identify any additional unique words for both outcomes. The performance of the machine learning approach was much higher than the performance of the keyword-based approach. However, whereas the machine learning approach requires annotated data, the keyword-based method does not.

6. Discussion and Conclusion

In this paper we have presented two experiments, one to identify case law having a certain topic, specifically judgements concerning evictions, and one to subsequently identify the outcomes of these eviction judgements. For both tasks, we have shown a high performance, being able to identify eviction-related cases with 88% accuracy, and correctly identifying the outcome in eviction-related cases in 92% of cases. While identifying this type of information may seem easy (as the information is available when reading the document), a keyword-based approach showed it is not straightforward when the information is provided as natural text. While in this paper we were not able to identify *all* eviction cases perfectly, our machine learning approach can suitably be used to identify cases which are *likely* to be eviction cases. Manually checking this smaller set of cases (at a rate of about 1 case per minute) is feasible, whereas checking the full set is not. With relatively little effort, a new database containing thousands of cases can therefore easily be created.

Such a more restricted subject-specific database is also useful in the context of increasing research focusing on categorising or forecasting court decisions (12; 16; 17; 18; 19). This type of research is mostly limited to only a few courts, such as the US Supreme Court or the European Court of Human Rights. This is partly due to the courts' publishing policies, even though more and more courts publish their case law. The dom-

inant focus on a few courts, however, is also caused by the relative large diversity of uncategorised cases in other courts. Therefore narrowing down the task, as we have done here, will likely help subject-specific machine learning systems for these courts (e.g., distinguishing between bankruptcy cases) to be developed.

References

[1] Vols M. Evictions in the Netherlands. In: Loss of Homes and Evictions across Europe. Edward Elgar Publishing; 2018. p. 214–238.

[2] Vol M. Legal Research. One Hundred Questions and Answers. Eleven; 2021.

[3] Vols M. European Law and Evictions: Property, Proportionality and Vulnerable People. European Review of Private Law. 2019;27(4).

[4] Dyevre A. Text-mining for Lawyers: How Machine Learning Techniques Can Advance our Understanding of Legal Discourse. Available at SSRN 3734430. 2020.

[5] Silveira R, Fernandes CG, Neto JAM, Furtado V, Pimentel Filho JE. Topic Modelling of Legal Documents via LEGAL-BERT. Proceedings http://ceur-ws org ISSN. 2021;1613:0073.

[6] Remmits Y. Finding the topics of case law: Latent dirichlet allocation on supreme court decisions [Bachelor's thesis]; 2017.

[7] Novotná T, et al. Document Similarity of Czech Supreme Court Decisions. Masaryk University Journal of Law and Technology. 2020;14(1):105-22.

[8] Barco Ranera LT, Solano GA, Oco N. Retrieval of Semantically Similar Philippine Supreme Court Case Decisions using Doc2Vec. In: 2019 International Symposium on Multimedia and Communication Technology (ISMAC); 2019. p. 1-6.

[9] Petrova A, Armour J, Lukasiewicz T. Extracting Outcomes from Appellate Decisions in US State Courts. In: Legal Knowledge and Information Systems: JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020. vol. 334. IOS Press; 2020. p. 133.

[10] Vacek T, Schilder F. A sequence approach to case outcome detection. In: Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law; 2017. p. 209-15.

[11] Tagny-Ngompé G, Mussard S, Zambrano G, Harispe S, Montmain J. Identification of Judicial Outcomes in Judgments: A Generalized Gini-PLS Approach. Stats. 2020;3(4):427-43.

[12] Medvedeva M, Vols M, Wieling M. Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law. 2020;28(2):237-66.

[13] Platt J, et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers. 1999;10(3):61-74.

[14] Basile A, Dwyer G, Medvedeva M, Rawee J, Haagsma H, Nissim M. N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017. In: CEUR Workshop Proceedings. vol. 1866; 2017. .

[15] Medvedeva M, Kroon M, Plank B. When sparse traditional models outperform dense neural networks: the curious case of discriminating between similar languages. In: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial); 2017. p. 156-63.

[16] Chalkidis I, Androutsopoulos I, Aletras N. Neural Legal Judgment Prediction in English. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 4317-23. Available from: https://www.aclweb.org/anthology/P19-1424.

[17] Katz DM, Bommarito II MJ, Blackman J. A General Approach for Predicting the Behavior of the Supreme Court of the United States. PloS one. 2017;12(4).

[18] Waltl B, Bonczek G, Scepankova E, Landthaler J, Matthes F. Predicting the outcome of appeal decisions in Germany's tax law. In: International Conference on Electronic Participation. Springer; 2017. p. 89-99.

[19] Strickson B, De La Iglesia B. Legal Judgement Prediction for UK Courts. In: Proceedings of the 2020 The 3rd International Conference on Information Science and System; 2020. p. 204-9.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210313

A Pragmatic Approach to Semantic Annotation for Search of Legal Texts - An Experiment on GDPR

Adeline NAZARENKO [a,1], François LÉVY [a] and Adam WYNER [b]
[a] *LIPN, University Sorbonne Paris Nord, France*
[b] *Department of Computer Science, Swansea University, United Kingdom*

Abstract.
Tools must be developed to help draft, consult, and explore textual legal sources. Between statistical information retrieval and the formalization of textual rules for automated legal reasoning, we defend a more pragmatic third way that enriches legal texts with a coarse-grained, interpretation-neutral, semantic annotation layer. The aim is that legal texts can be enriched on a large scale at a reasonable cost, paving the way for new search capabilities that will facilitate mining of legal sources. This new approach is illustrated on a proof-of-concept experiment that consisted in semantically annotating a significant part of the French version of the GDPR. The paper presents the design methodology of the annotation language, a first version of a *Core Legal Annotation Language* (CLAL), together with its formalization in XML, the gold standard resulting from the annotation of GDPR, and examples of user questions that can be better answered by semantic than by plain text search. This experimentation demonstrates the potential of the proposed approach and provides a basis for further development. All resources developed for that GDPR experiment are language independent and are publicly available.

Keywords. Text annotation, Semantic search, Annotation methodology, Semantic markup language, Law

1. Introduction

One of the main early objectives of AI and Law [1] has been to analyse legislation and regulations to allow for querying and facilitate reasoning. The European General Data Protection Regulation (GDPR)[2], imposes data protection by design and by default rules on all companies and organisations that collect and use personal data. This requires all organisations in all EU member states to evaluate their compliance with the GDPR, which is why it is so important to provide tools assisting legal analysis.

Taking the reasoning approach, researchers have designed ontologies adapted to privacy issues [2,3] combined an ontology with a deontic logic [4], formalized the rules of GDPR provisions [5], and encoded rules in a machine-readable form [6], amongst others. Despite these efforts, automating reasoning for compliance evaluation seems un-

[1]Corresponding Author: A. Nazarenko, LIPN, France; E-mail: adeline.nazarenko@lipn.univ-paris13.fr.
[2]Adopted in 2016 and entered into force the 25th May 2018.

likely in the near term or on a significant scale. The *knowledge bottleneck* in the process of translating from Natural Language to formal and exploitable logical representation introduces several difficulties: the process of interpretation resists formalization because of the ambiguity and vagueness of legal texts along with the plurality of interpretation; operationalization – from few articles to a whole law and a set of laws – is a challenge in itself; and legal practitioners have to be convinced that they can use the new services.

Even if such issues remain of great research interest, we consider with [7] that information retrieval represents a promising alternative approach, more practicable on a large scale and more flexible to address the diversity of drafters' and legal professionals' uses (*e.g.* retrieving, clustering, comparing and contextualizing provisions). While public services[3] are adapted for the browsing of multiple legal texts, they do not offer advanced semantic search functionalities, which could facilitate the daily work of legal professionals as well as contribute to the impact of digital and AI methods in the long term. Semantic annotation and search can usefully identify the named entities, other relevant textual segments, and relationships amongst them.

To navigate between the difficult analysis of statutory rules and needs of current legal practice, we propose a coarse-grained and interpretation-neutral approach to annotating legal texts with semantic information, enabling semantically-based information retrieval capabilities. Since the annotations apply to textual passages, leave out finer details, and constitute a simple language, people familiar with legal sources can annotate the text; the annotators do not need to be professional lawyers or logicians. Even if any annotation involves some interpretation, annotation in CLAL should remain "neutral" or consensual, in the sense that annotators must agree significantly (high inter-annotator agreement) as to the annotation of textual passages, leaving aside what does not meet with consensus.

This paper reports on a proof-of-concept experiment on the GDPR. It provides an annotation language, methodology, and annotated text. We demonstrate semantic search by answering user queries, such as *What are the obligations of a data controller?, What are the rights of the data subject?, What are the possible fines and sanctions issued in response to violations by a data controller? Who supervises a data controller?*[4]. This shows the practicability of the language, method, and corpus for semantic search.

The article is organized as follows. Sections 2-4 present the core of the work: the approach taken to defining annotation guidelines, the proposed vocabulary and the study on GDPR. Section 5 describes previous work on the semantic analysis and annotation of legal texts. Section 6 concludes and outlines the perspectives of this work.

2. GDPR annotation methodology

For this work, we followed the methodological recommendations inherited from the major annotation campaigns of the 1990s and 2000s [9].

Design of the annotation language Ideally, the annotation should identify elementary provisions so that they can be easily found and browsed by users. The *provision fragment* (hereafter *fragment*) is the main element of the annotation language: its text span

[3] Such as Legifrance (`www.legifrance.gouv.f`) or UK legislation (`www.legislation.gov.uk`).

[4] For demonstration purpose, we started with a small test set of user questions, some of which were derived from the competency questions of [8].

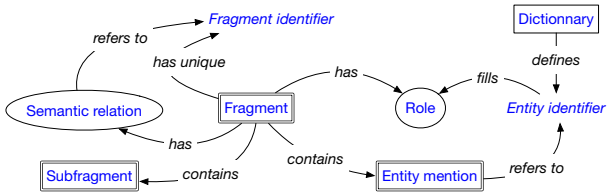


Figure 1. Relations between the CLAL components (in blue) encoded as XML elements with or without text content (double and simple square boxes), XML attributes (round boxes) and identifiers (no box).

corresponds to that of a sentence and it is a practical approximation of an elementary provision. Different types of fragments have been introduced, so that all sentences can be categorized, and the regulation be entirely annotated.

We also defined: some entity categories to account for the actors and concepts that play key roles in provisions; some relations to model the regulation as a semantic graph of entities and provision fragments; and a sub-fragment for encoding exceptions expressed not as independent sentences or fragments but as clauses or prepositional phrases.

Annotation tuning The tuning of the vocabulary and annotation instructions was done together with the annotation of the GDPR in an incremental and iterative way by two annotators working in parallel. We first selected the concepts on which there was a broad consensus for the description of elementary provisions, *e.g.* `obligation` and `permission`, and added more as necessary. The discussions on disagreements or misunderstanding led to adding, merging, splitting or redefining categories, until the whole process stabilized when no further language revision was needed after four months. We also had to give up some elements that could not be annotated in a consistent way, like conditional sub-fragments or actions. Approximately, one third of the GDPR was annotated by then, but the language may still evolve according to the needs dictated by the remaining text, for example to add a sanction type.

Writing the annotation instructions Annotation guidelines were written at the end of the annotation tuning phase. They include a presentation of the annotation vocabulary, the semantics associated with each element, a description of its syntax, some examples, and recommendations to facilitate the arbitration of difficult cases.

Evaluation of the annotation The quality of an annotation is first measured by its homogeneity and stability, using inter- and intra-annotator agreements. Comparing the annotations of parallel versions of the text (*e.g.* English/French) would also be interesting.

3. CLAL Semantic annotation vocabulary

This section presents and discusses the main components of the XML CLAL vocabulary, focusing on the most original or less obvious aspects. Figure 1 and Table 1 give an overview of the language, its vocabulary components, and how they relate to each other in the annotation.

3.1. Fragment categories

The language has two types of provision fragments, deontic and non-deontic ones.

Table 1. CLAL vocabulary: main elements and attributes

XML elements				XML attributes	
Entity mentions	Fragments		Sub-fragments	Roles	`bearer`
	Deontic	Non-deontic			`target`
`CONCEPT`	`OBLIGATION`	`DEFINITION`	`EXCEPT`		`obj`
`PERSON`	`PROHIBITION`	`LEGAL_PRECISION`		Semantic relations	`rel`
`LEGAL_ ENTITY`	`PERMISSION`	`QUALITY_ATTRIBUTION`			`except`
	`POWER`	`EXCEPTION`		Other	`type`
	`RIGHT`				

The deontic categories correspond to the traditional, familiar deontic concepts for which CLAL provides 5 types of XML fragment elements: `OBLIGATION`, `PROHIBITION`, `PERMISSION`, `RIGHT`, and `POWER`. The `POWER` case deserves an explanation. The GDPR gives the power to some institutions to specify the impact of rules according to various contextual parameters, as in Art 45 §3,[5] which delegates to the Commission the definition of an "adequate level of protection". The empowered decisions have statutory consequences, which differs from permissions.[6] A `POWER` is further specified as `ruling` or `execution` depending if it concerns a rule definition or an actual case.

A lot of attention was paid to fragments having no obvious deontic semantic value. Definitions and exceptions are usually easy to identify, but 48 out of 178 fragments appeared to pertain neither to a clear deontic category, nor to definitions or exceptions. After discussion and different tests, we included in the vocabulary two additional categories which are sub-typed with the help of an attribute. `QUALITY_ATTRIBUTION` labels fragments specifying that an entity has a given quality entailing legal consequences[7]. `LEGAL_PRECISION` labels fragments that complement other fragments. For instance, the fragment "*The information shall be provided in writing*" (Art. 12) complements a previous `OBLIGATION`.[8] In this case, the precision has a procedural type (`type="procedure"`) but there are also text specifications (`type="text_specification"` for provisions constraining the content of a text, such as contracts or statutory decisions) and underspecifed precisions (`type="default"`).

3.2. Entities

The annotation points out the entities that play key roles in the regulation. In the annotation perspective, the entity which is referred to is and must be distinguished from its mentions, the occurrences of (multi-)words that refer to it. We identified three main types of entities in the GDPR: concepts, persons, and legal entities, the latter two being actors.

3.3. Roles and semantic relationships

To relate the identified elements, the language allows for entity to fragment relations (roles) and fragment to fragment relations (semantic relationships), which are all en-

[5] *The Commission may decide, . . . , that a third country, . . . ensures an adequate level of protection*

[6] Transfers to the third country is allowed without further control.

[7] The fragment "*Any controller [. . .] shall be liable for the damage caused by processing which infringes this Regulation*" (Art 82 §2) is encoded as a `QUALITY_ATTRIBUTION` of a `responsibility` type.

[8] *The controller shall take appropriate measures to provide any information [. . .] to the data subject [. . .].*

coded as XML fragment attributes. The goal of annotation is not to provide a full semantic graph of the regulation provisions and entities but to encode the relations that are indisputable, useful for answering user queries, and otherwise inaccessible to users.

An entity to fragment relation denotes the role played by the entity in the provision. For instance, the `obj` attribute relates the concept to its DEFINITION fragment, the `bearer` attribute indicates the actor primarily concerned by a provision (*e.g.*, the one who is granted a permission or to whom is attributed a quality), and the `target` attribute specifies the actor who incurs an obligation with respect to the right of a person .

The annotation language provides two types of relationship between fragments. 1) The generic dependency one (`rel`) indicates that one fragment specifies the meaning or must be interpreted in the light of another. *E.g.*, it is used to relate a LEGAL_PRECISION to the fragment it complements. 2) The exception relationship is more specific.

Because of their importance for legal reasoning, special attention is given to exceptions. There are several ways to annotate an exception relation, allowing the annotation to follow the structure of the text, but all of them indicate that a piece of text *B* introduces an exception to the rule expressed in a fragment *A*. If *B* is a fragment, it has an `except` attribute with the identifier of *A* as value. *B* can be any type of fragment and it is annotated as EXCEPTION if it has no other identified semantic value.[9] Besides that, a specific sub-fragment element EXCEPT is introduced to mark-up exceptions when the rule and its exception are part of the same sentence.

4. GDPR experiment

The proposed pragmatic approach supporting semantic search has been implemented and tested on the text of the GDPR. This included the formalization of the language itself in XML (annotation vocabulary and rules), the actual annotation of a large part of the French version of the regulation, and the evaluation of the overall approach. The XML CLAL description and the annotated corpus are openly available together with guidelines for the annotation of other versions of GDPR or other regulations.[10] The Oxygen XML Editor[11] and an *ad hoc* semantic search tool were used for the development phase.

4.1. Implementation of the annotation language

The annotation language includes the vocabulary and the set of constraints that regulate the use of annotations. It is formalized as an XML schema which follows two design principles: independence of annotation layers and maximum control of annotations.

Since two different annotation layers – pre-existing structural and CLAL semantic annotations – enrich the same text, it is essential to preserve their independence. Each one has its own namespace and XML schema. An integration layer is added to articulate the structure and the semantics, so that CLAL annotations can be combined with various structural schemas and the integrity of the initial document be preserved if the semantic annotations are to be removed. Technically, a semantically annotated text is associated with 3 (sets of) schemas corresponding to the three mentioned layers.

[9] *E.g. In that case, Article 43 does not apply.*
[10] www.lipn.univ-paris13.fr/~fl/CLAL.
[11] www.oxygenxml.com

```
<xs:element name="PERMISSION">
  <xs:complexType >
    <xs:complexContent>
      <xs:extension base="leg:fragment_base_type">
        <xs:attribute name="IDENTIFIER" use="required" type="leg:fragment_identifier"/>
        <xs:attribute name="rel" type="leg:fragment_identifier_list"/>
        <xs:attribute name="except" type="leg:fragment_identifier_list"/>
        <xs:attribute name="bearer" use="required" type="xs:IDREFS"/>
      </xs:extension>
    </xs:complexContent>
  </xs:complexType>
</xs:element>
```

Figure 2. Definition of the PERMISSION element in the semantic schema. The type `fragment_base_type` specifies the elements allowed in the content of a fragment (namely entities and sub-fragments). `IDENTIFIER` and `bearer` are required attributes. Most attributes may have lists of identifiers as values.

```
<leg:PERMISSION IDENTIFIER="012.001.003" except="012.001.002" bearer="p_CONT">When
requested by the <leg:PERSON ref="p_DS">data subject</leg:PERSON>, the information
may be provided orally, provided that the identity of the <leg:PERSON
ref="p_DS">data subject</leg:PERSON> is proven by other means.</leg:PERMISSION>
```

Figure 3. Example of a PERMISSION fragment annotation: the identifier `012.001.003` indicates the fragment localisation (Art 12, §1); the `bearer` is a person identified as `p_CONT`; the optional attribute `except` indicates that this permission is an exception to the previous fragment (`012.001.002`). The bearer is not mentioned in the fragment text content, which only includes two mentions of another person `p_DS` whose role is not specified.

The CLAL XML schema has been defined so as to control the annotation, thus to guide and ease the annotators' work: at a given point in the text, only the elements/attributes authorized for insertion should be accessible in the annotation tool and the mandatory ones should be indicated. This also allows a strict validation of the annotation syntax. Among others, there are constraints for the verification of the existence and uniqueness of identifiers, and the differentiation of identifier types. Fragment signatures indicate which attributes (roles and semantic relations) are allowed and required.

Figures 2 and 3 show how the element PERMISSION is defined in the CLAL schema and used for annotating a text fragment.

4.2. GDPR annotation

31 of the 99 articles of the French version of the GDRP have been annotated in CLAL by two annotators, following the methodology described in section 2. The annotated corpus is made available as a gold standard. Figure 4 gives an example of combined structural and semantic annotation. Table 2 gives an overview of the resulting annotation.

The stability of the annotation from one annotator to another and from one period to another for a given annotator is a good indicator of its quality. To measure the quality of the CLAL annotation of the GDPR, we performed an additional experiment consisting in having the two trained CLAL annotators who initially developed the annotation guidelines annotate a new part of the RGPD after a break of several months after the design of the annotation language and the annotation of the gold standard. Concretely, they annotated the same 8 additional articles on their own, and we measured the inter-annotator agreement by comparing their two annotations. They agreed on 90% of the fragments

```
</ARTICLE>
<ARTICLE IDENTIFIER="041">
<TI.ART>Article 41</TI.ART><STI.ART>Monitoring of approved codes of conduct</STI.ART>
  <PARAG IDENTIFIER="041.001"><NO.PARAG>1.</NO.PARAG>
      <ALINEA><leg:POWER IDENTIFIER="041.001.001" type="execution" bearer="le_SA">
     Without prejudice to the tasks and powers of the competent <leg:LEGAL_ENTITY
     ref="le_SA">supervisory authority</leg:LEGAL_ENTITY> ..., the monitoring of
     compliance with a code of conduct...may be carried out by a <leg:LEGAL_ENTITY
     ref="le_BOD">body</leg:LEGAL_ENTITY> which has an appropriate level of
     expertise ... and is accredited for that purpose by the competent
     <leg:LEGAL_ENTITY ref="le_SA">supervisory authority</leg:LEGAL_ENTITY>
     </leg:POWER></ALINEA>
  </PARAG>
...
  <PARAG IDENTIFIER="041.006"><NO.PARAG>6.</NO.PARAG>
      <ALINEA><leg:EXCEPTION IDENTIFIER="041.006.001" except="041">This Article
     shall not apply to processing carried out by public authorities and bodies.
      </leg:EXCEPTION></ALINEA>
  </PARAG>
</ARTICLE>
```

Figure 4. Example of annotation: Article 41 of the GDPR includes 6 different paragraphs, the first one being a POWER and the last one introducing an exception to the article itself. The structural and semantic layers (annotations resp. with no and `leg:` prefix) are intertwined.

Table 2. Distribution of the most frequent types of the XML elements and attributes in the annotated part of the GDPR. Mind that this distribution is probably not representative of the full regulation annotation.

Elements (709)						Attributes (890)	
Fragments (178)		Entities mentions (515)		Sub-fragments (14)			
OBLIGATION	60	PERSON	290	EXCEPT	14	ref	515
LEGAL_PRECISION	31	LEGAL_ENTITY	217			bearer	136
POWER	23	CONCEPT	8			rel	34
QUALITY_ATTRIBUTION	17					except	18

(segmentation and typing), 94% of the roles and 60% of the semantic relations.[12] The good scores for fragments and roles give credibility to the proposed approach based on interpretation-neutrality and consensus. Unsurprisingly, the agreement is lower on `rel` attributes for which there is more annotation flexibility: the guidelines will probably have to be further specified. Annotators also reported that annotating these 8 articles took them 3 and 5 hours respectively, giving an average of 0.5 hour per article and approximately 50 hours for the entire GDPR.[13] These first figures show that large portions of text can be annotated quickly, without the need for legal experts.[14] This suggests that the proposed approach opens the way to a large-scale semantic search for legal texts.

[12]There is too few sub-fragments to give a reliable agreement measure for this category of elements. The figures for the annotation of entities are more difficult to interpret, due to the freedom left by the guidelines.

[13]This is certainly a very high estimate as the annotation task could be alleviated by using more user-friendly annotation tools and may be partly automated.

[14]The annotators should simply be familiar with the reading of legal documents.

4.3. Semantic search on GDPR

To show the benefit of semantic annotation for search, we designed a small experimental search engine based on an SQL-like querying language combining semantic and plain-text criteria and we tested our set of test questions.[15] We illustrate the semantic search on 3 of these user questions in an informal way.

Q1. What are the rights of the data subject? The translation of the question in a semantic query is straightforward (*Which fragments are annotated as* `RIGHT` *which* `bearer` *role is filled by the identifier of the data subject?*[16]). It returns 6 fragments.[17] On the other side, a search for sentences that contain the strings "right" and "data subject" provides the same fragments plus 17 additional ones which are noisy answers, except for one. The semantic search therefore appears to be more precise than full text search, which is an advantage for legal practitioners who have to browse large quantities of legal sources. It is always possible to broaden the search with a full text search to ensure one is not missing any information, at the cost of an important additional effort to analyse the results.

Q2. What are the obligations of a data controller? The question seems to translate directly into a semantic query: *Which fragments are annotated as* `OBLIGATION` *with the* `bearer` *role filled by the identifier of the data controller?* However, rights also express obligations in some cases, which leads to a second query: *Which fragments are annotated as* `RIGHT` *with a* `target` *role filled by the identifier of the data controller?*[18] This double query returns 26 obligations and 6 rights. In comparison, a plain text search (*Which sentences contain both "data controller" and "obligation—obligatory" keywords?*) gives 13 fragments among which only one right and one obligation are relevant. This is due to the many ways to express obligations in the text and the absence of constraint on the controller's role in plain text search.

Q3. What are the obligations of the controller in case of data breach? This question shows how semantic categories and strings can be combined in the same query and how fragment relationships, especially for exceptions, can be exploited. *Q*3 is similar to *Q*2 with an additional condition (*in case of data breach*) that does not translate into a semantic restriction because the term "data breach" is not marked-up as a concept. However, *Q*3 can be translated into a hybrid query combining semantic criteria and keywords (*Which fragments are annotated as* `OBLIGATION` *with the* `bearer` *role filled by the data controller identifier and containing the string "data breach"?*). This query directly returns a single fragment of Art. 34 §1[19] but this fragment happens to be related to two exceptions that are relevant for answering *Q*3. The first exception appears in the close context of the `OBLIGATION` statement (in the same article, two paragraphs apart) but the second one (in Article 23) would be difficult to spot for the user if it were not explicitly marked in the annotation.

[15] Note that semantic criteria can only be matched with the annotated part of the GDPR.

[16] `SELECT fragments WHERE name ~ "leg:RIGHT", bearer ~ "p_DS"`. The formalization of the two other queries is omitted due to space limitations.

[17] Such as *the right to obtain from the controller, without undue delay the rectification of inaccurate personal data concerning him or her.*

[18] The rights of third parties towards the controller are as many obligations imposed on the latter.

[19] *When the personal data breach is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall communicate the personal data breach to the data subject without undue delay.*

5. Previous works

It has long been understood that *enriching legal sources with metadata* is essential to make them more accessible. XML vocabularies [10] have been designed to account for document information, the structure of the legal sources, and cross-references (*e.g.* case law, citations, modification). Much effort has also been devoted to the semantic enrichment of legal documents to facilitate legal reasoning. This medatada ranges from tags associated to key textual elements (*e.g.* actors, dates) to rule annotations [11], possibly associating formal descriptions to textual fragments and allowing for associating multiple interpretations. Nevertheless, bridging between the natural language of the legal sources and the logical or rule formalizations remains an open challenge despite efforts to define a methodology for deriving formal rules from texts, relying on controlled or semi-formal languages, or providing interfaces to support human formalization [12,13].

The modeling of the semantic information has also been the subject of numerous works, especially for the design of legal ontologies, core legal ontologies [14] or domain ontologies, some of which focusing on privacy and the specific issues of the GDPR (*e.g.* ODRL [2], PrivOnto [15], PrONTO [3]). Another approach aims to develop machine-readable languages for formally representing legal rules, such as LegalRuleML [16], but with few examples from source texts.

A sound methodology has been developed over time for manual annotation of corpora [9]. The issues have mainly concerned the annotation format, the tagset choice, the underlying theory, the training of annotators, the complexity/cost of an annotation task, and the quality of the annotation (inter- and intra-annotator agreement).

It is well known that legal information retrieval has to meet specific requirements, such as the size and interdependence of legal sources or the needs for retrieval completeness [17]. These requirements favor semantic or hybrid approaches, which the present work shows the potential of.

6. Conclusion and future work

This paper advocates a new approach of legal text mining relying on semantic technologies, which represents an alternative and middle ground to the traditional statistical-based information retrieval methods – which are applicable on a large scale and have a good recall but low precision – and those aiming at formalizing the content of rules – which are quite ambitious but difficult to implement due to the plurality of interpretations and the complexity of translating natural language rules into logic.

Considering that the text (be it legislation, case law, decisions, contracts…) is the reference for any legal work, we propose to keep the text at the center of attention while enriching it with a semantic annotation layer, thus enabling access using semantic search services. Our approach of regulatory texts is based on a coarse-grained, interpretation-neutral annotation that nevertheless semantically enriches the text. The approach is illustrated by a proof of concept experiment of annotating the French version of the GDPR and searching over the annotations. The paper presents the methodology followed for the design of the annotation language, the language itself as well as its implementation in XML, the annotated GDPR that might serve as a gold standard for training automatic annotation tools, and finally queries that illustrate the benefit of annotation for semantic

search. The resources provided also include an annotation guide that allows the experiment to be extended to other texts. In the end, this GDPR experiment shows the added value of annotation for legal text mining and exploration.

Further work is required. To evaluate the robustness of the proposed language and annotation approach, the experiment must be extended to new annotators, to other versions of the GDPR, and to other legal texts. While it may be necessary to revise the annotation language, the gold standard, and the annotation guide, the balance between annotation granularity, cost, and reliability will have to be maintained. It would be interesting to use the manually annotated corpora as training data in a machine learning study. Finally, with regard to users, we must refine the requirements analysis and provide easy-to-use tools for querying the GDPR and annotating new texts.

References

[1] Stamper R. LEGOL: Modelling Legal Rules by Computer. Computer Science and Law. 1980:45–71.

[2] Group WOC. Open Digital Rights Language: Vocabulary & Expression 2.2. W3C Recommendation; 2018. Available from: www.w3.org/TR/odrl-vocab/.

[3] Palmirani M, Martoni M, Rossi A, Bartolini C, Robaldo L. PrOnto: Privacy Ontology for Legal Reasoning. In: Kő A, Francesconi E, editors. Electronic Government and the Information Systems Perspective. Cham: Springer International Publishing; 2018. p. 139–152.

[4] Palmirani M, Governatori G. Modelling Legal Knowledge for GDPR Compliance Checking. In: Palmirani M, editor. Proc. of the 31st JURIX. vol. 313 of Frontiers in A.I. and Applications. IOS Press; 2018. p. 101–110.

[5] Libal T. A Meta-level Annotation Language for Legal Texts and Argumentation. In: Dastani M, Dong H, van der Torre L, editors. Logic and Argumentation. CLAR 2020. Cham: Springer; 2020. p. 131–150.

[6] Robaldo L, Bartolini C, Lenzini G. The DAPRECO Knowledge Base: Representing the GDPR in LegalRuleML. In: Calzolari N, et al., editors. Proc. of the 12th LREC. Marseille: ELRA; 2020. p. 5688–5697.

[7] Maxwell KT, Schafer B. Concept and Context in Legal Information Retrieval. In: Oric. of the 21st JURIX. IOS Press; 2008. p. 63–72.

[8] Bartolini C, Muthuri R. Reconciling Data Protection Rights and Obligations: An Ontology of the Forthcoming EU Regulation. In: Language and Semantic Technology for Legal Domain; 2015. p. 8.

[9] Fort K. Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects. Wiley-ISTE; 2016.

[10] Barabucci G, Cervone L, Palmirani M, Peroni S, Vitali F. Multi-layer Markup and Ontological Structures in Akoma Ntoso. In: Casanovas P, Pagallo U, Sartor G, Ajani G, editors. AI Approaches to the Complexity of Legal Systems. Berlin, Heidelberg: Springer; 2010. p. 133–149.

[11] Wyner AZ, Peters W. On Rule Extraction from Regulations. In: Atkinson K, editor. Proc. of the 24th JURIX. vol. 235 of Frontiers in A.I. and Applications. Vienna: IOS Press; 2011. p. 113–122.

[12] Nazarenko A, Lévy F, Wyner A. Towards a Methodology for Formalizing Legal Texts in LegalRuleML. In: Bex F, Villata S, editors. Proc. of the 29th JURIX. vol. 294 of Frontiers in A.I. and Applications. IOS Press; 2016. p. 149–154.

[13] Libal T, Steen A. NAI: The Normative Reasoner. In: Proc. of the 7th ICAIL. New York: ACM; 2019. p. 262–263.

[14] Hoekstra R, Breuker J, Di Bello M, Boer A. The LKIF Core ontology of basic legal concepts. International Journal of High Performance Computing Applications (IJHPCA). 2007 01:43–63.

[15] Oltramari A, Piraviperumal D, Schaub F, Wilson S, Cherivirala S, Norton TB, et al. PrivOnto: A semantic framework for the analysis of privacy policies. Semantic Web. 2018;9:185–203.

[16] Athan T, Governatori G, Palmirani M, Paschke A, Wyner AZ. LegalRuleML: Design Principles and Foundations. In: Faber W, Paschke A, editors. Reasoning Web. Web Logic Rules - 11th International Summer School , Tutorial Lectures. Berlin: Springer; 2015. p. 151–188.

[17] Van Opijnen M, Santos C. On the Concept of Relevance in Legal Information Retrieval. Artif Intell Law. 2017 Mar;25(1):65–87.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210314

Accounting for Sentence Position and Legal Domain Sentence Embedding in Learning to Classify Case Sentences

Huihui Xu [a,1], Jaromir Savelka [b,2] and Kevin D. Ashley [a,c,d,3]
[a] *Intelligent Systems Program, University of Pittsburgh*
[b] *School of Computer Science, Carnegie Mellon University*
[c] *Learning Research and Development Center, University of Pittsburgh*
[d] *School of Law, University of Pittsburgh*

Abstract. In this paper, we treat sentence annotation as a classification task. We employ sequence-to-sequence models to take sentence position information into account in identifying case law sentences as issues, conclusions, or reasons. We also compare the legal domain specific sentence embedding with other general purpose sentence embeddings to gauge the effect of legal domain knowledge, captured during pre-training, on text classification. We deployed the models on both summaries and full-text decisions. We found that the sentence position information is especially useful for full-text sentence classification. We also verified that legal domain specific sentence embeddings perform better, and that meta-sentence embedding can further enhance performance when sentence position information is included.

Keywords. Information retrieval; Natural language processing; Annotation; Embedding

1. Introduction

As an initial step toward automatically generating comprehensible legal summaries, we have been exploring machine learning (ML) methods for classifying sentences of legal cases in terms of issues a court addresses, its conclusions of those issues, and its reasons for so concluding (IRCs). In previous work, we have experimented with different models—both traditional machine learning and deep learning—to identify these types of sentences in both summaries and full texts. While we demonstrated that those models can identify IRC types of sentences to some extent, the task remains challenging for machine encoding.

In this paper, we employ supervised ML based on a larger annotated dataset, 1049 pairs of full text cases and summaries in which sentences have been manually annotated in terms of IRCs. We also explore if two new techniques, sentence embeddings pretrained on large quantities of legal texts and taking account of sentence order, help machine annotation of legal cases.

[1] huihui.xu@pitt.edu
[2] jsavelka@cs.cmu.edu
[3] ashley@pitt.edu

In attempting to leverage the power of state-of-the-art sentence embeddings, pre-trained on legal texts, we hypothesize that the broader contextual information associated with the sentence embeddings will improve performance.

We also hypothesize that taking sentence ordering information into account will improve the classifier's performance. In regular meetings with our two third-year law student annotators to resolve differences concerning annotations, we noticed that they tended to rely on ordering information to mark up certain types of sentences. For example, annotators would look for conclusions following issues or at the end of a case. We wondered if ML could also employ such position information.

1.1. Extracting Issues, Reasons, and Conclusions

The ultimate goal of our work is to enable an intelligent system to help end users assess a case's potential relevance by effectively and efficiently conveying some important substantive information about the case. Human-prepared legal summaries are available through various on-line legal service providers. For example, the CanLII Connects website[4] of the non-profit Canadian Legal Information Institute,[5] features summaries of legal decisions prepared by members of Canadian legal societies.

Based on the experience of CanLII Connects, summaries as short as three sentences could be even more effective in a legal IR interface. This raises a practical question: "What can a three-sentence case summary provide?". Legal argument triples, IRCs, may be the answer. Issues, reasons, and conclusions form the skeleton of case briefs, a legal writing technique for summarizing cases that has long been taught in American law schools. Thus, the potential utility of summarizing cases in terms of issues, conclusions, and reasons seems clear.

Based on our annotation experience, the human-prepared CanLII summaries regularly include issues raised by the courts, the conclusions reached, and reasons connecting them. Those summaries also include some procedural information, descriptions of facts, statements of legal rules, case citations and explanations, and other information. Since the expert legal summarizers act as an intelligent and well-informed filter on importance, it made sense to leverage their expertise by annotating their summaries rather than the full texts. CanLII has provided 28,733 paired cases and human-prepared summaries for purposes of this research. The cases cover a variety of kinds of legal claims and issues presented before Canadian courts.

1.2. Hypotheses

We try to answer two long-standing questions in the Artificial intelligence and Law field: first, whether legal language is so unique that the legal pre-trained models would assist downstream legal natural language processing tasks and which tasks; second, whether sentence position information helps a model as it appears to help human annotators.

We investigate how well the classification models perform based on sentence embeddings, and annotate full texts of cases and summaries in terms of issues, reasons, and conclusions. As noted, we examine two hypotheses in this paper:

[4] https://canliiconnects.org/en
[5] https://www.canlii.org/en/

(1) A model would perform better when incorporating sentence position information.
(2) A model would perform better when incorporating specific legal domain knowledge.

2. Related Work

2.1. Word and Sentence Embeddings

Word embeddings, dense vector representations trained with neural language models, capture some linguistic relationships between words and assist with various natural language processing tasks. See, e.g., [1]. Researchers further explored word meta-embeddings for operations such as concatenation, SVD, and 1toN [2]. Experiments in [3] proved that averaging different sources of word embeddings has similar effects as concatenating those embeddings. Researchers in [4] used three types of autoencoders to learn meta-embeddings of words.

Similarly, sentence embedding is the dense vector representation of a sentence. Sentence embedding provides information about larger contexts of words. [5] introduced Sentence-BERT in 2019; they used siamese and triplet network structures to derive fixed-sized 768 dimensional vector representations for input sentences. Google Research developed the Universal Sentence Encoder in 2018 [6]. The encoder has two model architectures: one based on transformer architecture and the other on Deep Averaging Networks (DAN). Both transfer input sentences into fixed 512 dimensional sentence embeddings. Both Sentence-BERT and Universal Sentence Encoder are state-of-the-art sentence embeddings.

In the legal domain, words may have different semantic meanings than in other domains. For example, 'sentence' means the judgment that a court formally pronounces after finding a criminal defendant guilty.[6] In order to address this, we employed Legal-BERT, a BERT model trained on legal domain sentences [7]. Legal-BERT was pre-trained on the entire Harvard Law case corpus from 1965 to present, comprising 3,446,187 legal decisions across all federal and state courts [7]. [7]

2.2. Argument Mining and Summarization

Extracting propositions, premises, conclusions, and nested argument structures [8,9] is an active research topic in the legal argument mining field. Rhetorical and other roles that sentences play in legal arguments have been employed for legal argument mining [10]. Citing information and fact patterns [11,12] that effect the strength of a side's claim in special legal domains are also being explored. Segmenting legal text by functions [13,14], and by topic [15] or by linguistic analysis [16,17,18] are some initial steps for dissecting a legal document.

Researchers have applied legal argument mining to the task of summarizing legal cases. In [19], the authors propose an unsupervised algorithm that incorporates legal domain knowledge, such as rhetorical roles sentences play in a legal document. [20] have summarized Japanese judgments in terms of issues, conclusions, and framings. Our legal

[6] https://www.law.cornell.edu/wex/sentence
[7] The pre-trained Legal-BERT model can be found here: https://huggingface.co/zlucia/legalbert

argument triples have a similar structure but are more understandable types than those tailored to Indian or Japanese legal judgements. In addition, in our work a set of case summaries prepared by legal experts is used to extract argument triples from the full case texts.

3. Dataset

Our type system for labeling sentences in legal cases comprises:

1. **Issue** – Legal question which a court addressed in the case.
2. **Conclusion** – Court's decision for the corresponding issue.
3. **Reason** – Sentences that elaborate on why the court reached the Conclusion.

We treat all non-annotated sentences as non-IRC sentences.

Two hired third-year law school students annotated sentences from the human-prepared summaries to identify and annotate the issues, reasons, and conclusions. Both students have annotated 1049 randomly selected pairs from the 28,733 case/summary pairs available. The total number of sentences from the corresponding full texts is 215,080, which is significantly more than the corresponding summaries' 11,496 sentences.

Both annotators followed an 8-page detailed Annotation Guide prepared by the third author, a law professor, in order to mark-up instances of IRC sentence types in both the summaries and full texts of cases. The annotators worked on successive batches of summaries using the Gloss annotation environment developed by the second author. After annotating each batch, the annotators resolved any annotation differences in regular Zoom meetings attended by the first and third authors.

The procedure for annotating the full texts of cases differs from annotating the summaries. The Annotation Guide instructs annotators to search the full text of the case for those sentences that are most similar to the annotated summary sentences and to assign them the same labels (i.e., Issue, Conclusion, or Reason) as in the summaries. Annotators may pick terms or phrases from the annotated summary sentences as anchors to search for corresponding sentences in the full texts. Annotators do not need to read the full text of the case if they find the corresponding sentences. The Guide warns that there may not be an exact correspondence between the annotated sentences in the summary and those in the full text of the case. This is fairly common, because human summarizers tend to edit selected sentences in the full case texts. For example, a human summarizer may combine some shorter sentences into a longer one.

By using the summaries' annotations as anchors to target corresponding sentences in the full text, we attempted to leverage the summarizers' work in selecting important sentences and the annotators' work in marking up some of those full texts sentences as issues, conclusions, or reasons. We developed this strategy to expedite the full text annotation process, since it would be much more time-consuming and costly if annotators had to read the full texts of cases. The strategy is based on the observation that sentences of summaries stem from those in the full texts. The strategy also helps us to confirm the mapping relationship between summaries and full texts, which is a step towards generating summaries automatically.

Cohen's κ [21] is used to measure the degree of agreement between two annotators after their independent annotations. The mean of Cohen's κ coefficients across all

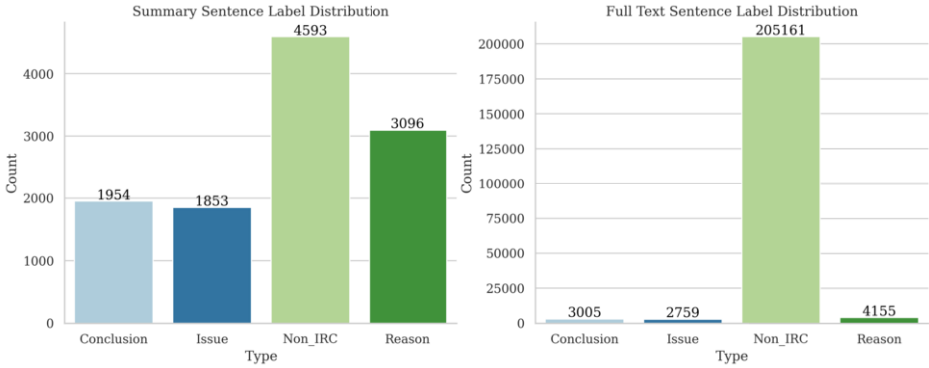


Figure 1. Distribution of annotated IRC type sentences in 1049 summaries (left) and full texts (right).

Table 1. Descriptive statistics of the resulting dataset. We report the basic descriptive statistics of each type in both summaries and full texts. The lengths of the summary and the full texts are also included in the table.

	Summary			Full text		
	Min.	Max.	Mean	Min.	Max.	Mean
Issue	3 tokens	140 tokens	27.96 tokens	3 tokens	427 tokens	36.80 tokens
Reason	3 tokens	257 tokens	26.61 tokens	3 tokens	229 tokens	31.36 tokens
Conclusion	2 tokens	289 tokens	20.40 tokens	2 tokens	314 tokens	28.55 tokens
Length of text	1 sents	90 sents	10.96 sents	9 sents	2411 sents	205.03 sents

types for summaries is 0.734, and the mean for full texts is 0.602. According to [22], both scores indicate substantial agreement between annotators about the sentence type. For the summary annotation, the mean of Reason agreement is the lowest among those three types. Annotating Reasons is more challenging since they are entwined with case facts. The agreement scores of full texts are lower than the summaries' scores, since sentences from summaries and full texts are not in a one-to-one mapping. This increases the difficulty of full text annotation.

Figure 1 reports the distributions of final consensus labels from summaries and full texts. The most frequent label is the non-IRC label for both summaries and full texts. The second most frequent label is the Reason label for both summaries and full texts. The label distribution is aligned with our observation: Reasons tend to be more elaborated than Issues and Conclusions.

The descriptive statistics of the processed dataset are shown in Table 1. The average number of sentences in a full text is 205.03, while the range of the full text length is quite large. Comparatively, the average number of sentences in summaries is 10.96 which, as expected, is much shorter than full texts. We also observe that the average length of Issues is the highest in both summaries and full texts.

4. Experiment

4.1. Models

We use Sentence-BERT [5], Universal Sentence Encoder(USE) [6], and Legal-BERT [7] to encode sentences from summaries and full texts into a semantic space. Each sentence

then becomes a fixed sized vector. Each document is comprised of a series of converted sentence vectors.

Sentence-BERT uses two BERT models with tied weights and adds a pooling operation to the output to derive fixed sized sentence embedding. We chose the 'all-mpnet-base-v2' model trained on a dataset of over 1 billion pairs.[8] This model encodes sentences into 768-dimensional vectors and has achieved competitive performance over different datasets. USE takes a tokenized string and outputs a fixed 512-dimensional vector as sentence embedding.[9] Legal-BERT was trained on the entire Harvard Law case corpus. In order to derive the fixed sized sentence embedding, we simply keep the output of the last pooling layer of this model as the sentence embedding. The dimension of the Legal-BERT sentence embedding is also 768.

The Long Short-Term (LSTM) neural network [23], a variant of a recurrent neural network (RNN), can deal with arbitrary lengths of input. A traditional RNN does not perform well on long sequences due to the problem of vanishing gradients. LSTM tackles the problem by incorporating different gates. Bidirectional LSTM consists of two separate LSTMs: one takes an input from right to left; the other one from left to right.

We also examined the effect of one of the meta-sentence embedding techniques. Averaging is one of the commonly used meta-embedding techniques. It simply requires averaging different sources of embeddings. According to [24], averaging has similar performance to concatenation while taking less time and resources in terms of meta-word embedding. We extend this idea to the sentence embedding. We construct two types of meta-sentence embeddings: Legal-BERT + USE and Legal-BERT + Sentence-BERT. Both types of meta-sentence embeddings are 768-dimensional.

4.2. Experimental Design

This work employs two designs which differ as to the way in which the sentence embeddings are fed into the bidirectional LSTM model: 1) Single time step is associated with a sentence, and no sentence position information is provided. 2) Fixed sized document matrices are input into the model; each time step is associated with a sentence, where sentence position information is provided. We refer to [14] for the padding procedure. For example, since the maximum length of full texts is 2411, we transferred each full text document into a 2411×768 matrix when using Sentence-BERT embedding. The shorter case will be padded to the maximum length. In this paper, we chose pre-padding over post-padding since [25] demonstrates that pre-padding for LSTM performs substantially better than post-padding. Figure 2 shows the structure of the models and the main difference between the two designs. Our rolled LSTM reads one sentence at each time step. The returning arrow (left) represents multiple time steps for "with position information". "Without position information" (right) involves only a single time step.

We split the dataset into training, validation, and test sets. The training set comprises 70% of 1049 cases; the validation and test sets each have 15% of the cases. The data is fed into the bidirectional LSTM model with 256 units and a dropout rate of 0.2. Categorical cross-entropy loss function and Adam optimizer are used for optimizing the model. The initial learning rate is set to $1e^{-3}$ and reduced at factor 0.1 if the validation loss has stopped decreasing with a patience of 20. The training procedure will be stopped when

[8] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[9] https://tfhub.dev/google/universal-sentence-encoder/4

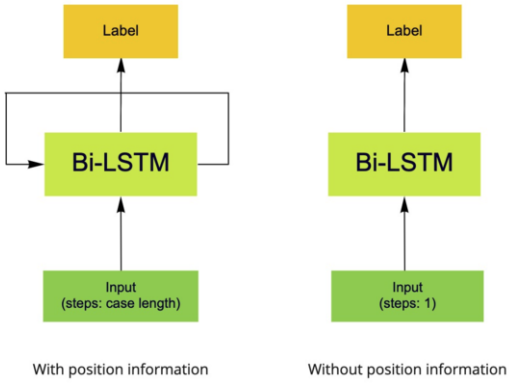


Figure 2. Main difference of two experiment designs: with and without sentence position information.

the validation accuracy has not been increased in 20 epochs. Validation accuracy is used to select the best model.

5. Results and Discussion

The results of the two experimental designs are shown in Table 2. The first 5 rows of the table show how well the model performs on the summary and the full text without position information; the next 5 rows show the performance of the model when incorporating sentence position information. All the numbers are reported as F_1 scores.

5.1. Without Position Information vs. With Position Information

For summaries, the model performs better on identifying Reasons with position information no matter which sentence embedding is used. However, Legal-BERT and Legal-BERT + USE sentence embeddings with position information do not have better performance in terms of Issue and Conclusion classification. The average F_1 scores of all sentence embeddings are higher when the model digests sentence position information at the same time.

For the full text sentence classification, the pattern is much clearer: Issue and Reason can be more easily identified by the model when including sentence position information. Sentence-BERT and Legal-BERT embeddings do not perform better with position information in terms of Conclusion classification. The average F_1 scores are better when the model is fed with sentence position information.

5.2. Domain Specific Sentence Embedding vs. General Purpose Sentence Embedding

Legal-BERT sentence embedding achieves the best performance on Issue, Reason and Conclusion on both summary and full text, if the model was not fed sentence position information. Legal-BERT sentence embedding performs better than other types of sentence embeddings when the model takes sentence position information into account except on classifying Issues in summaries.

Sentence-BERT sentence embedding is the second best embedding on most of the classification tasks, while USE is the second best on full text Reason classification.

Table 2. Results of classification on summaries and full texts with and without position information. All the results are reported as F_1 scores.

	Summary (without position information)					Full text (without position information)				
	Issue	Reason	Conclusion	Non-IRC	Ave.	Issue	Reason	Conclusion	Non-IRC	Ave.
SBERT	0.69	0.62	0.68	0.62	0.66	0.17	0.04	0.44	0.97	0.40
Legal-BERT	0.74	0.68	0.73	0.67	0.70	0.30	0.13	0.49	0.98	0.47
USE	0.55	0.58	0.61	0.61	0.59	0.15	0.06	0.36	0.97	0.39
Legal-BERT+USE	0.73	0.68	0.74	0.68	0.71	0.25	0.14	0.48	0.98	0.46
Legal-BERT+SBERT	0.71	0.70	0.69	0.67	0.69	0.29	0.12	0.47	0.98	0.46
	Summary (with position information)					Full text (with position information)				
	Issue	Reason	Conclusion	Non-IRC	Ave.	Issue	Reason	Conclusion	Non-IRC	Ave.
SBERT	0.73	0.69	0.69	0.65	0.69	0.30	0.06	0.40	0.97	0.43
Legal-BERT	0.69	0.75	0.75	0.66	0.71	0.36	0.14	0.47	0.98	0.49
USE	0.67	0.65	0.64	0.60	0.64	0.31	0.08	0.36	0.97	0.43
Legal-BERT+USE	0.71	0.72	0.72	0.67	0.71	0.41	0.18	0.49	0.98	0.51
Legal-BERT+SBERT	0.76	0.72	0.72	0.70	0.72	0.38	0.20	0.49	0.98	0.51

5.3. Meta-Sentence Embedding vs. Singular Sentence Embedding

For summaries, Legal-BERT + USE and Legal-BERT + SBERT improve model performance on Issue identification with position information. Those two sentence embeddings have tied or better performance on Reasons without position information.

For full texts, meta-sentence embedding substantially improves the performance on Issue, Reason and Conclusion when position information is included. Without position information, the meta-sentence embedding does not show improved performance.

Generally speaking, meta-sentence embeddings coupled with position information show higher increases in performance on full texts than on summaries.

5.4. Error Analysis and Limitations

We present a brief error analysis for comparing the errors between including position information and not including position information when using Legal-BERT sentence embedding. In the test set, the model has $F_1 = 0.30$ on Issues for the full texts without position information as opposed to $F_1 = 0.36$ when including position information. We read some examples that both experimental methods get right, and some instances that only the model fed with position information correctly classified. We noticed that without position information the model tends to select only Issue sentences with certain sentence structure, like "This is an appeal...". With position information, the model would be able to pick up Issue sentences relying less on sentence structure. For example, "The charge arises out of... " has an implicit semantic cue regarding the type of the sentence. The position information provides additional information to help the model to make the correct classification.

We verified the importance of position information in terms of the model classification performance on full texts. However, this pattern does not apply to some types of sentences in summaries, like Issue sentences. It seems like the position information in

summaries is not as reliable as in full texts. We found that the model ignores some Issue instances that appear in the middle of the summaries.

Compared to our prior work [26], the F_1 scores across all types decrease substantially. In [26], we obtained F_1 scores of 0.58, 0.15, and 0.53 on Issue, Reason and Conclusions, respectively. Our expectation that the performance would improve after training on more data was not confirmed. Several reasons could contribute to this result: first, the initial learning rates are different; this will lead to different performance. Second, noisy data also increase along with the increase of data.

6. Conclusion and future work

We analyzed the effect of sentence position information and legal domain specific sentence embedding in a task of labelling case sentences in terms of legal argument triples. We found that the sentence position information does assist the model to perform better, especially for full texts. We also verified that legal domain specific sentence embedding performed better on this legally intensive task than the other general purpose sentence embeddings. Meta-sentence embedding that inherits benefits from general purpose sentence embedding and legal sentence embedding can outperform its components when the position information is incorporated. The result suggests a promising path to annotate legal documents automatically. This is also a step towards automatically generating succinct legal summaries since the model can identify the important sentences.

This work is subject to certain limitations as well. As mentioned before, paradoxically, the overall performance on full texts tended to decrease with the larger training set. For future work, we will explore a more effective model to improve the performance, such as by introducing additional linguistic features and their semantic values.

Acknowledgement

This work has been supported by grants from the Autonomy through Cyberjustice Technologies Research Partnership at the University of Montreal Cyberjustice Laboratory and the National Science Foundation, grant no. 2040490, FAI: Using AI to Increase Fairness by Improving Access to Justice. The Canadian Legal Information Institute provided the corpus of paired legal cases and summaries. Computation resources are provided by the Center for Research Computing at the University of Pittsburgh.

References

[1] Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning. In: Proc. 48th Ann. Mtg. of the Association for Computational Linguistics; 2010. p. 384-94.

[2] Yin W, Schütze H. Learning word meta-embeddings. In: Proc. 54th Ann. Mtg. of the Association for Computational Linguistics (Volume 1: Long Papers); 2016. p. 1351-60.

[3] Coates J, Bollegala D. Frustratingly Easy Meta-Embedding – Computing Meta-Embeddings by Averaging Source Word Embeddings. In: Proc. 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, V. 2. New Orleans, Louisiana: Association for Computational Linguistics; 2018. p. 194-8. Available from: https://aclanthology.org/N18-2031.

[4] Bollegala D, Bao C. Learning word meta-embeddings by autoencoding. In: Proc. 27th Int'l Conf. on Computational Linguistics; 2018. p. 1650-61.
[5] Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:190810084. 2019.
[6] Cer D, Yang Y, Kong Sy, Hua N, Limtiaco N, John RS, et al. Universal sentence encoder. arXiv preprint arXiv:180311175. 2018.
[7] Zheng L, Guha N, Anderson BR, Henderson P, Ho DE. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. arXiv preprint arXiv:210408671. 2021.
[8] Feng V, Hirst G. Classifying arguments by scheme. In: Proc. 49th Ann. Mtg of the Association for Computational Linguistics: Human language technologies; 2011. p. 987-96.
[9] Mochales R, Moens M. Argumentation mining. Artificial Intelligence and Law. 2011;19(1):1-22.
[10] Saravanan M, Ravindran B. Identification of rhetorical roles for segmentation and summarization of a legal judgment. Artificial Intelligence and Law. 2010;18(1):45-76.
[11] Bansal A, Bu Z, Mishra B, Wang S, Ashley K, Grabmair M. Bex F, editor. Document Ranking with Citation Information and Oversampling Sentence Classification in the LUIMA Framework; 2016.
[12] Falakmasir M, Ashley K. Utilizing Vector Space Models for Identifying Legal Factors from Text. In: JURIX; 2017. p. 183-92.
[13] Savelka J, Ashley K. Segmenting U.S. Court Decisions into Functional and Issue Specific Parts. In: Proc. 31st Int. Conf. on Legal Knowledge and Information Systems, Jurix; 2018. p. 111-20.
[14] Savelka J, Westermann H, Benyekhlef K, Alexander CS, Grant JC, Amariles DR, et al. Lex rosetta: transfer of predictive models across languages, jurisdictions, and legal domains. In: Proc. 18th Int'l Conf. on Artificial Intelligence and Law; 2021. p. 129-38.
[15] Lu Q, Conrad J, Al-Kofahi K, Keenan W. Legal document clustering with built-in topic segmentation. In: Proc. 20th ACM int'l conf. Info. and knowledge management; 2011. p. 383-92.
[16] Grover C, Hachey B, Korycinski C. Summarising legal texts: Sentential tense and argumentative roles. In: Proc. HLT-NAACL 03 Text Summarization Workshop; 2003. p. 33-40.
[17] Farzindar A, Lapalme G. Legal text summarization by exploration of the thematic structure and argumentative roles. In: Text Summarization Branches Out; 2004. p. 27-34.
[18] Wyner A, Mochales-Palau R, Moens M, Milward D. Approaches to text mining arguments from legal cases. In: Semantic processing of legal texts. Springer; 2010. p. 60-79.
[19] Bhattacharya P, Poddar S, Rudra K, Ghosh K, Ghosh S. Incorporating domain knowledge for extractive summarization of legal case documents. In: Proc. 18th Int'l Conf. on Artificial Intelligence and Law; 2021. p. 22-31.
[20] Yamada H, Teufel S, Tokunaga T. Building a corpus of legal argumentation in Japanese judgement documents: towards structure-based summarisation. Art Int and Law. 2019;27(2):141-70.
[21] Cohen J. A coefficient of agreement for nominal scales. Educational and psychological measurement. 1960;20(1):37-46.
[22] Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics. 1977:159-74.
[23] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735-80.
[24] Coates J, Bollegala D. Frustratingly Easy Meta-Embedding–Computing Meta-Embeddings by Averaging Source Word Embeddings. arXiv preprint arXiv:180405262. 2018.
[25] Dwarampudi M, Reddy N. Effects of padding on LSTMs and CNNs. arXiv preprint arXiv:190307288. 2019.
[26] Xu H, Savelka J, Ashley KD. Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. In: Proc. 18th Int'l Conf. on Artificial Intelligence and Law; 2021. p. 250-4.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210315

Generation of Legal Norm Chains: Extracting the Most Relevant Norms from Court Rulings

Ingo GLASER [a,1], Sebastian MOSER [a] and Florian MATTHES [a]
[a] *Software Engineering for Business Information Systems, Department of Informatics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany*

Abstract. Various online databases exist to make judgments accessible in the digital age. Before a legal practitioner can utilize state-of-the-art information retrieval features to retrieve relevant court rulings, the textual document must be processed. More importantly, many verdicts lack crucial semantic information which can be utilized within the search process. One piece of information that is frequently missed, as the judge is not adding it during the publication process within the court, is the so-called norm chain. This list contains the most relevant norms for the underlying decision.

Therefore this paper investigates the feasibility of automatically extracting the most relevant norms of a court ruling. A dataset constituting over 42k labeled court rulings was used in order to train different classifiers. While our models provide F1 performances of up to 0.77, they can undoubtedly be utilized within the editorial publication process to provide helpful suggestions.

Keywords. natural legal language processing, norm chains, legal court rulings, multi-label classification

1. Introduction

Legal research constitutes a significant part of the daily work of a legal practitioner, particularly lawyers [1,2]. As the work of legal workers is not just knowledge-driven but also time-consuming, recent research activities and the industry try to support the legal research process. The focus here is often on legal information retrieval. That is why various online databases exist that provide convenient search functionalities. However, the path from a textual court ruling (in the remainder of this paper the terms ”verdict”, ”ruling”, and ”decision” all refer to the entire court ruling document), as it is created by a judge, into an online database is often tedious and involves a vast amount of human labor. This path includes typical processing tasks such as segmentation or information extraction and enrichment with semantic information. One type of such semantic information is the so-called norm chain.

[1]Corresponding Author: Ingo Glaser, Software Engineering for Business Information Systems, Department of Informatics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany; E-mail: ingo.glaser@tum.de

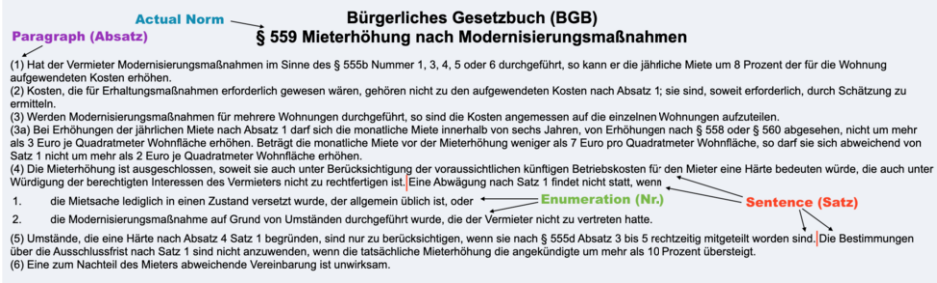


Figure 1. Annotated screenshot of BGB § 559 taken from www.gesetzte-im-internet.de

A norm chain is a series of legal norms that lead to the consequence of a verdict. Thus, with the help of legal norm chains, the practiced civil lawyers can save the time-consuming step of searching for the relevant and referenced legal norms. Therefore, a general definition of the legal norm chain is the following:

> The chain of norms is a combination of explicitly or implicitly referencing legal norms that extends from a legal consequence order to the lowest level of the facts.

Hereby, a legal norm is understood to be either a statutory regulation or a rule of a general abstract nature issued on a statutory basis or contained in the common law. Figure 1 provides an example of a norm from the German Civil Code (BGB). The norm would be referenced as "BGB § 559". The shown norm regulates the increase of rent after modernization measures. Depending on the context, a more granular reference can be made such as "BGB § 559 Absatz 3", which would refer to the paragraph after "(3)". Going one stop further, it is also possible to reference a precise sentence within a paragraph, e.g., "BGB § 559 Absatz 4, Satz 1". Last but not least, enumerations might be referenced as well, such as "BGB § 559 Absatz 4 Satz 2 Nr. 1". Now, a norm chain consists of one-to-many of such norm references of varying granularities. An example of such a norm chain would be "BGB § 535 Abs. 1 Satz 2, § 536 Abs. 1, § 536a Abs. 2 Nr. 1". This chain consists of three different norms. The first norm references a concrete sentence, the second norm constitutes a complete paragraph, and the final norm refers to a specific enumeration number. That example was taken from a verdict from the German Supreme Court (VIII ZR 271/17).

While particularly verdicts from higher courts, such as the German Supreme Court, usually contain the norm chain as the judge provides it, the vast majority of court rulings leave the court's internal publication process without it. That is why legal authors participating in the editorial process of a legal publisher are required to create the norm chain afterward.

Therefore, in this work, we want to investigate the feasibility of extracting relevant norms to automatically create norm chains utilizing natural language processing (NLP). The remainder of this paper is structured as follows: Section 2 describes the data set utilized in this work together with required pre-processing steps. The applied methods are discussed in Section 3. Detailed analysis and discussion of our results are provided in Section 4. Limitations of the presented work, along with a short overview of related work, are provided in Section 5 before Section 6 closes with a conclusion and outlook.

2. Data

For this research, we used a dataset of 42k German court rulings from various instance levels. The verdicts were given to us by a German legal publisher via an annotated XML format that contains the norm chain separated into its containing norms and the whole text of the verdict. Norm references within the text are annotated. Based on this, we extracted the text from the different verdict sections and segmented it into sentences via spaCy[2]. The norms from the norm chain were extracted, cross-referenced against a list[3] with all existing norms to validate the classification targets, and then cleaned (more precisely, special characters were stripped away, and additional information such as dates were removed). For each verdict, we also extracted all the referenced norms from the text sections utilizing the annotations, and regular expressions as some norms are not annotated. The referenced norms were then also validated against the list from *gesetze-im-internet.de*.

Based on this preprocessing process, our dataset contains 8,359 different, unique classification targets (e.g., *BGB 3*) extracted from a total of 111k norms from all norm chains with an average number of 2.6 norms per norm chain. When only considering the specific norm without any paragraph or section reference (e.g., *BGB*), there are 666 different targets for classification with on average 1.8 norms per norm chain. We will call this the reduced target set, and when referencing a specific norm from this set, we will call it the reduced norm. Overall, the distribution of norms is highly skewed as, on average, a norm only appears in 3.6 norm chains. This number is slightly better for reduced norms as they appear on average in 11.7 norm chains, but their occurrences still follow a power-law distribution. Some norms appear more frequently, such as *BGB* or *ZPO*, but others are only found once in our dataset. Around 20% of the target norms are from 14 different norm texts, and around 3.500 norms are only used once. We cannot expect reliable results for norms that are not used very often. Any prediction score will be skewed by this imbalance, which we want to quantify during our analysis stage. Based on this, our problem can be described as multi-label classification as each verdict can have multiple important norms assigned, and we need to select those from a large set of possible targets.

When looking at the referenced norms, we see that only 55% of the norms in norm chains are actually cited within the verdict itself (with an exact match for paragraph, section, etc.). For the reduced norms, a much more significant percentage (94%) is found within the content of the court ruling. However, some norms are not referenced while still being identified as an essential norm towards that decision (i.e., they appear in the norm chain.

Lastly, we randomly selected 10% of the dataset as a test set for our final evaluation. From the remaining verdicts, we again used 10% for a development set to select the best hyperparameters during training of the different classifiers.

3. Methods

We applied four different classification models with varying complexity levels to tackle this classification problem. We used one linear layer for the classification on top of the

[2]spacy.io

[3]The list was extracted from gesetze-im-internet.de

Table 1. Micro-F1 test set performance for the each model type and each classification target.

Method	Reduced Norm	Norm
word2vec	23.10	8.86
TF/IDF	77.05	52.57
Ref. Norms	71.48	49.15
BERT	53.88	31.04

featurization described next. First, we used 100-dimensional word2vec embeddings [3] which were trained on a different legal corpus with around 50k sentences. This corpus contains many different German laws and court rulings to have a wider variety in terms of textual content. The embeddings were trained with Gensim [4] using a word window of 5 while unknown words are replaced with an all-zeros embedding. We then average all word embeddings in a court ruling. Second, we used the occurrence of a norm in the court ruling text as features for classification. The occurrences are one-hot encoded, i.e., if a specific norm is referenced within the decision, its corresponding entry in the feature vector is set to 1. We assume that it is possible to determine the most important norms for the norm chain based on the mixture of referenced norms. Our third model uses TF/IDF as a feature with a minimum document frequency of 50 per word. The final model we tested is utilizing *bert-base-german-cased* a BERT-model [5] which is pre-trained on a German corpus which also contains legal documents. As the number of usable words is limited for the BERT-based classification, we used the first 512 tokens of a court ruling.

All models were implemented in PyTorch[4]. We used Adam as the optimizer for the Binary-Cross-Entropy loss. To determine the best learning rate for each model type, we first did manual testing to discover an appropriate learning rate range. We then randomly sampled three learning rates from those ranges per type and classification target (norm vs. reduced norm) and selected the best model through the micro-F1 score on the development set. We decided to use this optimization scheme, as a grid-search would have been too costly, but more importantly, each model needs slightly different learning rates to obtain optimal results. We trained each model for 100 epochs with early stopping based on the development set micro-F1 score and a patience of 10 epochs, except for the BERT-based model. Here we only trained for 25 epochs as the model converged faster, and no significant improvements were observed afterward. The micro-F1 score for each model on the test set can be seen in Table 1 for the norm set and the reduced norm set.

4. Model Analysis

As seen in Table 1, the averaged word2vec featurization has the worst performance by a wide margin. For that reason, we do not further investigate this model.

Surprisingly, TF/IDF and the referenced norms outperform the BERT-based classifier by almost 20% for both target sets. A possible reason for the lower performance of BERT could be the text type. German court rulings are relatively long, and we needed to make some simplifying assumptions to encode the court ruling. We could not identify any evidence towards this claim for the BERT-based model when investigating the relation between document length and prediction performance. Additionally, as the used BERT

[4] pytorch.org

model was pre-trained on German legal documents, a domain mismatch also seems unlikely. The most likely reason for the lower performance on our dataset seems to be the significant skew in the distribution of labels. Targets with many samples are usually easier to predict, but the BERT-based model seems to need many more samples per class in many cases. Still, many samples do not necessarily result in high prediction scores, as, for example, the trade law (HGB) holds 974 samples in the train set and is never predicted for the reduced test cases. For all norms that are not predicted (241 out of the 328 reduced norms in the test set), it has the highest number of samples, but there are also norms with a lower sample count and near-perfect predictions, e.g., the law on associations (VereinsG) with 13 train samples and an F1 score of 100%. More specific laws with more minor possible applications might be easier to predict, but we did not investigate this further for the BERT-based model. Next, we will closely look at our best model based on TF/IDF and discover why it can actually outperform the other models before pointing out possible problems.

Due to the number of possible targets it is not feasible to manually identify specific shortcomings for each particular target. Nevertheless, as our best model is rather simple, it is possible to reason about its predictions and performance. In our case, y_i is the prediction value for the i-th norm $n_i \in N$, $t \in T$ are the individual terms, b_i a bias term towards predictions for a specific norm and w_t^i is the weighting of the term t for predicting the i-th norm. With σ denoting Sigmoid function used for calculating a prediction value between $[0,1]$ and $f(t)$ as the TF/IDF value our prediction function can be written down as follows:

$$y_i = \sigma\left(\sum_{t \in T} w_t^i f(t) + b_i\right) \tag{1}$$

As the TF/IDF-based model assigns a positive or negative weight w_t^i to the terms used in a court ruling (which is weighted by importance via TF/IDF) and as those individual weightings are then additively aggregated, we can identify which terms have a positive or negative impact on a prediction. When looking at the magnitude of those weights $\{w_t^i | n_i \in N\}$, we can identify that a significant majority of all terms is negatively weighted. This makes intuitively sense as for all targets, there are more negative than positive samples, and the model needs to focus more on which targets not to predict.

The resulting follow-up question is which terms are the most negatively/positively weighted, and is there an overlap between the different targets. We only want to focus on the reduced target set for this analysis, as this is not feasible otherwise. When overlapping the five positive and negative weights with the highest magnitude, we can identify in which cases a specific cue word is a positive sign for one target and a negative for another. Thus we can identify which target clusters have a thematic overlap but need some differentiation via negative cue words. We can interpret the weights like this as a big negative weighting towards a term that will only happen due to a repeated misclassification towards a different target. Moreover, we can identify these different targets by looking at enormous positive weights as they have the most decisive influence on classification.

To visualize this, we build a network with the reduced norms, where two norms A and B are connected if A has at least one term in its top-5 positive terms, which is also in the top-5 negative terms of B. This network of connected norms can be seen in Figure 2. There are smaller components of two or three norms that are connected, such as the property tax law (VStG) and the law for taxation in foreign relations (AStG). As far as

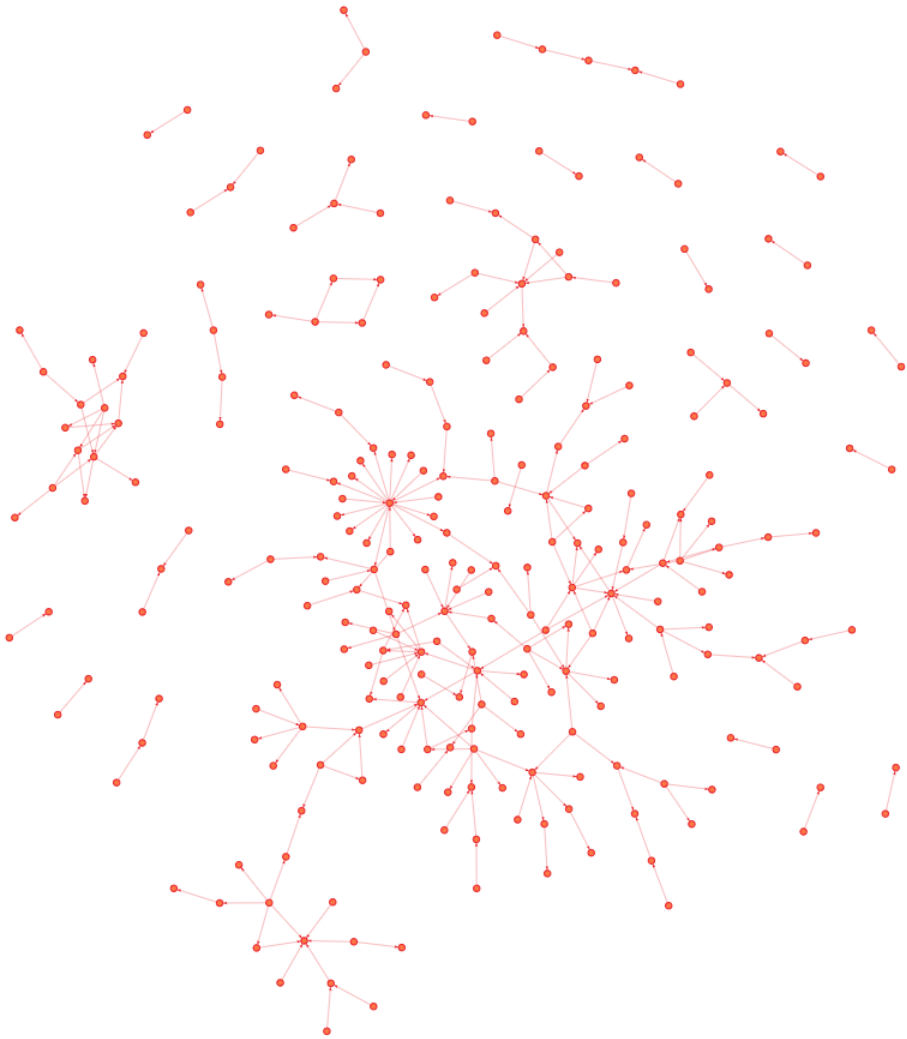


Figure 2. Connection graph of reduced norms which share at least one term in their top-5 positive and negative weights for the TF/IDF-based model. Unconnected norms were omitted.

we can tell, those connections denote exceptional cases for specific circumstances, such as, in this case, the location change. Furthermore, in those cases, one of the norms is not applicable anymore. More prominent connected components and especially star-shaped connections are interesting cases. For example, the bigger cluster on the left deals with taxation laws, with most of the norms specifically focus on energy taxes for natural gas or combined heat and power generation. In those cases, the prediction depends on precise textual details. However, by far, the most significant connected component is around the federal constitutional court law (BVerfGG), with the star in the middle of the picture. As this is only applicable at the highest instance level, our model needed to implicitly identify some information about the instance level of a court ruling to assign this law correctly. Furthermore, this becomes more evident when looking at the top-10 weights

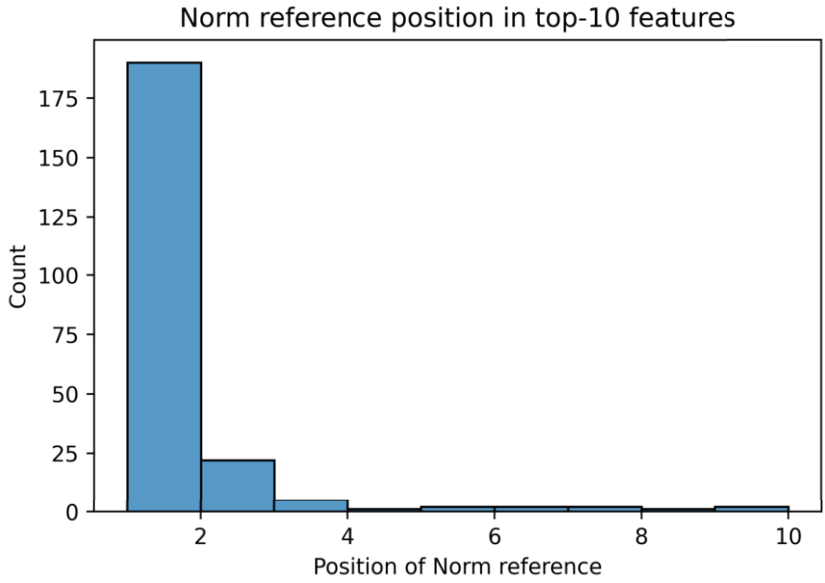


Figure 3. Position of the norm reference in the top-10 features for the reduced norms. Average position is 1.4 and only 84 norm targets do not have their reference in the top-10 positive terms.

with multiple dozen connections to this law.

We now want to look at the positively weighted terms with the highest magnitude, and surprisingly the reference to the norm itself within the text is possibly the most important term. As seen in Figure 3, it is on average the 1.4th highest weighted term, and it is for only 84 of the 666 reduced norms not in the top-10 features. This fact also explains the high performance of the model with the used norms as features.

We cannot look at each essential individual term, but we can identify if they are important in the sourcing norm or if there exists a different norm for which the top-10 positive terms are more critical/higher weighted. In order to extract the most relevant terms per reference norm A, we used TF/IDF on the source documents and then extracted the values for the top-10 positive weighted terms for that norm. After extracting the values for the same top-10 terms for every other norm, we ranked all the norms based on the sum of their TF/IDF values in decreasing order. Consequently, a norm will have a high ranking if those top-10 terms are essential in its text. We then calculated the rank for the reference norm A and could identify that in the majority of cases, the reference norm has the highest summed value as seen in Figure 4. Based on this, we can tell that the prediction is based on important terms from the source document.

Lastly, we want to look at the influence of the number of trainset samples on the test set F1 score to identify how many samples would be necessary for each target norm to get more reliable predictions. As the number of occurrences per target follows a power-law distribution, we plotted the test set F1 score based on the sorted, decreasing train set occurrences for each reduced norm in Figure 5. As seen in this figure, most reduced norms with at least 50 samples in the train set can be predicted to a certain extent, but there are also cases with much fewer samples and near-perfect predictions. We can also

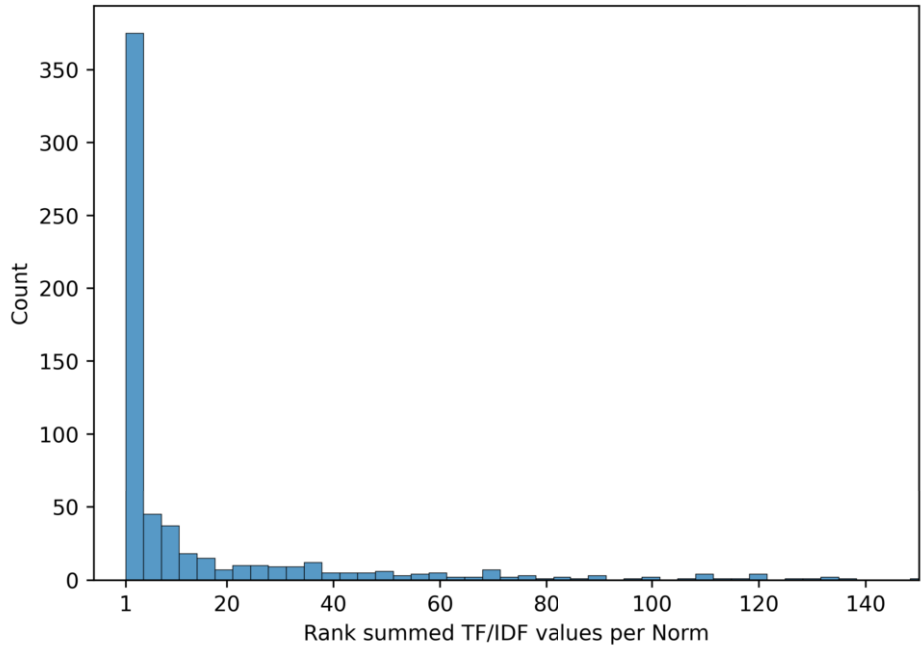


Figure 4. Ranks of the norm when extracting the TF/IDF values from the referenced norm text for the top-10 positive features. Those values are then summed and sorted in decreasing order. The rank of a norm is its position in this sorted list. Cutoff at rank 150 as the counts stay relatively similar. Maximal rank for German civil code (BGB) with rank 366.

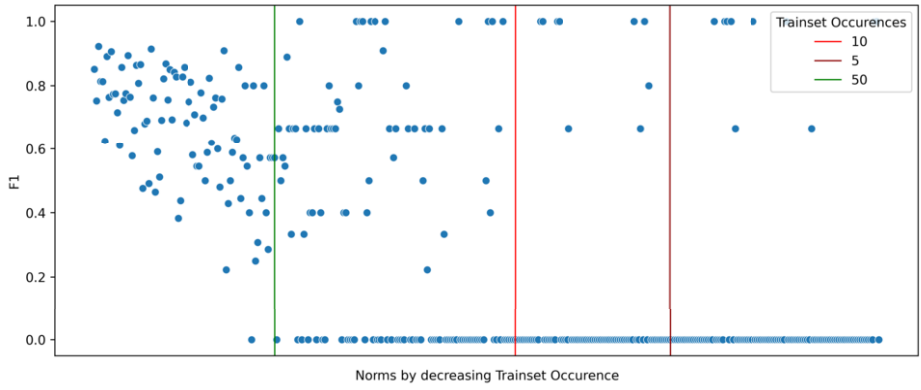


Figure 5. Number of occurrences of a reduced norm in the train set against its test set F1 score.

observe that the spread in F1 scores steadily increases with decreasing number of train samples, which was to be expected. There are again targets that are not used for any predictions, but the TF/IDF-based model uses 86 more norms in their predictions than the BERT-based model and thus at least tries to predict 53% of the possible reduced targets (173 out of 328 reduced norms in the test set). We also investigated prediction

errors for the combination of norms as targets, e.g., are norms mispredicted more often when a semantically similar norm is also a target label due to a thematic overlap. We could not identify any relationship. Consequently, in our opinion, the most considerable influence on low-performance metrics for a specific (reduced) norm is the low number of samples.

When doing the same analysis with the whole target set, the results are similar, although the influence of the number of trainset samples is more extreme. As the TF/IDF model depends on some particular textual cues, we hypothesize that BERT cannot extract those specific cues.

5. Limitations and Related Work

The framework we built around our task has some shortcomings, which we want to address in the following. First, as we phrased our problem as a classification task, we cannot predict norms or reduced norms that are not in our dataset. If we want to include new norms, we would need to retrain all of our models. This is particularly problematic for norms as there are many more targets. Second, a substantial number of samples is necessary per target to get reliable predictions. As the targets follow a power-law distribution, this is not practically feasible. Third, the law is constantly changing. To be as precise as possible with our predictions, we would need to keep track of all the changes in the past and then only predict viable norms based on the date of a court ruling. Initially, we tried to avoid the first two shortcomings by posing our problem in the framework of a recommendation system with different features representations for the court rulings and norms, different extraction granularities (document level, sentence level, etc.), and different extraction methods such as approximate k-Nearest Neighbor. Thereby, we could not achieve satisfactory performance levels even when drastically increasing the number of recommendations.

Looking at related research, many papers exist that deal with multi-label text classification. As the legal domain has longer and more complex texts, we want to focus on legal text classification papers. While the list of tremendous research activities within the AI&Law community is extensive, we tried to point towards the most relevant papers for our work.

Sulea et al. [6] try to predict the decisions of the rulings of the French Supreme Court as well as its area of law, while Soh et al. [7] used classification methods on Singapore Supreme Court judgments to identify their legal areas. While much current work focuses on various classification topics such as legal document segmentation and including metadata extraction [8,9,10,11,12], Chalkidis et al. [13] are the most closely related to ours, mainly because they encountered a problem with a considerable number of possible target labels as well. They apply different Deep Learning architectures to classify European legislative documents with over 4k labels. In contrast to our work, they could achieve their highest performance with a similar BERT-based model to ours. Most recently, Huang et al. [14] provided approaches to recommend legal citations based on deep learning. They trained four different types of machine learning models. In future work, it may be worth investigating whether their approach can be applied to court rulings and norms to identify possibly relevant norms. Such identification of potential norms could then be utilized as another feature in our classification models.

6. Conclusion

This paper examined the possibility of automating the legal norm chain creation process for the German legal domain. The problem was modeled as a multi-label classification task, utilizing a linear layer for the actual classification. Four different methods were applied for the underlying featurization: (1) word2vec, (2) one-hot encoded occurrences of the individual norms, (3) TF/IDF, and (4) BERT. We could show that it is feasible up to a certain degree to extract the relevant norms from court rulings with an F1 measure of 0.77.

Nonetheless, as stated in Section 5 this research contains some limitations. Most limitations arise from the multi-label classification setup. However, we already implemented different recommendation systems in an unsupervised fashion, utilizing external sources such as vectorized representations of the actual norm content. As a result, we believe that the task must be tackled as a supervised multi-label classification task. It would be inevitable to capture even more different norms in future work by utilizing a larger corpus. In doing so, it may be beneficial to train different models for verdicts from courts of different jurisdictions.

With this work, we laid down essential groundwork in the context of extracting the most relevant norms from court rulings. We not only provided methods that can extract such norms but also performed a detailed analysis of our models. Those insights can be utilized within the research community in future work. Moreover, our algorithms can already be integrated into the existing editorial process within legal publishers in order to provide their legal authors with adequate suggestions for norms that should be included in the norm chain.

References

[1] S. A. Lastres, "Rebooting legal research in a digital age," 2015.

[2] L. F. Peoples, "The death of the digest and the pitfalls of electronic research: what is the modern legal researcher to do," *Law Libr. J.*, vol. 97, p. 661, 2005.

[3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'13. Red Hook, NY, USA: Curran Associates Inc., 2013, p. 3111–3119.

[4] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[6] O.-M. Sulea, M. Zampieri, S. Malmasi, M. Vela, L. Dinu, and J. Genabith, "Exploring the use of text classification in the legal domain," 10 2017.

[7] J. Soh, H. K. Lim, and I. E. Chai, "Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments," in *Proceedings of the Natural Legal Language Processing Workshop 2019*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 67–77. [Online]. Available: https://aclanthology.org/W19-2208

[8] Q. Lu, J. G. Conrad, K. Al-Kofahi, and W. Keenan, "Legal document clustering with built-in topic segmentation," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 383–392.

[9] A. Lyte and K. Branting, "Document segmentation labeling techniques for court filings." in *ASAIL@ ICAIL*, 2019.

[10] E. Loza Mencía, "Segmentation of legal documents," in *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, 2009, pp. 88–97.

[11] B. Waltl, G. Bonczek, E. Scepankova, and F. Matthes, "Semantic types of legal norms in german laws: classification and analysis using local linear explanations," *Artificial Intelligence and Law*, vol. 27, no. 1, pp. 43–71, 2019.

[12] I. Chalkidis and D. Kampas, "Deep learning in law: early adaptation and legal word embeddings trained on large corpora," *Artificial Intelligence and Law*, vol. 27, no. 2, pp. 171–198, 2019.

[13] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, "Large-scale multi-label text classification on EU legislation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6314–6322. [Online]. Available: https://aclanthology.org/P19-1636

[14] Z. Huang, C. Low, M. Teng, H. Zhang, D. E. Ho, M. S. Krass, and M. Grabmair, "Context-aware legal citation recommendation using deep learning," *arXiv preprint arXiv:2106.10776*, 2021.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210316

Data-Centric Machine Learning: Improving Model Performance and Understanding Through Dataset Analysis

Hannes WESTERMANN [a,1], Jaromír ŠAVELKA [b], Vern R. WALKER [c], Kevin D. ASHLEY [d] and Karim BENYEKHLEF [a]
[a] *Cyberjustice Laboratory, Faculté de droit, Université de Montréal*
[b] *School of Computer Science, Carnegie Mellon University*
[c] *LLT Lab, Maurice A. Deane School of Law, Hofstra University*
[d] *School of Computing and Information, University of Pittsburgh*

Abstract. Machine learning research typically starts with a fixed data set created early in the process. The focus of the experiments is finding a model and training procedure that result in the best possible performance in terms of some selected evaluation metric. This paper explores how changes in a data set influence the measured performance of a model. Using three publicly available data sets from the legal domain, we investigate how changes to their size, the train/test splits, and the human labelling accuracy impact the performance of a trained deep learning classifier. Our experiments suggest that analyzing how data set properties affect performance can be an important step in improving the results of trained classifiers, and leads to better understanding of the obtained results.

Keywords. Classification, Evaluation, Data-centric Approach, Machine Learning, Legal Texts, Semantic Homogeneity

1. Introduction

Two fundamental components of a machine learning (ML) experiment are data and a model. The ML community appears to prefer putting more effort into tweaking the models while spending less time on important data considerations [3]. This means that researchers often invest considerable resources into developing novel models and approaches, achieving marginal improvements. At the same time, they pay much less attention to the properties of the data set (e.g., size, quality, train/test split), or to the effects these might have on the performance of the ML models. Potentially, this under-investigated area of research could lead to significant improvements of the models.

2. Related Work

The ML community has shown increased interest in exploring how data set properties affect trained classifiers. In [3], researchers investigated data-cascades, where issues with

[1]Corresponding Author: Hannes Westermann, E-mail: hannes.westermann@umontreal.ca

data labelling affected downstream systems. In [4], the authors estimated that ten of the most commonly used ML data sets contain an average of 3.4% errors in labelling. AI & Law researchers have investigated data set effects on model performance, including iterative masking of predictive sentences [11], ablating data about criminal charges or sentences [8] and enhancing lawsuit data with ODR data [12]. Researchers have investigated if models can transfer information from single or pooled data sets in different domains [6] or different contexts (languages, jurisdictions and domains) [5].

3. Experimental Design

We evaluated a trained classifier on three data sets annotated on a sentence level under three experimental settings. **Data** We used three publicly available data sets:

- 50 decisions by the U.S. Board of Veterans' Appeals **(BVA)**, containing 6153 sentences tagged with rhetorical roles [9].[2]
- 880 sentences from court opinions mentioning vague statutory terms **(StatInt)**, tagged with usefulness of sentences for statutory interpretation [7].[3]
- 50 opinions of the Supreme Court of India **(ISC)**, containing 9,380 sentences tagged with the rhetorical roles of the sentences [1].[4]

Model We embed each sentence using the Google Universal Sentence Encoder [2] (**GUSE**).[5] We input these embeddings into a two-layer dense neural network classifier (NN model). Full model specs and training procedure are available on github.[6]

Experiments *E1 - Sample-Size Sensitivity:* In this experiment we analyzed the impact of increasing the size of a data set, by first training a classifier on very little data, and then adding more data points each iteration. This allowed us to investigate how adding data to the training set impacts the performance of the classifier, and whether performance trends suggest that adding additional data could be beneficial.

E2 - Split Sensitivity: It is common practice to divide data sets into train and test splits. To investigate the impact of split selection, we split each data set into five folds. In each iteration, four of the folds were used as training split, and the remaining fold as test split. This allowed us to observe the particular split's impact on the performance of the trained NN model, and how much the scores vary on a per-label basis.

E3 - Error Sensitivity: The high-quality of the human labelling is an important concern in ML. To investigate the impact of labelling errors on classifier performance, we started with the original data and then replaced an increasing percentage of the labels with a randomly chosen incorrect label.

4. Results and Discussion

Experiment 1 - Figure 1 shows the evolution of F1-scores for each individual class when adding more and more data. The rates of improvement among the classes are not con-

[2]Dataset available at https://github.com/LLTLab/VetClaims-JSON
[3]Data set available at https://github.com/jsavelka/statutory_interpretation
[4]Data set available at https://github.com/Law-AI/semantic-segmentation
[5]https://tfhub.dev/google/universal-sentence-encoder/4
[6]https://github.com/hwestermann/jurix2021-data_centric_machine_learning

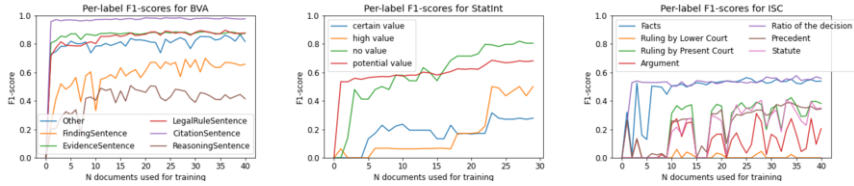


Figure 1. Evolution of per-label F1-score of classifier, as documents are added to training data one by one.

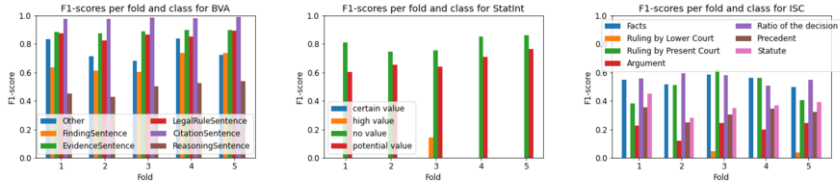


Figure 2. Variations in F1-score for individual classes per five different folds for training and test data.

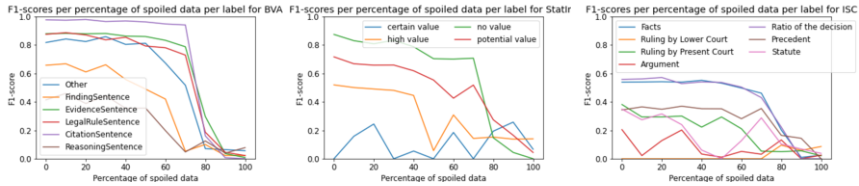


Figure 3. Evolution of per-class F1-scores when replacing set percentage of labels with random label.

sistent. Certain classes reach a high performance quickly, while others take considerably longer. The experiment shows the varying importance of larger data sets. The needed sample size, and whether adding data increases performance, can vary significantly depending on the class and the dataset.

Experiment 2 - Figure 2 shows the scores per class, across the training folds. Some classes' scores differ considerably across the folds. For some use cases, a single train-test split may not produce reliable results when working with legal data sets of limited sizes.

Experiment 3 - Figure 3 shows the effects of randomly mislabeling a portion of the training data set, on a per label basis. Some of the classes start to lose performance quite rapidly, while the others are more resilient to the random errors. In real-world scenarios, human annotators may more likely make systematic errors, increasing the impact.

Class difficulty and semantic homogeneity: It appears that certain labels are significantly more difficult for the classifier to learn than others. In Figure 1, some labels (such as "Citation" for BVA) quickly achieve a high level of performance and then improve more slowly, while other classes require more data to achieve high performance and continually improve (such as "Finding" for BVA). The latter classes also have a more variable performance across folds (Figure 2) and depend more upon high-quality data (Figure 3). Interestingly, this "difficulty" of classes does not fully correspond to the frequency of certain labels appearing in a data set. Rather, it seems related to what we refer to as the *semantic homogeneity* of a class, i.e., how semantically similar the sentences are within a particular class. In [10] we grouped sentences based on semantic similarity

in an embedding space (as determined by Euclidian distance in the GUSE embedding space). For each sentence in a certain class, we explored how many on average of the top 20 most similar sentences were also of that same class. Looking at the table presented in [10], it appears that classes with higher semantic homogeneity are easier to learn for the classifier, and vice-versa. The reason could be that the classifier can more easily find decision boundaries for sentences grouped into clear semantic clusters.

5. Conclusions and Future Work

We trained a classifier on three publicly available data sets, altering the size, training/test split and data labelling quality, to investigate the effects of these properties on ML classifier performance. We observe significant variations in performance over the experiments. These experiments could provide guidance in deciding to continue collecting data, and whether to focus on certain classes during data collection. Our work could represent the initial step in developing a methodology to assess properties of a data set.

Acknowledgments

We are grateful for support from the U. de Montréal Cyberjustice Laboratory, LexUM Chair on Legal Information, and Autonomy through Cyberjustice Technologies project.

References

[1] Bhattacharya, P., Paul, S., Ghosh, K., Ghosh, S. & Wyner, A. "Identification of Rhetorical Roles of Sentences in Indian Legal Judgments." *Jurix 2019*. pp. 3-12 (2019).
[2] Cer, D., Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. St John, N. Constant et al. "Universal sentence encoder." *arXiv preprint arXiv*:1803.11175 (2018).
[3] Sambasivan, N. et al. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. *2021 CHI Conference* pp. 1-15 (2021).
[4] Northcutt, C. G., Athalye, A. & Mueller, J. "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks." *arXiv:2103.14749* [cs, stat] (2021).
[5] Savelka, J., Westermann, H., Benyekhlef, K. et al. "Lex Rosetta: transfer of predictive models across languages, jurisdictions, and legal domains." In *ICAIL 2021*, pp. 129–138. (2021).
[6] Savelka, J., Westermann, H. & Benyekhlef, K. "Cross-Domain Generalization and Knowledge Transfer in Transformers Trained on Legal Data." *ASAIL@ Jurix.* (2020).
[7] Šavelka, J., Xu, H., & Ashley, K. "Improving Sentence Retrieval from Case Law for Statutory Interpretation." *Proc. 17th Int'l Conf. on Artificial Intelligence and Law*, pp. 113-122. (2019).
[8] Tan, H., Zhang, B., Zhang, H., & Li, R. "The Sentencing-Element-Aware Model for Explainable Term-of-Penalty Prediction". In *CCF Int'l Conf. on NLP and Chinese Computing*. pp. 16-27. (2020).
[9] Walker, V. R., et al. "Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning." *Proceedings of ASAIL 2019* (2019).
[10] Westermann, H., Savelka, J., Walker, V., Ashley, K. & Benyekhlef, K. "Sentence Embeddings and High-Speed Similarity Search for Fast Computer Assisted Annotation of Legal Documents." *Jurix*. (2020).
[11] Zhong, L., Zhong, Z., Zhao, Z., Wang, S., Ashley, K. D., & Grabmair, M. "Automatic summarization of legal decisions using iterative masking of predictive sentences." *ICAIL 2019*, pp. 163-172. (2019).
[12] Zhou, X., Zhang, Y., Liu, X., Sun, C., & Si, L. "Legal Intelligence for E-commerce: Multi-task Learning by Leveraging Multiview Dispute Representation." In *Proc. 42nd Int'l ACM SIGIR*, pp. 315-324. (2019).

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210317

The Unreasonable Effectiveness of the Baseline: Discussing SVMs in Legal Text Classification

Benjamin CLAVIÉ [a,1] and Marc ALPHONSUS [a]
[a] *Jus Mundi*

Abstract. We aim to highlight an interesting trend to contribute to the ongoing debate around advances within legal Natural Language Processing. Recently, the focus for most legal text classification tasks has shifted towards large pre-trained deep learning models such as BERT. In this paper, we show that a more traditional approach based on Support Vector Machine classifiers reaches competitive performance with deep learning models. We also highlight that error reduction obtained by using specialised BERT-based models over baselines is noticeably smaller in the legal domain when compared to general language tasks. We discuss some hypotheses for these results to support future discussions.

Keywords. Natural Language Processing, Text Classification, Machine Learning

1. Introduction

Recently, the state-of-the-art in many Natural Language Processing (NLP) tasks has been achieved by large pre-trained models such as BERT and its variants [1]. Specialised BERT-based models have been developed for many fields, establishing the state-of-the-art in domain specific tasks, as evidenced in the biomedical domain [2].

In legal NLP, recent work has focused on exploring the applications of BERT-based approaches on a variety of existing tasks and how to best adapt BERT to the legal domain [3,4]. These efforts, while successful at establishing state-of-the-art on a variety of tasks, also reveal an interesting trend: the performance gain between a general language BERT and a specifically legal-language trained BERT appears to be smaller than in other specialised domains [4].

A common application of legal NLP is text classification. Text classification tasks target various kinds of legal insight, such as predicting the outcome of a ruling from a decision's body [5], whether a given clause is likely to be unfair to a customer [6] or whether a sentence indicates the overruling of a precedent [4].

Little attention has been given to comparing these new BERT-based approaches to well-optimised baselines, such as Support Vector Machine (SVM)-based classifiers, which historically perform well on text classification tasks, opting instead for comparisons with other deep learning-based baselines.

[1]Contact: Benjamin Clavié, Jus Mundi, 30 Rue de Lisbonne, 75008 Paris, France; b.clavie@jusmundi.com.

Table 1. Best performing model on all tasks.

	ECHR (Both)	Overruling	Terms of Services
Best Approach	NBSVM + bigrams	NBSVM + bigrams	Linear SVM + trigrams

In this short paper, we aim to (A) highlight the very strong performance of optimised *baseline* classifiers on multiple legal text classification tasks compared to deep learning classifiers, (B) show that the gains from BERT-based approaches is noticeably smaller on legal-domain tasks than on general tasks and (C) discuss three hypotheses to explain the previous two phenomena.

2. Experimental Setup

2.1. General Domain

For all general domain tasks, we use results from BERT-ITPT-FiT [7], which optimises BERT for text classification, on four common benchmarks. For SVM results, we report the score of the best performing variant from a large scale comparison [8].

2.2. Legal Domain Experiments & Baselines

We compare SVMs to BERT-based results on four existing legal text classification tasks.

Terms of Services (ToS) is a task aiming to determine whether or not a clause found in a contract is likely to be unfair to the customer [6].

Overruling is a binary classification task to identify if a given sentence in a US court decision represents a reversal of precedent (*overruling* a previous decision) [4].

ECHR text classification tasks use the text of the *Facts* part of decisions from the ECtHR and exists in two variants. **ECHR (Binary)** aims to detect if any article of the Charter has been violated and **ECHR (Multi)** requires identifying specifically which article has been violated. We use the *Frequent* version of this last task, meaning we remove any label with fewer than 50 training examples from the data [5].

On each of the legal tasks, we train and evaluate SVM classifiers with modest optimisation. We also experiment with NBSVM[2], an SVM classifier using Naïve Bayes features to represent words [9]. The best approach for each task is listed in Table 1.

For BERT-based models, we report results from the literature. Results for BERT on both **ECHR** tasks are from the paper introducing the task [5] and results from Legal-BERT from the paper introducing it [3]. Results for all BERT-based models on **Overruling** and **Terms of Services** are from the paper introducing Legal-Bert-Custom [4].

2.3. Metrics and Evaluation

In line with the nature of this paper, all metrics reported follow the existing literature. For all General Domain tasks, the metric used is accuracy over the test set.

[2]Implementation available at https://gitlab.com/jusmundi/Legal-svm-baselines/

Table 2. Results for the best performing model of each kind on a variety of General Domain (GD) and Legal text classification tasks. *Error reduction is calculated between the Best SVM and the best BERT variant.*

	General Domain				Legal			
Model	**AGNews**	**IMDB**	**Yelp!**	**DBPedia**	**ECHR (Binary)**	**ECHR (Multi)**	**Over-ruling**	**ToS**
Best SVM	75.3	80.7	84.0	87.1	82.2	61.1	94.9	79.3
BERT	95.2	95.6	98.1	99.3	82.0	60.8	95.8	72.2
Legal-BERT [3,4]	n/a	n/a	n/a	n/a	88.3	65.2	97.4	78.7
Error Reduction	80.6%	77.2%	88.1%	94.8%	34.3%	10.5%	49%	-1.8%

We report macro-averaged F1 score for **ECHR (binary)**, micro-averaged F1 score over all classes for **ECHR (multi)** [5] and average F1 score over 10-fold cross-validation on both Overruling and ToS [4].

In all cases, we report the error reduction between BERT models and SVMs as the percentage decrease in error rate between models to simplify evaluating the impact of using a different model over multiple tasks. The error rate is calculated as $100 - Score$.

3. Classification Results

Table 2 gives an overview of the various classification results and presents the error reduction obtained by using the best BERT-based model over the relevant SVM classifier.

The error reduction between SVM and BERT models in the general domain is high, at **85.175%** on average over the four tasks, with the lowest reduction being 77.2%.

The difference is much less stark within the legal domain: on all but one of the tasks, the performance of the SVM classifier exceeds that of a general domain BERT[3] .

Legal-BERT models, optimised for legal texts, do reach the best performance on three out of four tasks and only slightly fall short of it on the fourth. However, in all cases, the performance increase is much less pronounced, with an average error reduction across all four tasks of **23%**.

4. Discussion

The results highlight an interesting phenomenon: despite impressive performance in both the general domain and other specialised domains without the need for domain adaptation [2], BERT falls short within the legal domain. Even after domain adaptation is performed to train specialised Legal-BERT models, the performance improvement remains modest and does not reproduce the very notable gains found in other applications.

On **ECHR (Multi)**, potentially the most complex task due to being multi-label, there is remarkably only an **10.5%** error reduction between SVM and Legal-BERT.

There is no clear explanation for this phenomenon, but we discuss multiple hypotheses. The first, initially proposed by Zheng et al. [4] to explain the mild improvements from Legal-BERT, is that the tasks on which we evaluate legal NLP algorithms are not suit-

[3]The results on ECHR are considerably better than the SVM approach reported in the original paper [5] as they use tf-idf weighting for feature generation, which performs notably worse than other methods.

able, either due to them being too simple or their language not being sufficiently domain-specific to take advantage of the models' pretraining. However, this does not provide an explanation for the overall weak improvement from deep learning over SVM classifiers.

A similar potential explanation could be that *simple* mono-lingual text classification is not enough to truly take advantage of the possibilities offered by more powerful BERT-based models. This would indicate that the powerful language representation of Legal-BERT models could be key to tackling more complex tasks which have started being explored, such as textual entailment [4] or legal rationale extraction [10].

However, this still does not fully address the weak performance gains on text classification. A final hypothesis we propose is that large language models, even when trained on legal language, still lack the ability to capture the depth of legal language and its specific vocabulary. These models could also fail to properly weigh the meaning of multiple legal concepts being mentioned together. This hypothesis would suggest the need to develop a way to integrate sources of legal information, such as knowledge-bases or ontologies, within deep learning models to truly take advantage of their potential.

5. Conclusion

We show that SVM classifiers perform well on multiple legal text classification benchmarks. SVM models can outperform general domain BERT models, but perform worse than BERT-based models adapted for legal text. We also show that the relative performance improvement between the BERT-based models and the SVM models is considerably smaller within the legal domain than on general domain classification tasks.

We propose and discuss three potential explanations for these results. Future work will focus on exploring the limits of BERT models within the legal field, both by exploring more complex tasks and integrating existing knowledge bases with them.

References

[1] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL 2019;. .
[2] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019 09.
[3] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: The Muppets straight out of Law School. In: Findings of EMNLP 2020;. .
[4] Zheng L, Guha N, Anderson BR, Henderson P, Ho DE. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. In: Proceedings of ICAIL2021. ACM;. .
[5] Chalkidis I, Androutsopoulos I, Aletras N. Neural Legal Judgment Prediction in English. In: Proceedings of ACL; 2019. p. 4317-23.
[6] Lippi M, Pałka P, Contissa G, Lagioia F, Micklitz HW, Sartor G, et al. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. Artif Intell Law. 2019;27(2):117-39.
[7] Sun C, Qiu X, Xu Y, Huang X. How to fine-tune bert for text classification? In: China National Conference on Chinese Computational Linguistics. Springer; 2019. p. 194-206.
[8] Riekert M, Riekert M, Klein A. Simple Baseline Machine Learning Text Classifiers for Small Datasets. SN Computer Science. 2021;2(3):1-16.
[9] Wang S, Manning CD. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In: Proceedings of ACL 2021;. .
[10] Chalkidis I, Fergadiotis M, Tsarapatsanis D, Aletras N, Androutsopoulos I, Malakasiotis P. Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases. In: Proceedings of NAACL 2021;. .

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210318

Assessing the Cross-Market Generalization Capability of the CLAUDETTE System

Agnieszka JABLONOWSKA [a], Francesca LAGIOIA [a,b,1], Marco LIPPI [a,c,2], Hans-Wolfgang MICKLITZ [d], Giovanni SARTOR [a,b] and Giacomo TAGIURI [a,e,3]

[a] *Law Department, European University Institute*
[b] *CIRSFID - Alma AI, University of Bologna*
[c] *DISMI, University of Modena and Reggio Emilia*
[d] *Robert Schuman Centre, European University Institute*
[e] *Polish Academy of Science*

Abstract. We present a study aimed at testing the CLAUDETTE system's ability to generalise the concept of unfairness in consumer contracts across diverse market sectors. The data set includes 142 terms of services grouped in five sub-sets: travel and accommodation, games and entertainment, finance and payments, health and well-being, and the more general others. Preliminary results show that the classifier has satisfying performance on all the sectors.

Keywords. Unfair clause detection, machine learning, cross-market analysis

1. Introduction

In the AI and Big data era, where suppliers' power is boosted by technologies, consumer-empowering technologies are needed to support the countervailing power of civil society [1,2,3]. The CLAUDETTE project contributes to this goal, by automating the assessment of standard terms' compliance with EU consumer law. It adopts a supervised machine learning approach, based on a corpus of contracts annotated by domain experts [4], where clauses are labeled either as fair or unfair.

The aim of this work is to study whether and to what extent CLAUDETTE is able to generalize the concept of unfairness across diverse market sectors. To address this question we extended the data set to two additional domains: Health and Finance. The present study allowed us to generate new insights from the pre-existing corpus. In particular, two further groups of companies displaying a degree of sector-specificity, i.e., Travel and Games, were isolated from the original data set.

The paper is organised as follows. In Section 2 we describe the extended corpus, the document annotation procedure, and we briefly introduce a new category of unfair

[1]F. Lagioia and G. Sartor have been supported by the H2020 ERC Project "CompuLaw" (G.A. 833647)
[2]M. Lippi has been supported by the SCUDO project, within the POR-FESR 2014-2020 programme of Regione Toscana.
[3]G. Tagiuri has been supported by the National Science Centre in Poland (G.A. UMO-2019/35/B/HS5/04444)

clauses. We also explain the adopted methodology and discuss the results. Section 3 analyses some possible causes of misclassifications, affecting the CLAUDETTE performance. Finally, Section 4 concludes and presents future research directions.

2. Data set and experimental setting

In our previous works [4,5] we produced a data set consisting of 100 relevant online Terms of Service (ToS) of online platforms, analysed by legal experts and marked in XML. In this research, we increased the data set by adding 42 new contracts, marked by two independent annotators, for a total of 142 ToS.[4] The new documents were selected among those offered by some of the major players in the health and finance market sectors. For the purpose of this study, we split the data set in five groups – i.e., (i) Finance and Payments, (ii) Health and Well-being, (iii) Games and Entertainment, (iv) Travel and Accommodations, and (v) the more general class Others, which contains the remaining contracts originally included in the CLAUDETTE data set. The division of documents into groups is unbalanced: 21 in Finance, 24 in Health, 18 in Games, 8 in Travel, and 71 in others.[5] Such an unbalance is due to two main reasons. On the one hand, the aim pursued in this study led us to collect and analyze new ToS in the Health and Finance domains, which present some peculiarities if compared to other online services. On the other hand, given the effort needed to increase the data set, we decided to reuse the pre-existing corpus.

The annotations reflect the methodology described in [4], where we identified eight different categories of unfair clauses, establishing (1) jurisdiction in a country different than consumer's residence (`<j>`); (2) choice of a foreign law (`<law>`); (3) liability limitations (`<ltd>`); (4) the provider's right to unilaterally terminate the contract/access to the service (`<ter>`); (5) the provider's right to unilaterally modify the contract/the service (`<ch>`); (6) the mandatory arbitration before the court proceedings can commence (`<a>`); (7) the provider's right to unilaterally remove consumer content (`<cr>`); and (8) the consumer consent to the agreement simply by using the service, downloading the app or visiting the website, (`<use>`). In this research we present an additional category of potentially unfair clauses, i.e., those stating (or implicitly assuming) that (9) the scope of consent granted to the ToS incorporates also the privacy policy, which forms part of the "General Agreement" (`<pinc>`). As reported by [6,4,7,8] such categories are widely

[4]The corpus is made freely available for research purposes at the following link: https://claudette.eui.eu/corpus_142_ToS.zip.

[5]In particular, the Finance group includes the following ToS: Bondora, ETFmatic, GoFundMe, Google Payments GB, Google Pay, Kickstarter, Klarna credit agreement, Klarna.com, Ledger.com, Ledger Live, Monzo, Paypal Italy, Revolut, Swanest, Transferwise, Trustly, Visa Solution, Wefox, Western Union for Italy, Xoom, YNAB. The Helath sector includes: 23andme, Ada, Ava, Betterpoints UK, Clue, Endomondo, Fitbit, Flo, Flow, Headspace, Hexoskin, idoc24, iHealth, Kardia, Kry, Lady Cycle, Muse, Mysugr, MyHeritage, Natural Cycles, Polar, Skinvision, Unmind, Woebot. Games includes: ElectronicArts, Epic Games, Habbo, Lindenlab, Masquerade, Nintendo, Oculus,Paradox, Pokemon Go, Rovio, Shazam, Sporcle, Spotify, Steam, Supercell, Ubisoft, World of Warcraft, Zynga. Travel includes: Airbnb, Booking.com, Couchsurfing, eDreams, Expedia, Ryanair, Skyscanner, Verychic. Finally, the Others group includes: 9gag, Academia, Alibaba, Amazon, Atlas, Badoo,Blablacar, Box, Crowdtangle, Dailymotion, Deliveroo, DeviantArt, Diply, Dropbox, Duolingo, eBay, Evernote, Facebook, Foursquare, Garmin, Goodreads, Google, Grammarly, Grindr, Groupon, Happn, HeySuccess, Imgur, Instagram, Lastfm, Match, LinkedIn, Microsoft, Moves, Mozilla, Musically, Myspace, Netflix, Onavo, Opera, Pinterest, Quora, Reddit, Skype, Slack customer, Slack user, Snap, Syncme, Tagged, Terravision, TikTok, Tinder, TripAdvisor, TrueCaller, Tumblr, Twitch, Twitter, Uber, Viber, Vimeo, Vivino, WeChat, Weebly, WeTransfer, WhatsApp, Yahoo, Yelp, YNAB, YouTube, Zalando, Zara, Zoho.

used in ToS for online platforms. To capture the different degrees of (un)fairness we appended a numeric value to each XML tag, with 1 meaning clearly fair, 2 potentially unfair, and 3 clearly unfair, according to the criteria defined in [4,5]. We consider a binary classification task: the positive class is made by all potentially or clearly unfair clauses, for all categories, indiscriminately, and the negative class by all the remaining clauses. We used each of the four sectors, in turn, as a test set, whereas the three remaining sectors, plus all the contracts that belong to none of those four sectors, constituted the training set. For each sector, we performed an additional inner 5-fold cross-validation on the training set to choose the best C hyper-parameter for the linear support vector machine.

Table 1 reports the values of precision (P), recall (R), and F_1 score for each sector considered as test set. P is the percentage of clauses predicted as positive which are really positive (thus accounting for false positives), R is the percentage of correctly detected positive clauses (thus accounting for false negatives) and F_1 is the harmonic mean between P and R. The results show that the classifier has satisfying performance on all the sectors, especially in terms of recall. The sector with the best performance is the Health sector, while the Travel sector has the lowest performance. Both Games and Travel sectors suffer in particular in terms of precision. Section 3 further discusses the reasons why certain types of clauses are wrongly detected as false positives or rather missed by the classifier.

Table 2 shows the percentage of detected potentially unfair clauses (i.e., the recall of the system) for each sector and for each category. To assess the weight of each category across sectors (and therefore on the overall performance), Table 3 shows the average number of clauses per ToS for each sector and category. This information clearly shows which are the most critical clause categories to detect across sectors. The `<pinc>` category (clauses including consent to data processing in the contract, see Section 2) has quite a low recall, especially for Finance and Travel. However, it has to be remarked that at most one clause for such category is usually encountered in a contract. On the other hand, a low recall in all sectors affects limitation of liability clauses, despite their high frequency. Additionally, we can note how unilateral termination and contract by using categories have a low recall in Finance, and arbitration clauses in Travel. More details on the false positives and false negatives are presented in the next section.

Table 1. Experimental results on the cross-sector analysis. For each test sector, we report the micro-averaged precision (P), recall (R), and F_1 as the harmonic mean between the first two metrics.

Sector	P	R	F_1
Finance	0.689	0.739	0.713
Games	0.629	0.849	0.723
Health	0.685	0.809	0.741
Travel	0.597	0.778	0.675

Table 2. Percentage of correctly detected clauses (recall) for each sector and category.

Sector	A	CH	CR	J	LAW	LTD	PINC	TER	USE
Finance	1.000	0.829	0.923	0.765	0.833	0.734	0.375	0.739	0.630
Games	0.833	0.902	0.854	0.920	0.885	0.796	0.769	0.919	0.875
Health	0.950	0.885	0.848	0.909	0.848	0.754	0.786	0.840	0.820
Travel	0.667	0.909	0.875	0.900	0.800	0.633	0.500	0.929	0.847

Table 3. Average number of clauses per contract, for each sector and for each category.

Sector	A	CH	CR	J	LAW	LTD	PINC	TER	USE
Finance	0.333	3.333	0.619	0.810	0.857	7.333	0.381	4.000	1.286
Games	1.333	3.389	2.667	1.389	1.444	6.278	0.722	4.778	2.667
Health	0.833	4.708	1.917	1.375	1.375	8.792	0.583	5.458	2.542
Travel	1.125	1.375	1.000	1.250	1.250	3.750	0.250	1.750	1.625

3. Error analysis

By performing an analysis of the classification errors, we identified three main issues that could cause wrong classifications: (1) the presence of rare linguistic patterns and lexical choices; (2) the specificity of the content of some clauses and (3) the existence of sector-specific regulations.

Linguistic and lexical patterns. Rare linguistic and lexical patterns present in online Terms of Service may be due either to the specificity of a service or to the country in which the service originates (e.g., not in English speaking countries, where most of the ToS in the data set originate). Even though uncommon semantic and lexical formulations can appear in all sectors, we empirically observe that they more frequently emerge in Finance and Health. Since ToS belonging to these sectors were added more recently in the corpus, peculiarities of and changes in recurrent linguistic expressions may also be due to the different times at which the ToS were drafted and analysed.

As an example, consider the following clause taken from the AVA health app ToS (updated May 27th, 2019):

> *Any use of the Products or Site other than as specifically authorized herein, without the prior written permission of Ava is strictly prohibited, and Ava may terminate the license granted herein with immediate effect.*

In this case, "terminate the *license*" slightly differs from the more typical expression "terminate the *contract*". This hypothesis of failure in correctly classifying the clause as `<ter2>` finds support in a similar formulation of a misclassified clause in the Flo ToS (updated February 5th 2020).We plan to examine whether ontologies my help in dealing with such terminological issues.

Content-specificity. The second cause of failure concerns the content–specificity of certain clauses, especially in relation to services dealing with both a digital and a physical component.

While most of the ToS in the original data set concern services that are only digital, others rely on physical devices, are integrated with a physical service, and/or allow to place orders for physical goods. This occurs more frequently in Travel and Health, where, for instance, fitness apps are usually associated to wearable devices. As an example, consider the following clause taken from the Lady Cycle ToS (updated July 8th, 2018):

> *By using Lady Cycle, you acknowledge that you have read and understood the tutorial and manual for its use.*

The clause above has been incorrectly classified by CLAUDETTE as unfair. We can speculate that this is due to the linguistic similarities with the typical consent by using (`<use2>`) clauses. Despite these similarities, it rather relates to consumer's acquaintance with product-related instructions. As noted in Section 2, the correct detection of limitation of liability clauses seems to be particularly problematic for all the analysed sectors.

Once again, these may be due to their domain-specific content. Consider for instance, the following clause, not recognized as potentially unfair by CLAUDETTE, and taken from the Hexoskin ToS (updated August 31st, 2019) in the Health sector:

> *We are not responsible for any health problems that may result from training programs, products, or events you learn about through the Hexoskin Services.*

Content-specificity also characterize a number of consent by using clauses in the finance and payments sub-set, where multiple kinds of agreement are mentioned in ToS in relation to multiple services. Consider the following clause taken from the Paypal ToS (updated July 30th, 2021):

> *If we offer you the new checkout solution service and you choose to use it, in addition to this User Agreement, you agree to the following further terms relating to the following capabilities: when you use our APM functionality as part of the new checkout solution, the PayPal Alternative Payment Methods Agreement; and when you use: our Advanced Credit and Debit Card Payments service as part of the new checkout solution; and our Fraud Protection as part of the new checkout solution; Our Fees for using the new checkout solution apply.*

Sector-specific regulations. The third type of cause of misclassification relates to the existence of sector-specific regulations. Such regulations may induce businesses to include in sectorial contracts certain clauses whose content is markedly different from the content of other clauses having a similar wording. For example, the Payment Services Directive II,[6] lays down quantitative limits to losses that the payer may be obliged to bear in case of unauthorised transactions. Specifically, the EUR 50 amount fixed by Art. 74 can recurrently be found in ToS of payment services providers, such as the Transferwise ToS (updated July 28th, 2020):

> *You will be liable for the first 50 EUR of any unauthorised payments if we believe you should have been aware of the loss, theft or unauthorised use.*

This clause concerns the limitation of consumers' liability (thus to their advantage), classified as unfair since its language is similar to clauses stating (unfair) limitation of liabilities of businesses.

Additional remarks. We further analysed the false positives of the Games and Travel sectors, which appear to be the most difficult ones for CLAUDETTE, in terms of specificity.

Regarding Games, we noticed how over 17% of false positives contain the word *arbitration* (or at least its root, like *arbitrate*), and around 16% include the terms *liable* or *liability*. As for arbitration, many of such clauses are sentences that describe the arbitration procedures. For limitation of liability, a significant number of false positives consists in statements confirming providers liabilities, that often use the same terms of potentially unfair clauses excluding such liabilities. One typical example of such clauses is the following one, from Supercell (updated on August 1st, 2017).:

> *Nothing in these terms of service shall affect the statutory right of any consumer or exclude or restrict any liability resulting from gross negligence or willful misconduct of Supercell or for death or personal injury arising from any negligence or fraud of Supercell.*

[6] Directive (EU) 2015/2366 of the European Parliament and of the Council of 25 November 2015 on payment services in the internal market.

Such a clause, stating that the provider is responsible for damages to the consumer, is very similar to unfair clauses stating that the provider is not responsible.

Regarding Travel, around 67% of false positives derive from the Airbnb (46%) and Expedia (21%) ToS. In some cases, such clauses are meant to inform parties on the possible behaviour by a third party, as in the following case taken from Expedia (updated on February 21st, 2018):

Airlines and other travel suppliers may change their prices without notice.

4. Conclusion

Our study investigated the ability of the original CLAUDETTE model to generalize the concept of unfair clauses regarding four market sectors, i.e., games and entertainment, travel and accommodation, finance and payments and health and well-being. The results show that the classifier has satisfying performance on all the sectors, especially in terms of recall. The analysis of false positives and false negatives revealed some possible types of causes of misclassification, including linguistic and lexical patterns, content-specificity, and sector-specific regulations.

In the future, our final goal is to enable CLAUDETTE to automatically detect unfair clauses in consumer contracts across diverse critical market sectors. To this end, we plan to increase the data set with more recent ToS and enlarge the number of market sectors under investigation. We also plan to apply more sophisticated NLP techniques, such as transformers or methods based on sentence embeddings [9]. Finally, we will also compare the current approach to a multi-class formulation, by adding the class of clearly fair clauses for each specific category as an additional output of our system.

References

[1] Lippi M, Contissa G, Lagioia F, Micklitz HW, Pałka P, Sartor G, et al. Consumer protection requires artificial intelligence. Nature machine intelligence. 2019;1(4):168-9.

[2] Lippi M, Contissa G, Jablonowska A, Lagioia F, Micklitz HW, Palka P, et al. The force awakens: Artificial intelligence for consumer law. Journal of artificial intelligence research. 2020;67:169-90.

[3] Thorun C, Diels J. Consumer protection technologies: an investigation into the potentials of new digital technologies for consumer policy. Journal of Consumer Policy. 2020;43(1):177-91.

[4] Lippi M, Pałka P, Contissa G, Lagioia F, Micklitz HW, Sartor G, et al. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. Artificial Intelligence and Law. 2019;27(2):117-39.

[5] Ruggeri F, Lagioia F, Lippi M, Torroni P. Detecting and explaining unfairness in consumer contracts through memory networks. Artificial Intelligence and Law. 2021:1-34.

[6] Loos M, Luzak J. Wanted: a bigger stick. On unfair terms in consumer contracts with online service providers. Journal of consumer policy. 2016;39(1):63-90.

[7] Dari-Mattiacci G, Marotta-Wurgler F. Learning in Standard Form Contracts: Theory and Evidence. NYU Law and Economics Research Paper. 2018;(18-11).

[8] Micklitz HW, Pałka P, Panagis Y. The empire strikes back: digital control of unfair terms of online services. Journal of consumer policy. 2017;40(3):367-88.

[9] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. In: Proc. 2020 EMNLP Conference. Online: Association for Computational Linguistics; 2020. p. 38-45.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210319

Hybrid AI Framework for Legal Analysis of the EU Legislation Corrigenda

Monica PALMIRANI[a1], Francesco SOVRANO[b], Davide LIGA[a], Salvatore SAPIENZA[a] and Fabio VITALI[b]

[a] *CIRSFID-ALMA-AI, University of Bologna*
[b] *DISI, University of Bologna*

Abstract. This paper presents an AI use-case developed in the project "Study on legislation in the era of artificial intelligence and digitization" promoted by the EU Commission Directorate-General for Informatics. We propose a hybrid technical framework where AI techniques, Data Analytics, Semantic Web approaches and LegalXML modelisation produce benefits in legal drafting activity. This paper aims to classify the corrigenda of the EU legislation with the goal to detect some criteria that could prevent errors during the drafting or during the publication process. We use a pipeline of different techniques combining AI, NLP, Data Analytics, Semantic annotation and LegalXML instruments for enriching the non-symbolic AI tools with legal knowledge interpretation to offer to the legal experts.

Keywords. Akoma Ntoso, Classification AI, NLP, legal drafting techniques.

1. Introduction: AI for legislative drafting process

The scope of the "Study on legislation in the era of artificial intelligence and digitization", promoted by the EU Commission Directorate-General for Informatics, is part of the digital transformation agenda supported by the EU Commission, particularly relevant in this historical moment where the Rules of Law changes quickly due to the COVID-19 special regulation. Companies and society require legal certainty and it is fundamental to implement policies such as "Better regulations" [2], "Fit for the Future"[3], in conjunction with the "evidence-based legislation" [4] methodology and the "digital-ready policymaking"[5] approach. With this study we intend to improve the quality of the law-making process and of the content of each legislative regulation by investigating the following features: i) text clarity supporting legal drafters and end-user presentation; ii) linguistic variants and temporal versions management; iii) law-making/policy development process in decision making of the Commission supporting also amendments and consolidation; iv) metadata integration (ELI, ECLI, AKN, CDM, etc.) in the different steps of the law-making process; iv) modelling legal norms expressed in the legislative document; v) facilitation of the implementation of law by the Member

[1] E-mail: {monica.palmirani, francesco.sovrano2, davide.liga, salvatore.sapienza, fabio.vitali}@unibo.it

[2] https://ec.europa.eu/info/sites/default/files/better_regulation_joining_forces_to_make_better_laws_en_0.pdf

[3] https://ec.europa.eu/info/law/better-regulation/have-your-say-simplify_en

[4] https://ec.europa.eu/info/sites/default/files/better-regulation-toolbox.pdf

[5] https://joinup.ec.europa.eu/collection/better-legislation-smoother-implementation/digital-ready-policymaking

States. The study has the following goals: 1) reducing manual/error-prone work using patterns (e.g., corrigenda) and best practices templates during the legal drafting, to automatize as much as possible consolidation and semantic annotation, using legal ontologies and thesauri (e.g., EuroVoc); 2) maximising the reuse of similar legal concept detected using Machine Learning and legal data analytics applied to the whole legal system (e.g., definition, derogation); 3) favouring the implementation of some policies in the legislation (e.g., digital-ready, gender neutrality); 4) increasing transparency up to publication, thus increasing the searchability. In this light we have isolated[6], three main use-case scenarios and this paper aims to present the preliminary results of the first use case. The first use-case focuses on corrigenda and provides a clustering of them to understand which patterns could help the informatics tools (e.g., LEOS editor) to develop new relevant features to minimize errors and to improve the quality of legislation. This use-case also provides more information to the legal drafter.

2. Corrigenda in the EU Legislation and preliminary taxonomy

Corrigenda is a special modification necessary due to an error occurred in the official publication process. Since under theory of law it is a material error, not substantial, it has immediate efficacy since the beginning of the legislative act. The modifications of corrigenda are thus inserted in the first emission of the text, as if it had never been published differently. Corrigenda involve Directives, Regulation, and Decisions. For this reason, corrigenda need an immediate publication of the modificatory instructions on the official EU Official Journal and they are immediately implemented in the original text. Making a query to CELLAR[7] we get about 24.000 triples that connect each corrigendum to the document corrected, involving all the 24 official languages of the EU institutions, but only about 8.500 of them are connected to the English language variant. The corrigendum actions can be numerous, sparse across different points of the destinations, and they can also play a different semantic role, not only textual. The aim of this study is to isolate better the portion of the text involved (more granularity), to understand the legal role of the modification (e.g., temporal modification), to evaluate why they are frequent. We have prepared a light taxonomy of the quality of the modificatory instructions (25 classes) grouped in five *macro-areas*:

i) ***Structure modifications*** (e.g, provisions, annexes, footnotes, recitals, preamble, etc.)

```
On page 1, footnote 1:
for: '(1)    OJ L 145, 13.6.1977, p. 1. Directive as last
amended by Directive 2006/98/EC (OJ L 221, 12.8.2006, p. 9).',
read: '(1)    OJ L 145, 13.6.1977, p. 1. Directive as last
amended by Directive 2006/98/EC (OJ L 363, 20.12.2006, p. 129).
```

ii) ***Legal temporal information*** (e.g., date of efficacy, date of adoption)

```
On the cover page, on page 11 and page 12, adoption date:
for: '15 March 2021',
read: '15 February 2021'.
```

iii) ***Qualified portion of text*** (e.g., definitions, references, modification of modifications)

```
   On page 257, point (b) of the first paragraph of Article 112:
for: '(b)    Article 10 and points (a) and (b) of Article 12(1)
of Directive 98/79/EC, and …',
```

[6] Two focus groups composed by EU Commission legislative drafting offices, Open Data Office, Parliament of EU, Publication Office of EU were held using a questionnaire.

[7] http://publications.europa.eu/webapi/rdf/sparql

```
read: '(b)   Article 10, points (a) and (b) of Article 12(1) and
Article 15(5) of Directive 98/79/EC, and ...'.
```

```
   On page 98, Article 2(1)(18):
   for: '(18)"competent authority" means a competent authority
as defined in Article 2(1)(26) of Directive 2014/65/EU;',
   read: '(18)"competent authority" means a competent
authority as defined in Article 4(1)(26) of Directive
2014/65/EU;'
```

iv) ***Entities*** (e.g., role, places, number, organization, etc.)

```
On page 10, in the column 'COUNTRY OF ISSUE':
   for:'CZECH REPUBLIC',
   read: 'CZECHOSLOVAKIA'.
```

v) ***Presentational information*** (e.g., images, punctuation, publishing information)

```
On page 89, in the Annex, on the 12th line 'Austria', in the
second column:
for: '343 405 392',
read: '343 473 407'.
```

3. Dataset

The first step of the experiment was the dataset selection: all the corrigenda files in Formex 4.0, in English language, with the corresponding original file. The total number of corrigenda files is 2.513 documents, 3.478 pairs of modifier and modified text. The words in the old text are 87.906 and the words in the new text are 100.416. The average of the modifications for each correcting document is 1,81, but even corrigenda with 77 instructions of modification can be found. The second step was to convert these files in Akoma Ntoso including the CELLAR RDF information inside of a unique XML file that, despite not perfectly marked-up, is valid against the AKN-XSD schema or matches perfectly the AKN4EU specifications. This second step allows to have context, normative references, temporal parameters, metadata (e.g., ELI), modifications annotation qualifications in a unique consistent XML format. Publication Office supported the team of University of Bologna with the extraction operations.

4. Methodology

The methodology used in this work combines unsupervised clustering K-means enriched with Akoma Ntoso annotation and light-taxonomy information. At the end it is a mix of annotated text and unsupervised classification. Differently to many other research in the same field, we want to foster the structure information of the legal document (e.g., articles) and the light-taxonomy extracted using classic NLP techniques. Machine Learning (ML) approaches can classify a part of the legal text as 'definition', or a 'modification', or detect the 'date' included in the sentence but connecting all this information in a meaningful manner is quite difficult. Additionally, the same corrigendum could be classified in different ways: it can be a temporal modification, a table modification, or a definition modification. We intend to go beyond a pure classification methodology and to group in cluster the corrigenda modifications using the destination type (table, annex, normative provision, footnote, etc.), the type of modification (substitution, insertion, repeal), the text modified in relation with the old

text (when it is present), the role of the text modified (e.g., definition) and the temporal parameters (e.g., date of application). For this reason, the methodology is called hybrid and it mixes annotated validated information and unsupervised AI techniques. The mix of the two could permit to obtain a more semantic clustering that can be closer to the legal needs of the domain. The clustering may help the end-user and the tools to avoid the mistakes that produced the corrigenda. For permitting the interpretation we used KNIME as Data Analytics tool for comparing the clustering with some parameters: similarity distance, typology, granularity of the text of destination involved in the modification, and the typology of the document.

5. Hybrid Pipeline

The pipeline uses a hybrid approach, and it is composed by following steps:
a) ***Preliminary light-taxonomy of the corrigenda***: legal experts have analysed a random sample of corrigenda with a good balance between years and then they have created an agnostic taxonomy of the main modificatory events that is used by the technical team as the light-taxonomy needed for the classification. Legal experts have identified also good signals in the text for classifying the corrigenda using regular expressions. We have identified 25 classes; b) ***Conversion in Akoma Ntoso***: we have converted corrigenda documents from Formex 4.0 in Akoma Ntoso using Python and RegEx; c) ***Classification of the Corrigenda***: using simple NLP signatures we have classified the corrigenda using a light-taxonomy and the metadata of Akoma Ntoso. In this way we have assigned the qualification of each modification (e.g., substitution, insertion, repeal); d) ***Clustering of the Corrigenda***: we have created clusters of the corrigenda using K-means algorithm techniques; e) ***Distance of the text calculation***: we have calculated the distance between the old text and the new text using the Levenshtein distance; f) ***Data Analytics***: this step combines the results of the previous ones with AKN information to explain by user interfaces some interpretations, statistics, analyses using KNIME; g) ***Evaluation***: we set up a legal expert team composed by three members: two members check, and the third supervises them and resolves conflicting interpretations. The goal of this step is to evaluate the results of the clustering and of the Data Analytics work; h) ***Legal interpretation***: the legal experts use the diff-text and the graphs of the user-interface for providing a legal interpretation. In this step we also refine the light-taxonomy adding legal meaning. The same error could have different meanings and semantics depending also on the topic, so the legal interpretation is a fundamental part of the research.

6. Related Work

We have already converted several pre-existing document collections [10][14] in Akoma Ntoso, developed different NLP tools using patterns and RegEx rules [6][7] for extracting legal knowledge from the text (e.g., normative citations), classified legal text using ML or Deep Learning (DL) techniques [8][15]. Other researches have demonstrated the effectiveness of the ML/DL in the legal documentation fields [3][16][17][18][11] but without including the necessary semantic information for completing the context. The innovative approach in this work is to use hybrid architecture that uses unsupervised approach adding semantics [4][5] to the clustering results using light-taxonomy, NLP extraction, Data Analytics. The aim is to interpret the

output with the legal knowledge information supported by other techniques. We use data analytics tool (KNIME platform[8]) for providing information necessary to detect some best practices to suggest to legal drafters and software designers.

7. Akoma Ntoso Conversion of the corrigenda

We have converted Formex 4.0 in Akoma Ntoso in order to reach the following goals: a) **to detect the granular citations of the destination**. In Formex 4.0 this information is not present, and we have parsed the normative citations for representing the correct destination (e.g., article 23, paragraph 3, point a). This is relevant in order to provide the context of the semantic action of corrigendum. b) **to convert the modifications** in metadata that are not represented in Formex 4.0. The attributes @period that qualifies the span of time when the modification is valid, @old and @new that are also present in Formex 4.0 and the @destination with a precise specification.

8. Unsupervised Corrigenda Clustering

The pipeline we adopted to analyse the corrigenda consists in three main phases: i) *Feature Identification*; ii) *Dimensionality Reduction*; iii) and *Clustering*.
During the *Feature Identification* phase, we selected the pieces of information to be considered for clustering. In the present case, the opted for the following features, deemed to bear enough information to (unsupervisedly) push the clustering algorithm towards the structure of our taxonomy: a) the difference between the embeddings (representative numerical vectors obtained via [2]) of the corrigens and the corrigenda. This is crucial to the clustering on the semantic contents of the modifications; b) a set of booleans that indicate whether the description contains the keywords 'table', 'annex', 'recital', 'title', 'note'. This is used for clustering on basis of the modification's description. Considering that the resulting number of features may be large, depending on the characteristics of the embedding function that is used, we then perform one step of *Dimensionality Reduction*. *Dimensionality Reduction* is quite commonly used in conjunction with further clustering techniques, to foster better clusters. In our specific case, the number of features (about 773 in total, between features *a* and *b*) is arbitrarily reduced (to 50), removing the less significant ones through Principal Component Analysis[9]. Finally, after the *Dimensionality Reduction* we perform one step of automated *Clustering*. In our case, K-Means[10] is applied in an attempt to extract 25 different classes. The number is 25 because our reference taxonomy consists of 25 classes and the goal is to extract a clustering that is possibly aligned to it (see Figure 1). Twenty-four clusters are detected, and the most numerous clusters are C4 and C19.

9. Levenshtein Distance

We noticed also that the corrigenda often use significant portions of text, usually structured in hierarchical normative provision (e.g., article, paragraph, point), even if the real modification is limited to a few characters. For this reason, we have calculated the

[8] https://www.knime.com/knime-analytics-platform
[9] PCA is not necessarily the best technique to use, other can be envisaged.
[10] DBSCAN, OPTICS and many other clustering techniques appeared to not work very well in our case, making impossible to specify the final number of clusters to extract.

Levenshtein Distance (LD)[11] and we have discovered that than 81,4% involves parts of the text in excess respect the real needs (between 0,6 and 1). To evaluate the correctness of the Levenshtein distance the legal experts checked the text using a naïf diff algorithm written in Python for permitting a correct visualization to the legal expert team in agnostic way and not influenced by the previous tool of classification. We have also taken the Levenshtein distance, and we have made a comparison with other parameters including the type of provision of the text modified. Ultimately, we have noticed that the big partitions like article, table, annex, recital have a high index of LD (higher than 80%) with respect to little portions of text such as heading, number, reference.

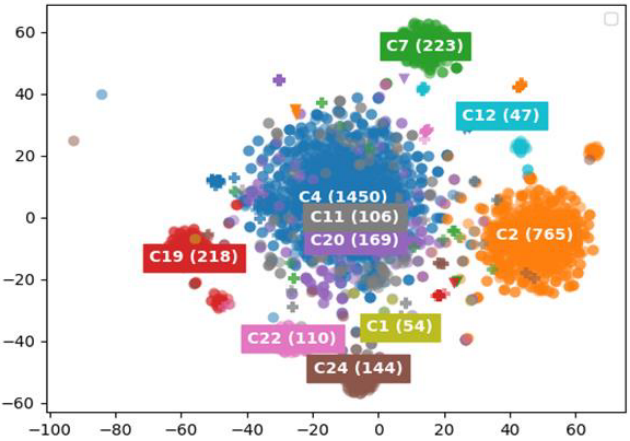


Figure 1 – Visualisation of the clusters we automatically extracted. This visualisation is obtained with tSNE.

10. Data Analytics

We added also other parameters of the data analysis with the goal to interpret the clustering made by the unsupervised algorithm. We have analysed the type of the modifications, and we have noticed a relevant concentration in the period 2004-2009, in correspondence of some of the most intensive period of the EU institutions (e.g., 2004 enlargement to ten new countries, 2009 Lisbon Treaty). We have also investigated the relationship between clusters, partition type and type of document and we have found a relevance between partition. It is contrariwise not influenced by the type of document even if "Regulation" is the higher for occurrences. For instance, cluster C01 seems to be aligned on footnotes. Since this interpretation was not entirely satisfactory, we opted to make the supervised follow-up annotation experiment.

11. Supervised Experiment

We built a dataset of 199 annotated corrigenda, according to the 25 identified classes. The corrigenda were randomly selected by one legal expert and then manually cross-annotated by two legal experts by relying on the 25 classes. The resulting dataset defines a multi-label text classification task. Considering the plethora of existing classifiers and the complexity of finding the right one, with the right configuration, we

[11] 0 means that old and new text diverge, 1 means that old and new are identical.

designed a tool that automatically searches for the best classifier within a pre-defined search space. This tool evaluates each possible classifier with a k-fold cross-validation (in our case, k=4). Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split. Each classifier is trained for a maximum of 10 epochs (it sees the same data a maximum of 10 times). The classifier with the highest F1-Macro (average calculated on k-folds) is kept as the best. The features considered by the classifier are: i) the difference between the embedding of the corrigenda and the corrigens; ii) a vector of 6 booleans that indicates if particular keywords can be found in the description: amend, recital, title, note, annex and table. The embeddings of the corrigenda and the corrigens were obtained through a deep language model called paraphrase-mpnet-base-v2 [19]. As classifier, we tried 1500 configurations of hyper-parameters of a shallow neural network with one regularised hidden layer of u units. We used a shallow neural network and a k=4 because the size of the dataset was small, therefore using a too large k would have resulted in very small test-sets, while a deep neural network would have clearly overfitted. These configurations were tested with an Async HyperBand Scheduler [20] performing a grid search on the following hyper-parameters: 1) batch size (2,3,4); ii) units (4,6,8,10,12); iii) activation function (None, relu, leaky_relu, selu, tanh); iv) learning rate (0.3, 0.1, 0.03, 0.01); v) regularisation strength (0.01, 0.003, 0.001, 0.0003, 0.0001). The best results were given by the following configuration: batch size: 3; units: 4; activation function: None; learning rate: 0.03; regularisation strength: 0.0001. This means that a linear classifier (activation function: None) suffices with the feature we used, and no complex deep learning models are needed. This linear classifier produced the following average results over the 4 folds:

1) F1-macro (the average F1-score for each class): 0.076 ± 0.001;
2) F1-weighted (the average F1-score for each class, weighted by its representativeness): 0.904 ± 0.007.

These results show that the dataset is unbalanced, meaning that some classes do not have enough datapoints, so the algorithm is not capable to recognize them. In fact, 12 of the 25 classes have less than 10 samples in the dataset, being significantly under-represented. Nonetheless, the algorithm can classify correctly the most represented classes.

12. Conclusions

Our conclusions[12] can be summarised as follows: 1) too much text is involved in the corrigenda that could produce new errors and it is then very difficult for the end-user to detect the new part of the text involved in the corrigendum. Also, the consolidated text offered by the EUR-LEX service is not granularly annotated and the legal expert needs to read in comparative manner the two texts; 2) the clustering operates on the basis of the type of provision involved in the modification and the type of modifications (e.g., C4 is mostly modifications at article level and with modification of the meaning); 3) the statistics detected an intense period of modifications between 2004 and 2009 and it is also natural considering the relative figures of the total number of legal documents emitted in this interval of time. We need to elaborate these findings to transform the

[12] See the dataset, the software, the output in https://gitlab.com/CIRSFID/AI4LegalDrafting

outputs in a policy to be provided to the legal drafters, decision-makers and to the technical team for improving the quality of the legislation. This work underlines also the difficulty to provide an interpretation and sound evidence of the meaning of the results coming from unsupervised ML and confirmed the hypothesis that a supervised hybrid architecture could help also in the task of explaining AI for a better transparency.

Acknowledgements. We thank Digit for Informatics of the European Commission - Study on 'Drafting legislation in the era of AI and digitization'- for supporting our research, and the European Publication Office for the extraction of data using CELLAR.

References

[1] AKN4EU https://ec.europa.eu/isa2/news/akoma-ntoso-eu-akn4eu-version-30-has-been-published_en
[2] Akoma Ntoso Version 1.0. Part 1: XML Vocabulary http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part1-vocabulary.html
[3] Branting K., B. Weiss, B. Brown, C. Pfeifer, A. Chakraborty, L. Ferro, M. Pfaff, and A. Yeh. (2019). Semi-Supervised Methods for Explainable Legal Prediction. In *ICAIL '19*. Association for Computing Machinery, New York, NY, USA, 22–31.
[4] Casanovas P. et al. (2016). Semantic Web for the Legal Domain: The Next Step. 1 Jan. 2016: 213-227.
[5] Francesconi E., Küster M., Gratz P., Thelen S. (2015). The Ontology-Based Approach of the Publications Office of the EU for Document Accessibility and Open Data Services. EGOVIS2015: 29-39.
[6] Gianfelice D., Lesmo L., Palmirani M., Perlo D., Radicioni D. P. (2013). Modificatory Provision Detection: a Hybrid NLP Apprach, in: ICAIL2013 Proceedings, ACM New York: 43-52.
[7] Liga D., Palmirani M. (2019). Classifying argumentative stances of opposition using Tree Kernels, in: ACAI 2019, New York, ACM: 17-22.
[8] Liga D., Palmirani M. (2020). Combining Tree Kernels and Tree Representations to Classify Argumentative Stances. ASLD@ISWC 2020: 12-23.
[9] Mandal A., Ghosh K., Ghosh S. *et al.* (2021). Unsupervised approaches for measuring textual similarity between legal court case reports. *Artif Intell Law* 29, 417–451 (2021).
[10] Palmirani M. (2018). Akoma Ntoso for Making FAO Resolutions Accessible. Law via the Internet 2018: 159-169.
[11] Robaldo L., Villata S., Wyner A. *et al.* (2019) Introduction for artificial intelligence and law: special issue "natural language processing for legal texts". *Artif Intell Law* 27, 113–115.
[12] Rossi A., Ducato R., Haapio H., Passera S., Palmirani M. (2019). Legal Design Patterns: Towards a New Language for Legal Information Design, in: Internet of Things. Proceedings of the 22nd International Legal Informatics Symposium IRIS 2019, Bern, Editions Weblaw, 2019, pp. 517 – 526.
[13] Song K, Tan X, Qin T, Lu J, Liu TY. Mpnet (2020). Masked and permuted pre-training for language understanding. arXiv preprint arXiv:2004.09297. 2020 Apr 20.
[14] Sovrano F., Palmirani M., and Vitali F. (2020) Deep learning based multi-label text classification of UNGA resolutions. In Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance (ICE-GOV2020).
[15] Sovrano F., Palmirani M., Vitali F. (2020). Legal Knowledge Extraction for Knowledge Graph Based Question-Answering. JURIX 2020: 143-153.
[16] Tagarelli A., Simeri A. (2021). Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code. *Artif Intell Law*.
[17] Waltl B., Bonczek G., Scepankova E. *et al.* (2019). Semantic types of legal norms in German laws: classification and analysis using local linear explanations. *Artif Intell Law* 27, 43–71.
[18] Huang Z., Low C., Teng M., Zhang H., Ho D. E., Krass M. S., and Grabmair M. (2021). Context-aware legal citation recommendation using deep learning. In *ICAIL21*. ACM, New York, 79–88.
[19] Reimers N., and Gurevych I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. arXiv preprint arXiv:2004.09813.
[20] Li L., et al. (2018). A system for massively parallel hyperparameter tuning. arXiv preprint arXiv:1810.05934.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210320

Improving Legal Case Summarization Using Document-Specific Catchphrases

Arpan Mandal [a], Paheli Bhattacharya [b], Sekhar Mandal [a], Saptarshi Ghosh [b]
[a] *Indian Institute of Engineering Science and Technology Shibpur, India*
[b] *Indian Institute of Technology, Kharagpur, India*

Abstract. Legal case summarization is an important problem, and several domain-specific summarization algorithms have been applied for this task. These algorithms generally use domain-specific legal dictionaries to estimate the importance of sentences. However, none of the popular summarization algorithms use *document-specific catchphrases*, which provide a unique amalgamation of domain-specific and document-specific information. In this work, we assess the performance of two legal document summarization algorithms, when two different types of catchphrases are incorporated in the summarization process. Our experiments confirm that both the summarization algorithms show improvement across all performance metrics, with the incorporation of document-specific catchphrases.

Keywords. Legal case document, Summarization, Catchphrases

1. Introduction

Summarization of legal case documents is an important problem and has been well-studied by researchers [1–4]. Summarization algorithms can be of different types, viz. *extractive* vs. *abstractive*, *unsupervised* vs. *supervised*. Most summarization algorithms developed for the legal domain are extractive and unsupervised in nature, mainly due to the lack of large training data in the legal domain.

These algorithms being extractive in nature, attempt to assign a likelihood-score to each sentence of a document, and choose the top-scoring sentences as the summary of the document. While measuring the score of a sentence (the likelihood of the sentence to be included in the summary), various summarization methods consider either or both of two factors – (1) *document-specific importance* of the sentence with respect to other sentences in the document, (2) *domain-specific importance* of the sentence, e.g., several legal domain-specific algorithms use an external set of legal terms (a legal dictionary) and consider the number of legal terms contained in a sentence [2,3].

Although these two factors are independently used to characterize the likelihood-score of a sentence, we hypothesize – *"an appropriate amalgamation of document-specific and domain-specific importance may provide new useful information to the summarization algorithms, which can subsequently improve*

their performance". This combined information can be provided by the use of ***document-specific catchphrases*** that are a set of short (one-word or multi-word) phrases that collectively provide a concise representation of a legal document [5–8]. These catchphrases are not only legal domain-specific important terms, but also terms or phrases that have document-specific importance. Although domain-specific dictionaries have widely been used in summarization algorithms [2,3], catchphrases are different from domain-specific dictionaries in that they also capture document-specific important terms (which may not be legal keywords).

In this work, we investigate whether using document-specific catchphrases can improve the performance of legal document summarizers. To this end, we use two different types of catchphrases – extracted from a case document by two methods *PSLegal* [9] and *D2V-BiGRU-CRF* [10] (details in Section 2) – to aid two legal-specific summarization techniques – *DELSumm* [3] and, *CaseSummarizer* [2] (details in Section 3). We conduct experiments over a set of case documents from the Indian Supreme Court (details in Section 4). Our experiments demonstrate that, the performances of both the summarization algorithms improve when document-specific catchphrases are incorporated.

2. Related Works

In this work, we propose to use document-specific catchphrases to improve legal case document summarization. In this section, we survey some catchphrase detection methods and case document summarization methods.

2.1. Legal catchphrase extraction

Several catchphrase detection methods have been developed for legal documents [9–11]. We briefly discuss two catchphrase extraction methods developed in our prior works, both of which provide meaningful catchphrases that agree with those chosen by law domain experts [9,10].

PSLegal [9] – an unsupervised method: Given the text of a document d, this method involves two major steps to extract the set of catchphrases:
Step 1: Some *candidate phrases* are extracted from d. These are actually *noun phrases* extracted using a customized set of grammatical rules (details in [9]).
Step 2: Next, an appropriate scoring function takes as input the text of d and a candidate phrase c, and computes the likelihood for c to be a catchphrase for the document d. The scoring function takes into account three factors – (1) document-specificity of the phrase c, (2) domain-specific importance of c, (3) presence of a predefined legal term within the phrase c. The final PSLegal score is the product of these three factors. All candidate phrases are scored using this scoring function and then 10% of the top-scored ones are chosen as the catchphrases for the given document. Further details of the method can be found in [9], and a ready-to-use implementation of this algorithm is available at `https://github.com/amarnamarpan/PSLEGAL`.

D2V-BiGRU-CRF [10] – a supervised method: This is a *neural sequence tagging model* that has the ability to be trained over a relatively small training dataset

(typically, a few hundred documents and their gold standard catchphrases) and then the trained model can be used to extract catchphrases from unseen case documents. It takes as input a sequence of words and identifies each word to be either a part of a catchphrase or not.

The D2V-BiGRU-CRF architecture (details in [10]) employs these layers – (1) a bidirectional language model [12] that extracts word and character embeddings. (2) a BiGRU layer that combines both the embeddings, (3) a fully connected layer that learns a representation of the outputs of the BiGRU layer and a Doc2vec [13] embedding of the input document, and (4) a Conditional Random Field (CRF) layer which predicts whether a word is part of a catchphrase or not. It was demonstrated in [10] that the D2V-BiGRU-CRF extracts catchphrases that match well with those selected by law domain experts. An implementation of this model along with our trained model is available at https://github.com/amarnamarpan/D2V-BiGRU-CRF.

2.2. Summarization of court case documents

Many general (domain-independent) text summarization algorithms have been used for summarization of legal case documents, e.g., in [3,14]. Additionally, many algorithms have been developed specifically for the summarization of legal case documents, such as LetSum [15], K-mixture model [4], CaseSummarizer [2] and DELSumm [3]. Out of these, in this work, we use document-specific catchphrases to improve the performances of CaseSummarizer [2] and DELSumm [3].

3. Incorporating catchphrases in legal case summarization algorithms

In this section, we describe two unsupervised, extractive summarization algorithms built for summarizing legal case documents – (1) DELSumm [3] and, (2) CaseSummarizer [2] – and how each of these algorithms can be modified to incorporate document-specific catchphrases.

3.1. **DELSumm**

DELSumm [3] is a recently developed algorithm that models the problem of summarizing a case document as maximizing an Integer Linear Programming (ILP) objective function that maximizes the inclusion of the most informative sentences in the summary. As input, DELSumm takes - (1) a case document where each sentence is marked with its rhetorical role, out of the eight roles (e.g., Facts, Arguments, Ratio of the decision, Ruling), and (2) the desired length L of the summary in words. The output is a summary whose length is at most L, and contains sentences from each of the rhetorical roles. The algorithm considers a set of guidelines suggested by law experts as to how the different rhetorical segments of a case document should be summarized. To judge the informativeness of a sentence, the algorithm considers, among other factors, a set of *content words* which are basically terms in a legal dictionary compiled from various sources. More details can be found in [3].

Incorporating catchphrases in DELSumm: We replace the legal content words (described above) by a set of catchphrases extracted from the input document

(that is to be summarized). In the original DELSumm, the set of content words remains the same for every document. Whereas, now the algorithm gets modified in a way whereby, while constructing the summary, it gives more importance to the sentences that contain catchphrases specific to each input document.

3.2. CaseSummarizer

CaseSummarizer (CaseSumm in short) [2] uses a specialized scoring function to score each sentence in a case document, and then chooses the highest scored sentences to build the summary. To build the scoring function for sentences, it considers three factors – (1) *the occurrence of known important entities in a sentence (the entities were marked by domain experts in the original work [2])*, (2) the occurrence of dates in a sentence, and (3) the proximity of a sentence to section headings. First, an initial score/weight is computed for each sentence by summing the TF-IDF scores of the constituent words and normalizing over the the sentence length; this score is called w_{old}. A new score w_{new} is then computed for a sentence as $w_{new} = w_{old} + \sigma(0.2d + 0.3e + 1.5s)$ where, d is the number of dates in the sentence, e refers to the number of entities, s is a boolean variable specifying whether the sentence is the first sentence in a section, σ is the standard deviation between the scores of all sentences, and the coefficients are selected through trial-and-error and feedback from experts.

Incorporating catchphrases in CaseSumm: We modify CaseSumm by incorporating document-specific catchphrases (e.g., those generated by D2V-BiGRU-CRF or PSLegal) in place of the *entities* in the document. More specifically, in the expression for w_{new} stated above, we replace the term e (which signifies the number of entities in a sentence, in the original algorithm) with the number of document-specific catchphrases contained in the sentence.

4. Dataset and Experimental Results

Dataset for evaluation of summarization performance: We reuse the dataset and evaluation setup from our recent work [3]. The dataset consists of 50 legal case documents from the Indian Supreme Court, along with their summaries (of length approximately one-third of the original document lengths) written by two domain experts (senior Law students from the Rajiv Gandhi School of Intellectual Property law, one of the most reputed Law schools in India).[1] As document summarization is a subjective task, the two experts wrote two separate summaries; all scores presented in the paper are averaged over the two gold standard summaries written by the two experts.

Metrics for summarization performance: We compare the match between the algorithmic summaries and the gold standard summaries, and report Rouge-2 (considers bigram matches) and Rouge-L (considers Longest Common Subsequence matches) Recall and F-scores. All scores for a particular document are averaged

[1] Also, each sentence in the documents is labeled with its rhetorical role by the same experts; these labels are used by DELSumm (details in [3]).

Table 1. Comparing the performance of original DELSumm (abbreviated as DLS) and DELSumm with different variations of catchphrases. DBC: catchphrases extracted by D2V-BiGRU-CRF; PSL: catchphrases extracted by PSLegal; Ldict: the Legal Dictionary used by the original summarization algorithm. The highest value for each metric is in bold-fonts.

	Rouge-2 R	Rouge-2 F	Rouge-L R	Rouge-L F
Original DELSumm (DLS)	0.4323	0.4217	0.6831	0.6017
DLS with DBC	0.4588	0.4411	0.6892	0.6102
DLS with DBC & Ldict	0.4557	0.4372	**0.6909**	0.6096
DLS with PSL	**0.4593**	**0.4435**	0.6763	0.6111
DLS with PSL & Ldict	0.4479	0.4343	0.6805	0.6105
DLS with PSL & DBC	0.4574	0.4422	0.6757	0.6103
DLS with PSL, DBC & Ldict	0.4509	0.4365	0.6828	**0.6118**

Table 2. Comparing the performance of original CaseSumm (CSM) and CaseSumm with different variations of catchphrases. The highest value for each metric is in bold-fonts. The terms DBC, PSL and Ldict are as explained in the caption of Table 1

	Rouge-2 R	Rouge-2 F	Rouge-L R	Rouge-L F
Original CaseSumm (CSM)	0.3198	0.3636	0.5415	0.5343
CSM with DBC	0.3258	0.3726	0.5490	**0.5426**
CSM with DBC & Ldict	**0.3265**	**0.3738**	**0.5493**	0.5425
CSM with PSL	0.3221	0.3690	0.5465	0.5397
CSM with PSL & Ldict	0.3221	0.3690	0.5465	0.5397
CSM with PSL & DBC	0.3220	0.3689	0.5463	0.5396
CSM with PSL & DBC & Ldict	0.3229	0.3702	0.5473	0.5411

over the two gold standard summaries written by the two experts for the document (as stated above).

Variations of the summarization methods tried: As was explained in Section 3, we shall compare the performance of a summarization algorithm (DELSumm / CaseSumm) when used in its original setting (with an in-built legal dictionary, referred to as 'LegDict'), and when used with document-specific catchphrases. We have two kinds of catchphrases – (1) those generated by PSLegal (referred to as PSL), and (2) those generated by D2V-BiGRU-CRF (referred to as DBC). Thus, for a summarization algorithm (say, S) that originally uses the legal dictionary 'LegDict', we experiment with the following variations: (1) Original S with the default LegDict, (2) S with only DBC, (3) S with DBC and LegDict, (4) S with only PSL, (5) S with PSL and LegDict, (6) S with both DBC and PSL, (7) S with all three – both type of catchphrases, and the LegDict.

In variations (2) and (4) stated above, the catchphrases identified by D2V-BiGRU and PSLegal respectively are used *in place of* the default LegDict in the original summarization algorithms. Whereas, in the variations (3), (5), and (7), the catchphrases identified by D2V-BiGRU and/or PSLegal are used in conjunction with LegDict.

Results: Table 1 shows the performance of the original DELSumm and DELSumm with different variations of catchphrases (as explained above). We see that, in all cases, use of document-specific catchphrases leads to better summarization than what is achieved by the original DELSumm. The best performance is achieved when DELSumm is used along with catchphrases identified by PSLegal – in this setting, the Rouge-2 F-score increases from 0.4217 (for the original DEL-

Summ) to 0.4435. Note that the original DELSumm already out-performs several other summarization methods on this dataset, as was shown in [3]. Incorporating document-specific catchphrases improves the summmarization even further.

Similarly, Table 2 shows the performance of the original CaseSummarizer and its variations with different sets of catchphrases. In all cases, use of document-specific catchphrases leads to better summarization than what is achieved by the original CaseSummarizer.

5. Conclusion

We show that using document-specific catchphrases can improve the performance of existing summarization algorithms while summarizing legal case documents.

While the present work considers only unsupervised and extractive summarization algorithms, in future, we plan to explore ways of improving *supervised* and *abstractive* summarization algorithms using catchphrases.

References

[1] Erkan G, Radev DR. LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization. J Artif Int Res. 2004;22(1):457–479.
[2] Polsley S, Jhunjhunwala P, Huang R. CaseSummarizer: A System for Automated Summarization of Legal Texts. In: Proceedings of COLING 2016; 2016. .
[3] Bhattacharya P, Poddar S, Rudra K, Ghosh K, Ghosh S. Incorporating Domain Knowledge for Extractive Summarization of Legal Case Documents. In: Proc. ICAIL; 2021. .
[4] Saravanan M, Ravindran B, Raman S. Improving Legal Document Summarization Using Graphical Models. In: Proc. of the Conference on Legal Knowledge and Information Systems: JURIX 2006. IOS Press; 2006. p. 51–60.
[5] Galgani F, et al. Towards Automatic Generation of Catchphrases for Legal Case Reports. In: Proceedings of Computational Linguistics and Intelligent Text Processing; 2012. .
[6] Witten IH, et al. KEA: Practical Automatic Keyphrase Extraction. In: Proceedings of the Conference on Digital Libraries; 1999. p. 254–255.
[7] Liu Z, et al. Clustering to Find Exemplar Terms for Keyphrase Extraction. In: Proc. of the Conference on Empirical Methods in Natural Language Processing; 2009. p. 257–266.
[8] Wu YFB, Li Q. Document Keyphrases as Subject Metadata: Incorporating Document Key Concepts in Search Results. Information Retrieval Journal. 2008;11:229–249.
[9] Mandal A, Ghosh K, Pal A, Ghosh S. Automatic Catchphrase Identification from Legal Court Case Documents. In: Proc. ACM Conference on Information and Knowledge Management (CIKM); 2017. .
[10] Mandal A, Ghosh K, Ghosh S, Mandal S. A sequence labeling model for catchphrase identification from legal case documents. Artificial Intelligence and Law. 2021 June.
[11] Tran VD, Nguyen ML, Satoh K. Automatic Catchphrase Extraction from Legal Case Documents via Scoring using Deep Neural Networks. CoRR. 2018;abs/1809.05219.
[12] Peters M, et al. Deep Contextualized Word Representations. In: Proceedings of the North American Chapter of the Association for Computational Linguistics; 2018. p. 2227-37.
[13] Le Q, Mikolov T. Distributed Representations of Sentences and Documents. In: Proceedings of International Conference on Machine Learning; 2014. p. 1188-96.
[14] Bhattacharya P, Hiware K, Rajgaria S, Pochhi N, Ghosh K, Ghosh S. A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments. In: Advances in Information Retrieval; 2019. p. 413-28.
[15] Farzindar A, Lapalme G. LetSum, an automatic Legal Text Summarizing system; 2004. .

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210321

Towards Reducing the Pendency of Cases at Court: Automated Case Analysis of Supreme Court Judgments in India

Shubham PANDEY [a], Ayan CHANDRA [a], Sudeshna SARKAR [a] and Uday SHANKAR [a]

[a] *Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India*

Abstract. The Indian court system generates huge amounts of data relating to administration, pleadings, litigant behaviour, and court decisions on a regular basis. But the existing Judiciary is incapable of managing these vast troves of data efficiently that causes delays and pendency of a large volume of cases in the courts. Some of these time-consuming tasks involve case briefing, examining the legal issues, facts, legal principles, observations, and other significant aspects submitted by the contending parties in the court. In other words, computational methods to understand the underlying structure of a case document will directly aid the lawyers to perform these tasks efficiently and improve the overall efficiency of the Justice delivery system. Application of Computational techniques (such as *Natural Language Processing*) can help to gather and sift through these vast troves of information, identify patterns, extract the document structure, draft documents and make the information available online.

Traditionally lawyers are trained to examine cases using the *Case Law Analysis* approach for case briefing. In this article, the authors aim to establish the importance and relevance of the automated case analysis problem in the legal domain. They introduce a novel case analysis structure for the supreme court judgment documents and define twelve different case law labels that are used by legal professionals to identify the structure. Finally the authors propose a method for automated case analysis, which will directly aid the lawyers to prepare speedy and efficient case briefs and drastically reduce the time taken by them in litigation.

Keywords. Law and Technology, AI in Law, Natural Language Processing (NLP), Legal Document Analysis, Case Analysis

1. Introduction

Case Analysis refers to the study of court cases and drawing essential conclusions on how courts have applied certain norms and how they interpret them in accordance with law. This is one of the most crucial parts of the training of all legal professionals to understand a judgment given by a court. Naturally, to analyse the cases heard at the Indian Courts, Case Law Analysis (CLA) technique is used by the lawyers. Thus, an automated approach to find different case analysis roles from a judgment document can directly help the lawyers in their study and preparing the case briefs. Furthermore, identifying the case analysis roles of the sentences in a judgment can also aid in many downstream tasks

such as semantic search, summarization etc. However, there is a set of reasons which make this task computationally difficult such as the length of the judgment text, lack of understanding of the domain knowledge incorporated in the computational approaches etc. We introduce and define different case analysis roles with examples in the upcoming sections.

In this work, we aim to achieve the following objectives: (a) *Establishing case analysis as an important problem in the legal domain*. (b) *Defining a case law analysis structure for the supreme court judgment documents*. (c) *Proposing an example automated case analysis system for the legal community*.

1.1. Importance of the Automated Case Law Analysis

The pendency of a large number of cases is a long ailing problem of the Indian Judiciary. The Indian court system consists of the Apex Court or the *Supreme Court of India*, the *High Courts* in the individual states and union territories, and the *lower courts* at the district level. The recent data estimates (as projected by National Judicial Data Grid [1] on August 21, 2021) show that a total of 36,780,460 cases are pending in the various courts. Even though the lower courts dispose of more than half of the new cases filed (56%) within a year, a significant number of these cases are either transferred to the higher courts, or the verdicts are challenged in the higher courts. Thus, it increases the overall volume of pending cases in the High courts and the Supreme court. Furthermore, the ongoing COVID-19 pandemic has impacted the judicial system severely causing further delays in the Justice delivery process.

Researchers, lawyers, government, and administration are working tirelessly to find ways to reduce the pendency of cases and improve the overall efficiency of the legal system. For every case, the advocates of the contending parties have to submit/put forward arguments on behalf of the respective parties, after going through all similar prior documents and the legal principles. Based on the available information on the current case and a set of similar prior documents (also known as *"precedents"*), they prepare a case briefing. Often, these case documents are thousands of pages long and it takes a considerable amount of time for them to prepare the briefing. Again, when the judge gives the verdict of the presented case, all the legal issues, facts, arguments provided by the advocates are examined by the court. For both processes, case analysis is one of the most used techniques or methods available to legal practitioners. Therefore, proposing approaches to solve the task of automated case analysis will directly improve the efficiency of the courts and help in mitigating the bigger problem by reducing the pendency of cases.

In this work, we define a case analysis structure for the supreme court judgment documents with the help of the legal professionals and propose a rule-driven system by which one can perform the task. This is a significant contribution towards the legal community to reduce the pendency and backlogs in the court-rooms.

2. Limitations of Prior Work: Comparison between Case Analysis and Rhetorical role labelling

Earlier works [2] [3] assumed that case analysis can be performed using Rhetorical Role labeling of the legal document. Rhetorical role labeling of a sentence from a legal text is

a means to understand the semantic function of the sentence (such as arguments of the parties, the final judgment, background, statute and so on).

When we compare the rhetorical roles and the case law analysis information that a legal professional requires, we observe various limitations. We note that: **(i)** Certain key information labels such as "Legal issue" and "Observation" are neither present in the existing rhetorical structure nor can be derived from any of those. **(ii)** The information whether a sentence is an argument is not sufficient to a lawyer, but whether the argument is submitted by the appellant, respondent, or amicus is important. Furthermore, the party or advocate is mentioned once in the text and thereafter referred to by appropriate pronouns in consecutive sentences. Hence, both the semantic and case-related auxiliary information (named entities) and their anaphora and coreferences [5] are equally important in case analysis. **(iii)** Any mention of established law (e.g, Acts, Sections, Articles, Rules, Notices, etc.) is treated as a statute. However, not all the mentions of established laws in a judgment text are relevant to the analysis of the given case. In particular, the set of legal principles which are discussed for the current judgment are deemed important. Legal principles may include not only the established laws but also the mention of enumerated rights and unenumerated rights. e.g. "Right to clean air", "Right to education" are not statutes but these are some of the examples of enumerated rights which can be applied in a judgment by the court, and mentioned with its reason for application. Hence, the understanding of legal principle is different from that of the rhetorical role "statute". **(iv)** For long-running cases, the court often provides various interim orders to provide relief to one of the contending parties. These relief prayers asked by the parties, interim orders given by the court from time to time, and whether parties have complied with earlier orders become a crucial subject of examination in later judgments. Thus, relief prayer, interim order, and compliance also become part of the reasoning, and therefore, demands a separate label in the case analysis process.

Above observations eventually lead us to establish the automated case analysis of judgment documents as an important task in the legal domain different from the rhetorical role labeling.

3. Case Analysis Labels

After going through a large volume of court documents, we have come up with the following twelve case analysis roles. We describe the roles in the following:

1. **Legal Issues**: The legal issue or question that the Court is adjudicating upon in the current case. e.g. *"The issue for our consideration today is fixing standards for 34 industries with regard to the SO2, NOX and SOX emissions."*
2. **Argument by Appellant:** Different arguments submitted by the appellants or the advocates on their behalf in the court. e.g. *"Mr. Sundaram finally submitted that since none of the grounds given by the High Court in the impugned judgment for directing closure of the plant of the Appellants are well-founded, ..."*
3. **Argument by Respondent:** Different arguments submitted by the respondents or the advocates on their behalf in the court. e.g. *"It was pointed out by Shri Mukul Rohatgi, learned senior Counsel appearing on behalf of Government of NCT of Delhi that by applying odd-even scheme with respect to cars alone cannot be said to be a wholesome*

solution."

4. **Argument by Amicus Curiae:** Sometimes, the court invites third party members or organisations to provide insights to a case that is being discussed. Such third parties are legally known as "Amicus Curiae". Amicus curiae may also submit its arguments through its advocates. e.g. *"Mr. Harish N. Salve, learned senior Counsel appearing as amicus, argued that imposition of ECC and the directions issued by this Court regarding diversion of commercial vehicles/trucks to alternative routes has made some difference.."*

5. **Relief Prayer:** If a case is continuing for long duration, one of the parties may seek for immediate relief from certain inconveniences, before the final verdict comes out. e.g. *"The Petitioners have approached this Court seeking emergent reliefs in relation to the extreme air pollution in the National Capital Region (hereinafter 'NCR')"*

6. **Observation Findings:** Observations made by the court while assessing the facts. There is no immediate relation between the observation and the conclusion of a case. But these observations may be important in understanding how the court weighs the arguments submitted by different parties. e.g. *"We must note that there has been no response from the States within the NCR giving the impression that air pollution is not a problem for the State Governments despite the ill-effects and health hazards of bursting fireworks"*

7. **Legal Principles:** In the judgment, the court may state different Enumerated and Unenumerated rights, established laws (statutes, acts, constitution) etc. These principles are crucial for establishing the conclusion of the judgment. Enumerated rights are the rights given by the established laws and constitution (e.g. fundamental rights). Unenumerated rights are the rights that are coming from case laws [earlier verdicts from different important cases.] , and may become part of established laws later. In general, a sentence that represents a legal principle does not include any mention of appellant, respondent or any parties involved. e.g. *"The polluter pays principle demands that the financial costs of preventing or remedying damage caused by pollution should lie with the undertakings which cause the pollution, or produce the goods which cause the pollution"*

8. **Fact:** Fact refers to the chronology of events that led to filing the case, and how the case evolved over time in the court system. Facts are the sentences that have pieces of information about appellant/respondent/other parties involved or the current court but it does not include any argument, legal principle, verdict, direction or causation in itself. e.g. *"The Appellants are the owners of Hotels, Beach Resorts and Beach Bungalows in Goa who have been facing the prospect of demolition of their properties for the last several decades"*

9. **Rationale:** Rationale is the reasoning on how legal principle and facts can have a causation relationship. A rationale has two parts in it: One part includes a Legal Principle, and the other part includes a Fact. e.g. *"Under the principle of 'delegatus non potest delegare', the delegatee (the Chairman of the Board) could not have further delegated the authority vested in him, except by a clear mandate of law."*

10. **Conclusion Verdict:** The decision or verdict, conclusion of the Court, or, texts that can be considered as part of the verdict. If the document is of type Order, then orders or directions mentioned in the text can be considered as parts of the final verdict. e.g. *"we direct the States of Punjab, Haryana and Uttar Pradesh to disburse the money and they should not wait for or write letters to the Central Government to give certain funds for this purpose."*

11. **Interim Order:** If a case is continuing for long duration, the court may issue interim

orders to grant relief to the parties temporarily if the parties seek relief. In general, interim order can be found along with relief prayer in an Order document. e.g. *"We accordingly direct the Government of NCT of Delhi to take immediate steps for repair of pavements and make pavements wherever the same are missing and also to take immediate steps for ..."*

12. **Compliance:** Compliance signifies if one of the parties involved in the case has complied to the directions given by court or any earlier order or direction given by other organisations. A non-abiding of compliance may become a part of the Observation which may influence the verdict. e.g. *"Unit has complied with the conditions and the consent order issued to the Unit."*

4. Supreme Court of India corpus of Air Pollution cases

We create a corpus of 28 Supreme Court of India corpus of Air Pollution cases annotated with the case analysis structure. We consider the legal judgments that are related to Air Pollution cases under the Environment category from the Supreme court of India cases gathered from Manupatra [6], a reputed online journal for the period January 2010-January 2021. We take the cases which are "Against the government" i.e. Union of India is one of the respondents in those cases. No proprietary information is taken in this process.

Initially one annotator with domain understanding annotated the documents. Each sentence from the judgment documents is labelled with a single argument label from the set of 12 case analysis roles and "Other" (if the sentence is not useful in the case analysis process). Then a panel of legal experts reviewed the annotations and rectified it wherever necessary based on consensus decision. As the review of annotation was conducted based on the unanimous decision of a panel of legal professionals from *Rajiv Gandhi School of Intellectual Property Law, IIT Kharagpur* [1] with varied academic and court experiences, the quality of the annotation may be considered to be high.

5. System Overview

For the identification of the case analysis roles, the judgment copy is required as input. A judgment copy contains the judgment text, list of appellants, respondents, amicus if any, advocates of the contending parties, name of the judges, file number of the case, etc. Our proposed automated case analysis system consists of the following of steps.

Step 1: Extraction of the case metadata - The case metadata includes the date of the decision, the document type (e.g. Judgment, Order, etc.), the name of the judges, party names, and the name of the advocates of the parties. We segregate the case metadata from appropriate places in the judgment copy using position-based rules and store the segregated components in a configuration dictionary. We also consider various abbreviations and short names for party names and advocates in the dictionary.

Step 2: Entity Identification - In the judgment text, named entities such as the name

[1] RGSOIPL, IIT Kharagpur: http://www.iitkgp.ac.in/department/IP

of the appellants or respondents, name of the advocates, etc. are mentioned once and thereafter referred by appropriate pronouns in consecutive sentences or before the entities in the same compound sentence. In linguistics, this is known as co-reference and if the pronoun that refers to the entity comes after the mention of the entity, then it is called an anaphora. In order to understand which named entity that pronoun refers to, anaphora and co-reference resolution [5] are required. e.g. in the sentence *"He submits that the depending upon the improvement and the extent of collection of ECC, post installation of RFID, this Court could issue appropriate directions suitably balancing the equity among the State and the stakeholders."* , "He" refers to "Mr. Salve, learned Amicus" who was one of the advocates of the amicus in the given example. We use Neuralcoref [4] for the co-reference resolution as it performed well on our corpus. We replace each mention of the parties and their advocates, court, etc. in the resolved text with corresponding identifier tokens using the configuration dictionary. These identifier tokens will be used in different rules to identify the case analysis role.

Step 3: Syntactic Analysis - Syntactic structure of a sentence (such as the nature of the verb, and the relation of different noun phrases with the verb) plays a key role in constructing various rules for role identification. So, we perform parts of speech (POS) tagging and dependency parsing on the sentences from the text. We use Spacy [8] for both the tasks because it showed better results which agrees with [7].

Step 4: Rules Identification - Based on the annotated examples and the domain knowl-

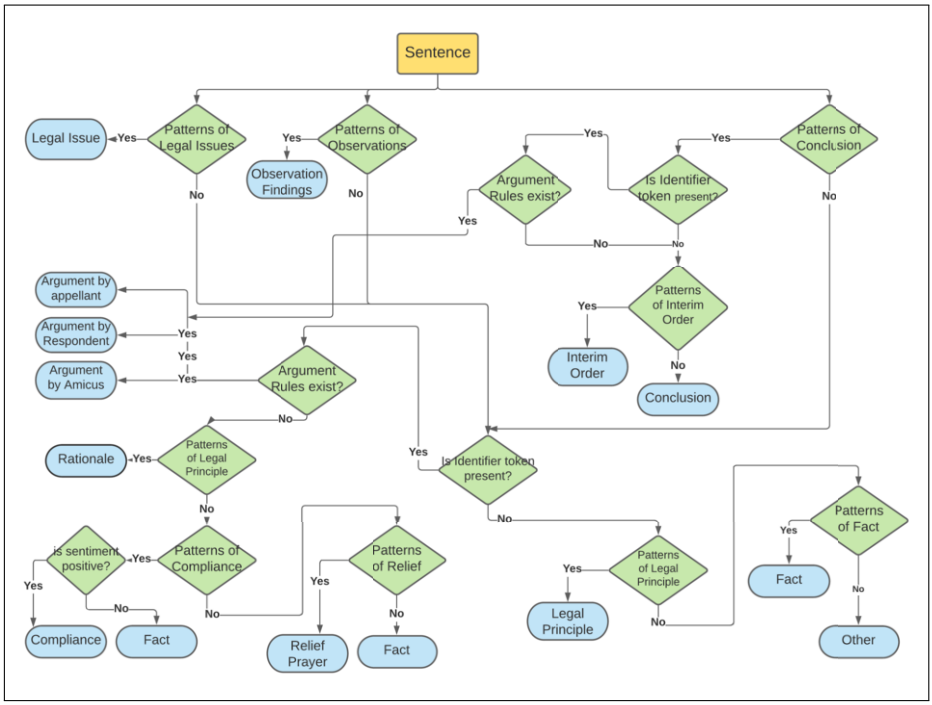


Figure 1. Flow chart for determining the case analysis role of a sentence from a given judgment text.

edge of the legal experts, we attempt to find the rules for each of the case roles. In Figure

Table 1. Precision, Recall, and F-score for individual case analysis role from law documents of test set predicted by our method

Case analysis role	P	R	F1	Support
Observation Findings	1.00	0.61	0.76	18
Conclusion Verdict	0.95	0.80	0.87	45
Relief Prayer	1.00	1.00	1.00	6
Argument by Amicus	1.00	0.62	0.77	8
Argument by Appellant	1.00	0.77	0.87	22
Argument by Respondent	1.00	1.00	1.00	10
Fact	0.97	0.83	0.90	132
Interim Order	1.00	1.00	1.00	1
Legal Principle	0.90	1.00	0.95	9
Legal Issue	1.00	1.00	1.00	3
Rationale	0.40	1.00	0.57	2
Compliance	1.00	1.00	1.00	1

1, we describe the process by which the case analysis role of a sentence is determined. We use a combination of various regular expressions and syntactic analysis in this step.

We find the rules after extensive study of 24 documents from the corpus with the help of the legal experts and report the results on 4 additional documents. The rules for the labels are made available for the interested readers [2].

6. Results

We report the Precision (P), Recall (R), and F1-score (F1) for each of the case analysis roles on the test set of 4 documents in Table 1. For some of the case roles, precision and recall are quite high. We discuss about the issues with the existing rules in the following subsection. A list of examples are given in the supplementary material for the readers[3].

6.1. Issues:

We went through the set of 24 legal documents and test set of additional 4 documents from the corpus and manually examined the case analysis roles of the sentences. Following issues are observed: **(i)** In some of the paragraphs, the third person pronouns are not properly co-referenced to the appropriate named entities. So, those sentences do not have any appropriate identifier token (e.g. appellants, etc.). As a result, the rules-identification module fails to identify these sentences properly. If we improve the co-reference resolution model, this issue can be mitigated for most of such examples. **(ii)** In case of interim orders, it was difficult to evaluate whether a sentence can be identified as an interim order or a conclusion, even though reliefs are asked. **(iii)** In certain documents, all the parties involved or heard in the case are not mentioned in the judgment copy beforehand. E.g. SDMC (South Delhi Municipal Corporation) and NDMC (North Delhi Municipal Corporation) are not mentioned as any party in the case MANU/SC/0609/2016 . When the rules identification module encounters the sentence *"These applications have been filed*

[2]Rules Description: https://github.com/legalArgMining/Case-Law-Analysis
[3]Same as above

by SDMC and NDMC seeking permission for registration of diesel vehicles used for collection and transportation of garbage on diesel based HCV and MCV vehicles.", it could not classify it into any existing roles and labels it as *"Other"*. However, a manual examination of the case conveys that both SDMC and NDMC are parties involved in the case. Therefore, this sentence should be treated as a "Fact". **(iv)** For certain sentences, "Fact" has been mis-predicted as "Rationale". The rules identification module rightly found the presence of both a legal principle and a Fact part in the sentence. However, the sentence conveys a generic statement and violation of certain legal principle and therefore, should be classified as a "Fact". For such cases, we need to incorporate specific exclusion rules. **(v)** In a few documents, the subject of the head word in the sentence is not explicitly mentioned. Such sentences are usually in passive voice. E.g. *"It is also submitted that.."* . In this sentence, the *agent* "by" of the head word verb "submitted" is not mentioned. As a result, the rules fail to identify these sentences properly.

7. Conclusion and Future Works

In this work, we introduce a case analysis structure for the Indian supreme court judgment documents with suggestions from the legal community, and show some initial reports for automated case analysis role identification to build the structure. Our methods of case analysis role identification are mostly a combination of different rules. However, many of these rules can be good indicators of possible features to be used in machine learning based methods. Exploring the task with more annotated examples using machine learning based models remains as a future direction of work. Also, our work centered around the Air Pollution cases. How the rules identification module performs across various categories of cases remains as a future task to be examined.

References

[1] National Judicial Data Grid (India) https://njdg.ecourts.gov.in/njdgnew/index.php

[2] Bhattacharya P, Paul S, Ghosh K, Ghosh S, Wyner A: *"Identification of Rhetorical Roles of Sentences in Indian Legal Judgments"* In: International Conference on Legal Knowledge and Information Systems (JURIX) 2019 ; Available from: https://arxiv.org/abs/1911.05405

[3] Walker VR., Pillaipakkamnatt K, Davidson AM., Linares M and Pesce DJ. *"Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning"*. In: Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Texts, Montreal, QC, Canada, June 21, 2019 2019 . CEUR-WS.org.

[4] Neuralcoref, Coreference Resolution tool from Huggingface; Available from: https://github.com/huggingface/neuralcoref

[5] Gupta A, Verma D, Pawar S, Patil S, Hingmire S, Palshikar GK, Bhattacharyya P *"Identifying Participant Mentions and Resolving Their Coreferences in Legal Court Judgements."* In: Text, Speech, and Dialogue - 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018, Proceedings 2018 (pp. 153–162). Springer.

[6] Manupatra, online legal journal https://www.manupatrafast.com/

[7] Ma M, Podkopaev D, Campbell-Cousins A, Nicholas A. Deconstructing Legal Text: Object-Oriented Design in Legal Adjudication. MIT Computational Law Report [Internet]. 2020 Nov 20; Available from: https://law.mit.edu/pub/deconstructinglegaltext

[8] Spacy, Industrial-Strength Natural Language Processing; Available from: https://spacy.io/

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210322

An Analytical Study of Algorithmic and Expert Summaries of Legal Cases

Aniket Deroy [a,1], Paheli Bhattacharya [a], Kripabandhu Ghosh [b], Saptarshi Ghosh [a]
[a] *Indian Institute of Technology, Kharagpur India*
[b] *Indian Institute of Science Education and Research, Kolkata*

Abstract. Automatic summarization of legal case documents is an important and challenging problem, where algorithms attempt to generate summaries that match well with expert-generated summaries. This work takes the first step in analyzing expert-generated summaries and algorithmic summaries of legal case documents. We try to uncover how law experts write summaries for a legal document, how various generic as well as domain-specific extractive algorithms generate summaries, and how the expert summaries vary from the algorithmic summaries. We also analyze which important sentences of a legal case document are missed by most algorithms while generating summaries, in terms of the rhetorical roles of the sentences and the positions of the sentences in the legal document.

Keywords. Case document summarization; extractive summarization; rhetorical roles; lead bias

1. Introduction

Summarization of legal case documents [1, 2] is a challenging problem, which has become important in recent years due to a massive increase in the amount of legal cases available online, and the huge length of most case documents. Several legal domain-specific algorithms as well as generic (domain-independent) algorithms have been used to generate summaries for legal case documents. Most prior works have compared the algorithmic summaries with expert-generated summaries for legal documents in terms of ROUGE scores [3]. However, there has not been much investigation on which parts (e.g., sentences) of a case document are actually selected by experts and by various algorithms for generating the summary of the document.

In this work, we take the first steps in this direction. Specifically, we seek answers to a set of Research Questions (**RQs**) that would help improve our understanding of how summarization algorithms perform in relation to summaries written by the legal experts, and thus give insights that may be valuable for the research fraternity. The Research Questions (**RQs**) that we address in this work are as follows:- **RQ1**: Which parts of a document are selected by summarization algorithms and domain experts while generating summaries? **RQ2**: Which *rhetorical roles* [4] (e.g., Facts, Issues, Ruling) are mostly selected by summarization algorithms and experts while generating summaries? **RQ3**: Are ROUGE scores (computed w.r.t. expert summaries) consistent with the rhetorical roles distributions selected by summarization algorithms? **RQ4**: Out of the sentences that are considered to be important by domain experts, which sentences are easier (or difficult) to identify for summarization algorithms?

[1]Corresponding Author: Aniket Deroy. Email: roydanik18@kgpian.iitkgp.ac.in

To answer these questions, we conduct a detailed analysis over 50 case documents from the Indian Supreme Court and their summaries written by two law experts (dataset obtained from our prior work [2]). We analyse the summaries generated by as many as 15 extractive summarization algorithms, including traditional unsupervised algorithms and supervised neural algorithms. We attempt to analyse how an algorithm or an expert chooses sentences for generating a summary in terms of (i) position of the sentences in the original legal document, and (ii) the rhetorical role [4] of the sentences. It can be noted that prior works have shown that rhetorical roles are necessary to detect important sentences to create good quality summaries of legal documents [1].

Based on the Research Questions stated above, we found the following insights.

1. Some summarization algorithms like BERTSUM [5], Luhn [6] and Letsum [7] tend to select sentences from the initial portions of the input document, which is termed as *lead bias* [8]. Supervised algorithms like BERTSUM which are *pretrained on news article corpora* especially suffer from the lead bias problem. On the other hand, supervised models like SummaRunner [9] and Chinese Gist [10] which can be trained from scratch do not suffer from lead bias problem (Sec. 4.2)
2. Some domain-specific algorithms like KMM [11], MMR [12] and DELSUMM [2] tend to focus on the *Ruling by present court* rhetorical role which is similar to what is done by the domain experts. The *Ruling by Present court* rhetorical role is significant because this rhetorical role includes the final judgement of a legal case (Sec. 4.3)
3. Most algorithms tend to include sentences from the *facts* rhetorical role (possibly due to *lead bias* as *facts* usually appear towards the beginning of a case). On the other hand, a large proportion of sentences belonging to the *Ratio of the decision* rhetorical role is missed by most summarization algorithms because these sentences occur mostly towards the latter portions of the document (Sec. 5.1)
4. The rhetorical role-wise ROUGE scores (computed w.r.t. expert summaries) is consistent with the rhetorical roles distribution selected by most of the summarization algorithms (though there are a few exceptions) (Sec. 4.4)

2. Related work

There have been several prior works on applying summarization algorithms to legal documents [1, 2]. In this section, we discuss about different categories of summarization algorithms that have been applied to legal case documents.

Unsupervised domain-independent: Lexrank [13] is a graph-based summarization technique that uses the idea of eigenvector centrality. Luhn summarizer [6] is a simple method for detecting the most important set of sentences in a document using the concept of TF-IDF vectors. LSA summarizer [14] uses Singular Value Decomposition to project the singular matrix from a higher dimensional plane to a lower dimensional plane to select the most important sentences in the document. Reduction summarizer [15] attempts to condense a long document into the most important parts by creating a rich semantic graph. DSDR [16] is an algorithm which works on the principle of data reconstruction, thereby minimizing the reconstruction error.

Supervised domain-independent: SummaRunner [9] is an algorithm which uses hierarchical Recurrent Neural Networks (RNNs) to learn sentence representations from the input document. To select the sentences for the summary, this algorithm uses relative and absolute position importance, salience, content and novelty. SummaRunner has 3

variations which are SummaRunner/RNN_RNN (which consists of two layers of RNNs), SummaRunner/CNN_RNN (which consists of one layer of Convolutional Neural Network and one layer of RNN), and SummaRunner/Attn_RNN that consists of an attention mechanism with a RNN layer. BertSum [5] is an algorithm which has been initially trained on large amount of news article data. The pre-trained model can be fine-tuned with document-summary pairs from a target domain (e.g., legal document-summary pairs in our case).

Unsupervised legal domain-specific: Letsum [7] divides the entire legal document into four parts namely Introduction, Context, Judicial analysis and Conclusion, and then takes portions of these four parts to form the summary. Case summarizer [17] uses parameters like TF-IDF values, number of dates in a sentence, number of named entities and whether a sentence is in the starting section of the document to select candidate sentences for the summary. MMR algorithm [12] is designed for legal cases related to post-traumatic stress disorder from the US Board of veterans appeal court. This method uses a pipeline consisting of a CNN classifier to select sentences for the summary. Delsumm [2] chooses sentences from the input legal document using a set of rules based on Integer Linear programming. KMM [11] stands for K-mixture model and this K-mixture model is used for selecting sentences to create the summary from the original document.

Supervised legal domain-specific: Chinese Gist [10] is a legal domain-specific supervised algorithm that uses several deep learning and machine learning methods (such as LSTMs) to create various classifiers that are together used with necessary features to generate summaries of legal documents.

Evidently, many prior works have applied summarization algorithms on legal case documents. However, there has not been any prior attempt toward analysis of the summaries generated by summarization algorithms as well as of gold standard summaries generated by experts. This work aims to fill this gap.

3. Dataset

We reuse the dataset from our prior works [2, 4] which consists of 50 case documents from the Indian Supreme Court. To improve the generalizability of the study, the 50 case documents are drawn from 5 different domains – (i) Criminal - 16 documents, (ii) Land and property – 10 documents, (iii) Constitutional – 9 documents (iv) Labour and Industrial – 8 documents, and (v) Intellectual Property Rights – 7 documents. Two senior law students from the Rajiv Gandhi School of Intellectual Property Law, India (one of the most reputed law schools in India) annotated the legal documents with rhetorical labels [4] for each sentence, as well as summarized the legal documents.

Annotation with rhetorical roles: Every sentence in every document has been annotated with one of the following 8 rhetorical labels (the annotation process is detailed in [4]):- **(1) Facts** (abbreviated as **FAC**) are the chronology of events which lead to the filing of the legal case (it includes events like doing FIR at the police station, filing of the case at the court, etc). **(2) Issues** (abbreviated as **ISS**) refer to the legal questions on which the legal case is based. **(3) Ruling by Lower court** (**RLC**) is the judgement given by a lower court on a case which is being contended in a higher court. Since here the legal cases that we are considering are Supreme Court cases so the cases have already being contended in the lower court(s) and the lower court's have passed a decision on that

	FAC	ARG	Ratio	PRE	RLC	RPC	ISS	STA
Original Document	**0.261**	0.082	0.364	0.141	0.033	0.033	0.013	0.069
Expert 1	**0.269**	0.091	0.380	0.074	0.001	0.070	0.032	0.079
Expert 2	**0.289**	0.078	0.371	0.088	0.002	0.067	0.026	0.075

Table 1. Distribution of the rhetorical roles in the original documents, and the summaries written by expert 1 and expert 2, averaged across the 50 documents. Each value is the fraction of sentences of a particular rhetorical role, out of the total number of sentences in the document / expert summaries, averaged over the 50 documents. Blue-underlined represents the rhetorical role with highest fraction of sentences. Violet-bold represents the rhetorical role with second highest fraction of sentences.

case. **(4) Arguments** (**ARG**) are presented by the lawyers of the parties involved in the case. **(5) Precedents** (**PRE**) are the past legal cases which are cited in the present case. **(6) Statutes** (**STA**) are the laws that are referred to, including orders, acts, notifications, articles, sections, rules, etc. **(7) Ratio of the decision** (abbreviated as **Ratio**) refers to the legal reasoning due to which the specific judgement is given. **(8) Ruling by present court** (abbreviated as **RPC**) is the final judgement given by the judge of the present court (the Supreme court of India, in our case).

Summaries of the documents: The two domain experts wrote summaries for each of the 50 documents. The summaries created by the experts are mostly extractive in nature; however, some sentences were slightly modified by the experts to improve the readability / grammatical flow of the summary The length of the summaries written by the experts was around 30% of the original legal document length. Specifically, the collection has on an average 5387.36 words per document, and 1648.76 and 1710.66 words on average in the summaries written by the two experts. Similarly, there are on average 172.86 sentences per document, and 54.06 and 56.72 sentences on average in the summaries written by the two experts. All these values are averaged across the 50 documents. Details of the summarization process can be found in [2].

Table 1 shows the distribution of rhetorical labels across the documents and the expert summaries. Each value is the fraction of sentences of a particular rhetorical role, out of the total number of sentences in the document / expert summaries, averaged over all the 50 documents / their expert summaries. We see that, for both the original documents as well as the expert-written summaries, *Ratio of the decision* is the largest class (these sentences occur most frequently across all documents/summaries) followed by the *Facts* and *Precedent*.

Training data for supervised algorithms: The supervised summarization algorithms (SummaRunner, GIST, BERTSUM) are trained/fine-tuned over a training set consisting of 7,100 Indian Supreme Court case documents and their headnotes (short abstractive summaries); further details can be found in [2]. We ensured that there was no overlap between this training set and the evaluation set of 50 documents.

4. Analyzing the Algorithmic and Expert Summaries

We applied all the summarization algorithms described in Section 2 on the 50 case documents. *For a particular document, all the algorithms were made to generate summaries of the same maximum word count as the average number of words in the two expert-summaries for the same doc.* This section compares the summaries generated by the summarization algorithms and the summaries written by the legal experts (for the same document).

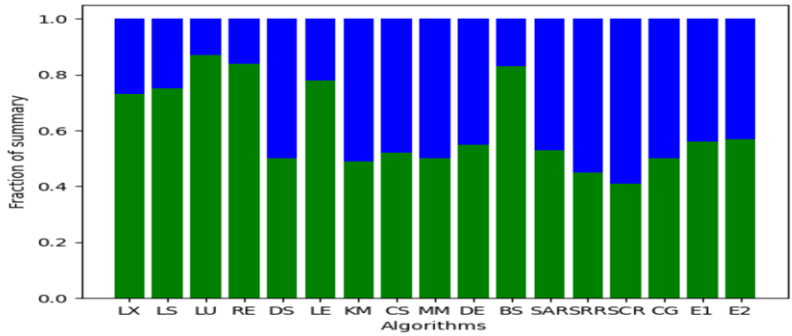


Figure 1. Fraction of the algorithmic summaries and expert summaries taken from the first half and second half of the original legal document, averaged for 50 legal documents. Green colour represents the fraction of the algorithmic / expert summaries taken from the first half of the original document, while blue colour represents the fraction of the summaries taken from the second half of the original document. The symbols on the X-axis are as follows. LX: Lexrank, LS: LSA summarizer, LU: Luhn summarizer, RE: Reduction summarizer, DS: DSDR, LE: Letsum, KM: KMM, CS: CaseSummarizer, MM: MMR summarizer, DE: Delsumm, SAR: SummaRunner/Attn_RNN, SRR: SummaRunner/RNN_RNN, SCR: SummaRunner/CNN_RNN, BS: BERTSUM, CG: Chinese Gist, E1: Expert 1 and E2: Expert 2.

4.1. Finding the closest matching sentence in the document for every sentence in the expert summary

While writing the summaries, the two legal experts mostly copied sentences directly from the document (extractive summarization), but sometimes they combined multiple sentences from the document and/or edited the text of some sentences to improve the fluency and grammatical structure of the summary. For various analyses reported later, we intend to find, for every sentence in the expert summaries, the *closest matching sentence in the document*. To this end, we proceed as follows. We take a sentence s_e from an expert summary and compare the sentence with every sentence in the corresponding original document. If we get an exact match between s_e and some sentence s_d in the document, then s_d is taken as the closest matching sentence for s_e. If we do not get an exact match for s_e, we perform an approximate matching – we calculate ROUGE-2 F1-score (that considers bigram overlap) of s_e with every sentence in the corresponding document. The closest matching sentence in the document for s_e is taken to be that sentence s_d in the document that has the highest ROUGE-2 F1-score match with s_e.

*4.2. **RQ1** : Which parts of a document are selected by summarization algorithms?*

We start by checking the location of the sentences (in a document) that are selected by various algorithms and the experts, for inclusion in the summary. To this end, we consider a document to be partitioned into *two equal halves*, and check what fraction of sentences selected by an algorithm / an expert lies in which half of the corresponding document. Figure 1 shows the fraction of sentences in the algorithmic summaries and expert summaries that are taken from the first half and second half of the original documents, averaged over all the 50 legal documents.

From Figure 1 we can observe the two experts write well-balanced summaries including approximately 55% of sentences from the first halves of the documents and around 45% of the sentences from the second halves of the documents.

In contrast, some summarization methods like BERTSUM, Luhn summarizer, LetSum, Lexrank, LSA summarizer and Reduction summarizer choose most sentences from the first half of the original legal document. This property, where a summarization algorithm tends to choose text mostly from the initial parts of the input document, is known as lead bias [8]. Algorithms such as BERTSUM that are *pre-trained on news article summarization corpus* (where the first few sentences of a news article is known to usually be a good summary of the article), is not able to come out of lead bias due to initial training on news articles, even after they are finetuned on legal documents. In constrast, SummaRunner is trained fully (from scratch) on legal documents and their summaries, and is hence able to avoid lead bias. Letsum is a domain-specific algorithm which divides the document into four parts namely Introduction, Context, Judicial analysis and Conclusion and picks up 10%, 25%, 60% and 5% from each of these individual parts of the document to form the summary. So a large fraction of the summary sentences are picked up from the initial portions of the documents because the Introduction and Context primarily occurs in the initial portions of the document.

It is observed that most unsupervised domain-independent algorithms like Lexrank, LSA, Luhn, Reduction summarizer display significant lead bias. On the other hand, most of the domain-specific algorithms like KMM, Case Summarizer, MMR, DELSUMM and Chinese Gist tend to pick up sentences uniformly from both halves of the document.

4.3. ***RQ2*** *: Which rhetorical roles are selected by summarization algorithms?*

For every sentence from an algorithmic summary, we find the closest matching sentence in the original document and also the rhetorical role of the sentence in the original document. In this way we detect the rhetorical roles that are being selected by the summarization algorithms. Table 2 shows the fraction of each rhetorical role captured by every algorithm and by the two experts, out of the total number of sentences of a rhetorical role present in the original document, averaged over all 50 documents.

From Table 2 we can observe that the experts focused most on the *Ruling by present court (RPC)* and *Issues (ISS)* though these classes are present in small proportions in the original documents (see Table 1). DELSUMM and Chinese Gist are domain-specific algorithms which also focus on *Ruling by present court*. DELSUMM gives highest weight to *Ruling by present court* followed by *Issues*. On the other hand, some domain-specific algorithms like Case Summarizer and Letsum focused most on *Facts* and less on *Ruling by present court*. Supervised algorithms like BERTSUM focused most on the initial portions of the document and picked up *Facts* which are mostly present in the initial portions of the document. LSA summarizer has focused most on *Facts* and *Arguments*. Chinese Gist has also focused well on *Facts*.

Table 2. Fraction of sentences of each rhetorical role captured by every algorithm and by the two experts, out of the total number of sentences of that rhetorical role present in the original text, averaged out of 50 documents. Blue-underlined represents the rhetorical role with highest value. Violet-bold colour represents the rhetorical role with the second highest value.

Algorithm	FAC	ARG	Ratio	PRE	RLC	RPC	ISS	STA
Unsupervised, Domain Independent								
Lexrank	**0.310**	0.319	0.162	0.105	0.292	0.027	0.307	0.183
LSA	0.333	**0.291**	0.145	0.133	0.256	0.015	0.213	0.147
Luhn	0.354	0.264	0.100	0.096	**0.284**	0.015	0.230	0.201
Reduction	0.285	**0.284**	0.108	0.085	0.274	0.013	0.265	0.190
DSDR	**0.333**	0.264	0.330	0.270	0.221	0.456	0.195	0.285
Unsupervised, Domain specific								
Letsum	0.568	0.230	0.190	0.201	0.220	0.029	**0.381**	0.280
KMM	0.245	0.317	0.260	0.235	0.243	0.274	**0.283**	0.250
Case Summarizer	0.298	0.268	0.293	0.124	0.181	0.115	**0.298**	0.194
MMR algorithm	**0.351**	0.317	0.343	0.271	0.266	0.427	0.299	0.243
Delsumm	0.422	0.543	0.239	0.300	0.0	**0.688**	0.739	0.319
Supervised, Domain Independent								
SummaRunner/Attn_RNN	0.389	0.326	0.285	0.300	**0.329**	0.182	0.296	0.141
SummaRunner/RNN_RNN	0.283	0.240	0.355	**0.345**	0.233	0.274	0.274	0.123
SummaRunner/CNN_RNN	0.305	0.257	0.335	**0.335**	0.278	0.594	0.331	0.156
BERTSUM	0.665	0.335	0.149	0.118	0.220	0.040	**0.356**	0.212
Supervised, Domain Specific								
Chinese Gist	**0.461**	0.274	0.432	0.348	0.280	0.608	0.365	0.255
Expert								
Expert 1	0.388	0.465	0.377	0.197	0.014	0.734	**0.665**	0.390
Expert 2	0.432	0.406	0.380	0.224	0.015	0.764	**0.583**	0.390

Table 3. Rhetorical role-wise and entire document-wise performance of all the summarization methods in terms of ROUGE-L F1-scores, averaged over the 50 documents. Values which are < 0.3 are represented in red underlined. Blue bold represents the best value for each rhetorical role.

Algorithm	Entire document	Final judgement	Issue	Facts	Statute	Precedent +Ratio	Argument
Unsupervised, Domain Independent							
Lexrank	0.5392	0.0619	0.3469	0.4550	0.2661	0.3658	0.4284
LSA	0.5483	0.0275	0.2529	0.5217	0.2268	0.3527	0.3705
Luhn	0.5521	0.0358	0.2754	0.5408	0.2662	0.2927	0.3781
Reduction	0.542	0.0352	0.3153	0.5064	0.2579	0.3059	**0.4390**
DSDR	0.5725	0.4987	0.1982	0.4501	0.3174	0.4631	0.3490
Unsupervised, Domain Specific							
LetSum	0.5846	0.0423	0.3926	0.6246	0.3469	0.3853	0.2830
KMM	0.5385	0.3254	0.2979	0.4124	0.3415	0.4450	0.416
Case Summarizer	0.5349	0.2474	0.3537	0.4500	0.2255	0.4461	0.4184
MMR	0.568	0.4378	0.3548	0.4442	0.2763	0.4647	0.3705
DELSUMM	0.6017	**0.7929**	**0.6635**	0.5539	**0.4030**	0.4305	0.4370
Supervised, Domain Independent							
SummaRunner/RNN_RNN	0.5821	0.4451	0.2990	0.5231	0.1636	**0.5215**	0.3090
SummaRunner/CNN_RNN	0.5757	0.5893	0.3586	0.5069	0.1998	0.5026	0.2765
SummaRunner/Attn_RNN	0.5877	0.3633	0.3176	0.6072	0.1869	0.4933	0.4191
BERTSUM	0.5529	0.0662	0.3544	**0.6376**	0.2535	0.3121	0.3262
Supervised, Domain Specific							
Chinese Gist	0.5501	0.5844	0.3856	0.4621	0.2759	0.4537	0.2132

4.4. *RQ3: Are ROUGE scores (computed w.r.t. expert summaries) consistent with the rhetorical role distributions selected by summarization algorithms?*

ROUGE scores are widely considered as a standard way for measuring the quality of an algorithmic summary with respect to the gold standard (expert) summary. Here we examine whether the rhetorical role-wise ROUGE scores calculated for a particular algorithm tally with the algorithmic distribution of rhetorical roles selected by that algorithm. Note that, rhetorical role-wise ROUGE scores are more suitable for this analysis, than ROUGE scores for the entire document.

Table 3 shows the rhetorical role-wise ROUGE-L F1 scores of the algorithms. For most algorithms, the results given by the rhetorical role-wise ROUGE scores (computed w.r.t. expert summaries) are similar to the results given by the algorithmic distribution of rhetorical roles (that were discussed as part of RQ2). For instance, methods like Lexrank and CaseSummarizer get lower ROUGE-L F1-scores; this agrees with the observations in Table 2 where *Ruling by present court* and *Issues* have been selected extensively by the experts but selected in much less proportions by these algorithms. On the other hand, algorithms such as LetSum and DELSUMM show higher ROUGE-L F1-scores because the rhetorical distributions chosen by these algorithms are closer to the experts' rhetorical distributions.

5. Which important parts do most summarization algorithms miss?

We now attempt to characterize which important parts (sentences) of a legal case document are missed (i.e., not included in the summary) by most summarization algorithms.

5.1. ***RQ4:** Out of the sentences that are considered to be important by domain experts, which sentences are easier / difficult to identify for summarization algorithms?*

Using the method described in Section 4.1, we found the closest matching sentence in the document for every sentence in the expert summaries. Now we focus on those sentences from the original document, that were selected by the experts for inclusion in the gold standard summaries. For each such sentence, we check how many algorithms have included that particular sentence in their summary. Table 4 states the number and percentage of sentences in the expert summaries that are selected by less than 3 algorithms, sentences selected by 3-9 algorithms, and sentences selected by 10 or more algorithms. To better understand which parts of the documents are being selected (or missed) by the summarization algorithms, we focus on the following two sets of sentences:

Frequently selected sentences: The set of sentences which appear in at least one expert summary and are chosen by 10 or more algorithms for inclusion in the summaries. There are 155 such sentences in total across all the 50 documents.

Frequently missed sentences: The set of sentences which appear in at least one expert summary but are chosen by less than 3 algorithms. There are 529 such sentences in total across all the 50 documents.

Note that, both these sets contain important sentences that are chosen by the Law experts while writing their summaries. We analyze these two sets of sentences with the objective of gaining a better understanding of the frequently missed sentences which are important sentences being missed by most summarization algorithms.

Characterizing the location of frequently selected/missed sentences: Out of the set of *frequently missed sentences*, 35% come from the first halves of the documents, while 65% sentences come from the second halves of the documents (numbers averaged over

Table 4. Number and percentage of sentences in the expert summaries that are selected by less than 3 algorithms (frequently missed sentences), 3-9 algorithms and 10 or more algorithms (frequently selected sentences). Blue-underlined cell represents the *frequently missed sentences* for a particular expert. Violet-bold cell represents the *frequently selected sentences* for a particular expert.

Number of algorithms →	less than 3 algorithms (Frequently missed sentences)	3-9 algorithms	10 or more algorithms (Frequently selected sentences)
Expert 1	456 (17.7%)	1968 (76.6%)	**142 (5.5%)**
Expert 2	478 (18.4%)	2063 (76.0%)	**140 (5.2%)**
Union of both experts	529 (17.8%)	2280 (76.9%)	**155 (5.2%)**

Table 5. Comparison of the *frequently missed sentences* and the *frequently selected sentences* in terms of their rhetoric distribution. Blue-underlined colour represents the rhetorical role which is present in the highest proportion in a row.

Rhetorical roles →	FAC	ARG	Ratio	PRE	RLC	RPC	ISS	STA
Frequently missed sentences	0.061	0.067	0.575	0.096	0.0	0.080	0.013	0.105
Frequently selected sentences	0.619	0.140	0.112	0.014	0.0	0.0	0.077	0.035

Table 6. Examples of frequently missed sentences that are selected by Law experts in their summaries, but are not selected by most of the summarization algorithms.

Sentence	Rhetorical role
the appeals are disposed of accordingly without any order as to costs	RPC
order was a legislative activity and therefore not subject to any principle of natural justice	ARG
no vested right as to tax holding is acquired by a person who is granted concession	Ratio
what the order does contemplate however is such enquiry by the government as it thinks fit	Ratio

all 50 documents). On the other hand, as many as 93.6% of the *frequently selected sentences* come from the first halves of the documents, and only 6.4% come from the second halves. These numbers re-confirm the lead bias of several algorithms, as was discussed in Section 4.2 (**RQ1**) – most of the *frequently missed sentences* come from the second half of the documents.

Characterizing the rhetorical labels of frequently selected/missed sentences: Table 5 shows the rhetorical role distribution of the frequently missed and frequently selected sentences. We see that for the *frequently missed sentences*, most sentences belong to the *Ratio of the decision* rhetorical role. For the *frequently selected sentences*, most number of sentences belong to the *Facts* rhetorical role. This observation can also be ascribed to the prevalence of lead bias of these algorithms, as discussed in Section 4.2 (**RQ1**), since *Facts* usually appear at the beginning of a case document and *Ratio of the decision* usually occurs at the latter portions of a document. Table 6 gives some examples of the *frequently missed sentences* and their rhetorical labels.

Characterizing the length and legal keywords content of frequently selected/missed sentences: We found that frequently missed sentences have a similar distribution of length (number of words) as frequently selected sentences. We checked the number of legal keywords contained in the two types of sentences, using terms from a legal dictionary provided by [18]. The frequently selected sentences contain 3.30 legal terms on average, while frequently missed sentences contain 2.89 legal terms on average. The fact that frequently selected sentences contain more legal terms (than frequently missed sentences) may be a potential reason why most summarization algorithms choose them.

6. Conclusion and Future Work

In this work, we compare the algorithmic and expert summaries of legal case documents to unearth the nature and position of the sentences chosen to create algorithmic summaries and expert summaries. This work gives us several insights that can help in improving the existing summarization algorithms that are capable of creating summaries aligning more with the notion of legal experts – potential end-users of such algorithms.

Our current work considers rhetorical role-wise ROUGE scores to analyse the quality of legal document summaries. In future, we can apply metrics other than ROUGE scores to evaluate the quality of legal summaries, since there are limitations of quantitative metrics like ROUGE scores. Also, we plan to generalize our observations through similar experiments on legal documents of other jurisdictions and countries.

Acknowledgements: The authors thank the Law domain experts from the Rajiv Gandhi School of Intellectual Property Law, India who annotated the legal documents and wrote the summaries. The research is partially supported by the TCG Centres for Research and Education in Science and Technology (CREST) through a project titled "Smart Legal Consultant: AI-based Legal Analytics".

References

[1] P. Bhattacharya, K. Hiware, S. Rajgaria, N. Pochhi, K. Ghosh, and S. Ghosh, "A comparative study of summarization algorithms applied to legal case judgments," in *ECIR*, 2019.
[2] P. Bhattacharya, S. Poddar, K. Rudra, K. Ghosh, and S. Ghosh, "Incorporating domain knowledge for extractive summarization of legal case documents," in *ICAIL*, 2021.
[3] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.
[4] P. Bhattacharya, S. Paul, K. Ghosh, S. Ghosh, and A. Wyner, "Identification of rhetorical roles of sentences in Indian legal judgments," in *JURIX*, 2019.
[5] Y. Liu, "Fine-tune bert for extractive summarization," *ArXiv*, vol. abs/1903.10318, 2019.
[6] A. Nenkova, S. Maskey, and Y. Liu, "Automatic summarization," in *ACL*, 2011.
[7] A. Farzindar and G. Lapalme, "Letsum, an automatic legal text summarizing system," in *JURIX*, 2004.
[8] M. Grenander, Y. Dong, J. C. K. Cheung, and A. Louis, "Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses," in *EMNLP*, 2019.
[9] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents," in *AAAI*, 2017.
[10] C.-L. Liu and K.-C. Chen, "Extracting the Gist of Chinese Judgments of the Supreme Court," in *ICAIL*, 2019.
[11] J. Vermunt, "K-means may perform as well as mixture model clustering but may also be much worse: Comment on steinley and brusco (2011)," *Psychological methods*, vol. 16, 2011.
[12] L. Zhong, Z. Zhong, Z. Zhao, S. Wang, K. D. Ashley, and M. Grabmair, "Automatic summarization of legal decisions using iterative masking of predictive sentences," in *ICAIL*, 2019.
[13] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Intell. Res.*, vol. 22, pp. 457–479, 2004.
[14] J.-Y. Yeh, H.-R. Ke, W.-P. Yang, and I. Meng, "Text summarization using a trainable summarizer and latent semantic analysis," *Information Processing and Management*, vol. 41, pp. 75–95, 2005.
[15] I. Moawad and M. Aref, "Semantic graph reduction approach for abstractive text summarization," in *International Conference on Computer Engineering and Systems*, pp. 132–138, 11 2012.
[16] Z. He, C. Chen, J. Bu, C. Wang, L. Zhang, D. Cai, and X. He, "Document summarization based on data reconstruction," in *AAAI*, 2012.
[17] S. Polsley, P. Jhunjhunwala, and R. Huang, "Casesummarizer: A system for automated summarization of legal texts," in *COLING*, 2016.
[18] A. Mandal, K. Ghosh, A. Pal, and S. Ghosh, "Automatic catchphrase identification from legal court case documents," in *ACM CIKM*, 2017.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210323

Semantic Search and Summarization of Judgments Using Topic Modeling

Tien-Hsuan WU [a], Ben KAO [a], Felix CHAN [b], Anne SY CHEUNG [b], Michael MK CHEUNG [b], Guowen YUAN [a], and Yongxi CHEN [b]
[a] *Department of Computer Science, The University of Hong Kong*
[b] *Faculty of Law, The University of Hong Kong*

Abstract. Online legal document libraries, such as WorldLII, are indispensable tools for legal professionals to conduct legal research. We study how topic modeling techniques can be applied to such platforms to facilitate searching of court judgments. Specifically, we improve search effectiveness by matching judgments to queries at semantics level rather than at keyword level. Also, we design a system that summarizes a retrieved judgment by highlighting a small number of paragraphs that are semantically most relevant to the user query. This summary serves two purposes: (1) It explains to the user why the machine finds the retrieved judgment relevant to the user's query, and (2) it helps the user quickly grasp the most salient points of the judgment, which significantly reduces the amount of time needed by the user to go through the returned search results. We further enhance our system by integrating domain knowledge provided by legal experts. The knowledge includes the features and aspects that are most important for a given category of judgments. Users can then view a judgement's summary focusing on particular aspects only. We illustrate the effectiveness of our techniques with a user evaluation experiment on the HKLII platform. The results show that our methods are highly effective.

Keywords. Topic modeling, Semantic search, Judgment summarization

1. Introduction

In common law jurisdictions, prior judgments (a.k.a. *precedents*) are important parts of the law. Retrieving relevant judgments is an important task in legal research. To find existing judgments, one may resort to legal databases such as the World Legal Information Institute (WorldLII) [1]. Although existing legal database systems provide search functions that facilitate judgment retrieval, they are mostly limited to simple keyword search. It is well known that keyword-based search suffers from poor query expressiveness.

In this paper we address the judgment retrieval problem by applying topic modeling techniques to perform semantic search and judgment summarization. Specifically, our approach consists of the following three components.

[Semantic Search] Existing search engines deployed in legal database systems such as HKLII mostly retrieve judgments based on keyword matching. This is ineffective especially when the search intent involves abstract concepts that can be expressed in various wordings or in technical terms that the query issuer is not familiar with. We achieve semantic search by applying *topic modeling*. In a nutshell, topic modeling is a

Figure 1. Word cluster that expresses a topic on "*finger injury*"

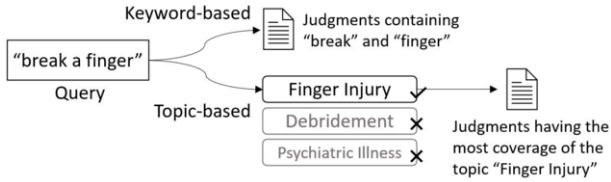


Figure 2. Topic-based semantic search

statistical framework that analyzes a document corpus to identify distinguishing words that have strong associations (e.g., based on co-occurrences of words in documents). A "*topic*" can be considered as a cluster of words based on their associations, which, collectively, express certain abstract concept. Figure 1 shows an example word cluster that expresses the concept (or topic) of "finger injury". Our method of semantic search is to first analyze court judgments to discover topics (in topic-modeling sense) and identify the topics that are covered by each specific judgment. When given a query, the topic that is most relevant to the query is found and the judgments that have the best coverage of the topic are retrieved. Figure 2 illustrates our idea.

[Query-driven Summarization] A search engine often returns search results as a list of hypertext links, each with the corresponding document title displayed. It is difficult for the user to determine if a document is really relevant to his/her query by inspecting the title only. As we observed in analyzing the HKLII search log, very often a user would click and read many returned documents that turned out to be irrelevant to the search intent. This is a major source of inefficiency in judgment search as users toil through long and complex judgments. Our approach to ameliorating unproductive search results filtering is to perform *automatic judgment summarization*. Specifically, a small fraction (such as 5%) of a judgment's paragraphs are selected by the machine based on their relevancy to the given user query. These paragraphs are highlighted in the judgment and they serve as a query-specific summary of the judgment. By reading this small (5%) summary, the query issuer gets to know why the machine thinks that the judgment is relevant to the query, obtains a basic understanding of the judgment's content, and thus is able to quickly determine if the judgment should be filtered or collected.

[Aspect-driven Summarization] After a user accepts a judgment as relevant with the help of a query-specific summary, the user typically needs to know more about the judgment with respect to different aspects of the case concerned. To address this, our system provides aspect-specific summaries of judgments. Our approach is to first consult legal experts on the most important features and aspects of each judgment category. For example, for personal injury cases, aspects of interests include a plaintiff's *background*, *treatments*, *losses*, and *compensations*. The machine would then summarize a judgment based on each aspect by finding a small number of paragraphs in the judgment that are most relevant to the chosen aspect. Figure 3 shows an example.

2. Algorithms

In this section we discuss how topics are generated (Section 2.1) and how we use the generated topics in semantic search and judgment summarization (Section 2.2).

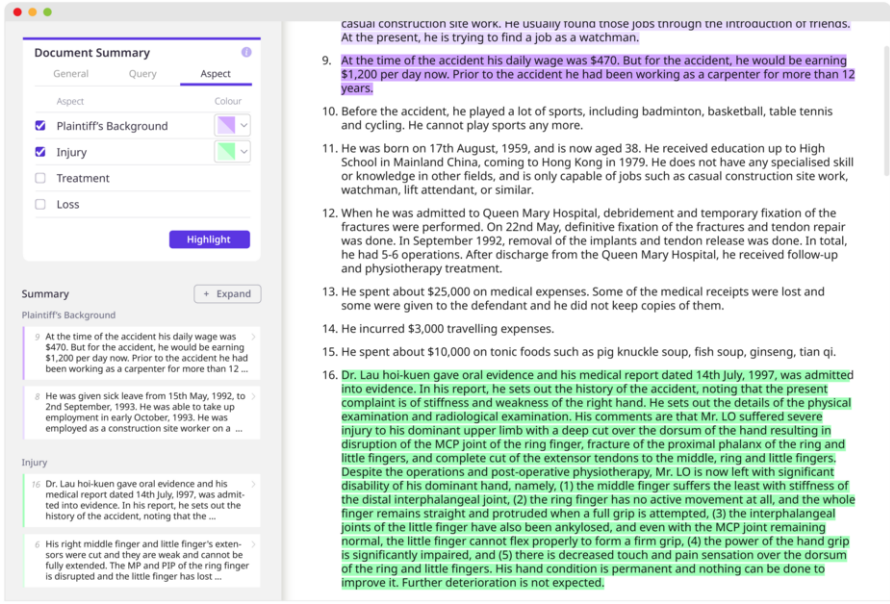


Figure 3. Interface for aspect-driven summarization

2.1. Topic Generation

We adopt *Latent Dirichlet Allocation* (LDA) [2] as the topic modeling method. Given a collection of documents, LDA generates *topics* by computing two sets of distributions: (1) A *word distribution* for each topic. The word distribution captures the words that express the topic. An example is shown in Figure 1. (2) A *topic distribution* for each document. The topic distribution reflects the probability of each topic occurring in the document. In the following discussion, we use *personal injury compensation cases in Hong Kong* to illustrate our methods. We consider three ways of applying LDA to judgment data. They differ in how judgment documents are processed and whether expert knowledge on specific judgment category is taken into account. Figure 4 illustrates the three approaches. Next, we give details of the approaches.

[No Domain Knowledge (NoDK)] Judgments are first preprocessed by removing numbers and stop words. Then, we run LDA on the judgments using MALLET's implementation [3] to generate topics. This is illustrated in Figure 4(a).

[Feature Domain Knowledge (DK-F)] We consult legal experts to obtain a list of features that are important for the specific category of judgments in the corpus. For example, for PI judgments, these features include "*age at the time of incident*" and "*whether the injury is permanent*". Judgments are then manually annotated to identify spans of text that contain information related to the features. We call these spans of text "*labeled text*". We strip each judgment of its unlabeled text; Only labeled text is retained to which we apply LDA. The idea is to remove unimportant details so that the topics generated are related to the more important contents of judgments. Figure 4(b) illustrates this approach.

[Aspect Domain Knowledge (DK-A)] We further consult legal experts to group features into *aspects*. For example, for PI cases, four aspects are given, namely, (plaintiff's) *background*, *injury*, *treatment*, and *loss*. We perform topic modeling on the labeled text under each aspect separately. For example, with respect to the background aspect, we

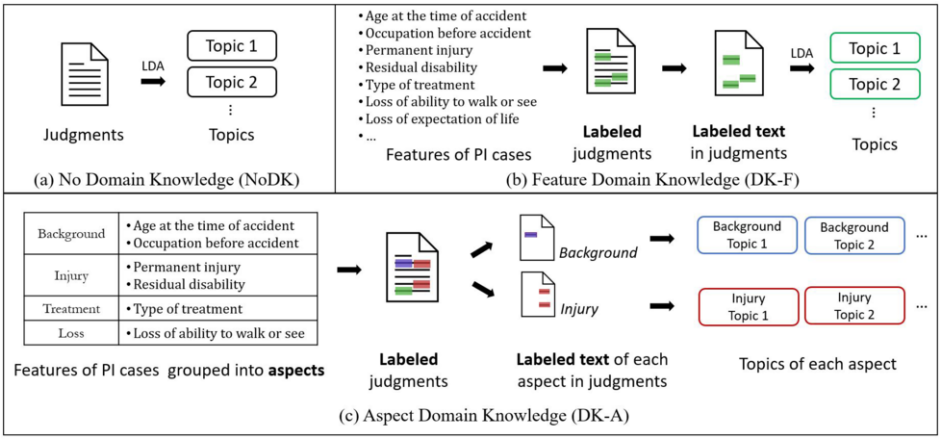


Figure 4. Topic Modeling Approaches

retain only the labeled text for background-related features in judgments before applying LDA. The idea is to generate aspect-specific topics so that judgments can be summarized based on a desired aspect. Figure 4(c) illustrates this approach.

2.2. Applications

In this section we discuss how we make use of the generated topics to perform semantic search and judgment summarization.

2.2.1. Semantic Search

Let $\mathcal{T} = \{T_1, ..., T_N\}$ be the set of N topics obtained from LDA. Given a query q and a judgment J, we evaluate the relevancy of q and J w.r.t. the N topics. The similarity of q and J is then measured by their overlapping topics. Specifically, a topic T_i is represented by a word vector $[w_{i,1}, ..., w_{i,k}, ...]$, where each $w_{i,k}$ is a word with probability $p_{i,k}$ of being relevant to topic T_i. For each word $w_{i,k}$, we apply word2vec [4] to obtain its word embedding vector $\mathbf{w}_{i,k}$. We then compute the average of all word embedding vectors $\mathbf{w}_{i,k}$'s weighted by their probabilities $p_{i,k}$'s. We call the resulting embedding vector the *topic semantic vector* $\mathbf{v}_{T_i}$ of topic T_i. Next, we process the query q in a similar fashion: we first obtain the word embedding vector of each word in q and then compute the vectors' average. We call the resulting embedding vector the *query semantic vector* $\mathbf{v}_q$ of query q. The *query-topic similarity score*, $s_{q,i}$, between query q and topic T_i is measured by the cosine similarity of the semantic vectors, i.e., $\mathbf{v}_{T_i} \cdot \mathbf{v}_q / \|\mathbf{v}_{T_i}\| \|\mathbf{v}_q\|$. We collect the similarity scores over all topics into a *query-topic probability vector* $\mathbf{p}_q = [s_{q,1}, ..., s_{q,i}, ...]$ of query q. This vector summaries the relevancy of each topic to the query q. For a judgment J, LDA produces a *judgment-topic probability vector* $\mathbf{p}_J = [t_{J,1}, ..., t_{J,i}, ...]$ where $t_{J,i}$ is the probability that judgment J is relevant to topic T_i. Finally, we compute the similarity between query q and judgment J by taking the dot product $\mathbf{p}_q \cdot \mathbf{p}_J$. Given a query q, we return the judgments with the highest similarities as the search results.

2.2.2. Query-Driven Summarization

Given a query q and a judgment J, our objective is to find a small fraction (e.g., 5%) of the paragraphs in J that are the most relevant to q. These selected paragraphs serve as a query-specific summary of J to q, which helps the user understand whether J is indeed desired. To achieve that, we first find the most relevant topic T_q, which is the topic that gives the highest query-topic similarity score, i.e., $T_q = \arg\max_{T_i}(s_{q,i})$. Next, for each paragraph G in J, we compute a *paragraph semantic vector* $\mathbf{v}_G$ by averaging the word2vec embedding vectors of the words in G. The similarity between paragraph G and topic T_q is then measured by the cosine similarity of their semantic vectors, i.e., $\mathbf{v}_{T_q} \cdot \mathbf{v}_G / \|\mathbf{v}_{T_q}\| \, \|\mathbf{v}_G\|$. Paragraphs with the highest similarities are selected as the summary.

2.2.3. Aspect-Driven Summarization

In Section 2.1 we discussed three ways of generating topics. In particular, with DK-A, topics are grouped into aspects (see Figure 4(c)). Given a judgment J and an aspect A, our objective is to find a small number of paragraphs in J that best describe the case w.r.t. aspect A. For simplicity, we explain our approach assuming that *plaintiff's background* is the aspect of interest. Our method can be generalized to cover any other given aspect. Let $\mathcal{T}_\mathcal{B} = \{T_1, ..., T_M\}$ be a set of M topics generated by the DK-A model under the "plaintiff's background" aspect. For a judgment J, we consider its *judgment-topic probability vector* $\mathbf{p}_J$ (see Section 2.2.1) and find the topic in $\mathcal{T}_\mathcal{B}$ that gives the highest probability among those in $\mathbf{p}_J$. We denote this top-ranked topic T_J. Formally, $T_J = \arg\max_{T_i \in \mathcal{T}_B} t_{J,i}$. Next, we measure the similarity between each paragraph G in judgment J and the topic T_J in the same way as we did in query-driven summarization, i.e., by the cosine similarity of $\mathbf{v}_G$ and $\mathbf{v}_{T_J}$. Paragraphs with the highest similarities are selected as the summary.

3. Evaluation

In this section, we present the evaluation of our topic modeling appraoches. In particular, we give experimental results comparing different topic generation methods using query-driven summarization as the target application.

We collected 832 judgments on personal injury (PI) compensation cases handed down in Hong Kong from 1999 to 2021. The judgments contain 606 to 11,257 words each, with an average length of 4,552 words. Our legal experts suggested 79 features for PI cases, among which 48 are of interests to this study. These 48 features are grouped into four aspects, namely, "*background*" (11 features), "*injury*" (9 features), "*treatment*" (8 features), and "*loss*" (20 features). We hired 10 law students to manually label these features in the judgments. The data is used to derive topic models under the NoDK, DK-F, and DK-A approaches (see Figure 4).

To evaluate the quality of the query-specific summaries provided by each method, we prepared a set of 20 *test queries* that search for PI judgments. These queries are real user queries extracted from the HKLII search log. We submitted each query q to HKLII search engine and retrieved the top-ranked PI judgment J after filtering out those in the search results that were irrelevant to q. This gave us a query-judgment (q-J) pair. We then applied query-driven summarization (Section 2.2.2) to determine a similarity score of each paragraph in J w.r.t. the query q. The paragraphs were then ranked based on their

similarity scores and the top 5% of the paragraphs were selected as the summary of J. We considered 3 approaches to generate topics, namely, NoDK, DK-F, and DK-A. Each of them resulted in a summary, which might differ from those of others. Hence, we got three summaries for each q-J pair corresponding to the three methods.

We recruited 8 legal experts (who either have the *Postgraduate Certificate in Laws* qualification or are currently practising law) to evaluate the summaries. Given a q-J pair, we merged the three summaries obtained from the methods into a single collection and presented the paragraphs to an expert for "grading". The expert was asked to read the query and the judgment, and then assign a score of '0' (not relevant), '1' (somewhat relevant), or '2' (relevant) to each paragraph in the collection. During the process, the expert was totally blind to which method was used to select the paragraphs. Each q-J pair was graded by 1 to 3 experts.

We evaluate a summary's quality by comparing it with the *optimal summary* using the *normalized discounted cumulative gain (nDCG)* metric [5]. Specifically, let $S_X = \{G_1, G_2, ..., G_k\}$ be a summary of k paragraphs taken from a judgment J by method X (X = NoDK, DK-F, or DK-A), such that the paragraphs G_i's are sorted in decreasing order of their similarities with the identified topic (i.e., $\mathbf{v}_{T_q} \cdot \mathbf{v}_G / \left\|\mathbf{v}_{T_q}\right\| \|\mathbf{v}_G\|$, see Section 2.2.2). Let $s(G_i)$ be the average relevancy score of G_i given by the human assessors. The DCG score of summary S_X is given by $DCG(S_X) = \sum_1^k s(G_i)/\log_2(i+1)$. The (theoretical) optimal summary, denoted by $\tilde{S}$, is constructed by collecting paragraphs in J that are given the highest average relevancy scores by the assessors until k paragraphs are collected. The nDCG score of summary S_X is then given by $DCG(S_X)/DCG(\tilde{S})$. Note that nDCG scores of summaries range from 0 to 1, with 1 indicating that the summary matches the optimal one perfectly in selecting paragraphs and assessing their relevancy to the query.

Table 1 shows the average nDCG scores of the summaries obtained by the three different methods. Moreover, we consider summaries with nDCG ≥ 0.75 (< 0.5) to be of good (poor) quality. Table 1 shows the number of good/poor summaries for each method. From the table, we see that DK-A, which considers domain knowledge of different PI aspects, significantly outperforms NoDK and DK-F. First, DK-A has a very high average nDCG score (0.86) compared with NoDK (0.66) and DK-F (0.64). Secondly, DK-A produces 18 good summaries for the 20 query-judgment pairs and no poor summaries. The reason for DK-A's excellent performance is that it generates topics with respect to different aspects. That allows DK-A to generate more topics than other methods and the topics are more precise and focused. Table 1 further shows that NoDK and DK-F have comparable performance in terms of average nDCG scores. On closer inspection, we find that there is not a clear advantage of one over the other; For some q-J pairs, NoDK gets better scores, while DK-F is better for other q-J pairs. As we mentioned in Section 2.1, DK-F ignores unlabeled text in generating topics. That helps remove unimportant content in judgments and improve topic modeling. Occasionally, however, DK-F is too aggressive and some content that is useful in generating topics is inadvertently removed, resulting in poor summary quality.

Table 1. Quality of query-driven summaries

	NoDK	DK-F	DK-A
Average nDCG	0.66	0.64	0.86
# of good summaries (nDCG $\geq$ 0.75)	9	9	18
# of poor summaries (nDCG $<$ 0.50)	7	9	0

Figure 5 shows a screenshot of our query-specific summary design. A user types a query in a search box (top of left panel). The retrieved judgment is shown in the right panel with paragraphs in the summary highlighted. Excerpts of the summary paragraphs

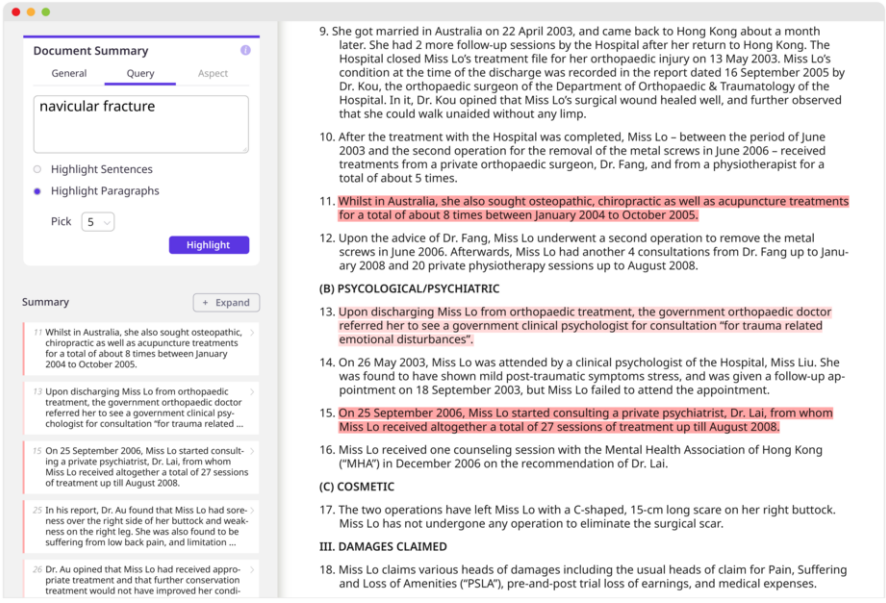


Figure 5. Screenshot for query-driven summarization

are collected and displayed in the lower part of the left panel. The user can read the paragraphs excerpts in the summary to determine if the retrieved judgment is relevant to his/her search intent. By clicking on a paragraph excerpt, the system will display the corresponding paragraph in the judgment in the right panel. This allows the user to read the context of the summary paragraphs for further details.

4. Conclusion

In this paper we studied the problem of effective semantic search and judgment summarization in digital legal library systems. We proposed a general framework to achieve the tasks through topic modeling. We considered three approaches (NoDK, DK-F, and DK-A) of generating topics. We also proposed algorithms for generating query-specific and aspect-specific judgment summaries, and algorithms for performing semantic search.

Acknowledgement

This work is supported by HKU-TCL Joint Research Center for AI and Innovation and Technology Fund (Grant ITS/234/20).

References

[1] World Legal Information Institute. WorldLII Website; 2021. https:/www.worldlii.org/.
[2] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. JMLR. 2003;3:993–1022.
[3] McCallum AK. MALLET: A Machine Learning for Language Toolkit; 2002.
[4] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013. p. 3111–3119.
[5] Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS). 2002;20(4):422–446.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210324

Analyze the Usage of Legal Definitions in Indonesian Regulation Using Text Mining Case Study: Treasury and Budget Law

Bakhtiar AMALUDIN [a] Fitria Ratna WARDIKA [a]
Putu Jasprayana MUDANA PUTRA [a] and I Gede Yudi PARAMARTHA [b]
[a] *Legal Bureau, Ministry of Finance of Indonesia*
[b] *Inspectorate General, Ministry of Finance of Indonesia*

Abstract. Legal definitions are an integral part of legal drafting practice to understand legal documents easily and prevent ambiguity. This research aims to describe how legal definitions are used among regulations in the domain of Indonesian Treasury and Budget. Simple text mining techniques are used to perform and deliver the process. We extracted definitions from more than 1.362 related regulations enacted through the period 2003-2020. We found that legal definitions were used in many variations which may lead to inconsistencies.

Keywords. legal definition, legal term, consistency, harmonization, text mining.

1. Introduction

Do the definitions in regulations need to be consistent? Gauci points out that there are situations where regulation has defined a legal term, but in another, the legal term is given a different definition [4]. This situation could trigger different interpretations and thus, it is not a mistake when Lucius Priscus said that every definition in law is dangerous [4]. Hence, as an essential part of legal drafting practice, having harmonized legal definitions is not merely for precise and effective communication [2]. In this situation, the challenge is how legal drafters formulate harmonized definitions and, more importantly, do not potentially contradict each other.

In Indonesia, there are several rules in drafting legal definitions, including consistency in defining terms, particularly in similar fields; and definitions in lower regulations must be in line with higher regulations [7]. However, learning from Gauci's findings [4], definitional inconsistencies seem unavoidable in Indonesia. For example, since State Finance Law[1] and State Treasury Law enacted [2], there have been numerous implementing regulations comprises of Government Regulations, Presidential Regulations, and Minister of Finance Regulations. Thus, the legal definitions also increase as the number of regulations grows. We often found that the definition of some legal terms within two or more regulations are varied and lead to confusion.

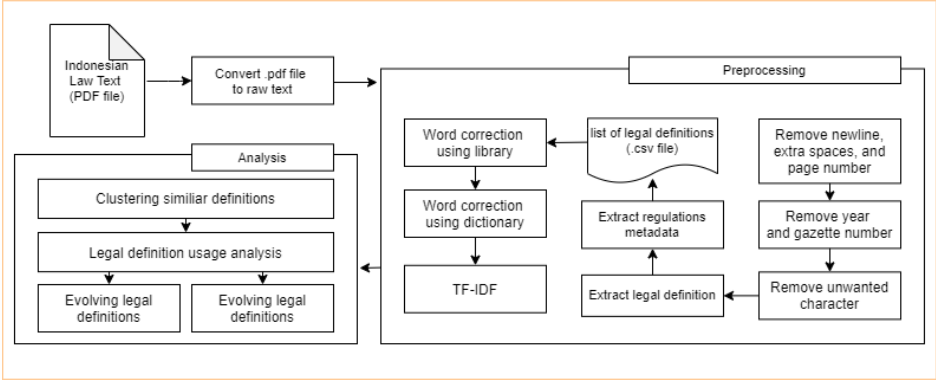
[1] Indonesian Law No. 17 of 2003 on State Finance
[2] Indonesian Law No. 1 of 2004 on State Treasury

To perform an in-depth analysis of this issue, we propose text mining to explore the use of definitions across regulations in the domain of Indonesia's treasury and budget. Finally, the results of this study are expected to be an input for legal drafters to make a consistent definition to prevent legal problems due to the existence of a term that is defined differently.

2. Methodology

Figure 1 presents an overview of the proposed framework for analyzing legal definition usage in Indonesian regulations. In Step 1 (Data Collection), regulations in the domain of treasury and budget law are collected. Step 2 (Preprocessing) emphasizes noise removal and data transformation. Step 3 (Analysis) aims to analyze legal definition usage in the regulations and find potential mismatch.

Figure 1. Legal Definition Usage Analysis Framework



2.1. Data Collection

We gathered 1362 regulations in the treasury and budget domain from Indonesian state gazette stored as PDF files. It comprises 4 Laws (UU), 15 Government Regulation (PP), 2 Presidential Regulation (PERPRES), and 1,313 Regulation of Minister of Finance (PMK). We transformed all the PDFs documents into machine-readable form (i.e. raw text) using the pdfminer library in Python.

2.2. Pre-processing

The purpose of this step is to extract information we need for data analysis (i.e. legal terms contained in regulation and its metadata). Initial text cleaning performed on the raw text to remove noises such as new lines, extra spaces, page number, year and gazette number, and unwanted character in the raw text.

2.3. Extract Legal Definitions

To extract the legal definitions from regulations, our legal drafting experts analyzed examples of legal definitions appearing in regulations. These patterns then transformed into regular expressions that form specific kinds of text patterns for a faster searching [5].

Indonesian regulations have a standardized structure that legal terms are always defined in the general provision of the regulations. Thus, we first identified the general provision part of each regulation (i.e. in the first article of the regulation). After that, we split general provision text into the segmentation of sentences using a sentence tokenizer. We analyzed each sentence to see whether it meets the legal definition pattern (Table 1) and extracted three parts from each of these (terms, alias, and definitions).

Table 1. Legal definition pattern in general provision.

Type	Sentence Pattern	Regular Expression
Direct definition	*term* is [...]	^(.+)(adalah)(.+)
Acronym	*term* hereinafter referred to as *alias* is [...]	^(.+)(yang selanjutnya disebut)(.+)(adalah)(.+)
Abbreviation	*term* hereinafter abbreviated as *alias* is [...]	^(.+)(yang selanjutnya disingkat)(.+)(adalah)(.+)

After automatically extracting legal keywords from regulation documents, we were able to extract 8,202 legal terms and the number of unique legal terms is 2,546 which means some legal terms appear in more than one regulation. However, not all legal definitions are extracted correctly by these regex patterns. From manual inspection we found the 117 incorrect legal definitions were captured and 11 records must be discarded because they are not considered as legal definitions. For the rest, although the regex were able to identify legal definitions component (terms, alias and definition) correctly, many of them have misspelled words, missing letters, incorrect word order, and mixed words. Thus, further data was cleaning performed to handle this problem. We used symspell library complemented with Bahasa Indonesia Frequent Words Dictionary[3] and dictionary-based spelling correction to fix several misspell words.

2.4. Extract Regulation Metadata

We also extracted some relevant information in the header part of each regulation. It was related to the source of legal definitions such as the regulation number, the type of regulation and year of enactment. The regular expression pattern for this was "REPUBLIK\s+INDONESIA \s+NOMOR\s+(\d+)\s+TAHUN\s+(\d{4})". We then captured the first group as the regulation number and the second group as the year of enactment.

2.4.1. Cluster Legal Definition

In this step, each legal term from the previous process clustered according to its similarity in definition. However, to work with clustering algorithms, we need to transform text into numerical representation. In this case, we implemented TF-IDF to transform legal definition text into number of matrix [5]. Nothing excessive in this transforming

[3]https://github.com/hermitdave/FrequencyWords

process, the only intervention is regarding tokenization. Indonesian cases are different where special cases occur such as not treating hyphen (-) as signs of word segmentation.

We used the most popular density based clustering [9] in particular DBSCAN clustering algorithm with euclidean distance to group similar words in separate clusters [9]. We set the best parameter that is given by silhouette score 0.56418 (i.e. epsilon=0.01 and min_samples=1). It produced 4,691 clusters which were then used for labeling each legal definition. The final result of these processes described in Table 2.

Table 2. Final dataset.

Terms	Alias	Description	Source File	Year	Reg.Type	Label
General Allocation Fund	DAU	General Allocation Fund, [...]	68/PMK.02/2016.pdf	2016	PMK	809

2.5. Legal Definition Variation Analysis

Based on the dataset illustrated in Table 2 above, we identified some potential conditions that may cause variation on legal definitions as follows.

- **Evolving Definitions**: a condition when the same legal terms appear as different cluster labels but used sequentially in time order according to the number and year of enactment.
- **Potential Mismatch** : a condition when same legal terms appears as different cluster in same or different type of regulations

These condition will be used as baseline to subset legal definitions for further analysis. However variation here can not be judged as unlawful practice since our approach in detecting variation is limited only on syntactical differences in definition text.

3. Analysis

3.1. Legal Definition Usage and Variety

Initial analysis goes into an insight how legal definitions are used repeatedly in several regulations. As depicted in Figure 1, the more frequent the legal terms used, the more varied they are defined across regulations. For example, terms ”DIPA” has more than 40 different definition that spread within in two different type of regulations (i.e. PP and PMK) totaling more than 100 documents.

3.2. Evolving Definitions

We found that there are 403 definitions that can be considered as evolving definitions. Evolving definition is a common aspect that causes the diversity of definitions which legal drafter make enhancement or improvement carried out in accordance with the current situation and conditions faced by policymakers. Laws are often revised several times and it is a necessary part of the legal process that may be modified or extended [1]. Legal rules are more general in the present and for the future scenarios such rules must be applied [11].

Figure 2. Legal Definition Usage and Variety

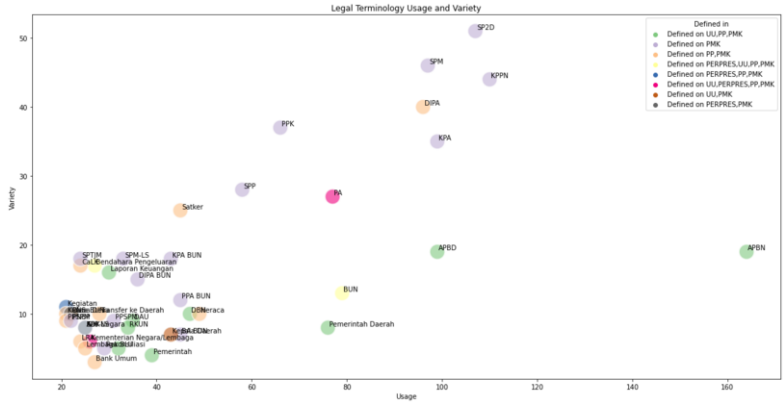


Table 3. Example of Term "SIKD" as evolving definitions

Definition	Year	Appear In
Regional Financial Information System, hereinafter abbreviated as SIKD, is a system that documents, administers, ***and processes other related data*** into information that is presented to the public and as [...]	2016-2017	93/PMK.07/2016; 18/PMK.07/2017;
Regional Financial Information System, hereinafter abbreviated as SIKD, is a system that documents, administers, and ***processes regional financial management data and other related data*** into information that is presented to the public and as [...]	2018-now	189/PMK.05/2018; 24/PMK.07/2020; 231/PMK.07/2020; 233/PMK.07/2020;

3.3. Potential Mismatch

Potential mismatch legal definition occurs when legal terms are defined differently across regulations on different level of regulations. We found 135 legal definitions that meet this condition. For instance the term "DIPA" has different definitions between those listed in the PP and the PMK as depicted in Table 4. This difference should not have occurred as it is clearly stated in Legal Drafting Guidance [7] that the general definitions in lower regulations cannot contradict with the definition of the higher regulation particularly in the same domain.

Table 4. Example of Potentialy Mismatch Definitions of Term "DIPA"

Definition	Appear In
DIPA is a ***budget execution (allotment) document*** which is used as ***reference*** for Budget Users in carrying out government activities as the implementation of the state revenue and expenditure budget	PP 45 of 2013;
DIPA is a ***budget execution (allotment) document*** prepared by the Budget User or Budget User Proxy	36/PMK.02/2015; 94/PMK.05/2016; etc.

Another example regarding potential conflict is when legal terms are defined differently across regulation on the same level. We found 658 legal definitions that meet this condition. For instance, inconsistency occurred in the definition of the term "SPTJM" (Table 5). This is not considered as evolving regulation since initially they used definition

A, then used definition B and using definition A again. Based on Legal Drafting Law [7] if a definition is reformulated in a new regulation, the formulation must be the same with the definition of the previous enforced regulation.

Table 5. Example of potential mismatch in definition of term "SPTJM"

Definition	**Appear In**
A. SPTJM is a statement letter which among other things contains a statement that all consequences of an official/person's actions that may result in state losses are the full responsibility of the official/person who took the action.	212/PMK.05/2020 156/PMK.05/2019
B. SPTJM is a statement of responsibility from the official for all expenses for payment of meal allowances and to return it to the state when overpayment and state loss.	110/PMK.05/2020

Therefore, to make regulations harmonized with each other as well as reduce risk of misinterpretation, a legal term must have consistent definition in every regulation, regardless of its type/level.

4. Conclusions

We presented a framework for exploring the consistency and to find potential mismatch of the use of legal definitions in regulations. The use of text mining for this purpose can be extended not only to other legal domains in Indonesia, but also by other jurisdictions or non-governmental organizations. Furthermore, the result also can be used as a baseline for building a legal terms dictionary that can be used by legal analyst/drafter.

References

[1] Ajani G, Boella G, Di Caro L, Robaldo L, Humphreys L, Praduroux S, Rossi P, Violato A. The European legal taxonomy syllabus: a multi-lingual, multi-level ontology framework to untangle the web of European legal terminology. Applied Ontology. 2016 Jan 1;11(4):325-75.

[2] Chiocchetti E. Harmonising legal terminology. EURAC research; 2008.

[3] Culy C, Chiocchetti E, Ralli N. Visualizing conceptual relations in legal terminology. In 2013 17th International Conference on Information Visualisation 2013 Jul 16 (pp. 333-338). IEEE.

[4] Gauci, Gotthard MarkİS It a Vessel, a Ship or a Boat, Is It Just a Craft, Or Is It Merely a Contrivance? Journal of Maritime Law and CommerceÖctober 2016, 47(4): 479

[5] Goyvaerts J. & Levithan S. Regular Expression Cookbook. BIM Publishing Servies. 2012.

[6] Hwang R.H., Hsueh YL, Chang YT. Building a Taiwan law ontology based on automatic legal definition extraction. Applied System Innovation. 2018 Sep;1(3):22.

[7] *Law No. 12 of 2011 on the Legislation Making as amended by Law No. 15 of 2019* (Indonesia).

[8] M. F. L. Schmitt and E. J. Spinosa, "Outlier detection on semantic space for sentiment analysis with convolutional neural networks," in 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Jul. 2018, ppİ-8, doi:10.1109/IJCNN.2018.8489200

[9] Mohammed S. M., Zeebaree S. R. M., & Jacksi K. A state of the art survey on semantic similarity for document clustering using GloVE and density based algorithms. Indonesian Journal of Electrical Engineering and Computer Science. April 2021.

[10] Mommers L, Voermans W. Using Legal Definitions to Increase the Accessibility of Legal Documents. In JURIX 2005 May 15 (pp. 147-156).

[11] Rawls J. Two concepts of rules. The philosophical review. 1955 Feb 1;64(1):3-2.

[12] Šavelka J, Ashley KD. Legal information retrieval for understanding statutory terms. Artificial Intelligence and Law. 2021 Jul 8:1-45.

[13] Winkelsm R, Hoekstra R. Automatic Extraction of Legal Concepts and Definitions. In Legal Knowledge and Information Systems: JURIX 2012: the 25th Annual Conference 2012 (Vol. 250, p. 157). IOS Press.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210325

Few-Shot Tuning Framework for Automated Terms of Service Generation

Ha Thanh NGUYEN [a,1], Kiyoaki SHIRAI [a] and Le Minh NGUYEN [a]
[a] *Japan Advanced Institute of Science and Technology*

Abstract. In this paper, we introduce BART2S a novel framework based on BART pretrained models to generate terms of service in high quality. The framework contains two parts: a generator finetuned with multiple tasks and a discriminator finetuned to distinguish the fair and unfair terms. Besides the novelty in design and the implementation contributions, the proposed framework can support drafting terms of service, a growing need in the digital age. Our proposed approach allows the system to reach a balance between automation and the will expression of the service provider. Through experiments, we demonstrate the effectiveness of the method and discuss potential future directions.

Keywords. few-shot tuning, terms of service, generation

1. INTRODUCTION

Natural language generation comes along with the development history of NLP. The first generation of these systems are simple rule-based systems, typically represented by ELIZA (1). These systems have a complex set of rules but can only generate language in a very limited context. Later systems with knowledge bases and statistical methods can perform this task better in more problems such as weather forecasts (2), storytelling (3), and dialogue system (4). With breakthroughs in hardware and architecture in recent years, transformer-based systems like GPT-2 (5), GPT-3 (6) and BART (7) are recently received great attention from both the industry and the research community.

These models have been very successful with destructive or teasing applications such as fake news, fake images, fake videos, but that does not guarantee they can generate high-quality content like terms of service. Compared to a naive copy-paste mechanism, a generative model generates not only the learned examples but also the synthetic sample from them. As a result, it gives editors more flexibility in drafting the documents. However, this problem contains two main difficulties. First, it requires a balance between automation and the will of the editor. Second, the meaning of the content that the system generates should be of high quality and fairness.

Finding a solution for terms of service generation problem, this paper proposes BART2S, a novel generator-discriminator framework for *terms of service generation*.

[1]Corresponding Author: Nguyen Ha Thanh, Japan Advanced Institute of Science and Technology; E-mail: nguyenhathanh@jaist.ac.jp

The generator is designed to solve the problem called *title-based generation* in order to both express the will of the editor and reduce their drafting effort. The editor provides a title and the framework completes the content according to the patterns it learned from the data. To implement this paradigm, we pretrain the generator with multiple sequence-to-sequence tasks. The discriminator is trained on the same vocabulary to classify the generated terms as fair or not. Our experiments show interesting results and prove the effectiveness of the approach.

2. BART2S Framework

2.1. Title Based Generation

Each content in terms of service has a title reflecting it. We recognize this feature as an opportunity for editors to participate in editing with minimal effort. The title is usually a sentence that describes the topic of the content and can even reflect the editor's point of view. Therefore, we use the title as information for the editor to guide the system. With a title as input and content as output, we propose the title-based generation problem. This problem can be considered as a conditional generation problem that the generated content must reflect the topic mentioned in the title.

The problem brings a challenge in signal recovery, usually, the information in the title is often much more concise than its content. To be able to fulfill the ideas from a short sentence or even a word of the title, the model needs to understand the patterns of idea development in a particular domain. For meaningful generated content, the pretraining stage needs to be done with the appropriate tasks and the appropriate data domain. We propose three tasks that help to train the generator for the desired goal: writing the next sentence, writing content from the title, and paraphrasing. In addition, we propose using a pretrained discriminator to evaluate and adjust the output of the generator.

2.2. Pretraining Encoder

The first task is the next sentence generation. We prepare the noised input similar to how BART (7) is trained. Given a noised sentence, the model needs to generate the sentence right after it. Our goal in training the model on this task is for the model to learn how to use words in the field of law. We assume that this skill can bring a better generation for title based generation problem. In addition, the title that the editor input can be an incomplete sentence. Trained by this task, the model is able to complete the idea from the input.

The second task that the model needs to learn is to generate content from the title. This task is directly related to the title-based generation problem. The title of a paragraph is usually a summary of it or the topic it covers. Generating content from the title demonstrates the model's ability to understand the title as well as find out the content representing that topic. This task also serves as a model guide in generating the desired output as the content from the title as input.

The third task is about paraphrase generation. The skills required in this task enable the model to understand the text and represent it in a different way. This task is useful to train the model not only for the flexibility of the model but also for the coherence of the

generated content. In essence, the content is a paraphrase of the title with ancillary information. Our assumption is that learning to paraphrase will help the model to generate the content from the title better.

2.3. Pretraining Discriminator

The discriminator proposed in this paper is used for regulating generated content. It is pretrained to distinguish between fair and unfair terms. The discriminator is fed by the input having the same format as the output of the generator. Let $C = [wc_1, wc_2, ..., wc_n]$ be the content and $L = [0, 1]$ be its corresponding fairness label. The discriminator is trained to map the content with their fairness label. This component makes our approach different from other systems; it enables us to build a system toward a constructive goal of generating high-quality content.

2.4. Cross-model Few-shot Tuning

The models can be represented as differentiable functions $G(x, \theta_g)$ and $D(x, \theta_d)$ with x, θ_g, and θ_d are the input, generator's parameters and discriminator's parameters, respectively. Tuning the generated output by the generator, we minimize $log(1 - D \circ G(x, \theta_g))$ using gradient descent process. In the backpropagation, for the loss to be able to pass through the two models, we replace *ArgMax* function at the last layer of the generator with *SoftArgMax* function represented in Equation 1.

$$\text{SoftArgMax}(x) = \sum_i \frac{e^{\beta x_i}}{\sum_j e^{\beta x_j}} i \tag{1}$$

where $x = [x_1, x_2, x_3, ..., x_n]$ and $\beta \geq 1$.

The discriminator's weights are frozen during the tuning process, which guarantees that this component is an independent observation. It is only based on the knowledge learned during the pretraining process to make the assessment. The loss reduces when the generator adjusts the generated content to make it fairer. This ability creates a bold difference in our framework compared to naive copy/paste systems and other non-regulated systems.

3. Experiments

3.1. Experimental Setup

Generator. Task 1's data is formed from the content of terms of service documents we collect on the Internet. Data for Task 3 is extracted from MSRP dataset (8), we only keep the sentence pairs with positive labels for paraphrasing. Data for Task 2 and data for evaluation are crawled from *Law Insider*[2], an online corpus that contains contract terms with their title. For each input, we use Token Masking, Token Deletion, and Token Infilling to transform the input in the same way that BART is pretrained. Sentence Permutation and

[2]https://lawinsider.com

Document Rotation are not applicable in this case. With this transformation, the model must learn the output based on the incomplete input. After being processed as above, our training data has 5,323 samples for Task 1; 901 samples for Task 2; and 3,728 samples for Task 3.

The terms of service generation problem is a multiple ground truth problem. In fact, there are many terms with the same name but different content. Therefore, we designed the test set to fit that characteristic. The ground truth of each title includes 1,000 most popular corresponding content according to statistics of Law Insider. Accordingly, the BLEU score is the measure we use to evaluate the performance of the model according to different training strategies.

Discriminator. For the discriminator, we use labelled samples provided by *ToS;DR project*[3] to finetune and evaluate the model. There are in total 4,152 samples in which 2,308 samples are fair terms with positive labels and 1,844 samples are unfair terms with negative labels. For both components, we use the configuration of *BART large* to initialize the models.

BART2S Framework. After finetuning the generator and the discriminator, we verify the BART2S framework presented in Section 2. The generator generates temporary output, and this output is assessed by the discriminator. The generator's weights are updated until the output meets the condition of fairness verified by the discriminator.

We design human-based metrics to evaluate and compare the tuned model with the BART2S framework and other candidates with the same configuration. In terms of ToS generation, we consider 4 aspects of good content as *Grammar*, *Readability*, *Relevance*, and *Fairness*. The content needs to be written in human language with good grammar, readable, and relevant to the title. Most importantly, the model needs to generate a fair term.

Among the release model of BART (7), we only consider models with BART Large configuration. Besides, for each classification and generation tuning task, we choose one candidate with the best performance reported by the authors. Finally, the three candidates to compare with the outputs of BART2S are as follows:

- BART Large w/o Ft: BART Large without finetuning on any task.
- BART Large MNLI: BART Large finetuned on MNLI dataset (9).
- BART Large CNN: BART Large finetuned on CNN-DM dataset (10).

These models are used as an end-to-end generator without the discriminator part as proposed in the BART2S framework. We create a collection of 30 short titles with an average length of 23 characters, feed them in the models and invite 10 evaluators to assess the generated content with the 4 metrics mentioned above. For each metric, we use a binary evaluation, the evaluators only need to check whether the content is acceptable or not. To avoid biases in the assessments, we only provide the evaluators with the title and the corresponding generated content. The evaluators do not know about the models and the process of generating the content.

The final score of each model in each aspect is calculated as in Formula 2.

$$score_a(\mathrm{M}) = \frac{1}{n}\sum_{i=1}^{n}\frac{p_a^i}{s} \tag{2}$$

[3] https://tosdr.org/

In which, $score_a(\mathrm{M})$ is the evaluation score of model M in aspect a, s is the total of sentences, n is the number of evaluators, p_a^i is the number of sentences evaluated as possitive by i^{th} evaluator in the aspect a.

3.2. Experimental Results

Approach	BLEU Score
All tasks	**60.07**
W/o Task 1	59.85
W/o Task 2	57.29
W/o Task 3	56.34

Table 1. Performance of generator trained with different approaches.

Generator. Table 1 summarizes our experimental results on training the generator with different settings. The model training with all tasks achieved the best performance. The surprising thing about the experimental results was that the model trained without Task 2 was not the model with the worst performance. From that result, we assume that Task 2 can be learned indirectly through the next sentence generation task and paraphrasing task. This once again confirms the idea of using multi-task learning for this problem is appropriate.

System	Grammar	Readability	Relevance	Fairness
BART Large w/o Ft	0.34	0.31	0.41	0.43
BART Large MNLI	0.32	0.33	0.37	0.37
BART Large CNN	0.69	0.73	0.72	0.86
BART2S	**0.80**	**0.82**	**0.87**	**0.94**

Table 2. Evaluation results on grammar, readability, relevance, and fairness of each system. The underlined line indicates our proposed system.

Discriminator. With the early stopping setting, the discriminator training process ends when the loss value on the validation set stops to decrease after 10 epochs. The accuracy on the training set is 66.6% and the accuracy on the validation set is 66%. These values reflect the difficulty of the fairness classification problem. It's not straightforward to detect an unfair term provided only its content. However these values are significantly greater than a random guess, which proves that there are latent patterns that support the model to do the task.

BART2S Framework. We feed the models with the short titles as described in Section 3.1. The max length of the generated content for every model is set to 512 subwords. With the given 30 titles, BART2S needs at most 2 epochs to tune the generator for generating desired content. Since we do not provide any ground truth of the data, BART2S is solely based on pretrained knowledge to adjust the outputs. Table 2 presents the evaluation results on grammar, readability, relevance, and fairness aspect of BART2S Framework and the controls. The BART2S framework leads all evaluation aspects, followed by the BART Large CNN model. BART Large w/o Finetuning model and the BART Large MNLI perform worst in the ranking.

4. Conclusions

This paper proposed BART2S, a regulated generative framework for generating terms of services automatically using the generative models. To ensure a balance between automation and expression of will, the framework is based on the title-based generation problem. The framework contains two sub modules as a generator and a discriminator. We use a custom pretrained model trained on 3 different tasks as the generator and a pretrained classification model with the same configurations as the discriminator. We also propose a novel tuning process to adjust the fairness of the generated content. The experimental results show that our approach is appropriate and the framework can produce high-quality results. Although this framework was proposed in the terms of service generation problem, the idea is general and can be applied to many different fields. Replacing the fairness discriminator with another set of constraints could be a potential direction.

References

[1] Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM. 1966;9(1):36-45.

[2] Angeli G, Liang P, Klein D. A simple domain-independent probabilistic approach to generation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing; 2010. p. 502-12.

[3] Holtzman A, Buys J, Forbes M, Bosselut A, Golub D, Choi Y. Learning to write with cooperative discriminators. arXiv preprint arXiv:180506087. 2018.

[4] Wolf T, Sanh V, Chaumond J, Delangue C. Transfertransfo: A transfer learning approach for neural network based conversational agents. arXiv preprint arXiv:190108149. 2019.

[5] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI blog. 2019;1(8):9.

[6] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv preprint arXiv:200514165. 2020.

[7] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:191013461. 2019.

[8] Dolan WB, Quirk C, Brockett C. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In: COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics; 2004. p. 350-6.

[9] Williams A, Nangia N, Bowman SR. A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:170405426. 2017.

[10] Hermann KM, Kočiskỳ T, Grefenstette E, Espeholt L, Kay W, Suleyman M, et al. Teaching machines to read and comprehend. arXiv preprint arXiv:150603340. 2015.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210326

An Information Retrieval Pipeline for Legislative Documents from the Brazilian Chamber of Deputies

Ellen SOUZA [a,b,1], Douglas VITÓRIO [a,d], Gyovana MORIYAMA [b], Luiz SANTOS [b], Lucas MARTINS [b], Mariana SOUZA [b], Márcio FONSECA [e], Nádia FÉLIX [b,c], André C. P. L. F. CARVALHO [b], Hidelberg O. ALBUQUERQUE [a,d], and Adriano L. I. OLIVEIRA [d]

[a] *MiningBR Research Group, Federal Rural University of Pernambuco, Brazil*
[b] *Institute of Mathematics and Computer Sciences, University of São Paulo, Brazil*
[c] *Institute of Informatics, Federal University of Goiás, Brazil*
[d] *Centro de Informática, Federal University of Pernambuco, Brazil*
[e] *Chamber of Deputies, Brasilia, Brazil*

Abstract. This work investigates information retrieval methods to address the existing difficulties on the *Preliminary Search*, part of the law making process from the Brazilian Chamber of Deputies. For such, different preprocessing approaches, stemmers, language models, and BM25 variants were compared. Two legislative corpora from Chamber were used to build and validate the pipeline. All texts were converted to lowercase and had stopwords, accentuation, and punctuation removed. Words were represented by their stem combined with word unigram and bigram language models. Retrieving the bill that was originated from a specific job request, the BM25L with Savoy stemmer reached a R@20 of 0.7356. After removing queries with inconsistencies or which made reference exclusively to attachments, to other job requests, or to bills, the R@20 increased to 0.94.

Keywords. Legal Information Retrieval, Legislative Document Retrieval, Brazilian Portuguese, BM25

1. Introduction

The Brazilian Chamber of Deputies was founded over two hundred years ago and has more than 20 thousand employees, including citizen representatives from all over the country. Since its founding, the Chamber has processed more than 144 thousand bills [1]. Each bill needs to be formalized as an initial legislative document *draft* and an optional justification document, which are submitted for discussion and voting. For a typical bill, a large number of documents is produced and aggregated in different stages of processing. This content, generated by the members of the parliament, is massive and keeps increasing. Besides, the unstructured nature of these documents makes their organization, access, and retrieval a challenging task [1].

[1]Corresponding Author: ellen.ramos@ufrpe.br.

A bill is submitted to the Legislative Consulting (CONLE), an advisory body of the House, whose main role is to provide the necessary support to the law making process. The CONLE has an internal team of specialists and researchers in 22 legal subjects, including economics, technology, and transportation. With the increasing demand for legislative production, a remarkable amount of legislative consulting requests is redundant, regarding other proposals already under analysis by the CONLE, and even existing laws. As consequence, a large deal of effort from the consulting team is devoted to this process, called *Preliminary Search*.

This work investigates the use of information retrieval (IR) methods to address these legislative production issues. Given a set of legislative documents and a query document (i.e., a job request), the system filters and ranks the documents according to their relevance to the query. The research is conducted in the context of the *Ulysses* project, an institutional set of artificial intelligence initiatives with the purpose of increasing transparency, improving the Chamber's relationship with citizens, and supporting the legislative activity with complex analysis [2]. This paper is organized as follows: Section 2 presents the major related studies. Section. 3 details the IR pipeline for Brazilian legislative documents. Section. 4 presents and discusses the obtained results. Section. 5 brings the conclusion and highlights future works.

2. Related Work

The only study found by the authors performing legislative document retrieval with data written in European Portuguese was [3]. In this study, a unsupervised document similarity algorithm is presented using sets of synonyms.The author's goal was to rank legislative documents based on their relevance to a query, regardless of the language used. Using the English, Spanish, French, and European Portuguese editions of the *JRC-Acquis* dataset they compared their unsupervised synset-based approach to a semi-supervised category-based one, reaching inferior results. The algorithm's performance was evaluated in terms of P@k (Precision at k documents): P@3, P@5, and P@10; achieving the results of 0.78, 0.75, and 0.71, respectively, for the Portuguese dataset.

Gomes and Ladeira [4] empirically evaluated the framework for case-law retrieval of the Brazilian Superior Court of Justice (STJ), comparing its legacy system to approaches based on text similarity: the TF-IDF traditional retrieval model, BM25, and four Word2Vec models. The STJ's system uses Boolean queries and the authors wanted to use free text as queries without any operator. The results reported, using NDCG@25 (Normalized Discounted Cumulative Gain with a cut off of 25 documents), demonstrated the superiority of BM25 based systems in this task, with a mean NDCG@25 equal to 0.752. Although the paper explored information retrieval in the real-world legal domain, it used a jurisprudence scenario, while, here, we are using legislative documents.

Another work investigating jurisprudence document retrieval and the impact of Stemming on the retrieval of real documents from the Court of Justice of the State of Sergipe (TJSE), in Brazil [5]. The authors compared four radicalization algorithms (Porter, RSLP, RSLP-S, and UniNE) to evaluate: 1) their gain in dimensionality reduction; 2) their predictive performance regarding legal document retrieval. the Okapi BM25 was used and evaluated by MAP (Mean Average Precision), MPC (Mean of Precision@10), and MRP (the average of R-Precision). RSLP obtained the largest dimen-

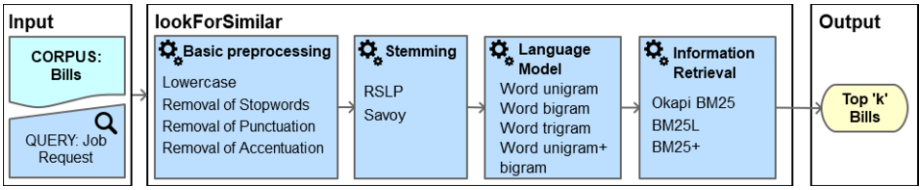


Figure 1. Brazilian Portuguese legislative information retrieval pipeline.

sionality reduction, while RSLP-S and UniNE were the best Stemming algorithms for IR, with the best MAP results, 0.87 and 0.88, respectively. According to the experimental results, the use of radicalization deteriorated the BM25 performance.

Chalkidis et al. [6] investigated regulatory compliance in EU and United Kingdom (UK) legislation using IR. They proposed a new approach, called Regulatory Information Retrieval (REG-IR), for document-to-document IR, in which a query is an entire document. The authors used two groups of legislation: EU directives and UK laws. REG-IR uses a neural IR system with a two-step pipeline: first, an IR algorithm (pre-fetcher) retrieves the top-k documents related to a query; next, a neural model re-ranks the documents. As pre-fetching algorithms, the authors evaluated Okapi BM25, W2V-CENT, BERT, S-BERT, LEGAL-BERT, C-BERT (BERT fine-tuned to predict EUROVOC concepts), and an ensemble of C-BERT and BM25; alongside six re-ranking techniques. Using R@100 (Recall at 100 documents) as metric to evaluate the pre-fetchers and R@20, NDCG@20, and R-Precision for the re-ranks, C-BERT was the best pre-fetcher for the datasets used, while the neural re-ranks failed to improve the retrieval performance.

Cantador and Sánchez [7] proposed a new approach for IR of parliamentary content, such as debate transcripts and laws proposals. The authors present a case study, in the Spanish Congress of Deputies, where they integrate their approach into *Parlamento2030*, an online platform that monitors parliamentary activity. They investigated the application of the Generalized Vector Space Model (GVSM) to the *Parlamento2030* dataset. The GVSM incorporates a semantic relatedness measure into the Vector Space Model (VSM), combined with an ontology-based document representation model. The authors used average P@5, P@10, P@15, and P@20. The results obtained (0.733, 0.683, 0.656, 0.600) were better than those obtained using just the matches of query and document key terms (0.633, 0.483, 0.422, 0.358).

3. The Method Used

Figure 1 presents the Brazilian Portuguese legislative IR pipeline[2]. The *job requests* are the queries and represent the user's input to the system. While the bills are the output answer, ranked according to a matching rate between the documents and the query (Subsection 3.1). We also evaluated basic preprocessing techniques (Subsection 3.2), two stemmers for the Portuguese language (Subsection 3.3), four word n-gram language models (Subsection 3.4), and three BM25 variants (Subsection 3.5).

[2] https://github.com/Convenio-Camara-dos-Deputados/BM25-Experiments

3.1. Corpora

Two legislative corpora from the Brazilian Chamber of Deputies were used to build and validate this pipeline: the *Bills* and the *Job Request* corpora. The former is available [3], while the latter has confidential information and cannot be made available.

The three most common types of bills were selected for the *Bills* Corpus: Law Project (Projeto de Lei - PL), Complementary Law Project (Projeto de Lei Complementar - PLC), and Constitutional Amendment Proposal (Proposta de Emenda Constituicional - PEC). The final corpus has 48,555 proposals. The attribute *imgArquivoTeorPDF*, which is the bill itself, was used in the experiments. It has an average of 300 words.

The *Job Request* corpus represents the user's query and contains 295 anonymized *Job Requests*. Data identifying the parliamentarian who made the request to CONLE were removed. This corpus has two attributes. The former contains the number of the bill that was originated from the *Job Request* specified in the latter attribute. Table 1 shows examples of parliamentarians' *Job Requests* (i.e. queries). Most requests have between 10 and 40 words.

Table 1. Samples from anonymized Job Request corpus.

Originated bill	**Job Request (user's query)**
PL XXXX/2019	Projeto para restabelecer na CLT a proibição de terceirização para atividade fim (Project to prohibit the outsourcing of core activity in the CLT)
PL XXXX/2019	Criação de PL, com base nos dois esboços encaminhados anexo. (Make of bill based on the two sketches sent in the attachment)
PL XXXX/2019	Solicito parecer pela aprovação de acordo com a solicitação XXXX/2019. (Request an opinion for the approval according to job request number XXXX/2019.)
PL XXXX/2019	Complementar parecer em função da apensação do PL XXXX/19 ao mesmo (Complementary opinion according to the PL XXXX/19)
PL XXXX/2019	Parlamentar solicita aprovação (Parliamentarian requests approval)

3.2. Basic Preprocessing

Both corpora presented in previous subsections had their texts converted to lowercase and had stopwords, accentuation, and punctuation removed. We evaluated each technique separately and all techniques together. The preprocessing techniques were performed using the Python NLTK. For the stopword removal, we used a Portuguese stopword list.

3.3. Stemming

The main purpose of stemming is to reduce the inflected words into its root form or stem. Thus, words can be mapped to the same concept, improving the process of IR, regarding its ability to index documents and to reduce data dimensionality [5]. RSLP and Savoy algorithm were chosen because of their effectiveness in the retrieval of documents [12,5, 13,14].

[3] https://drive.camara.leg.br/s/c3p2nLgLRcMz6eX

- RSLP (Removedor de Sufixos da Lingua Portuguesa): a rule-based algorithm developed by [9] and improved by [10]. Like Porter, it applies successive steps to remove the suffixes. As it was developed specially for Portuguese, it has more rules than Porter. It has 8 steps and a list of exception which prevents the algorithm from removing suffixes of words that have endings that are similar to suffixes.
- Savoy (UniNE): developed by Jacques Savoy in 2006, it presents stemmers for various languages, including Portuguese. The algorithm is simpler than the others, as it has less rules. It removes inflections attached to both nouns and adjectives, based on rules for the plural and feminine form. Our implementation is based on [11].

3.4. Language Model

An n-gram language model predicts the probability of a given n-gram within any sequence of words in the language. It is widely used in text mining [15,16], including in the legal domain [19]. An n-gram is a contiguous sequence of n items from a given sequence of text. These items can be phonemes, characters, words, and others. Unigram refers to n-gram of size 1, bigram refers to n-gram of size 2, and so on. In this work, we evaluated four different word n-gram combinations [17,15,18]

3.5. Information Retrieval

BM25 [20] is the most well-known scoring function for "bag of words" document retrieval [21]. It is derived from the binary independence relevance model to include within-document term frequency information and document length normalization in the probabilistic framework for IR [22]. The algorithm has also been used successfully in the retrieval of legal documents [5,4,6,23]. We implemented the variants presented in [24].

Okapi BM25 [20] scoring function estimates the relevance of a document d to a query q, based on the query terms appearing in d, regardless of their proximity within d: where q_i is the *i-th* query term, with $idf(q_i)$ inverse document frequency and $tf(q_i,d)$ term frequency. The formula for the Okapi BM25 is presented below:

$$score(q_i,d) = \frac{IDF(q_i)\cdot TF(q_i,d)(k_1+1)}{TF(q_i,d)+k_1(1-b+b\cdot\frac{|d|}{L})} \tag{1}$$

where $TF(q_i,d)$ is the frequency of term q_i in document d, $IDF(q_i)$ is the inverse document frequency of term q_i, $\mid d \mid$ is the number of terms in document d and L is the average number of terms per document. The effectiveness of BM25 is highly dependent on properly selecting the values of k_1 and b. In traditional *ad hoc* IR, k_1 is typically evaluated in the range [0, 3] (usually $k_1 \in [0.5, 2.0]$); b needs to be in [0, 1] (usually $b \in [0.3, 0.9]$) [24]. We defined the following parameters in our experiments: $k_1 = 1.5$, $b = 0.75$, and $\varepsilon = 0.25$.

BM25L [25] is built on the observation that Okapi penalizes more longer documents compared to shorter ones. It *shifts* the term frequency normalization formula to boost scores of very long documents. Finally, BM25+ encodes a general approach for dealing with the issue that ranking functions unfairly prefer shorter documents over longer ones. The proposal is to add a lower-bound bonus when a term appears at least one time in a document [26]. The difference with BM25L is a constant δ to the TF component.

3.6. Evaluation

We have only one relevant document for each query(see Table 1, because of this, we are evaluating the results in terms of *Recall* (R), which is the fraction of relevant documents that are retrieved. We are analyzing the results with R@20 (Recall at 20 documents).

4. Experimental Results

Table 2 presents the experimental results. We checked if the bill which was originated by a specific job request appears in the top-20 relevant documents retrieved by the BM25 algorithms. BM25L achieved the best results in almost all experiments, outperforming the Okapi variant which has been widely used and performed better in previous works [5, 4,6]. This may be due to the size of the documents used in our experiments.

Table 2. Experimental results with R@20 (Recall at 20 documents).

No.	Originated Bill	R@20		
	basic preprocessing	**Okapi**	**BM25L**	**BM25+**
1	no preprocessing	0,6441	0,6678	0,6542
2	lowercase	0,6542	0,6983	0,6814
3	lowercase + punctuation removal	0,6678	**0,7153**	0,6847
4	lowercase + punctuation and acetuation removal	0,6780	**0,7153**	0,6847
5	lowercase + punctuation, acetuation, and stopword removal	0,7085	**0,7153**	0,6847
	stemming	**Okapi**	**BM25L**	**BM25+**
6	stemming (RSLP)	0,6271	0,6847	0,6508
7	stemming (Savoy)	0,6203	0,6712	0,6441
8	lowercase + punctuation, acetuation, and stopword removal + stemming (RSLP)	0,7085	**0,7288**	0,6915
9	lowercase + punctuation, acetuation, and stopword removal + stemming (Savoy)	0,6949	0,7186	0,6881
	word n-gram	**Okapi**	**BM25L**	**BM25+**
10	bigram	0,5898	0,5864	0,5729
11	trigram	0,4881	0,4881	0,4983
12	unigram + bigram	0,6542	**0,6712**	0,6441
	word n-gram + basic preprocessing	**Okapi**	**BM25L**	**BM25+**
13	lowercase + punctuation, acetuation, and stopword removal + bigram	0,5932	0,5898	0,5932
14	lowercase + punctuation, acetuation, and stopword removal + trigram	0,4712	0,4712	0,4712
15	lowercase + punctuation, acetuation, and stopword removal + unigram and bigram	**0,7085**	0,7051	0,6983
	word n-gram + basic preprocessing + RSLP	**Okapi**	**BM25L**	**BM25+**
16	lowercase + punctuation, acetuation, and stopword removal + stemming (RSLP) + bigram	0,6373	0,6305	0,6305
17	lowercase + punctuation, acetuation, and stopword removal + stemming (RSLP) + trigram	0,4881	0,4847	0,4847
18	lowercase + punctuation, acetuation, and stopword removal + stemming (RSLP) + unigram and bigram	0,7220	**0,7322**	0,7017
	word n-gram + basic preprocessing + Savoy	**Okapi**	**BM25L**	**BM25+**
19	lowercase + punctuation, acetuation, and stopword removal + stemming (Savoy) + bigram	0,6237	0,6237	0,6237
20	lowercase + punctuation, acetuation, and stopword removal + stemming (Savoy) + trigram	0,4780	0,4780	0,4746
21	lowercase + punctuation, acetuation, and stopword removal + stemming (Savoy) + unigram and bigram	0,7288	**0,7356**	0,7051

For the BM25L, analyzing the basic preprocessing techniques, there was no difference between the removal of punctuation, accentuation, and stopwords. In order to reduce data dimensionality, two Stemming algorithms were evaluated, improving the pipeline result. RSLP performed better with basic preprocessing tecniques (Table 2, line 8), but Savoy performed slightly better in combination with unigram and bigram (Table 2, line 21). This was not observed by Oliveira and C. Junior [5], in whose study radicalization deteriorated the Okapi BM25 performance. Although Savoy showed a slightly better result than RSLP when combined with unigram and bigram, RSLP obtained the largest dimensionality reduction in the retrieval of legal documents [5]. Therefore, the use of

the word n-gram alone did not improve the results, but in combination with basic prepossessing (Table 2, line 5) and stemming (Table 2, line 21) the technique improved the pipeline result.

Considering our best result (Table 2, line 21), the algorithm failed to retrieve 55 queries from a total of 295 job requests (queries). The analysis of these queries showed the following problems with our *Job Request* corpus: 7 queries made reference only to attachments; (Table. 1, line 2); 6 queries made reference only to other job requests (Table. 1, line 3); 10 queries made reference only to a bill name (Table. 1, line 4); and 11 queries did not refer to any subject (Table. 1, line 5). For those 34 job requests, the BM25L needs more information in addition to the text presented in the query. Therefore, analyzing the remaining 21 failed job requests, it was possible to observe also that, for seven requests, the text presented in the query did not refer to the bill associated to it, increasing the BM25L R@20 to 0.94.

5. Conclusion and Future Work

This paper explored IR for the legislative domain in a real-world scenario. Our preprocessing approach converts text to lowercase, removes stopwords, accentuation, and punctuation. We evaluated RSLP and Savoy Stemming algorithm to reduce dimensionality, improving the performance of the IR pipeline. A combination of unigram and bigram also improved BM25 results. We compared different BM25 algorithms and the L outperformed the Okapi and Plus variants.

We plan to use word embedding language models to capture semantic knowledge. As highly relevant documents are more valuable than marginally [28], we parented to perform a rank evaluation in our pipeline as in [6], which have applied neural models to improve ranking ordering. Currently, we are evaluating Named Entity Recognition to expand those queries, as well as considering the user relevance feedback to improve the performance of the whole IR pipeline.

References

[1] Brandt MB. Modelagem da informação legislativa: arquitetura da informação para o processo legislativo brasileiro. Faculdade de Filosofia e Ciências da Universidade Estadual Paulista (UNESP); 2020.

[2] Almeida PGR. Uma jornada para um Parlamento inteligente: Câmara dos Deputados do Brasil. Red Información. 2021;24. Available from: https://www.redinnovacion.org/revista/red-información-edición-nº-24-marzo-2021.

[3] Badenes-Olmedo C, García JLR, Corcho Ó. Legal document retrieval across languages: topic hierarchies based on synsets. CoRR. 2019;abs/1911.12637.

[4] Gomes T, Ladeira M. A New Conceptual Framework for Enhancing Legal Information Retrieval at the Brazilian Superior Court of Justice. In: Proceedings of the 12th International Conference on Management of Digital EcoSystems; 2020. p. 26-29.

[5] Oliveira RA, C Junior M. Experimental Analysis of Stemming on Jurisprudential Documents Retrieval. Information. 2018;9(2).

[6] Chalkidis I, Fergadiotis M, Manginas N, Katakalou E, Malakasiotis P. Regulatory Compliance through Doc2Doc Information Retrieval: A case study in EU/UK legislation where text similarity has limitations. arXiv preprint arXiv:210110726. 2021.

[7] Cantador I, Sánchez LQ. Semantic Annotation and Retrieval of Parliamentary Content: A Case Study on the Spanish Congress of Deputies. In: Proc. of the Joint Conference of the Information Retrieval Communities in Europe. vol. 2621; 2020.

[8] Hotho A, Nürnberger A, Paaß G. A Brief Survey of Text Mining. Journal for Computational Linguistics and Language Technology. 2005:1-37.
[9] Orengo VM, Huyck C. A stemming algorithm for the Portuguese language. Proceedings Eighth Symposium on String Processing and Information Retrieval. 2001:186-193.
[10] Orengo VM, Buriol LS, Coelho AR. A study on the use of stemming for monolingual ad-hoc Portuguese information retrieval. In: Workshop of the Cross-Language Evaluation Forum for European Languages. Springer; 2006. p. 91–98.
[11] Savoy J. Light Stemming Approaches for the French, Portuguese, German and Hungarian Languages. In: SAC '06: Proceedings of the 2006 ACM symposium on Applied computing. New York, NY, USA: Association for Computing Machinery; 2006. p.1031–1035.
[12] de Oliveira RA, Colaço Júnior M. Assessing the Impact of Stemming Algorithms Applied to Judicial Jurisprudence-An Experimental Analysis. In: International Conference on Enterprise Information Systems. vol. 2. SCITEPRESS; 2017. p. 99–105.
[13] Flores FN, Moreira VP, Heuser CA. Assessing the impact of stemming accuracy on information retrieval. In: International Conference on Computational Processing of the Portuguese Language. Springer; 2010. p. 11–20.
[14] Flores FN, Moreira VP. Assessing the impact of Stemming Accuracy on Information Retrieval – A multilingual perspective. Information Processing & Management. 2016;52(5):840–854.
[15] Ravi K, Ravi V. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems. 2015;89.
[16] Castro DW, Souza E, Vitório D, Santos D, Oliveira ALI. Smoothed n-gram based models for tweet language identification: A case study of the Brazilian and European Portuguese national varieties. Applied Soft Computing. 2017;61:1160–1172.
[17] Pang B, Lee L. Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval. 2008;2(1–2):1–135.
[18] Tripathy A, Agrawal A, Rath SK. Classification of sentiment reviews using n-gram machine learning approach. Exp Sys with App. 2016;57:117 – 126.
[19] Katz DM, Bommarito MJ, Seaman J, Agichtein E. Legal n-grams? A simple approach to track the evolution of legal language. Frontiers in Artificial Intelligence and Applications. 2011;235(Vienna):167–168.
[20] Robertson S, Walker S, Jones S, Hancock-Beaulieu M, Gatford M. Okapi at TREC-3. In: TREC; 1994.
[21] Kamphuis C, de Vries AP, Boytsov L, Lin J. Which BM25 Do You Mean? A Large-Scale Reproducibility Study of Scoring Variants. In: Advances in Information Retrieval; 2020. p. 28–34.
[22] Robertson S, Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval. 2009;3:333–389.
[23] Bansal A, Bu Z, Mishra B, Wang S, Ashley KD, Grabmair M. Document Ranking with Citation Information and Oversampling Sentence Classification in the LUIMA Framework. In: JURIX; 2016.
[24] Trotman A, Puurula A, Burgess B. Improvements to BM25 and language models examined. ACM International Conference Proceeding Series. 2014;27-28-Nove:58–65.
[25] Lv Y, Zhai C. When Documents Are Very Long, BM25 Fails! In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2011. p. 1103–1104.
[26] Robertson S, Zaragoza H, Taylor M. Simple BM25 Extension to Multiple Weighted Fields. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. CIKM '04; 2004. p. 42–49.
[27] Manning CD, Raghavan P, Hinrich S. An Introduction to Information Retrieval Draft. c; 2009. Available from: http://www-nlp.stanford.edu/IR-book/.
[28] Järvelin K, Kekäläinen J. IR Evaluation Methods for Retrieving Highly Relevant Documents. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; 2000. p. 41–48.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210327

Signal Phrase Extraction: A Gateway to Information Retrieval Improvement in Law Texts

Michael VAN DER VEEN [a] and Natalia SIDOROVA [a]
[a] *Eindhoven University of Technology, The Netherlands*

Abstract. NLP-based techniques can support in improving understanding of legal text documents. In this work we present a semi-automatic framework to extract signal phrases from legislative texts for an arbitrary European language. Through a case study using Dutch legislation, we demonstrate that it is feasible to extract these phrases reliably with a small number of supporting domain experts. Finally, we argue how in future works our framework could be utilized with existing methods to be applied to different languages.

Keywords. information retrieval, legislative texts, signal phrase extraction

1. Introduction

Legislative texts are complex and information-dense by their nature. Automatic analysis of these texts is an active research field with applications in different areas ranging from document annotation and legal text generation [1] to rule extraction [2] and verdict prediction [3]. One of the complicating factors in a wider employment of NLP-techniques in the legal domain arises from the fact that all countries have their legislation in their national language(s). Many NLP-based techniques, like rule extraction or event mining [4], can potentially be implemented as multi-language tools. They are however based on the use of signal words or linguistic patterns [5], which are language specific. There are many efforts to provide support both for specific languages [6,7] and across multiple languages [8]. Still multiple gaps need to be filled to achieve this goal.

In this paper we focus on the problem of automated generation of categorized lists of signal words and linguistic patterns used in the legal domain and indicating causal or temporal relationships. These signal words and phrases are necessary for e.g. legal text annotation and rule extraction. Signal words and phrases used in the legal area often differ from the ones in regular language usage. For example, “mits” (provided that) is rarely used in modern spoken and written Dutch, but it is very common in legislative texts. Our goal is to develop a general framework for extracting signal words and phrases from legislative texts in a given language using language-independent techniques. We demonstrate the use of our approach on the example of the Dutch language.

In Section 2, we introduce our semi-automatic framework. In Section 3 we apply and evaluate our framework on Dutch legislation. We draw conclusions and discuss future work in Section 4.

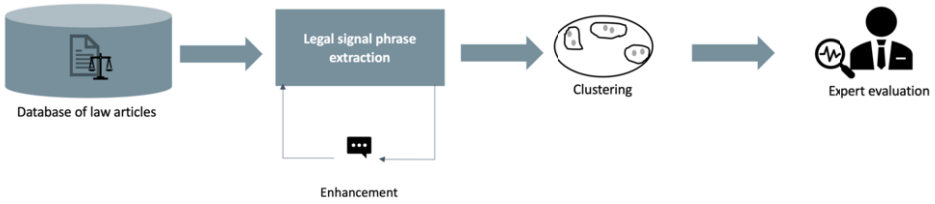


Figure 1. Framework to extract relevant signal words and phrases

2. Methodology and Framework

Our framework aims at the semi-automatic identification and categorisation of words and phrases indicating conditions and causal or temporal relationship between activities in legislation documents. We call these words and phrases *signal phrases*. Figure 1 illustrates the steps of the framework.

Step 1 First, we extract potential signal phrases. The extraction is based on thresholds for the total *frequency* of *n-grams* with $n \leq 3$ in all the laws and for the *coverage*, which is here the percentage of laws in which the signal phrase occurs. Sufficiently high coverage ensures that signal phrases are general for legal texts and span across multiple legal domains and laws. This reduces the number of false positives in the form of expressions frequent within certain legal areas but not used in other areas and therefore not carrying any causal or temporal meaning. Stop words are discarded from the search results. The choice of thresholds and the parameter 3 for n-grams was based on the input obtained from domain experts, who were enquired to deliver a list of typical signal phrases they expect to see in legal documents. Their lists were bundled and analysed on their total frequency and coverage and on their properties such as POS-tag. Our search strategy consisted of first selecting thresholds that would guarantee that all the key phrases provided by experts would be included in the search results. We do that to minimize the number of false negatives. Then, with each step we added more constraints on POS tags (based on our expert curated list' properties), e.g., including VERBS with coverage > 0.9. This process is repeated, until not too many phrases (< 400) are selected, while maintaining an adequate recall score.

Step 2 Phrases extracted in Step 1 are embedded and then clustered to their respective category. In our framework, we use the Universal Sentence Encoder introduced in [9], as it is able to handle multiple languages and n-grams with $n > 1$. We define a number of categories for the Dutch language based on [10], e.g., conditional, temporal and opposing. In further analysis, conditional phrases could be translated to different forms of implications, e.g., $A \longrightarrow B$, $\neg A \longrightarrow B$. We predefined a centroid per category to make sure that clustering leads to interpretable results. The embedded phrases are clustered around these pre-defined centroids using the cosine distance.

Step 3 Finally, interview sessions with domain experts were conducted. The main purpose of these sessions is to remove false-positive phrases. We also check the consistency of answers. Each interviewee receives a set of phrases consisting of two parts: one is the same for all of them (in order to check the consistency of answers) and the other is distinct. Additionally, each phrase is to be evaluated by two interviewees. Phrases are to be presented to interviewees in the context of their usage, in order to facilitate the work of experts.

3. Evaluation and Results

For this case study 1413 Dutch laws were utilized, scraped from the Dutch government website[1]. In the first step of our framework we initally set *frequency*: 1000 and *coverage*: 0.25 to include the expert curated phrase list of 36 items. After our first search we found 1453 phrases. Finally, after selecting phrases with POS-tags ADP, ADV, VERB(coverage$>$ 0.9) and SCONJ(coverage$>$ 0.35), 322 phrases were selected for our next step.

After the extraction step, we embedded the phrases and clustered them. To evaluate whether our embeddings worked correctly, we made a subset of phrases for which synonyms that originate from an online database[2] exist. Using the purity measure described in [11], we checked whether synonym phrases were assigned to the same cluster. This resulted in a perfect score of 1.0, which indicates an adequate embedding quality.

We conducted interviews with 5 experts. Each of them received 72 or 73 phrases from all clusters found in [10]. In the subset creation, we ensured that the consistency amongst the interviewees could be measured by including the same 10 randomly selected phrases to the set of each interviewee. Due to time constraints, we were not able to ensure that each phrase was evaluated twice. To check the consistency of evaluation of the 10 overlapping phrases we used the lower bound on the error relative to the (unknown) ground truth [12]. When the error rate is lower than 0.10, we can assume that the results consistently propagate to the non-overlapping phases [12]. The results of our interviews show that for classifying true positive (TP) phrases, we have an error rate of 0.08. The error rate for categorization was 0.24. This means that the selection of TP phrases can be considered as reliable. However, the evaluation of clusters is less sound. In future work, an experiment setting where at least 2 experts evaluate each phrase is required. Whenever they are in conflict, more analysis on context and semantics could prove useful One of the phrases where the experts were in conflict was ”op basis van” (based on). Some experts denoted this phrase as an explaining phrase, while others state that it is a referencing phrase. Both explanations are possible, depending on the context in which this phrase occurs and this shows that context information should be included in the analysis.

The experts selected $204/322(0.634)$ phrases as TPs. Several TPs were close to our predefined thresholds, which indicates that potentially there could be several false negatives. False positives were mostly phrases that are commonly used, but not specific enough for this research. Examples of such phrases are ”door” (by) and ”bedoeld” (meant). In the example for ”door”, we found that this phrase indicates a resource, which is not considered in the current setting of our research, as we focus on causal and temporal relations. It could be considered in future work since it maybe be important to extract such information. In the example ”bedoeld”, we found that this n-gram is too short to be recognized as relevant, ”als bedoeld” was considered a TP by experts.

From the selected TPs $104/204(0.510)$ were assigned automatically to the correct cluster. The clusters indicating examples and conditions were misclassified most often. This is probably due to the context-dependent nature of typical phrases in these clusters. It could also be caused by the fact that the embeddings used were trained on a regular corpus rather than a corpus specific for the Dutch legal domain. Such phrases sometimes have different meanings in regular language than in legislation.

[1] https://wetten.overheid.nl
[2] https://synoniemen.net

4. Conclusion

In this work we proposed a framework to semi-automatically mine signal phrases from legislative texts. This method combines automated processes with domain knowledge provided by experts. Furthermore, our case study demonstrated that a relatively small number of domain experts is required to filter out false positives consistently. Classification of clusters into categories requires more domain experts and further analysis. The quality of the language model used to generate embeddings is critical for successful automatic clustering of signal phrases. Identifying nuances in logical and temporal structures inside each cluster of signal words requires a collaboration of experts in law and in logics.

In future works we plan to integrate our technique with several others, namely [13]. We also plan to enhance our framework by using EU legislative texts, which are published in all 24 official languages of the EU. We expect that this will allow us to reduce the number of false positives and false negatives, as well as facilitate categorisation and interpretation of signal phrases.

References

[1] Dale R. Law and Word Order: NLP in Legal Tech. Natural Language Engineering. 2019;25(1):211-7.
[2] Dragoni M, Villata S, Rizzi W, Governatori G. Combining NLP approaches for Rule Extraction from Legal Documents. In: 1st Workshop on MIning and REasoning with Legal texts (MIREL 2016); 2016. .
[3] Medvedeva M, Vols M, Wieling M. Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law. 2020;28(2):237-66.
[4] Filtz E, Navas-Loro M, Santos C, Polleres A, Kirrane S. Events Matter: Extraction of Events from Court Decisions. In: Legal Knowledge and Information Systems. IOS Press; 2020. p. 33-42.
[5] van der Aa H, Di Ciccio C, Leopold H, Reijers HA. Extracting Declarative Process Models from Natural Language. In: International Conference on Advanced Information Systems Engineering. Springer; 2019. p. 365-82.
[6] Koeva S, Obreshkov N, Yalamov M. Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents. In: Proceedings of The 12th Language Resources and Evaluation Conference; 2020. p. 6988-94.
[7] Waltl B, Landthaler J, Scepankova E, Matthes F, Geiger T, Stocker C, et al. Automated Extraction of Semantic Information from German Legal Documents. In: IRIS: Internationales Rechtsinformatik Symposium; 2017. .
[8] Doncel VR, Ponsoda EM. LYNX: Towards a Legal Knowledge Graph for Multilingual Europe. Law in Context A Socio-legal Journal. 2020 Dec;37(1):175-8.
[9] Yang Y, Cer D, Ahmad A, Guo M, Law J, Constant N, et al. Multilingual Universal Sentence Encoder for Semantic Retrieval. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations; 2020. p. 87-94.
[10] Bovenhoff M, Zeijl W. Basisboek taal. Pearson Education; 2009.
[11] Sanderson M. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. ISBN-13 978-0-521-86571-5, xxi+ 482 pages. Natural Language Engineering. 2010;16(1):100-3.
[12] Smyth P. Bounds on the mean classification error rate of multiple experts. Pattern Recognition Letters. 1996;17(12):1253-7.
[13] Ferraro G, Lam HP, Tosatto SC, Olivieri F, Islam MB, van Beest N, et al. Automatic Extraction of Legal Norms: Evaluation of Natural Language Processing Tools. In: JSAI International Symposium on Artificial Intelligence. Springer; 2019. p. 64-81.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210328

Human Evaluation Experiment of Legal Information Retrieval Methods

Tereza NOVOTNÁ [a,1]
[a] *Institute of Law and Technology, Masaryk University, Brno, Czech Republic*

Abstract. In this article, I present the results of the human evaluation experiment of three commonly used methods in legal information retrieval and a new "multilayered" approach. I use the doc2vec model, citation network analysis and two topic modelling algorithms for the Czech Supreme Court decisions retrieval and evaluate their performance. To improve the accuracy of the results of these methods, I combine the methods in a "multilayered" way and perform the subsequent evaluation. Both evaluation experiments are conducted with a group of legal experts to assess the applicability and usability of the methods for legal information retrieval. The combination of the doc2vec and citations is found satisfactory accurate for practical use for the Czech court decisions retrieval.

Keywords. human evaluation, court decisions retrieval, doc2vec, citation analysis, LDA, multilayered approach

1. Introduction and Related Work

In this article, I summarize the results of two year long research on the application of different NLP methods to the Czech court decisions and human evaluation of these methods. In the first phase, I use semantic similarity doc2vec algorithm, citation network analysis and two topic modelling methods (Latent Dirichlet allocation as "LDA", non-negative matrix factorization as "NMF") to tackle different court decisions retrieval tasks. Afterwards, I evaluate all of the methods in the human evaluation experiment. The results of the first part of the research are rather average, therefore in the second phase I develop a new *multilayered approach* to achieve more accurate results. I again evaluate this approach in the human evaluation experiment and compare the results with the first phase evaluation results. I present here the results of the human evaluation experiments and their comparison.

The general research question that I try to answer is how accurate different commonly used methods for processing court decisions are for lawyers, who frequently perform court decisions research. The second (and more specific) research question is whether the combination of these methods leads to more accurate results. The third question is whether these results are good enough so that these methods could be the basis for practical court decisions search tools, which is a long-term goal of my research.

For the court decisions processing, I use the doc2vec method which was introduced by Le and Mikolov in [6]. In combination with the cosine similarity measure, it was

[1] E-mail: tereza.novotna@law.muni.cz.

successfully used in [7] to retrieve similar statutes or precedents to an in-hand document. Secondly, I use citation network analysis which examines the role of references among the set of legal documents, such as statutes, regulations or court decisions from which it creates the network. It is often applied to case law to observe citation patterns in [8], to improve the performance of a legal information retrieval system in [9] or for ranking of the importance of court decisions for court decision retrieval in [10]. LDA is a topic modelling algorithm introduced by Blei et al. in [1]. Legal documents clustering and summarization is a common application of topic modelling, as was shown in [2,3].

This article is structured as follows. Section 2 briefly describes the source data, the methods and a multilayered application of the methods. Section 3 contains the description of the evaluation experiment design and the evaluation group of lawyers. In Section 4, I summarize the most important results of the evaluation experiment and I discuss and compare the results with the first phase evaluation results. I conclude the article in Section 5.

2. Methodology

I conducted a two-phase evaluation experiment of three legal information retrieval methods. In the first phase, I used doc2vec, citation network analysis metrics and the topic modelling methods LDA and NMF. Afterwards, I asked legal experts to evaluate different tasks performed by these methods. Based on the evaluation results, I conducted a second phase evaluation experiment of the multilayered approach of these three methods.

2.1. Data

I used a dataset of the Czech Supreme court decisions available in the frame of the Czech Court Decisions Corpus (CzCDC 1.0) from [4]. This corpus is the only freely available set of court decisions of the Czech Supreme, Supreme Administrative and Constitutional Court. It contains raw texts of court decisions with basic metadata (date of publication, docket number, court). The Supreme Court subset of decisions contains 111 977 court decisions dated from 1994 to 2018. Nevertheless, I used the subset of the Supreme Court decisions related to the Czech Copyright Act from [11] to narrow the set of decisions to choose from in the evaluation.

2.2. Semantic Similarity - doc2vec

The first of the methods is the doc2vec model for semantically similar documents retrieval. The algorithm was used in standard settings and the model was trained for the whole dataset of the Czech Supreme Court decisions as described in [5]. The model provides for vector representations of court decisions and the similarity is computed as a cosine similarity measure between two vector representations. This method was used to retrieve semantically similar court decisions based on the *cosine similarity* measure and the similarity was evaluated in the evaluation experiment.

2.3. Citation Network Analysis

I used citation data from the freely available dataset of citation data of the Czech courts described in [12]. This data was used to explore several theoretical legal institutes, such as the precedent binding of court decisions of the Czech highest courts or citation practice of the Czech courts. I used *authority score* to indicate the domain importance of decisions and this importance was evaluated by legal experts in the following experiment.

2.4. Topic Modelling - LDA and NMF

I used LDA and NMF methods in the third experiment. Both methods are based on the assumption that the whole dataset consists of a set of latent topics and each document in the dataset is represented by these topics and their probabilities. These topics are characterized by a distribution over words. We again applied them to the dataset of the Supreme Court decisions and used the automatic coherence score metric to select the number of topics that the model should retrieve. The best models were the 30-topic LDA model and the 20-topic NMF model as described in [13]. I used the *three most probable topics* assigned by both models to court decisions and the relevance of the topics to the legal issues in presented decisions was evaluated by legal experts.

2.5. Multilayered Approach

Based on the evaluation results from the first phase of our research, I concluded that none of the three methods is simply applicable as such since the accuracy is not high enough. At the same time, mainly the doc2vec model and citation network analysis measures have the potential to be used when refined. Therefore, I applied the methods in a multilayered approach in the second phase of this research. The assumption behind the idea is that if the methods are applied in sequence, retrieved decisions (or metrics related to them) are refined and the strengths of the methods should be emphasized. I used doc2vec model, as it had the highest evaluation results, as a basic method, and I combined it with *1) citation network analysis* and *2) the 30-topic LDA model* in two partial experiments:

Ad 1) In the first partial experiment, the existence of a citation link between the decisions is used as a subsequent method to refine the doc2vec model. It is assumed that a pair of semantically similar decisions connected with a citation is more similar than a pair of semantically similar decisions without a mutual citation.

Ad 2) In the second partial experiment, the 30-topic LDA model is used as a method for the refinement of results because it is more accurate then NMF (Section 3.2.). It was assumed that a pair of semantically similar decisions with the same topic assigned to them is more similar than a pair of decisions with different topics assigned. Here, it is assumed that refining the most semantically similar decisions with the data on the same assigned topic should lead to a higher rating in evaluation.

As the doc2vec model is the basis of the multilayered approach, therefore the similarity of legal issues and background of court decisions is the evaluated characteristics.

3. Evaluation Experiment Design

The methods and data described in the previous Section are evaluated by the group of legal experts in the evaluation experiments. I look for data on the accuracy of all of the methods and potential improvement of the results of the multilayered approach compared to the other three methods. The general evaluation experiment design is based on asking legal experts to evaluate the accuracy of the methods via evaluation questionnaires. The questions on accuracy of different methods targets the specific goal of the individual experiment. That means, the similarity of retrieved decisions is evaluated for doc2vec model, the domain importance of decisions is evaluated for the citation analysis and the relevance of topics assigned to decisions is evaluated for topic modelling methods. For the multilayered approach, the similarity of decisions is evaluated because this approach is based on the doc2vec model.

3.1. Evaluation Group

The evaluation group in this experiment consists of 46 experts. Legal experts here are practicing lawyers from different legal fields as a high expertise in law is one of the key requirements for the evaluation participants. I asked judges (and court assistants) and lawyers as both of these categories work intensively with court decisions. Although both of the categories are not represented equally, I find it important to evaluate the methods by experts from different legal fields. In the first phase of our research, 26 legal experts participated in the evaluation. In the second phase, 20 legal experts participated in the evaluation.

3.2. Methodology of Evaluation

The evaluators were presented with court decisions to read through and evaluate in the form of online Google Form questionnaires. The evaluation experiment was conducted in two phases in accordance with the schedule described in the Section 2.

In the first phase, legal experts were asked to evaluate the doc2vec model, citation analysis and the two topic modelling method (LDA, NMF). For the doc2vec model, legal experts evaluated the similarity of legal issue and the factual background of the pairs of court decisions. They evaluated 26 pairs of the decisions with the smallest cosine distance (the highest similarity) and 26 pairs with the 10th smallest cosine distance (the 10th highest similarity) for comparison. The results are in the third and fourth column in Table 1. The evaluation scale was from 1 to 6 (1 means the least similar, 6 means the most similar).

Secondly, they evaluated the domain importance of 13 decisions with the highest and 13 decisions with the lowest (zero) authority scores. The evaluation scale was again from 1 to 6 (1 means the least important, 6 means the most important). The decisions with the highest authority score have an average rating of **3.42** and zero authority score decisions have an average rating of **2.73**, which is a significant difference.

Thirdly, they evaluated the relevance of assigned topics to the decisions. They evaluated a totally of 76 decisions with the set of three most probable topics retrieved by each model (LDA and NMF) for each decision. The evaluation scale was again from 1 to 6 (1 means the least relevant, 6 means the most relevant). The mean rating results of

both topic modelling methods were rather poor: **2.38** for the LDA model and **2.32** for the NMF model (on the scale from 1 to 6, where 6 means the most relevant).

In the second phase, legal experts evaluated the combination of the doc2vec model and citation analysis data and the doc2vec model and the LDA topics. They were asked to evaluate the similarity of legal issue and the factual background of the pairs of court decisions. Firstly, they evaluated 40 pairs of the decisions with the smallest cosine distance connected with mutual citation. Secondly, they evaluated 20 pairs of the decisions with the smallest cosine distance and with the same topic assigned by the LDA model and 20 pairs with the smallest cosine distance and with the different topic assigned by the LDA model for comparison. The evaluation scale was from 1 to 6 (1 means the least similar, 6 means the most similar).

4. Results and Discussion

I present the results of the evaluation experiments here and I compare the results of first phase experiments (doc2vec, citation analysis, topic modelling) with a multilayered approach (doc2vec and citation analysis, doc2vec and LDA topic model).

4.1. Means and Frequency of Ratings of doc2vec Model and Citations

Firstly, I consider the difference of the mean rating value of the two most similar decisions (third column of Table 1) and of the two most similar decisions connected with a citation (second column of Table 1) significant. Secondly, it is necessary to take into consideration the fact, that the evaluation group was high in legal expertise, however very domain diverse. Evaluated court decisions were decisions related to the Czech Copyright Act. The evaluation group on the other hand consists of lawyers from different legal fields and legal professions. That means that an expert in Copyright law will probably evaluate the same pair of presented decisions differently than a judge of criminal law. Regarding these reasons and regarding the fact, that the model shouldn't generally serve only a domain limited group of lawyers, but it should be used as widely as possible, these results are found sufficient.

The frequency of different rating values in Figure 1 only supports this conclusion. A vast majority of higher similarity ratings leads to the conclusion, that a vast majority of retrieved decisions with a mutual citation link are somehow relevant, even though lawyers find some differences either in the legal issue or in the background. Therefore, I find these results sufficient enough to create a base for the Czech Supreme Court decisions retrieval tool. The possible future steps and limitations of this idea will be discussed in the last Section.

	The most similar with a citation	The most similar	The 10^{th} most similar
Mean value	**4.4**	3.58	3.12

Table 1. Means of evaluation ratings of similarity of court decisions with/without a citation link

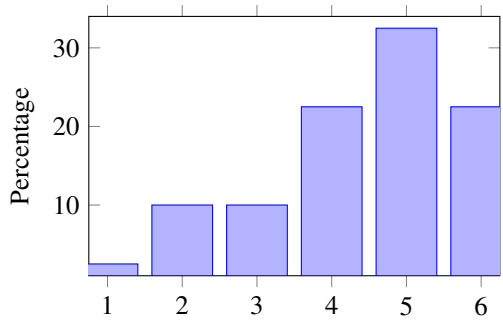


Figure 1. Frequency of evaluation ratings of similarity of court decisions with a citation link

4.2. Means and Frequency of Ratings of doc2vec Model and Topics

The topic modelling methods were the weakest in the first phase of research, LDA had better result than NMF. On the other hand, their potential is great in case the retrieved topics would be accurate enough. As the assigned topics could potentially mean another metadata layer for court decisions or even a very simple summarization of the text. Therefore, I decided to try to apply it in combination with the more accurate doc2vec model to see whether this combination could mean a way forward with LDA. The assumption here is that information on the most probable topic assigned by the 30-topic LDA model could discard potential false positives retrieved by the doc2vec model, i.e. court decisions retrieved with the highest cosine similarity but not relevant. This way, the combination could make the doc2vec model more accurate.

Generally, the results are better, but not great. The mean rating values are in Table 2, the results of the similarity of court decisions for comparison are in Table 1. The combination of methods is even slightly less accurate than the doc2vec applied solely. This combination of methods does not make the retrieval more accurate. On the other hand, when compared to the results of the LDA method itself in Section 3.2., the mean of evaluation ratings is significantly higher. However, this conclusion only supports the accuracy of the doc2vec model rather than the usability of the LDA topic model algorithm.

	The most similar with a same topic	The most similar with a different topic
Mean value	**3.25**	2.4

Table 2. Means of evaluation ratings of similarity of court decisions with assigned topics

5. Conclusion

The multilayered approach - the combination of the doc2vec model and a citation link - showed decent results when compared to the stand-alone application of the methods. The doc2vec model is a generally applicable algorithm with satisfactory results also in different domains, thus it is not a surprise. On the other hand, the citation data are originally created by judges and court assistants, i.e. subjectively and by highly qualified le-

gal experts. Therefore, it is again not a surprise that the citations make the text processing algorithm such as the doc2vec model more accurate and more relevant. On the other hand, the subjectivity, the context of a citation and last but not least, the time relevance of such citations need to be taken into consideration. Secondly, in line with expectations, the LDA method does not show sufficient results to be used in practice. The assigned topics, either alone or even in combination with the doc2vec model, were assessed as significantly less accurate in relation to the presented decisions.

Nevertheless, I find the presented results satisfactory as another step forward to a court decisions retrieval system in the Czech Republic. Additionally, I also consider these results to be important in terms of understanding how the methods used in this article work in practice when applied to legal sources. Although it is still a long way to go, I hope that this paper will lead to the practical application of methods and bridge the gap between legal informatics and daily legal practice in the Czech Republic.

6. Acknowledgment

I acknowledge the support of the ERDF project "Internal grant agency of Masaryk University" (No. CZ.02.2.69/0.0/0.0/19_073/0016943). I would like to thank Jakub Harašta for consultations and ideas for this research.

References

[1] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research. 2003; 3(4–5): p. 993–1022.

[2] Kumar VR, Raghuveer K. Legal Document Summarization using Latent Dirichlet Allocation. International Journal of Computer Science and Telecommunications. 2012 July; 3(7): p. 114-117.

[3] Lu Q, Conrad JG, Al-Kofahi K, Keenan W. Legal Document Clustering with Built-in Topic Segmentation. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management; c2011; New York, NY, USA; p. 383–392.

[4] Novotna T, Harasta J. The Czech Court Decisions Corpus (CzCDC): Availability as the First Step. 2019. arXiv preprint arXiv:1910.09513.

[5] Novotna T. Document Similarity of Czech Supreme Court decisions. Masaryk University Journal of Law and Technology. 2020; 14(1): p. 105-122.

[6] Le, Q., Mikolov, T. Distributed Representations of Sentences and Documents. International Conference on Machine Learning; 2014; p. 1188–1196.

[7] Renjit, S., Idicula S. M. CUSAT NLP@AILA-FIRE2019: Similarity in Legal Texts using Document Level Embeddings. Overview of the FIRE 2019 AILA track: Artificial Intelligence for Legal Assistance. Proc. of FIRE; 2019; p. 12-15.

[8] Fowler, J. H., Johnson, T. R., Spriggs, J. F., Jeon, S., Wahlbeck, P. J. Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. Political Analysis; 15(3); 2007; p. 324–346.

[9] Kumar, S. Similarity analysis of legal judgments and applying 'Paragraph-link'to find similar legal judgments (Doctoral dissertation, Ph. D. thesis, International Institute of Information Technology Hyderabad); 2014.

[10] Geist, A. The Open Revolution: Using Citation Analysis to Improve Legal Text Retrieval. European Journal of Legal Studies; 2008; 2(3); p. 137–145.

[11] Harašta, J. Srovnávací studie právních informačních systémů: Rozdíly mezi systémy při využití různých vyhledávacích strategií. Revue pro právo a technologie; 2020; 11(22); p. 219–260.

[12] Harašta, J., Novotná, T., Šavelka, J. Citation Data of Czech Apex Courts. arXiv:2002.02224; 2020.

[13] Novotná, T., Harašta, J., Kól J. Topic Modelling of the Czech Supreme Court Decisions. In: ASAIL 2020 Automated Semantic Analysis of Information in Legal Text; 2020; p. 1.5.

This page intentionally left blank

3. Logical and Conceptual Representations

This page intentionally left blank

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210330

A Kelsenian Deontic Logic

Agata CIABATTONI [a] Xavier PARENT [a] and Giovanni SARTOR [b]
[a] *Vienna University of Technology*
[b] *Cirsfid-Alma AI, University of Bologna; EUI, Florence*

Abstract. Inspired by Kelsen's view that norms establish causal-like connections between facts and sanctions, we develop a deontic logic in which a proposition is obligatory iff its complement causes a violation. We provide a logic for normative causality, define non-contextual and contextual notions of illicit and duty, and show that the logic of such duties is well-behaved and solves the main deontic paradoxes.

Keywords. Deontic Logic, Kelsen theory, Causality, Violations

1. Introduction

We develop a framework for deontic logic that combines violation and causality. Roughly speaking, an action is obligatory if refraining from performing it causes a violation. Our approach is inspired by the theory of norms developed by H. Kelsen, one of the most important legal scholars of the 20th century, in his "Pure Theory of Law," introduced in [1] and expanded in [2] (English translations in [3] and [4], respectively). According to Kelsen, obligations and prohibitions are mere reflexes of sanction-norms: "the legal order [...] prohibits a certain behavior by attaching to it a sanction or [...] it commands a behavior by attaching a sanction to the opposite behavior" [4, p. 55]. This idea may be connected to the reduction of deontic logic to alethic modal logic as proposed by Anderson [5], though the latter does not refer to the work of Kelsen (which at that time was only available in German, and usually only known to legal theorists).

Here we use a reduction *à la* Anderson, but depart from it in two respects. First, we model the connection between a sanction (actually, a violation) and its triggering condition as a causal relationship, rather than as a necessity relationship, as proposed in, e.g., [6]. The necessity connection is usually understood as a strict implication which is known to generate counter-intuitive inferences, such as Ross's paradox [7]. It also satisfies the reflexivity postulate ("If A then A"), which is often regarded as inappropriate for causal reasoning. Furthermore, Anderson's reduction is usually worked out within a possible worlds semantics, which is not compatible with Kelsen's view that norms do not bear a truth-value (see [8]). Second, in Anderson's perspective the same sanction is the consequent of each norm. In our approach instead different unlawful facts may lead to different violations. In this regard our approach corresponds to the working of legal and moral systems, where distinct unlawful or immoral acts lead to distinct sanctions or disvalues. This feature of our framework enables us to address contrary-to-duty CTD obligations, i.e., obligations which are applicable only if other obligations are violated. We can represent the original obligation by a norm linking the (prohibited) fact f_1 to a violation v_1, and the CTD obligation by a second norm linking the accomplishment of f_1 in combination with a further fact f_2 to an additional violation v_2. The first norm

expresses the obligation of $\neg f_1$ (e.g., the obligation not to kill, in Forrester's famous paradox [9]), while the second norm expresses the prohibition of f_2, when f_1 is the case (the prohibition to be cruel, when killing). Hence we model CTD obligations by making the complement of their content into an aggravating circumstance, as in legal codes.

Before presenting our formal framework, we clarify some ideas about sanction and violation. Neither Kelsen nor Anderson claimed that every action for which a sanction is foreseen will necessarily be followed by an action of coercive enforcement by the state (forced execution, fine, detention, etc.), nor even by the pronouncement of a sanction by a competent authority. Once the condition for the sanction is realised, what necessarily happens is that (for Kelsen) the sanction is authorised and thus can legitimately be applied through the appropriate legal process, or (for Anderson) that something unwanted has happened. Hence we will model legal norms as connecting unlawful facts to violations, rather than to sanctions.

The paper is organised as follows. Sect. 2 presents the norms we deal with, which causally link (unlawful) states of affairs to violations, and the logic to reason about them. The latter is a simplified version of input-output logic [10], where part of the normative system (regulative norms) has only a propositional constant on the right hand side. In Sect. 3 we discuss the notion of illicit, which is used to define contextual and non-contextual duties. In accordance with Kelsen's view that a normative system can be conflicting, Sect. 4 introduces the notion of (un-)obeyable system. The resulting logic of duties is analysed in Sect. 5, using as benchmarks well known properties and paradoxes from the deontic logic literature. Sect. 6 pinpoints a selection of topics for future research.

2. The violation logic

We introduce the base violation logic that will be used in this paper, starting with its language.

Definition 1 (Language)**.** *Let L be an ordinary propositional language containing the classical connectives and constants $\{\wedge, \vee, \neg, \rightarrow, t, f\}$, and a set V of violation atoms.*

Each violation atom denotes a particular violation or offence (or, following Kelsen, the authorisation to enact a specific sanction [4, 108ff]). Norms consist in causal-like connections, denoted by the $\Rightarrow$ symbol. They link factual circumstances to violations (regulative or violation norms) or Boolean antecedents to conclusions other than violations (constitutive or counts-as norms).

Definition 2 (Norms and Norm codes)**.** *A* norm code *is a finite set of:*

- Violations norms*: $A \Rightarrow v$, where each A is formula from $L \setminus V$ and $v \in V$.*
- Constitutive norms *(count-as norms): $A \Rightarrow B$, where A, B are formulas from $L \setminus V$.*

Example 1 (Auto code)**.** *A norm code capturing simple road traffic rules is:*

$\text{Speed} \Rightarrow v_1$	$\text{Red} \wedge \neg\text{Stop} \Rightarrow v_2$
$\text{Dark} \wedge \neg\text{LightsOn} \Rightarrow v_3$	$\text{BrokenLights} \Rightarrow v_4$
$\text{Phone} \wedge \text{Drive} \Rightarrow v_5$	$\text{BrokenFrontLights} \Rightarrow \text{BrokenLights}$
$\text{BrokenBackLights} \Rightarrow \text{BrokenLights}$	$\text{Fog} \wedge \neg\text{LightsOn} \Rightarrow v_6$

Note that in our example different norms establish different violations. This feature is meant to capture the fact that unlawful or immoral situations may trigger distinct re-

sponses by a legal and moral system. Such responses have to be added up to determine how the situation is assessed by the system (expanding the generated violations makes things worse). Our approach does not exclude that distinct facts may lead to the same violation. However, in this case the normative system would generate a single violation, whenever one, some, or all such norms are triggered. For instance, the violation v_4 may be triggered in two ways (via *BrokenFrontLights* or via *BrokenBackLights*). Should distinct fines be applied for broken front lights and broken back lights (to be added up when both are the case), different violations would be triggered by each of them.

We introduce our violation logic, that specifies a causal-like entailment for norms, in the spirit of Kelsen (who calls this entailment "imputation", see [4, 76ff]).

Definition 3 (Violation Logic)**.** *A violation inference relation is a binary relation* $\Rightarrow$ *between the set of propositions in L satisfying the following rules* ($\models$ *is the semantical consequence in classical propositional logic):*

(Truth) $t \Rightarrow t$
(Strengthening) If $A \models B$ *and* $B \Rightarrow C$*, then* $A \Rightarrow C$*;*
(Weakening) If $A \Rightarrow B$ *and* $B \models C$*, then* $A \Rightarrow C$*;*
(And) If $A \Rightarrow B$ *and* $A \Rightarrow C$ *then* $A \Rightarrow B \wedge C$*;*
(Cut) If $A \Rightarrow B$ *and* $A \wedge B \Rightarrow C$*, then* $A \Rightarrow C$*.*
(Or) If $A \Rightarrow C$ *and* $B \Rightarrow C$*, then* $A \vee B \Rightarrow C$*.*

Although causal relations satisfy most of the rules for classical entailment, their distinctive feature is that they are irreflexive, that is, they do not satisfy $A \Rightarrow A$. Actually, the above relation corresponds to the "basic reusable" input-output logic (out_4) from [10]. It is also closely related to Bochman's causal calculus [11], see Remark 1. The semantics for the violation logic is essentially the one for out_4. It is "operational", and takes the form of a set of procedures yielding outputs for inputs. Roughly speaking, to determine if formula A is in the output set, one considers in turn each maximal consistent extension of the input set that is closed under the norms, and checks if it contains A. If the answer is "yes", then A is in the output. This semantics fits Kelsen's idea that norms do not bear a truth-value and that logic cannot add new norms to a code N [12, Ch. 50], but rather identifies the input-output connections established by N, which are the object of "rules (propositions) of law" (in German *Rechtssätze*)—see [4, p. 72] and [12, Ch. 49].

3. From illicits to duties

In Kelsen's legal theory sanction norms (violation norms, in our framework) have a foundational status. Other normative notions are derivative. Every behaviour that may trigger a sanction against its author is a delict (*Unrecht*) and every delict is the content of the obligation that the delict does not take place [2, p. 39]. Here we prefer to speak of an illicit, rather than of a delict, to cover all elements (not only actions) that contribute to the triggering of a sanction. Elementary illicits represent the minimal conditions that lead to a violation, and elementary duties apply to their negations.

Definition 4 (Elementary Illicit and Duty)**.** *A conjunction of literals* $\bigwedge\{l_1, \ldots, l_n\}$ *is an* elementary illicit *relatively to a norm code N if and only if (iff) there is a* $v \in V$ *such that: 1.* $\bigwedge\{l_1, \ldots, l_n\} \Rightarrow v$*; 2. no proper subset of* $\{l_1, \ldots, l_n\}$ *satisfies condition 1.*

A proposition A is (the content of) an elementary duty*, relatively to a norm code N, iff* $A \equiv \neg B$ *and B is an elementary illicit relatively to N.*

Example 2. *In Ex. 1,* $\{\text{Speed}\}$, $\{\text{Red}, \neg\text{Stop}\}$, $\{\text{Dark}, \neg\text{LightsOn}\}$, $\{\text{BrokenFrontLights}\}$, $\{\text{BrokenBackLights}\}$ *and* $\{\text{Phone}, \neg\,\text{LightsOn}\}$ *are elementary illicits ; their negations* ($\neg\text{Speed}$, $\neg\text{Red} \vee \text{Stop}$, *etc.) are the content of elementary duties.*

To develop a deontic logic we introduce the idea of generalized illicit, which covers *all* possible conditions that may minimally lead to a disjunction of violations.

Definition 5 (Generalised illicit)**.** *A boolean formula A is a generalised illicit relatively to a norm code N if there exists a set of violations* $S \subseteq V$ *such that:*
(1) $A \Rightarrow \bigvee S$, *and* (2) *there is no B such that* $A \models B$, $B \not\models A$ *and* $B \Rightarrow \bigvee S$.

The rationale behind this notion is to ensure that the generalised illicit is not only a sufficient, but also a necessary condition for the disjunction of the violations to take place. Condition 2. makes A the weakest formula leading to the violations in question. This will be the key element to the solution of the deontic paradoxes in Sect. 5.2.

Example 3. *Consider again Ex. 1. The generalized illicits w.r.t.* $\{v_4\}$ *and* $\{v_1, v_5\}$ *are:* $\text{BrokenFrontLights} \vee \text{BrokenBackLights}$ *and* $\text{Speed} \vee (\text{Phone} \wedge \text{Drive})$, *respectively. Note that* BrokenFrontLights *alone is not a generalized illicits as it is not the only possible way of triggering the violation* v_4 *(i.e., condition 2. in Def. 5 fails).*

Generalised illicts have the following logical properties: they are reducible to a disjunction of elementary illicits, and their disjunction constitutes a new generalised illict.

Proposition 1. *If A is a generalised illicit relatively to a norm code N, then there exists a disjunction of elementary illicits* $\{L_1, \ldots, L_n\}$ *such that* $A \equiv L_1 \vee \ldots \vee L_n$.

Proof. Let $L_1, \ldots L_n$ be *all* possible elementary illicits relative to N w.r.t. the violations $v \in S \subseteq V$. $L_1 \vee \ldots \vee L_n$ is a generalized illicit w.r.t. $\bigvee S$. Condition 1 of Def. 5 is satisfied. Indeed, if $L_i \Rightarrow v \in S$, then by $(Weakening)$ $L_i \Rightarrow \bigvee S$, and by (Or) $L_1 \vee \ldots \vee L_n \Rightarrow \bigvee S$.

As for condition 2, assume there is a boolean formula B such that $L_1 \vee \ldots \vee L_n \models B$ and $B \Rightarrow \bigvee S$. Assume w.l.o.g. that B is in disjunctive normal form, say $B := N_1 \vee \cdots \vee N_m$. By $(Strenghtening)$ and the assumption $N_1 \vee \cdots \vee N_m \Rightarrow \bigvee S$ it follows that $N_i \Rightarrow \bigvee S$, for all $i = 1, \ldots, m$. Being the L_i's all the elementary illicits triggering the violations in S (and hence representing minimal conditions to trigger them), for each N_i there are some literals $\{L_1, \ldots L_n\}$ which are included in the literals of N_i. Hence $B \models L_1 \vee \ldots \vee L_n$. □

Corollary 1. *If* A_1 *and* A_2 *are generalized illicit relatively to N, so is* $A_1 \vee A_2$.

The concept of generalised illicits leads us to noncontextual duties, i.e., states of affairs whose non-realisation lead to a violation. More formally:

Definition 6 (Noncontextual Duty)**.** *A Boolean formula A is (the content of) a noncontextual duty relatively to a norm code N, denoted as* $\mathbf{O}_N A$, *iff* $A \equiv \neg B$ *and B is a generalised illicit relatively to N.*

Example 4 (Ctd. from Ex. 3)**.** $\mathbf{O}_N(\neg\,\text{BrokenFrontLights} \wedge \neg\text{BrokenBackLights})$ *and* $\mathbf{O}_N(\neg\text{Speed} \wedge (\text{Fog} \rightarrow \text{LightsOn}))$ *are duties.*

Corollary 2. *If* $\mathbf{O}_N A$ *then* $A \equiv \neg L_1 \wedge \ldots \wedge \neg L_n$, *for* $L_1, \ldots, L_n$ *elementary duties.*

Remark 1. *Out_4 [10], and hence our violation logic, differ from the causal calculus in [11] by the presence in the latter of axiom $f \Rightarrow f$. This axiom would create the following counter-intuitive situation when considering the corresponding obligations. From $A \Rightarrow B$ follows $A \wedge \neg B \Rightarrow f$.*[1] *Now, by (Weakening), for any violation v, $A \wedge \neg B \Rightarrow v$ (using the fact that $f \models v$), which leads to $\mathbf{O}_N(\neg A \vee B)$, for any constitutive norm $A \Rightarrow B$.*

3.1. Putting illicits and duties in context

We consider how norms operate relatively to contexts, i.e., in circumstances considered to be settled. We will regard contexts as kind of restrictions of the set of possible world, which are limited to those satisfying the context. These restrictions may depend on different considerations such as natural necessity (laws of nature), temporal necessity (the immutability of the past), or even the settled choices of the agent.

Definition 7 (Context)**.** *A context for a norm code is a consistent set of literals.*

Here we focus on the illicits that are not settled by the context (entailed by it), so that their happening is contingent on the choice (deliberation) of the involved agent.

Definition 8 (Contextual Illicit)**.** *A Boolean formula A is a contextual illicit (c-illicit) relative to a norm code N and context C iff there is a set of violations $S \subseteq V$ s.t.*

1. $\bigwedge(C \cup \{A\}) \Rightarrow \bigvee S$
2. *there is no B such that $A \models B$, $B \not\models A$ and $\bigwedge(C \cup \{B\}) \Rightarrow \bigvee S$*
3. $\bigwedge C \not\Rightarrow \bigvee S$ *and* $C \not\models \neg A$

Establishing that no weaker formula generates the considered violations, condition 2, is useful to resolve a number of deontic paradoxes, in particular those following from the assumption of closure of the deontic operator under logical consequence. Condition 3 tells us that A is "needed" to generate the violation in question, and also that the truth or falsity of A is not settled by the context, as shown in Remark 2.

Remark 2. *$\bigwedge C \not\Rightarrow \bigvee S$ in Def 8 (3) implies that $C \not\models A$. By condition 1 in Def 8, $\bigwedge(C \cup \{A\}) \Rightarrow \bigvee S$. Suppose $C \models A$. By propositional logic $\bigwedge C \models \bigwedge(C \cup \{A\})$. By (Strengthening), one gets $\bigwedge C \Rightarrow \bigvee S$. Contradiction.*

Example 5. *(Ctd. from Ex 1) In context $C = \{BrokenFrontLights\}$, $BrokenBackLights$ is not a c-illicit (due to the first half of condition 3 in Definition 8). The intuition is that when it is settled that one the two requirements leading to the same violation v_4 is met, meeting the other becomes irrelevant.*

On the basis of c-illicts we define contextual duties.

Definition 9 (Contextual Duty)**.** *A Boolean formula A is a contextual duty relatively to a norm set N and to a context C, denoted as $\mathbf{O}_{(N,C)}(A)$ iff $A \equiv \neg B$ and B is a c-illicit relatively to N and C (we omit N and C, when no ambiguity occurs).*

We apply below our approach to two well-known deontic paradoxes pertaining to contrary-to-duty (CTD) scenarios (see Sect. 5.2 for more paradoxes).

[1] Proof: If $A \Rightarrow B$, then $A \wedge \neg B \Rightarrow B$ by (Strengthening). By $f \Rightarrow f$ and (Strengthening), $A \wedge \neg B \wedge B \Rightarrow f$. By (Cut), $A \wedge \neg B \Rightarrow f$.

Example 6 (Forrester paradox [9])**.** *Consider the following premises: (1) You should not kill (2) If you kill, you should kill gently, (3) You kill. In Standard Deontic Logic SDL [13] (1)-(3) entail that you should both kill and not kill. By considering under what circumstance obligations (1) and (2) would be violated we have the violation norms*

$\text{Kill} \Rightarrow v_1 \qquad \text{Kill} \wedge \neg\text{KillGently} \Rightarrow v_2$

where $\text{KillGently} \Rightarrow \text{Kill}$. *In context* $\{\}$, $\mathbf{O}\neg\text{Kill}$ *holds, in context* $\{\text{Kill}\}$, $\mathbf{O}\text{KillGently}$.

Example 7 (Chisholm paradox [14])**.** *It consists of: (1) You ought to go to the assistance of your neighbours; (2) If you go you ought to tell them that you are coming; but (3) If you do not go then you ought not to tell them that you are coming; and (4) You do not go. SDL [13] entails that both you ought and you ought not to tell your neighbours that you are coming. In our framework the norms involved in this scenario are formalised as:*

$\neg\text{Go} \Rightarrow v_1 \qquad \text{Go} \wedge \neg\text{TellGo} \Rightarrow v_2 \qquad \neg\text{Go} \wedge \text{TellGo} \Rightarrow v_3$

In context $\{\}$, $\mathbf{O}\text{Go}$ *holds, in context* $\{\neg\text{Go}\}$, $\mathbf{O}\neg\text{TellGo}$.

Remark 3. *Some c-illicit A relative to a norm code N and context C_1, might not be a c-illicit relative to a superset C_2 of C_1 that is consistent with A and such that $C_2 \not\models \vee S$, for any (sub)set S of the violations in N. E.g., put $N = \{A \wedge B \Rightarrow v_1, D \wedge E \Rightarrow v_2\}$; $A \wedge B$ is a c-illicit in context $C_1 := \{E\}$, but not in $C_2 := \{A, E\}$ (condition 2 in Def. 8 fails).*

The following property, connecting violations entailed by contexts and duties, will be useful in Sect.5.2.

Lemma 1. *Let N be a norm code. If $C \not\models \neg A_1 \vee \cdots \vee \neg A_n$, for all duties $\mathbf{O}_N A_1, \ldots, \mathbf{O}_N A_n$ then $\bigwedge C \not\Rightarrow \bigvee S$, for every $S \subseteq V$.*

Proof. By Def. 6 each $\neg A_i$ is a generalized illicit relative to N and by Prop. 1 a disjunction of elementary illicits (minimal formulas that trigger violations). If there exists S s.t. $\bigwedge C \Rightarrow \bigvee S$, there are generalized illicits $\neg A_1, \ldots \neg A_m$ such that $C \models \neg A_1 \vee \cdots \vee \neg A_m$. □

4. (Un)Obeyability of normative codes

Kelsen pointed out that a normative code may establish requirements (cf. [15, p.25]) that cannot be jointly complied with: : "within [...] a normative order the same behaviour may be [...] commanded and forbidden at the same time [...]. This is the case if a certain conduct is the condition of a sanction and at the same time the omission of this conduct is also the condition of a sanction." We formalize this intuition through two notions, absolute and contextual unobeyability.

Definition 10 (Absolute (un)obeyability)**.** *A code N is absolutely unobeyable iff $t \Rightarrow \bigvee V$, and it is absolutely obeyable otherwise.*

Example 8. *An absolutely unobeyable code is $\{speed \Rightarrow v_1; \neg speed \Rightarrow v_2\}$. Indeed by (Weakening) and (Or) we get $t \Rightarrow v_1 \vee v_2$.*

Absolutely unobeyable codes are rare, as they involve norms that always establish sanctions. A weaker and more common notion is that of contextual unobeyability. A code N is unobeyable in a context C iff C entails that that a disjunction of alternative violations will be committed, not specifying which ones will be. Thus, the agent faces a predicament: possible violations can only be avoided by incurring in other violations.

Definition 11 (Contextual unobeyability)**.** *A code N is unobeyable in a context C iff there is a set of violations $V_i \subseteq V$ such that:* (1) $\bigwedge C \Rightarrow \bigvee V_i$, *and* (2) *for each $v \in V_i$, $\bigwedge C \not\Rightarrow v$.*

Example 9. *The code $N := \{\text{speed} \Rightarrow v_1; \text{motorway} \wedge \neg\text{speed} \Rightarrow v_2\}$ is unobeyable in context $\{\text{motorway}\}$ while being obeyable in context $\{\}$ (through $\neg$Speed). Since for no set of violations $V_i \subseteq V$, $t \Rightarrow_N \bigvee V_i$, N is absolutely obeyable.*

When a code N is unobeyable in a context C, in that context violations cannot be avoided and the addressees of N are forced to deliberate on which not yet settled violation to commit. They will have to face a "tragic dilemma", as in the biblical story below.

Example 10 (from the Book of the Judges)**.** *Jephthah promised to God that if he was given victory in a battle he would sacrifice (kill and dedicate to God) the first human being that he encountered coming home. After winning the battle, he first encountered his daughter (rather than an animal, as he may have assumed). Thus, he faced a hard choice: either violate his promise to God, or violate the moral prohibition to kill his daughter. We can model this situation through the code $N = \{\text{Win(j)} \wedge \text{Encounter(j,d)} \wedge \neg\text{Kill(j,d)} \Rightarrow v_1; \text{Kill(j,d)} \Rightarrow v_2\}$, which is unobeyable in context $C = \{\text{Win(j)}, \text{Encounter}(j,d)\}$.*

Note that in a context of unobeyability, the agent has the duty to prevent each fact causing a new violation. However, he has no contextual duty to prevent the tautological disjunction of all such facts (i.e. to realise the contradictory conjunction of them). In fact such a disjunction is not a c-illicit, being entailed by the context (Remark 2). Thus, in the above example, in context C Jephthah has duties $\mathbf{O}Kill(j,d)$ and $\mathbf{O}\neg Kill(j,d)$, but no duty $\mathbf{O}(Kill(j,d) \wedge \neg Kill(j,d))$.

5. The logic of duties

In this section we analyse the logic(s) of the $\mathbf{O}_N$ and $\mathbf{O}_{(N,C)}$ modalities. We consider the main principles discussed in the deontic logic literature, and check whether they hold. We also use some of the best known deontic paradoxes as benchmarks.

5.1. Properties of duties

For ease of readability, we consider first the version of a given property for noncontextual duty (i.e. unary obligation), when available.

Extensionality For non-contextual duty it takes the form (RE) "If $A \equiv B$, then $\mathbf{O}_N(A) \equiv \mathbf{O}_N(B)$". There are two versions for conditional duty: "If $\wedge C \equiv \wedge C'$, then $\mathbf{O}_{(N,C)}(A) \equiv \mathbf{O}_{(N,C')}(A)$" and "If $A \equiv B$, then $\mathbf{O}_{(N,C)}(A) \equiv \mathbf{O}_{(N,C)}(B)$". All versions trivially hold.

And introduction If $\mathbf{O}_N A$ and $\mathbf{O}_N B$, then $\mathbf{O}_N(A \wedge B)$. Assume $\mathbf{O}_N A$ and $\mathbf{O}_N B$. So $A \equiv \neg A'$, and $B \equiv \neg B'$ for some A', B' generalised illicits. We have $(A \wedge B) \equiv (\neg A' \wedge \neg B') \equiv \neg(A' \vee B')$. Furthermore, $A' \vee B'$ is a generalised illicit by Corollary 1. This suffices for $\mathbf{O}_N(A \wedge B)$. The analog principle for contextual duties does not hold unrestrictedly. E.g., assume that $\mathbf{O}_{(N,C)}(A)$ and $\mathbf{O}_{(N,C)}(\neg A)$. $\mathbf{O}_{(N,C)}(A \wedge \neg A)$ is not a contextual duty as the negation of its content is equivalent to t, which is not a c-illicit (cond. 3 in Def. 8 fails)

Remark 4. *The logic of $\mathbf{O}_N$ duties contains the non-normal modal logic EC [16]. EC is obtained by adding the C axiom $(\mathbf{O}_N A \wedge \mathbf{O}_N B) \rightarrow \mathbf{O}_N(A \wedge B)$ to the system E of so-called classical modal logic, consisting of the sole rule of extensionality (RE).*

Monotonicity wrt context It is the analog of the "Strengthening of the antecedent" principle in conditional logic. It fails in its usual form "If $\mathbf{O}_{(N,C)}(A)$, and $C \subseteq C'$ then $\mathbf{O}_{(N,C')}(A)$", as hinted in Remark 3. The following counter-example shows that its failure is in line with the idea that in the context of deliberation one puts aside the moral status of the facts which are settled. Let $N = \{\neg A \Rightarrow v_1\}$. We have $\mathbf{O}_{(N,\emptyset)}A$, while $\mathbf{O}_{(N,\{\neg A\})}A$ no longer holds due to the failure of (the first half of) condition 3 in Def. 8.

Factual detachment (= detachment via modus ponens [17]) In our framework it might be expressed as: If $\mathbf{O}_{(N,C)}(A \to B)$ then $\mathbf{O}_{(N,C\cup\{A\})}B$, for A a conjunction of literals. It holds under the hypotheses: $(*)$ $C \cup \{A\} \not\models \neg A_1 \vee \cdots \vee \neg A_n$, for all duties $\mathbf{O}_N A_1, \ldots, \mathbf{O}_N A_n$, and $(**)$ there is no formula D weaker than $\neg B$ s.t. $A \wedge D \models \neg B$. We show that $\neg B$ is a c-illicit in context $C \cup \{A\}$ so that $\mathbf{O}_{(N,C\cup\{A\})}B$. By $\mathbf{O}_{(N,C)}(A \to B)$ follows that there is $S \subseteq V$ s.t. $\bigwedge(C \cup \{A \wedge \neg B\}) \Rightarrow \bigvee S$, that is $\bigwedge(C \cup \{A\} \cup \{\neg B\}) \Rightarrow \bigvee S$. Hence Def. 8.1 holds for $\neg B$ relative to $C \cup \{A\}$. Condition 2 follows from the fact that if there is a D' weaker than $\neg B$ such that $\bigwedge(C \cup \{A\} \cup \{D'\}) \Rightarrow \bigvee S$, then $A \wedge \neg B$ is not a c-illicit ($C \cup \{A \wedge D'\} \Rightarrow \bigvee S$ and from $\neg B \models D'$ follows $A \wedge \neg B \models A \wedge D'$, and from $(**)$ that $A \wedge D' \not\models A \wedge \neg B$) contradicting the hypothesis. Condition 3 also holds: the hypothesis $(*)$ guarantees by Lemma 1 that $\bigwedge(C \cup \{A\}) \not\Rightarrow \bigvee V$, where V is the set of all violations, and hence $\bigwedge(C \cup \{A\}) \not\Rightarrow \bigvee S$; $C \cup \{A\} \not\models B$ follows from $C \not\models A \to B$.

No conflict $\neg(\mathbf{O}_N A \wedge \mathbf{O}_N \neg A)$ holds for a normative system N that is absolutely obeyable in the sense of Definition 10. Suppose N has this property. Assume by contradiction that both $\mathbf{O}_N A$ and $\mathbf{O}_N \neg A$ hold. Hence $A \equiv \neg B_1$, and $\neg A \equiv \neg B_2$ for some B_1, B_2 generalised illicits. Hence $B_1 \vee B_2 \Rightarrow \bigvee(V_1 \cup V_2)$, and so $t \Rightarrow \bigvee(V_1 \cup V_2)$, since $B_1 \equiv \neg A$ and $A \equiv B_2$. By (*Weakening*), $t \Rightarrow \bigvee V$, where V is the set of all violations. Contradiction.

The contextual version of the principle $-\neg(\mathbf{O}_{(N,C)}A \wedge \mathbf{O}_{(N,C)}\neg A)-$ holds for normative systems that are contextually obeyable in C. Assume by contradiction that so is N and that $\mathbf{O}_{(N,C)}A$ and $\mathbf{O}_{(N,C)}\neg A$ hold. Being A and $\neg A$ contextual illicits w.r.t. C, there are V_1 and V_2 s.t. $(*) \bigwedge(C \cup \{A\}) \Rightarrow \bigvee V_1$ and $\bigwedge(C \cup \{\neg A\}) \Rightarrow \bigvee V_2$, and $(**) \wedge C \not\Rightarrow \bigvee V_1$ and $\wedge C \not\Rightarrow \bigvee V_2$. From $(*)$ it follows that $\bigwedge C \Rightarrow \bigvee V_1 \cup V_2$ (i.e., condition (1) of Def. 11), and by $(**)$ that for each $v \in V_1 \cup V_2$, $\bigwedge C \not\Rightarrow v$ (i.e. condition (2) of Def. 11), thus establishing that N is unobeyable in context C. This property echoes the fact that for Kelsen normative systems can be conflicting. Thus, for him, and in our framework, the "no conflict" axiom (also known as D axiom in modal logic) only holds for normative systems that are either absolutely or contextually obeyable.

5.2. Solutions to deontic paradoxes

We analyse the behaviour of our logic w.r.t. the main deontic paradoxes that have beset SDL, which corresponds to the deontic logic obtained by Anderson using his reduction schema. Deontic paradoxes are intended here in a broad sense as (un)derivable theorems that are counter-intuitive in a common-sense reading.

Recall that unconstrained I/O logic [10], which is at the base of our framework, is unable to handle many of the considered paradoxes. Two paths have been followed in order to fix it: using constraints to filter excess outputs [18], or suitably weakening the logic. In particular, (Weakening) and axiom (Truth) go, and at the same time a consistency check restrains the application of some rules (like AND introduction). Although based on the strongest system of unconstrained I/O logic (Def. 2), our framework eliminates those paradoxes due to condition 2. in Definitions 5 and 8.

Contrary-to-duty obligation (Forester's and Chisholm's) paradoxes As seen before, in our framework we can overcome such paradoxes. In Forrester's case (see Ex. 6) in context $\{kill\}$, $\mathbf{O}KillGently$ holds, and the opposite obligation $\mathbf{O}\neg KillGently$ does not hold. The treatment of Chisholm's scenario is similar (see Ex. 7).

Ross' paradox The paradox consists in the derivation of (a) You should mail the letter or burn it, from (b) You should mail the letter. Introduced in 1941 by Ross [19], this paradox has been a discussion topic ever since. It does not appear in our logics. E.g., relative to a normative system $N = \{\neg PostLetter \Rightarrow v_1\}$, we have the obligation $\mathbf{O}PostLetter$, but we do not have $\mathbf{O}(PostLetter \vee BurnLetter)$, due to condition 2. in Def. 5.

Deontic detachment [17] It fails in its usual form: If $\mathbf{O}A$ and $\mathbf{O}(A \rightarrow B)$ then $\mathbf{O}B$. This can be illustrated with Broome [20]'s counter-example: let $N = \{\neg Exercise \Rightarrow v_1; Exercise \wedge \neg EatMore \Rightarrow v_2\}$ (one should exercise and if one exercises, one should eat more). We have $\mathbf{O}Exercise$ and $\mathbf{O}(Exercise \rightarrow EatMore)$. But it should not be the case that $\mathbf{O}EatMore$, since, intuitively, the obligation to eat more holds only if the first obligation is fulfilled. The correct inference is captured by the following form of deontic detachment, called "aggregative" in [21], as one keeps track of what has been previously detached: from $\mathbf{O}A$ and $\mathbf{O}(A \rightarrow B)$, infer $\mathbf{O}(A \wedge B)$, which holds in our logic due to *AND introduction* and *Extensionality*. Hence from N we can infer $\mathbf{O}(Exercise \wedge EatMore)$. The contextual version of deontic detachment, with $\mathbf{O}(A \rightarrow B)$ replaced by $\mathbf{O}_{(N,A)}B$, does not hold either. For a counter-example, let $N = \{\neg a \Rightarrow v_1, a \wedge \neg b \Rightarrow v_2\}$. We have $\mathbf{O}a$ and $\mathbf{O}_a b$, but not $\mathbf{O}b$, as $\neg b$ is not a generalized illicit (condition 1 in Def. 5 fails).

The alternative service paradox This scenario, proposed by Horty in 1994 [22], is handled similarly to the previous case, by changing the disjunctive obligation in the original formulation into a conditional obligation: from (1) You should fight in the army or perform alternative service ($\mathbf{O}(\neg A \rightarrow B)$), and (2) You should not fight in the army ($\mathbf{O}\neg A$), we derive that (3) You should not fight and perform alternative service ($\mathbf{O}(\neg A \wedge B)$).

Necessitation $\mathbf{O}t$ does not hold. Take any non-empty set of norms, e.g., $N = \{Speed \Rightarrow v_1\}$. $\mathbf{O}(Speed \vee \neg Speed)$ does not hold, because $Speed \wedge \neg Speed$ is not a generalised illicit, due to condition 2. in Def. 5.

Reflexivity This is the law $\mathbf{O}_{(N,C)} \wedge C$. It fails, because of condition 3 in Def 8 ($C \models \wedge C$, see Remark 2) and so $\neg \wedge C$ cannot be a c-illicit relative to C.

And elimination It is the principle "If $\mathbf{O}(A \wedge B)$ then $\mathbf{O}A$", known in the non-normal modal logic literature as axiom M (see [16]). As suggested by a number of authors, when A and B are not separable, such a principle is counter-intuitive, so that (to quote Hansen) "failing a part [of the order] means that satisfying the remainder no longer makes sense. E.g. if I am to satisfy the imperative 'buy apples and walnuts', and the walnuts [...] and the apples [are meant to] land in a Waldorf salad, then it might be unwanted and a waste of money to buy the walnuts if I cannot get the apples" [23, p. 91]. Put $N = \{\neg Apples \vee \neg Walnuts \Rightarrow v\}$ (or equivalently $\{\neg Apples \Rightarrow v, \neg Walnuts \Rightarrow v\}$). We have $\mathbf{O}(Apples \wedge Walnuts)$, but not $\mathbf{O}(Apples)$, due to condition (2) in Def. 5.

6. Future work

We have developed a Kelsenian deontic logic that defines obligation in terms of violation and causality. The behaviour of the resulting obligation has been analyzed using as benchmarks well known properties and paradoxes from the deontic logic literature.

It has often been argued that CTDs involve different senses of "ought", and that a fully adequate treatment of them must be able to capture those nuances. In our analysis of contextual duties we have focused on deliberative duties, where one puts aside the moral or legal status of the facts which are taken as settled and one must decide what to do. Our framework allows, however, a finer-grained analysis of duties –to be left to future work– which could not only shed more light on the various paradoxes, but also capture different types of "ought"-statements. Among them, we plan to investigate contextual (evaluative) duties that are appropriate for what is commonly referred to as the "context of judgement" [24,25], where one assesses the moral or legal status of settled facts through backward looking or *post-eventum* judgments [26, p. 157].

Furthermore, our logic is an ought-to-be deontic logic, as obligations may cover any cause of violations. We plan to develop an ought-to-do deontic logic, reflecting Kelsen's notion of a delict, by carving out agentive elements within illicits and "plugging" in a suitable logic of action in our base logic. Other topics for future research include the question of extending the framework to support defeasible reasoning, reasoning about exceptions, and the question of how to axiomatise and automatise the logic.[2]

References

[1] Kelsen H. Reine Rechtslehre Einleitung in die rechtswissenschaftliche Problematik. Mohr Siebeck; 1934.
[2] Kelsen H. Reine Rechtslehre. Franz Deuticke; 1960.
[3] Kelsen H. Introduction to the Problems of Legal Theory. Clarendon; 1992.
[4] Kelsen H. The Pure Theory of Law. University of California Press; 1967.
[5] Anderson AR. The Logic of Norms. Logique et Analyse. 1958;1:84–91.
[6] Wang PH. Anderson's reduction and Kelsen's Normativism. In: Pluralism and Law. Proceedings of the 20th IVR World Congress 2001. Vol. 4: Legal Reasoning. Franz Steiner Verlag; 2001. p. 104–110.
[7] Ross WD. Foundations of Ethics. Clarendon; 1939.
[8] Kelsen H. General Theory of Norms. Oxford University Press; 1990.
[9] Forrester JW. Gentle Murder, or the adverbial Samaritan. J of Philosophy. 1984;(81):193–197.
[10] Makinson D, van der Torre LWN. Input/Output Logics. J Phil Log. 2000;29(4):383–408.
[11] Bochman A. On Laws and Counterfactuals in Causal Reasoning. In: Proc. 16 Intern. Conf. on Principles of Knowledge Representation and Reasoning (KR 2018); 2018. p. 494–503.
[12] Kelsen H. General Theory of Norms. Clarendon; 1991.
[13] von Wright GH. Deontic logic. Mind. 1951;60(237):1–15.
[14] Chisholm RM. Contrary-to-Duty Imperatives and Deontic Logic. Analysis. 1963;24:33–6.
[15] Kelsen H. On the pure theory of law. Israel Law Review. 1966:193–215.
[16] Chellas B. Modal Logic. Cambridge University Press; 1980.
[17] Greenspan PS. Conditional Oughts and Hypothetical Imperatives. Journal of Philosophy. 1975;72(10):259–276.
[18] Makinson D, van der Torre LWN. Constraints for Input/Output Logics. J Phil Log. 2001;30(2):155–185.
[19] Ross A. Imperatives and Logic. Theoria. 1941;7(1):53.
[20] Broome J. Rationality through Reasoning. West Sussex, U: Wiley-Blackwell; 2013.
[21] Parent X, van der Torre LWN. "Sing and Dance!" - Input/Output Logics without Weakening. In: Cariani F, et Al, editors. DEON 2014, Proceedings. vol. 8554 of LNCS. Springer; 2014. p. 149–165.
[22] Horty J. Moral Dilemmas and Nonmonotonic Logic. Journal of Philosophical Logic. 1994;23(1):35–65.
[23] Hansen J. Imperatives and Deontic logic. University of Leipzig; 2008.
[24] Thomason RH. In: Hilpinen R, editor. Deontic Logic and the Role of Freedom in Moral Deliberation. Dordrecht: Springer Netherlands; 1981. p. 177–186.
[25] Asher N, Bonevac D. *Prima Facie* Obligation. Stud Logica. 1996;57(1):19–45.
[26] Hare R. The Language of Morals. Clarendon; 1991.

[2]G. Sartor has been supported by the H2020 ERC Project "CompuLaw" (G.A. 833647), A. Ciabattoni and X. Parent by the WWTF project MA16-028.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210331

Identification of Contradictions in Regulation

Michał ARASZKIEWICZ [a,1], Enrico FRANCESCONI [b] and Tomasz ZUREK [c]

[a] *Department of Legal Theory, Faculty of Law and Administration, Jagiellonian University, Ul. Gołebia 24, 31-007 Krakow, Poland*

[b] *Italian National Research Council (Florence, Italy) and European Parliament (Luxembourg)*

[c] *Institute of Computer Science, Maria Curie Sklodowska University, Ul. Akademicka 9, 20-033 Lublin, Poland*

Abstract. This paper presents a Semantic Web–based model for detecting contradictions in regulations. We introduce a conceptual model of contradictions and, on the basis of this model, a knowledge representation–based model is used, which is able to represent the semantics of provision types and related properties. The usefulness of the model is shown through an example.

Keywords. Frames, Legal Interpretation, Legislative errors, Knowledge Representation

1. Introduction[2]

Legislative drafting is a complex task. It requires not only linguistic competence, but also thorough knowledge of the regulated domain, as well as of the legal/theoretical assumptions concerning the legal system, including its completeness, consistency, and lack of redundancy [1,2]. One type of error occurs when the regulation is encumbered with contradictions. Legislators should ensure not only the internal consistency of the regulation but also a lack of contradiction with hierarchically higher regulations. While some contradictions may be eliminated through legal reasoning, this is not always possible or desirable. The lack of consistency in regulation is typically not straightforward. Therefore, it is worthwhile to provide a legislative tool representing the structure and semantic content of the drafted provisions - to ascertain whether an inconsistency actually exists. Knowledge representation tools have been fruitfully used to represent legislation ([3,4,5]). However, the development of such a model is a complex task, also because it requires expertise in the regulated domain of law. In this paper, we do not develop a formal representation of legal regulation but, rather, a conceptual framework that may be used by a legislator to formalize a drafted text and analyze its features, in particular to detect legislative errors. Our proposal fits well with the idea of designing systems that

[1] Corresponding Author: michal.araszkiewicz@uj.edu.pl

[2] The paper presents the results of the project UMO-2018/29/B/HS5/01433 entitled Legislative Errors and Comprehensibility of Legal Text.

assume human–machine interaction, taking into account the strengths and weaknesses of each party.

We present an ontology-based model to describe the structure and the semantic content of legal provisions. The purpose of this model is to support the process of legislative drafting. Let us assume that a legislator has already prepared a draft of a set of legal provisions in natural language. The tool enables the legislator to:

- Represent the content of legal provisions in a systematized manner, taking into account the defined categories of entities, such as legal subjects, legal objects, different types of deontic relations, etc.
- Use ontological reasoning facilities to detect errors, in particular contradictions, in regulation;
- Evaluate whether the regulation under analysis is acceptable with regard to the assumed criteria or if it needs amending.

We argue that this approach enables us to identify certain non-trivial types of legislative errors, including potential and actual contradictions.

2. Conceptual Framework

2.1. Introductory Remarks

The main part of any normative piece of legislation is composed of legal provisions, namely elementary, sentence-like linguistic expressions systematized as sections, articles, or points. We focus on the representation of the structure and semantic content of legal provisions. Each representation of a legal provision in a knowledge-based model is an interpretation of this provision. For the purposes of our project, we understand the interpretation as a result of the reasoning process performed by a human agent.

The model, based on our earlier work ([6,7]) enables the representation of the structure and semantic content of legal text for the purpose of identifying legislative errors. A legislative error may be understood as a violation of one or more criteria of rational lawmaking, such as aiming at consistency, completeness, lack of redundancy or axiological coherence of the created regulation. The legislator should detect potential violations of these criteria during the process of legislative drafting and consider whether these violations can be eliminated through legal reasoning in the first place. The application of different interpretive arguments, including linguistic, systemic, and functional ones, [8] may lead to the resolution of apparent problems and, therefore, the drafted text may not need amending. However, not all the results of such remedial interpretation are acceptable, as there exist some interpretive rules that constrain the space of acceptable interpretations, discussed in [6]. In some cases, the identification of a legislative error does not involve complex interpretive reasoning because some violations are ascertainable through a literal interpretation of the considered provisions, and no other interpretation is conceivable. This does not mean that detecting a violation is always a trivial task. In some cases, the structure of the drafted provision needs to be clearly represented for a violation to become apparent. In this paper, we focus on one particular type of legislative error: contradictions in regulations.

2.2. The notion of contradiction between provisions

We argue that the nature of contradiction between legal provisions requires in-depth analysis of the features and scope of the provisions because it is relatively seldom that two provisions lead to strictly opposite conclusions (p vs $\neg p$ or $perm(p)$ vs $proh(p)$).

First, following [6], we assume that every regulation creates a certain legal relation between particular entities. For example, by *Bearer*, we denote an agent which is an addressee of the regulation, by *Object*, we denote the object of a regulation, by *Action*, we denote the action regulated by the provision, and by *Counterpart*, we denote the agent functioning as a counterpart of the regulation. Second, we have to notice that every concept has its own scope and that scopes of different concepts can be in different relations, e.g., inclusion. We have to assume also that the analyzed sets of provisions can contradict an existing regulation not only in their whole scope but also in a subset of the regulated entities. To represent the idea of the broader and narrower scopes of two concepts, we introduce operator $\sqsubseteq$, where $X \sqsubseteq Y$ denotes that every entity represented by X is also represented by Y.

To represent the relation defined by a legal provision (or set of legal provisions), we introduce a predicate $rel()$ representing such a relation. The first argument of this predicate will be a type of legal relation created by the analyzed provision; other arguments will represent the concepts discussed above used in the wording of a legal provision (for example, the second argument will represent *Bearer*, the third *Object*, etc.).
If we introduce:

- a set of types of legal provisions ($Type = \{perm, proh, obl, ...\}$);
- a relation of opposition represented by pairs of opposite types of provisions $Opposite = \{< perm, proh >, < obl, proh >, ...\}$ (relations containing those pairs are potentially contradictory);
- a predicate $rel(type, X_i, X_j, ...)$, where $X_i, X_j, ...$ represent various concepts;

then the conflict will appear if:

- one set of provisions defines $rel(type_l, X_i, X_j, ...)$,
- second set of provisions defines $rel(type_k, X_m, X_n, ...)$ [3]
- there will be a set of concepts $x_i, x_j, ...$ s.t. $x_i \sqsubseteq X_i$, $x_i \sqsubseteq X_m$, $x_j \sqsubseteq X_j$, $x_j \sqsubseteq X_n$, ... and
- relations ($type_l$, $type_k$) are opposite i.e. $< type_l, type_k > \in Opposite$ (for example $type_l = perm, type_k = proh$)

The above model can be illustrated by a very simple example. Suppose we have two provisions: (1) vehicles are forbidden to enter the park and (2) ambulances are allowed to enter the park. Additionally, we know that *vehicles* is a broader concept than *ambulance* (i.e., every ambulance is vehicle). To represent the example, we assume that predicate $rel()$ has the following arguments: 1) type, 2) *Bearer*, 3) *Object*, and 4) *Action*. The example can be modelled as follows: $rel(proh, vehicles, park, enter), rel(perm, ambulance, park, enter), ambulance \sqsubseteq vehicle$ Since *ambulance* is a narrower concept than *vehicles* and $< proh, perm > \in Opposite$, then ambulances are simultaneously prohibited (on the

[3] Note that in order to correctly detect contradictions, the arguments in both relations should be in the same order, i.e. *Bearer* should be the second argument in both relations

basis of provision 1) and permitted (on the basis of provision 2), which is an obvious contradiction[4].

3. The description of the ontology and tools

The presented conceptual model for detecting errors and inconsistencies in legislative draft bills can be implemented in the Semantic Web using the Provision Model, introduced by [9]. The Provision Model aims to provide a formal representation of textual objects, subject to a given interpretation. For this reason, it represents a model of legal rules that can be effectively used to detect legislative errors. The Provision Model has been used in the literature to provide advanced legal information retrieval and reasoning services based on the semantics of legal rules [10], but primarily, it has been targeted at implementing model-driven legislative drafting facilities [11]. The aim of this approach is to improve the quality of legal texts and ensure the maintenance of legal information by monitoring the impacts of new regulations on the legal system (including the consistency and completeness of new provisions within the same text or of different texts within the same legal order as well as between different legal orders, as in the case of a European directive and its transposition in national legislation).

3.1. The Provision Model

The Provision Model represents a knowledge model of the rules conveyed in legislative texts. It is organized into provision types and properties. It describes constitutive and regulative rules independently of the domain in which they operate, as well as rules on rules, namely, amendments (which may be related to the textual content, the timing of the enactment of the rule, or the scope within which the rules operate).

Examples of regulative rules, in terms of provision types, are shown in Fig. 1; they are represented as classes of the Provision Model ontology. In particular, the provision types prv:Obligation and prv:Permission are reported. They are associated with specific properties; for example, for prv:Obligation, the specific properties are prv:hasObligationBearer, prv:hasObligationAction, prv:hasObligationCondition, and prv:hasObligationCounterpart (see Fig. 1). This model can be used to provide semantic annotation of legal texts and is able to reflect the lawmakers' directions in a machine-readable format.

Note that the properties of the provision types represent the arguments of relation *rel*() described in section 2.4. In particular, prv:hasObligationBearer and prv:hasPermissionBearer represent the argument *Bearer*; prv:hasObligationAction, and prv:hasPermissionAction represent the argument *Action*, etc.

As the Provision Model describes legislative rules independently of the domain in which they operate, a complete representation of a legal rule instance typically contained in a textual paragraph can be obtained by combining the Provision Model with the controlled vocabularies represented in knowledge organization systems capable of providing additional information on the entities of the regulated domain [12,13]. Controlled vocabulary terms can be used as provision property values to be used for semantic annotation of legal provisions. This can be useful to disambiguate and harmonize different possible literal interpretations of textual wording in the law and minimize the risk of interpretive doubts.

[4]This example can be easily solved by using *lex specialis*..., but this is outside of the scope of this paper

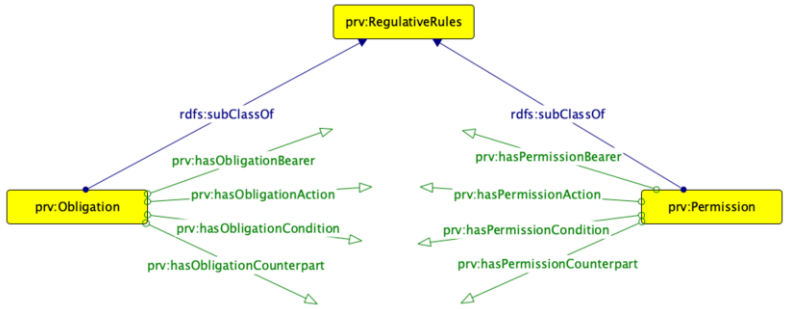


Figure 1. Provision properties in Permission and Obligation

3.2. Logical and technical relations in the Provision Model

The Provision Model is also endowed with axioms that are able to describe logical relations between provisions. *Logical relations* are relations between provisions that are necessary from a logical point of view, such as the classical Hohfeldian relations between Right and Duty as well as No-right and Privilege and the opposite relations between Right and No-right as well as Duty and Privilege [10].

In our conceptual model of contradiction (see section 2.4), logical relations are represented by pairs from set $Opposite$, which provides a set of relations that may constitute potential contradictions. *Technical relations* between provisions, on the other hand, are relations that are not necessary from a logical point of view, but they derive from considerations of legislative techniques. This means that they are possible and can be identified in legislative texts, provided that the legislative drafter follows specific legislative techniques in expressing these provisions. An example of such relations is the one existing between a *Definition*, introducing a concept identified by the attribute *Definiendum*, and all the other provisions having, as an attribute value, the value of such *Definiendum*. Another example is the relation between the *Duty* of a specific *Bearer* to accomplish a specific *Action* toward a given *Counterpart* as well as the *Procedure* to fulfill it. In the conceptual model of contradiction, technical relations are represented by arguments of predicate $rel()$. While *logical relations* can be described by axioms at the level of the Provision Model and are inherited by all the related instances (see [10] and [7]), *technical relations* can be identified and described at the level of specific provision instances only, because they are linked to their content. As reported in [9], technical relations between provisions can be established directly by the legislator through references or can be deduced by reasoning over provisions content, expressed by provisions attribute values.

Therefore, contradictions can be detected by checking *technical relations* between provisions, namely, relations that can be revealed only by reasoning about the semantics and content of the provision instances (including axioms if needed) and not by reasoning about provisions semantics and related axioms only.

4. Example of provision semantic representation

We apply the presented model to an example involving a contradiction between a European Union directive, on the one hand, and a Member State law, on the other hand.

Directives are instruments which bind the Member State as to the result to be achieved but leave to the national authorities the choice of form and methods (TFEU Art. 288)[5]. Therefore, directives need to be properly transposed into the legal system of the Member State. Incorrect transposition of a directive may lead to diverse legal consequences. Due to the general wording of directives, it is sometimes difficult to determine whether a particular Member State regulation is a proper transposition. The model developed here may be fruitfully used to detect contradictions between national and EU regulations by taking into account *technical relations*. The following example is based on the Judgment of the Court of 4 May 2006, case C-508/03. Let us consider Art. 2 Par. 1 in the Council Directive of 27 June 1985:

> *Member States shall adopt all measures necessary to ensure that, before a planning permission is given, urban development projects likely to have effects on the environment are made subject to an assessment with regard to their effects.*

From the Provision Model viewpoint, this paragraph can be classified as an prv:Obligation for Member States "to adopt all measures necessary to the purpose that before a planning permission is granted, environmental impact assessment of a project is performed." The prv:Obligation to be implemented can be expressed in an attribute–value pair notation, according to the Provision Model, as follows:

prv:Obligation	
prv:hasObligationBearer =	'Competent authority'
prv:hasObligationAction =	'to assess effects on environment'
prv:hasObligationCondition =	'before planning permission'
prv:hasObligationObject =	'project'

Note that, in this example, property values are reported as literals for the sake of readability. However, as previously discussed, they are typically terms in controlled vocabularies[6]. Domain terms from a controlled vocabulary are important in the context of detecting technical relations between provision instances as they contribute to disambiguation and to comparison between domain entities.

To provide a Provision Model representation of the above EU directive paragraph, let's consider that the relevant concepts introduced by this directive are organized in a controlled vocabulary called ex:EnvironmentalImpactAssessementVoc, described in SKOS[7], with a namespace ex: (which just stands for "example") that includes concepts like ex:MemberState, ex:PlanningPermission, ex:Project, and related skos:prefLabel.The model graph is provided in Fig. 2; the graph also shows the relations skos:inScheme between the vocabulary terms and the related vocabulary top-concept ex:EnvironmentalImpactAssessementVoc (instance of a skos:ConceptScheme).
The RDF representation of such a graph related to the obligation in the cited directive, having [docURI] as URI, is therefore the following:

```
<rdf:Description  rdf:about="[docURI]">
   <rdf:type rdf:resource="prv:Obligation"/>
   <prv:hasObligationBearer rdf:resource="ex:CompetentAuthority"/>
   <prv:hasObligationAction rdf:resource=
```

[5]Treaty on the Functioning of the European Union (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:4301854)

[6]Examples of controlled vocabularies can be found at https://op.europa.eu/en/web/eu-vocabularies/controlled-vocabularies

[7]Simple Knowledge Organization System (https://www.w3.org/2004/02/skos/)

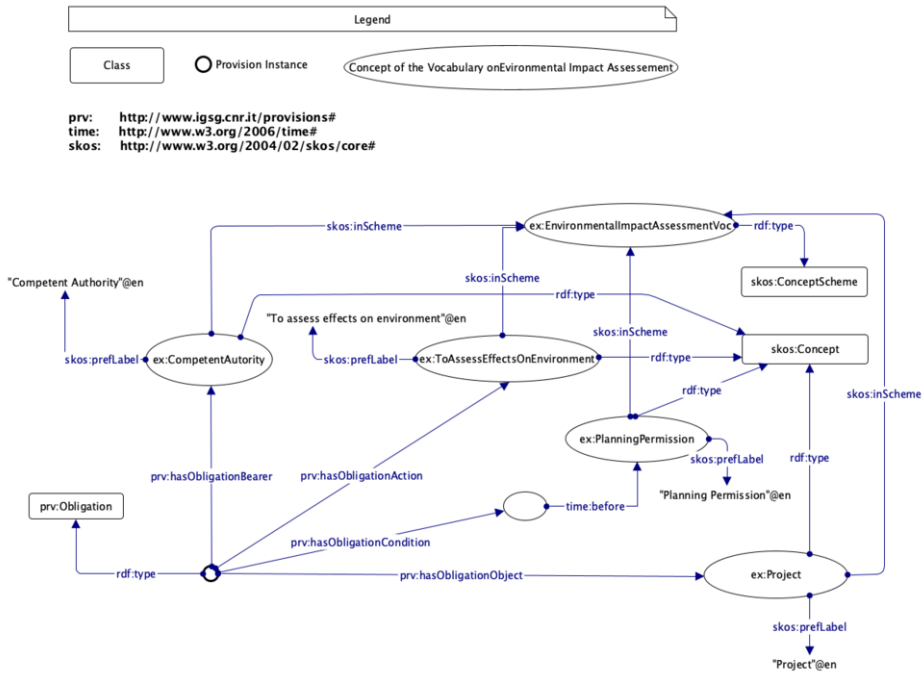


Figure 2. Semantic representation of EU Directive 27 June 1985 art. 2 par. 1

```
                                    "ex:ToAssessEffectsOnEnvironment"/>
   <prv:hasObligationCondition>
                 <time:before rdf:resource="ex:PlanningPermission"/>
   </prv:hasObligationCondition>
   <prv:hasObligationObject rdf:resource="ex:Project"/>
</rdf:Description>
```

Note that:

- the Provision Model imports the W3C Time Ontology[8] allowing the expression of time-wise relations between domain entities.
- this formal representation of legal provisions is used in our approach to identify errors, particularly contradictions/inconsistencies in a set of legal rules. Errors may be revealed according to the interpretative canon used to describe provision semantics, in this case, a literal interpretative canon.

5. Approach for detecting contradictions

To show the approach to detecting errors/inconsistences in legal rules, let's consider the case of a set of provisions in the U.K. legal order, aimed to implement the provision of the EU Directive of 27 June 1985 art. 2 par 1, introduced in Section 4. The UK provisions are as follows:

[8] http://www.w3.org/2006/time#

(1) "outline planning permission" means planning permission granted, in accordance with the provisions of a development order, with the reservation for subsequent approval by the local planning authority or the Secretary of State of matters not particularized in the application ("reserved matters").

(2) the competent authority cannot grant planning permission unless it has first taken the environmental information into consideration and states in its decision that it has done so.

(3) In the case of outline planning permission, an environmental impact assessment can be carried out only at the initial stage of granting such permission and not at the later stage of approval of the reserved matters.

Paragraph (1) introduces the expression "outline planning permission," which represents a narrower concept than the concept "planning permission" that is introduced by the transposed EU directive. In paragraph (3), the implementing measure reproduces the obligation to assess the effects on the environment before planning permission is given but makes reference to the concept "outline planning permission," opening the possibility of not assessing the environmental impact at a later stage when "reserved matters" are approved.

This represents a violation of the EU directive, which we intend to detect automatically. To do that, first, we provide a semantic representation of the implementing provisions. Here, just the representation of paragraph (3) is reported because it is sufficient to detect the inconsistency. The attribute–value pair notation of the paragraph is as follows:

prv:Obligation	
prv:hasObligationBearer =	'Competent authority'
prv:hasObligationAction =	'to assess effects on environment'
prv:hasObligationCondition =	'before outline planning permission'
prv:hasObligationObject =	'Project'

The knowledge graph of both legal provisions (the provision of the EU directive described in Section 4 and the implementing U.K. provision) are reported in Fig. 3, including the concept ex:OutlinePlanningPermission in the vocabulary of the domain concepts. The RDF representation of paragraph (3) of the U.K. implementing measure, having [docURI-UK] as URI is, therefore, the following:

```
<rdf:Description rdf:about="[docURI-UK]#par3">
   <rdf:type rdf:resource="prv:Obligation"/>
   <prv:hasObligationBearer rdf:resource="ex:CompetentAuthority"/>
   <prv:hasObligationAction rdf:resource=
                    "ex:ToAssessEffectsOnEnvironment"/>
   <prv:hasObligationCondition>
        <time:before rdf:resource="ex:OutlinePlanningPermission"/>
   </prv:hasObligationCondition>
   <prv:hasObligationObject rdf:resource="ex:Project"/>
</rdf:Description>
```

Given the U.K. provision semantic representation, it is possible to select the provisions concerning the obligation of a competent authority to assess the effect on the environment for urban development project. This leads to a comparison of the provisions reported in Fig. 3, which spots the existence of two different conditions to carry out such an assessment: one before "planning permission" (ex:PlanningPermission) is given and the other before "outline planning permission" (ex:OutlinePlanningPermission) is given.

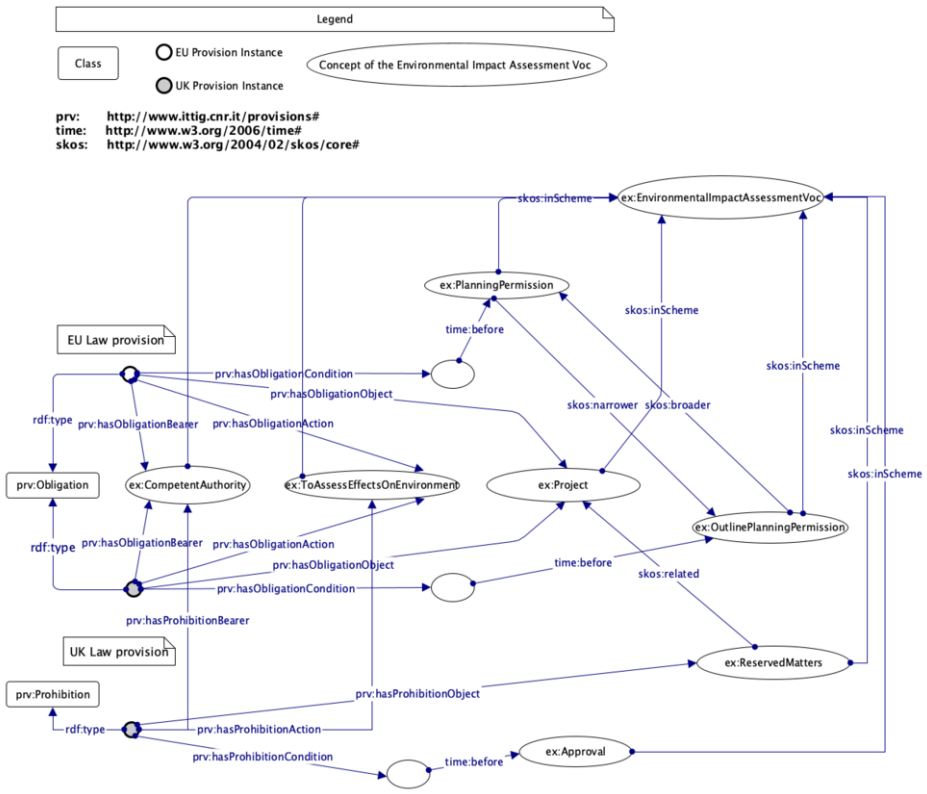


Figure 3. Semantic representation of EU and UK legal provisions

Given that these two concepts are in a skos:narrower relation, the inconsistency at the level of the condition of the obligation in the U.K. implemented provision with respect to the EU provision can be automatically detected. Moreover, the relation (skos:related[9]) between the concept of "urban development project" (ex:Project) and the "reserved matters" (ex:ReservedMatters) is able to automatically reveal the existence of a prohibition, introduced in the U.K. implemented legislation, to "assess the effect on the environment" (ex:AssessEffectsOnEnvironment) on "reserved matters," which represents an inconsistent exception with respect to what is prescribed by the EU directive.

It is worth emphasizing that the contradictions revealed are detected at the level of provision content by the relation between provision property values, as discussed in Section 3.2.

[9]Note that the relation between ex:ReservedMatters and ex:Project is rather a part-of relation (the specialization of the skos:related or skos:broader/narrower relation to capture partitive relations has been discussed but not approved by W3C yet)

6. Conclusions

In this paper, we introduced a model of contradictions between legal provisions and a practical implementation with the use of knowledge representation tools. The analysis of legal provisions requires a consideration not only of logical contradictions ($proh(p)$ vs $perm(p)$) or (p vs $\neg p$) but also of the scope of the terms used. We argue that, to detect a contradiction in sets of provisions, two conditions must be satisfied: 1) the analyzed provisions regulate at least a partially overlapping set of entities and 2) the analyzed provisions impose conflicting modalities on those sets.

On the basis of this assumption, we introduced a conceptual model of contradictions in legal text, as well as a Semantic Web-based implementation of such model, allowing a mechanism for automated detection of contradictions. This model enables the human expert to evaluate the drafted regulation and detect the contradictions therein, thus enabling the correction of the draft regulation and avoiding the necessity of amending the regulation after it enters into force. The presented model assumes human–computer interaction.

References

[1] Alchourron CE, Bulygin E. Normative Systems. Wien-New York; 1971.
[2] Araszkiewicz M, Pleszka K. The Concept of Normative Consequence and Legislative Discourse. In: Araszkiewicz M, Pleszka K, editors. Logic in the Theory and Practice of Lawmaking. vol. 2 of Legisprudence Library. Springer; 2015. p. 253-97.
[3] Van Kralingen R. Frame-based Conceptual Models of Statute Law. Kluwer; 1995.
[4] Visser PRS, van Kralingen RW, Bench-Capon TJM. A Method for the Development of Legal Knowledge Systems. In: Proceedings of ICAIL 1997. ACM; 1997. p. 151–160.
[5] van Doesburg R, van Engers T. The False, the Former, and the Parish Priest. In: Proceedings of ICAIL 2019. ACM; 2019. p. 194–198.
[6] Araszkiewicz M, Zurek T. Identification of Legislative Errors through Knowledge Representation and Interpretive Argumentation. In: Rodríguez-Doncel V, Palmirani M, Araszkiewicz M, Casanovas P, Pagallo U, Sartor G, editors. AI Approaches to the Complexity of Legal Systems. Cham: Springer International Publishing; 2021. p. 1-16.
[7] Francesconi E. A Description Logic Framework for Advanced Accessing and Reasoning over Normative Provisions. International Journal on Artificial Intelligence and Law. 2014;22(3):291-311.
[8] MacCormick N, Summers R. Interpreting Statutes. A Comparative Study. Dartmouth: Ashgate; 1991.
[9] Biagioli C. Modelli Funzionali delle Leggi. Verso testi legislativi autoesplicativi.. vol. 6 of Legal Information and Communications Technologies Series. Florence, Italy: European Press Academic Publishing; 2009.
[10] Francesconi E. Semantic Model for Legal Resources: Annotation and Reasoning over Normative Provisions. Semantic Web journal: Special Issue on Semantic Web for the legal domain. 2016;7(3):255-65.
[11] Biagioli C, Cappelli A, Francesconi E, Turchi F. Law Making Environment: perspectives. In: Proceedings of the V Legislative XML Workshop. European Press Academic Publishing; 2007. p. 267-81.
[12] Antoniou G, Billington D, Governatori G, Maher MJ. On the modeling and analysis of regulations. In: Proceedings of the Australian Conference Information Systems. Victoria University of Wellington, New Zealand; 1999. p. 20-9.
[13] Hoekstra R, Breuker J, di Bello M, Boer A. Lkif core: Principled ontology development for the legal domain. In: Breuker J, Casanovas P, Klein M, Francesconi E, editors. Law, Ontologies and the Semantic Web. vol. 188 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2009. p. 21-52.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210332

A GDPR International Transfer Compliance Framework Based on an Extended Data Privacy Vocabulary (DPV)

David HICKEY[ab], Rob BRENNAN[a]
[a]*ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland*
[b]*Datanet International, Dublin, Ireland*

Abstract. This paper describes a tool using an extended Data Privacy Vocabulary (the DPV) to audit and monitor GDPR compliance of international transfers of personal data. New terms were identified which have been proposed as extensions to the DPV W3C Working Group. A prototype software tool was built based on the model plus a set of validation rules, and synthetic use-cases created to test the capabilities of the model and tool (together a compliance framework). This framework was created because the rules around international transfer compliance are complex and changing, there is an absence of a common approach to ensuring compliance, few tools exist to assist, and those that do lack interoperability. Evaluation results demonstrate that the proposed model improves compliance identification and standardisation. The tool received positive feedback from the data protection practitioners who participated in the evaluation, and an initial version of is now in use in one financial services organisation. While currently the tool only addresses international transfers, in theory the framework can be extended through further work to the broader area of compliance of other aspects of the GPDR.

Keywords. GDPR, International Transfer, Compliance, DPV, Data Protection, Privacy.

1. Introduction

The General Data Protection Regulation (GDPR) requires organisations to demonstrate compliance with data protection law. Whether needed to create standardised and reliable internal processes, or to respond to an external regulatory audit, there are few models or tools available to address this compliance requirement. Those that are available tend to focus on recording overall compliance by asking the user to confirm they are compliant with specific GDPR Articles, through guided questions, and then logging this. The shortcomings of this approach are that it generally requires legally trained staff to answer the questions effectively, there is little or no support for automated checking or validation of the answers as little data is collected on the transfer process itself, and form filling may become habitual (checking 'yes', 'yes', 'yes', repeatedly) leading to a degradation of accuracy.

In terms of available tools, some regulators such as the Irish DPC [1] have introduced checklists, others have provided templates such as the Accountability Framework [2] from the ICO in the UK, and some like the French CNIL [3] have developed open-source compliance software. And there are also commercially available compliance tools such as

OneTrust[1], PrivIQ[2] and others. These tools all seem to be limited in a number of ways: (a) rather than assisting the user in determining compliance, they focus on recording the user's view or declaration (b) they require an expert knowledge of data protection law (c) the tools are standalone and not interoperable and (d) they are not well suited for ongoing management of compliance.

For compliance of international transfers of personal data, the tools are even more limited. The tools referenced here typically record the fact of an international transfer, the legal basis stated by the user as evidence of compliance as single entries, but little else. Compliance determination for transfers of personal data is complex. There is a need for a transfer legal basis, country-specific transfer safeguards, requirements for Transfer Impact Assessments, an evaluation of whether supplementary measures need to be put in place and if so, a determination of which ones to implement. Added to this is the changing landscape where additional countries such as UK [4] can be granted GDPR Art. 45 Adequacy, and in other cases legal bases for transfers such as EU-US Safe Harbor [5] have been invalidated by the Court of Justice of the European Union (CJEU).

This paper focuses on International Transfers of personal data to third countries, and seeks to provide a framework (model and tool) to assist in ensuring such transfers are compliant with EU law. The main research question this paper asks is: *to what extent can a model based on DPV ensure regulatory-compliant international transfers of personal data ?* Related to this two additional sub-questions are asked: (i) how can we currently identify cross-border transfers and privacy issues in existing business processes ? and (ii) what are the requirements in developing a model and related tool to audit, report on, and monitor transfers ?

The approach taken in this paper to solving this problem is to:

- gather requirements for international transfers from domain experts
- identify extensions to DPV needed to model transfers
- develop a set of validation rules
- develop a prototype using Flowfinity Actions, the extended DPV and rules
- evaluate the model and this tool using synthetic use cases and external test users

The contributions of this paper are:

- identifying gaps, proposing 5 new DPV concepts and 6 properties for transfers
- creating a software tool based on the model (extended DPV and rules)
- developing a new framework for transfers based on the model and tool
- evaluating the tool and model against sample use cases

The rest of this paper is structured as follows: section 2 describes related work in this area; section 3 explores the DPV as a basis for the model; section 4 provides a model design leading to a tool design in section 5; section 6 and 7 describe the testing of the framework and results respectively; section 8 provides an analysis and discussion; and section 9 draws conclusions from the work.

[1] https://www.onetrust.com

[2] https://priviq.com/solutions-gdpr-software/

2. Related Work

Reasons for examining this area are three-fold: (a) limited work done to date on modelling compliance of transborder data transfers; (b) such transfers have been a focus of European regulators in the recent past – particularly since the introduction of the GDPR; (c) concerns about a need to evaluate the adequacy of data protection in the country in which an organisation processes personal data.

GDPR Art. 5(2) introduced new accountability obligations on Controllers and Processors to demonstrate compliance, with some specific examples being Art. 30 Record of Processing Activities (ROPA) and Art. 35 Data Protection Impact Assessment (DPIA). Since then, organisations have approached compliance in non-standard ways, typically using paper records, spreadsheets or a range of commercial tools. Even in 2018, there was a recognition of the gap in this approach. Ujcich et al. [6] comment on how increasingly complex personal data workflows create challenges for GDPR compliance. Research since has further highlighted the need for a standardised model for compliance. In [7], a year after the introduction of the GDPR, Torre et al. comment on the lack of an automated solution and the need for manual audits.

In 2020, Ryan et al. [8] reviewed tools in use for GDPR compliance and determined that there are gaps – particularly in standardization and interoperability. They suggest a RegTech approach to explore a prototype for GDPR compliance. At the same time, Pandit et al. [9] recommended the creation and adoption of standards and a common language for the exchange of GDPR compliance data. This theme aligns with the creation and development of the Data Privacy Vocabulary [10] or DPV. Early work by Pandit et al. [9,11] on interoperability and consent gives us a rich set of concepts and an extended (common) vocabulary relating to personal data processing. Their research references related work by the W3C Community Group for Data Protection Vocabularies and Controls (DPVCG) who have published the Data Privacy Vocabulary (DPV) specification. Ryan et al. [12] look at how the DPV might be used to model compliance with the GDPR requirement for Records of Processing Activities (ROPA), and proposes extensions to the vocabulary. Research on OWL2/DPV in the context of privacy policy language by Bonati et al. [13] shows how it can encode consent, business processes, and regulatory obligations. It also highlights the need for automated compliance checking. In [14], Leoni and Di Caro use the DPV to look at natural language processing of privacy policies.

The research work to date has identified a gap and a need for an automated compliance tool. GDPR requirements and new European Data Protection Board (EDPB) guidance [15] and [16] are creating an even stronger demand for a suitable standardised compliance framework.

3. Extending the DPV for International Transfers

To address the research question, it was necessary to understand (a) what the scope of such a model would need to be and (b) if the DPV could address the necessary scope. A baseline survey of domain experts was carried out in March 2021. This survey posed 20 questions, covering two areas: (a) how the organisation identifies and monitors compliance and (b) the nature of those transfers (in terms of the data subjects, personal data transferred, purposes and legal basis). Questions were deliberately open, to identify as many free-text vocabulary terms as possible in relation to transfers.

The survey was sent to students in the DCU Masters Programme in Data Protection and Privacy Law (with expert knowledge). 13 responses were received from respondents in various industry sectors helping to gather a representative view. Participants in this survey stated that spreadsheets are the most common way of tracking compliance, there are few other tools available. They also stated that identification of transfers typically take place through consultation with the business, or discovered from the outputs of impact assessments, ROPAs or breaches. Analysis of the responses (see Table 1) showed a great variation in the terminology used, without an agreed and defined vocabulary.

Table 1. Variation in Responses to Survey 1

Category	Different answers	Common answers
Legal basis for transfer	8	11
Data subject types	22	13
Purposes	21	2
Overall Common Responses		33%

A dictionary of terms commonly used when describing international data transfers was then built from the survey results combined with a review of transfer-related terms in the GDPR, to determine whether the DPV could usefully provide a potential basis for standardising compliance of international transfers. A total of 114 relevant concepts were listed, and compared to the DPV.

The results were promising, as 44% of the terms needed were fully provided by the DPV, and a further 40% partially matched. In building the initial model it was determined that these partial matches were usable. Of the remaining 18 terms not in the DPV, 4 related to Purposes/Measures and were not critical, and 4 were safeguards that had been replaced, leaving 10 which were proposed to the W3C-DPVCG in Aug 2021. In October 2021, 7 of these terms were adopted into the DPV[3]. These terms are shown in **Table 2**.

Table 2. Transfer compliance rule requirements requiring additions to DPV

Validation Rule Requirement	DPV Concept Proposed	Adopted in DPV v0.3
Transfer Start and Date	**dpv:hasStartDate, dpv:hasEndDate**	(alternative = dpv:Duration)
Data Exporter	**dpv:DataExporter**	**YES**
Data Importer	**dpv:DataImporter**	**YES**
Transfer to Country	**dpv:TransferCountry**	(Under discussion)
Legal basis for transfer	**dpv:DataTransferLegalBasis** **dpv:DataTransferTool,**	**YES** **YES**
Safeguard in use	**dpv:Safeguard** **dpv:SafeguardForDataTransfer**	**YES** **YES**
Supplementary Measures	**dpv:SupplementaryMeasure.**	**YES**

[3] DPV and DPV-GDPR v0.3 release can be found at https://w3.org/ns/dpv and https://w3.org/ns/dpv-gdpr

4. Framework Design

This now facilitated the initial design of the model, which seeks to identify international transfers, their legal basis, the levels of safeguards in use, and to show whether the transfer is compliant. For consistency and future interoperability, the model was designed to draw from the terms and concepts in the (extended) DPV. The final stage of design was to map the information capture requirements, the workflow and the associated DPV concepts and properties as shown in Figure 1.

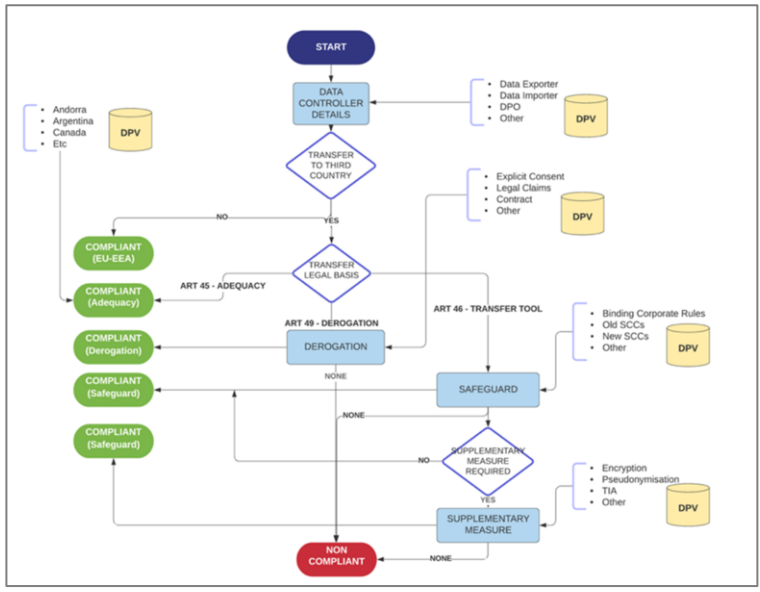


Figure 1. International transfer framework design and workflow

5. Prototype Design

A prototype software tool was designed with a user interface, guided workflow, a database (DPV extract), validation rules and an output (compliance report). The tool (Figure 2) takes user input, with intelligent workflow and conditional logic and is built on a schema based on a snapshot of the extended DPV. Natively using the DPV Linked Data rather than real-time queries (for no real gain) avoids possible versioning problems.

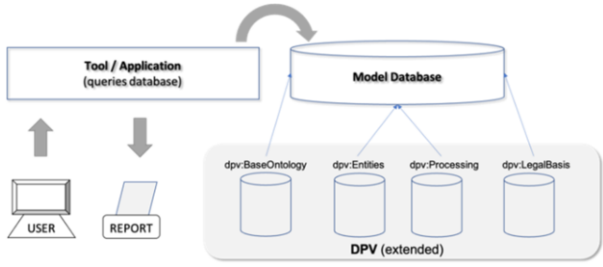


Figure 2. Transfer compliance prototype tool design

In building the tool, the primary purpose was to allow the framework to be evaluated. Rather than coding and development of a feature-rich tool, the focus became delivery of an MVP (Minimum Viable Product) to put a usable tool in the hands of test users (external participants) quickly[4]. The application development platform chosen to develop the prototype was Flowfinity Actions[5]. This provided a web-based portal, security infrastructure, workflow, a customisable user-interface, and integration capability.

The model's data capture requirements, logical rules and DPV datasets were then developed in the application. To assist in building the tool, approx. 200 test cases for international transfers were created based on the authors' own experience. The prototype Transfer Compliance tool was built and deployed in about six weeks, rather than potentially many months of development.

Two types of user or personas were envisaged while building the tool: (i) business user (without any particular domain knowledge); (ii) practitioner (someone with data protection knowledge). The tool therefore allows for separate (or common) data entry and review. In Figure 3, the user is presented with workflow driven conditional questions. At the back-end, relevant concepts from the DPV act as an internal data model schema for the tool. These DPV terms are then presented to the end user.

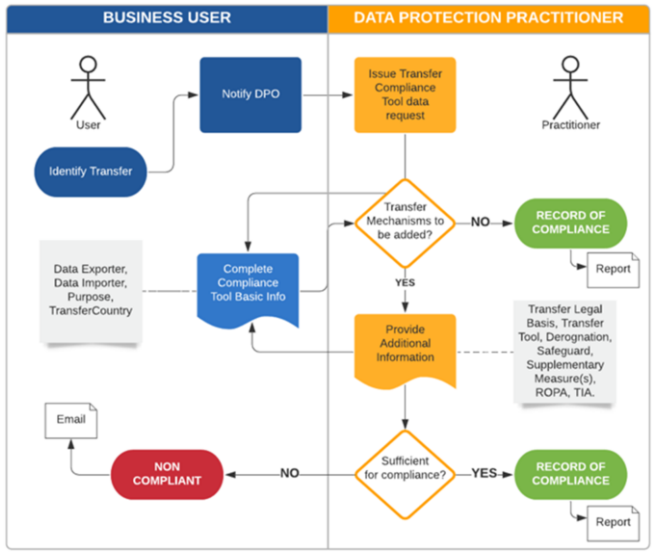


Figure 3. Tool user personas

The Transfer Compliance Tool collects a minimum of 34 data fields, of which only 3 are free-text input. The remaining fields are conditional or calculated which all combine to provide significant automation in the tool. Selection of a transfer country (see Figure 4) auto-populates compliance fields thus ensuring a high degree of consistency, and making inter-operability with other (DPV-based) tools easier.

[4] An initial version of the tool has now been deployed in a financial services organisation in Ireland.

[5] https://www.flowfinity.com/apps/

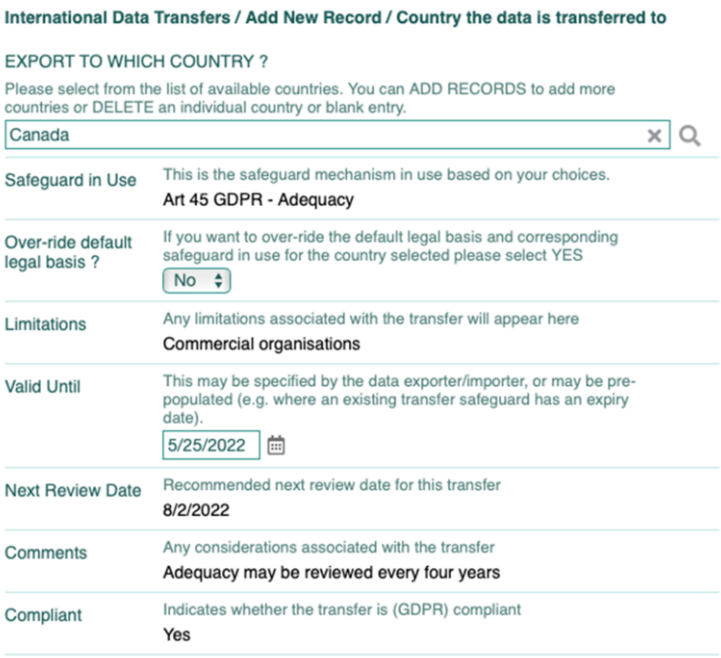


Figure 4. User interface from the prototype tool

When data entry is complete, compliance is automatically determined and a report is available within the tool (typically for the Practitioner) and an email report with details of compliance or deficiencies sent (typically to the Business user).

6. Evaluation

The hypotheses under test were whether a framework can be developed based on an extended DPV to model international transfers and ensure compliance; whether using a prototype tool based on this model will result in greater consistency in information gathered, conditions for compliance, and more informed decisions being made; and whether such a tool would achieve greater usability than current compliance methods.

A 'Gold Standard' of expected answers was created and externally validated by a domain expert. To measure accuracy, responses with expected answers were assigned a score of 1 and unexpected answers a score of 0. Tasks completed were evaluated against the Gold Standard. Consistency was measured as a percentage of expected out of 59.

Experimentation revolved around two synthetic use cases requiring the most common types of safeguards, which participants had to analyse for transfer compliance. The first (Green Bank) was a relatively simple, transfer of data to just one country (UK). The second (Childcare International) was more complex involving transfers to multiple countries (UK, Luxembourg, Singapore, USA) and some non-compliance.

A total of 13 participants was recruited from Ireland, UK and Germany based on personal contacts. Participants were classified as: Practitioners (10) with at least 3 years' experience or Business Users (3) with little or no previous data protection experience.

Participants used the Transfer Compliance Tool to complete the two use cases. Following this, a SUS questionnaire with additional open questions was completed by participants to gather feedback on the tool and model. In the evaluation, participants were expected to complete 59 fields in the Transfer Compliance tool.

7. Results

Thirteen sets of results were collected, each set covering the five international transfers. For each participant, 59 fields were evaluated against the Gold Standard answers for consistency. The results gave a high degree of correlation between actual and expected.

Table 3 Percentage of correct tasks vs Gold Standard

USE CASE	Business User Correct Tasks	Practitioner Correct Tasks	OVERALL Correct Tasks
Green Bank	93%	96%	95%
Childcare International	87%	92%	91%
OVERALL	89%	93%	**92%**

As seen in Table 3, the first use case gave an average of 95% consistency when compared to expected results. The second use case gave a slightly lower average of 91% consistency against expected results. And business users performed almost as well as practitioners. This is perhaps an indication that the tool helps more informed decision making by non-experts. Greatest discrepancies between actual and expected results were seen (Table 4) in open questions or when there were a large number of choices.

Table 4. Most common failed tasks vs Gold Standard

		Respondents	Failed %
1	Failure to correctly identify all processing purposes (C)	13	100%
2	Failure to correctly identify supplementary measures (C)	12	92%
3	Lack of standard naming for processing activity (O)	8	62%

(O) = open question (C) = over 40 choices

To evaluate the usability of the tool, a survey questionnaire based on the System Usability Scale (SUS) was used, with ten standard questions and an additional four optional open questions to gather less structured feedback about the potential value of the tool. All participants completed the SUS survey after using the tool for the two test cases. The resulting mean **SUS score was 82**.

One quote from a respondent that is representative of the responses received to the open questions reads: *"I have not found any specific tools that deal with international transfers. I think this tool could really help me with compliance."*

8. Discussion

Results show that a DPV-based framework can improve identification and compliance of international transfers. There was a clear improvement in standardisation of compliance responses (33% at the outset of this work to 92% using the framework). This resulted (i) from a linkage between the model and DPV terms (ii) from constraining data input in the tool to those terms available in the DPV and (iii) automation.

Modelling based on the DPV is not without its challenges, in particular as the DPV is still evolving. The failures shown in Table 4 mainly arise as the terms are not an exact match. Improving on the residual error rate may be possible with further development of the DPV and extensions to its concepts. Relating the results to the research question:

- The tool provides automated measurement of compliance of transfers
- Business and expert users can use the tool to identify and improve compliance
- Using the framework (DPV-based model and prototype tool) led to over 90% adherence to a gold standard for compliance validated by a domain expert.
- The prototype achieved a mean SUS score of 82, which is an 'A' grade score and where respondents are "*likely to recommend the product to a friend*" [17]

The DPV is a suitable vocabulary for such modelling, but extensions are needed. In developing the model, new concepts were identified, proposed as additions, and most adopted into the DPV. Compliance rules and guidance around international transfers change regularly. The model addresses this by importing the latest information from the DPV, and the tool allows for this to be extended to dynamic queries in the future.

9. Conclusions

By achieving a high consistency of standardised results, the framework developed has helped to answer the research question and demonstrate that the DPV, once extended and complemented with validation rules, can be used as a basis for ensuring compliance of international data transfers. The prototype tool received positive feedback on identification of transfers and helping improve compliance and accountability.

A limitation of the research, particularly for the quantitative results, was the limited dataset of respondents and use cases. Further research might revisit this and further extensions to the DPV to fully describe international transfers, while also looking at integration with other DPV tools currently being developed e.g. CSM-ROPA [12].

Acknowledgments

A word of thanks to Dr Julio Hernandez-Torres and Paul Ryan for their encouragement and sharing their own experiences.

This research has received funding from the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106_P2), co-funded by the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] Ireland DPC. GDPR Readiness Checklist Tools [Internet]. Self-Assess. Checkl. 2021 [cited 2021 Feb 2]. Available from: https://www.dataprotection.ie/en/organisations/resources-organisations/self-assessment-checklist.

[2] ICO UK. Introduction to the Accountability Framework [Internet]. ICO; 2020 [cited 2020 Nov 15]. Available from: https://ico.org.uk/for-organisations/accountability-framework/introduction-to-the-accountability-framework/.

[3] France C. The open source PIA software helps to carry out data protection impact assesment [Internet]. PIA Softw. 2021 [cited 2021 Jan 15]. Available from: https://www.cnil.fr/en/open-source-pia-software-helps-carry-out-data-protection-impact-assesment.

[4] Commission adopts adequacy decisions for the UK [Internet]. Eur. Comm. - Eur. Comm. [cited 2021 Aug 14]. Available from: https://ec.europa.eu/commission/presscorner/detail/en/ip_21_3183.

[5] Court of Justice declares that the Commission's US Safe Harbour Decision is invalid [Internet]. Off. J. 215 25082000 P 0007 - 0047. OPOCE; [cited 2021 Aug 14]. Available from: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32000D0520:EN:HTML.

[6] Ujcich BE, Bates A, Sanders WH. A Provenance Model for the European Union General Data Protection Regulation. In: Belhajjame K, Gehani A, Alper P, editors. Proven Annot Data Process. Cham: Springer International Publishing; 2018. p. 45–57.

[7] D. Torre, G. Soltana, M. Sabetzadeh, et al. Using Models to Enable Compliance Checking Against the GDPR: An Experience Report. 2019 ACMIEEE 22nd Int Conf Model Driven Eng Lang Syst MODELS. 2019. p. 1–11.

[8] Ryan P, Crane M, Brennan R. Design Challenges for GDPR RegTech: Proc 22nd Int Conf Enterp Inf Syst [Internet]. Prague, Czech Republic: SCITEPRESS - Science and Technology Publications; 2020 [cited 2021 Mar 7]. p. 787–795. Available from: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0009464507870795.

[9] Pandit HJ, O'Sullivan D, Lewis D. GDPR Data Interoperability Model. 2018 [cited 2021 Mar 7]; Available from: https://zenodo.org/record/3246438.

[10] Pandit HJ, Polleres A, Bos B, et al. Creating a Vocabulary for Data Privacy: The First-Year Report of Data Privacy Vocabularies and Controls Community Group (DPVCG). In: Panetto H, Debruyne C, Hepp M, et al., editors. Move Meaningful Internet Syst OTM 2019 Conf [Internet]. Cham: Springer International Publishing; 2019 [cited 2020 Nov 23]. p. 714–730. Available from: http://link.springer.com/10.1007/978-3-030-33246-4_44.

[11] Pandit HJ, Debruyne C, O'Sullivan D, et al. GConsent - A Consent Ontology Based on the GDPR. In: Hitzler P, Fernández M, Janowicz K, et al., editors. Semantic Web [Internet]. Cham: Springer International Publishing; 2019 [cited 2021 Mar 7]. p. 270–282. Available from: http://link.springer.com/10.1007/978-3-030-21348-0_18.

[12] Ryan P, Pandit HJ, Brennan R. A Common Semantic Model of the GDPR Register of Processing Activities. In: Villata S, Harašta J, Křemen P, editors. Front Artif Intell Appl [Internet]. IOS Press; 2020 [cited 2021 Mar 7]. Available from: http://ebooks.iospress.nl/doi/10.3233/FAIA200876.

[13] Bonatti P, Kirrane S, Petrova I, et al. Machine Understandable Policies and GDPR Compliance Checking. KI - Künstl Intell. 2020;34.

[14] Leone V, Di Caro L. The Role of Vocabulary Mediation to Discover and Represent Relevant Information in Privacy Policies. In: Villata S, Harašta J, Křemen P, editors. Front Artif Intell Appl [Internet]. IOS Press; 2020 [cited 2021 Mar 7]. Available from: http://ebooks.iospress.nl/doi/10.3233/FAIA200851.

[15] Olbrechts A. Recommendations 01/2020 on measures that supplement transfer tools to ensure compliance with the EU level of protection of personal data [Internet]. Eur. Data Prot. Board - Eur. Data Prot. Board. 2020 [cited 2021 Mar 7]. Available from: https://edpb.europa.eu/our-work-tools/public-consultations-art-704/2020/recommendations-012020-measures-supplement-transfer_en.

[16] Olbrechts A. Recommendations 02/2020 on the European Essential Guarantees for surveillance measures [Internet]. Eur. Data Prot. Board - Eur. Data Prot. Board. 2020 [cited 2021 Mar 7]. Available from: https://edpb.europa.eu/our-work-tools/our-documents/preporki/recommendations-022020-european-essential-guarantees_en.

[17] Sauro PhD J. Does Better Usability Increase Customer Loyalty? – MeasuringU [Internet]. 2010 [cited 2021 Aug 16]. Available from: https://measuringu.com/usability-loyalty/.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210333

Computability of Diagrammatic Theories for Normative Positions

Matteo PASCUCCI [a], Giovanni SILENO [b,1]
[a] *Slovak Academy of Sciences, Bratislava, Slovakia*
[b] *University of Amsterdam, Amsterdam, The Netherlands*

Abstract. Normative positions are sometimes illustrated in diagrams, in particular in didactic contexts. Traditional examples are the Aristotelian polygons of opposition for deontic modalities (squares, triangles, hexagons, etc.), and the Hohfeldian squares for obligative and potestative concepts. Relying on previous work, we show that Hohfeld's framework can be used as a basis for developing several Aristotelian polygons and more complex diagrams. Then, we illustrate how logical theories of increasing strength can be built based on these diagrams, and how those theories enable us to determine in a computably efficient way whether a set of normative positions can be derived from another set of normative positions.

Keywords. Computable Normative Theories and Diagrams and Hohfeldian relationships and Normative Positions and Polygons of Opposition

Introduction

Diagrams are acknowledged to be effective instruments for didactic purposes: they are commonly used to illustrate in an intuitive and accessible way the most various conceptual models, procedures, systems. Examples abound in engineering and theoretical sciences, but archetypes of diagrams include also the *Aristotelian square of opposition* [1], used in philosophical, linguistic, literary, and semiotic studies; and, within legal studies, the *two squares of fundamental normative relationships* due to Hohfeld [2,3]. Reference to regular shapes may be not by chance; cognitive studies show that symmetries facilitate perception of structure, memorization and recall [4]. Indeed, the motivating intuition behind this research is that diagrams may be useful to create user-friendly interfaces for the analysis of legal/contractual constructs. Rather than inspecting hundreds of sentences in the text of a contract, a subject may more easily figure out her normative relations (duties, rights, etc.) with the other parties by navigating or exploring a diagram-construed model. One may also ask whether these diagrams have similarly interesting computational properties; if this is the case, unveiling diagrammatic theories may benefit on multiple levels. For this reason, the present work focuses on the computational treatment of *Aristotelian diagrams* to represent normative positions, building upon previous interpretations of Hohfeld's squares relating these two types of diagrams [5,6,7,8,9].

[1]Corresponding Author: `g.sileno@uva.nl`. Matteo Pascucci was supported by Štefan Schwarz Fund (project "A fine-grained analysis of Hohfeldian concepts") and VEGA Grant No. 2/0117/19; he thanks Jonas Raab for discussion. Giovanni Sileno was supported by NWO for the DL4LD project (628.009.001) and the HUMAINER AI project (KIVI.2019.006). This article is the result of a joint research work of the two authors.

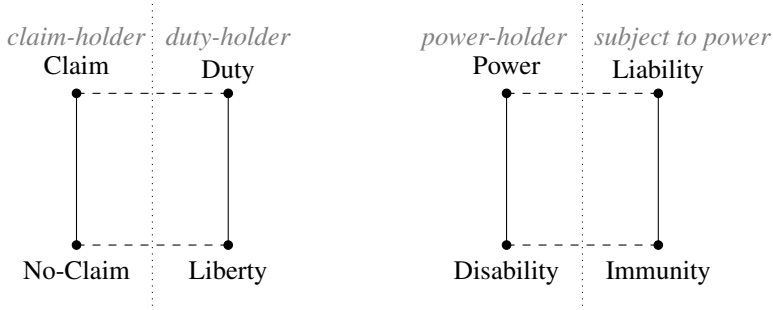


Figure 1. The two Hohfeldian squares: (left) the deontic square, (also obligative or of the first-order), and (right) the potestative square (or of the second-order).

The paper proceeds as follows. Section 1 introduces the notion of a normative position, the two Hohfeldian squares, and the notion of an Aristotelian diagram. Section 2 provides a formal language to encode normative positions and, reorganizing and extending previous results in [8], analyses the representation of Hohfeld's deontic square and of three versions of Hohfeld's potestative square (change-centered, force-centered and outcome-centered) in terms of Aristotelian diagrams. Section 3 shows how to develop logical theories over a diagram relying on the syntactic notion of an *inference tree*. These will be called *diagrammatic theories*. The analysis of a complex diagram which combines the three potestative squares of opposition is used as an example. Section 4 focuses on the computational dimension of diagrams, many aspects of which have recently gained attention in the literature (see, e.g., [10]). In particular, it shows that diagrammatic theories enable one to compute in polynomial time whether a finite set of normative positions can be derived from another.

1. Normative positions on diagrams

A *normative position* is described by a statement involving a normative concept, one or more normative parties related to that concept and a certain type of behaviour of one of those parties. For instance, consider the following:

- it is obligatory for the company to pay an annual fee;
- borrowers have a duty towards lenders to bring back the relevant goods;
- by accepting a sale offer, a buyer creates a duty upon the seller to deliver.

Two families of normative positions have been extensively analysed by Hohfeld [2,3]. He proposed to graphically visualize their relations via two diagrams, a *deontic* and a *potestative* square, which are reproduced in Fig. 1.

Many formal accounts of Hohfeld's analysis have been proposed (see, e.g., Lindahl [11], Makinson [12], and, more recently, Markovich [13] or Sileno and Pascucci [14]). Furthermore, some works [5,6,7] have shown that relationships on Hohfeldian squares can be used to construct Aristotelian polygons of opposition. The fundamental advantage of the latter over Hohfeld's squares is that they *unambiguously express logical relations* of a certain kind between normative statements. Aristotelian polygons can be combined to form more complex figures, whence we will generally speak of *Aristotelian diagrams*. At the basis of an Aristotelian diagram are four types of logical relations [10].

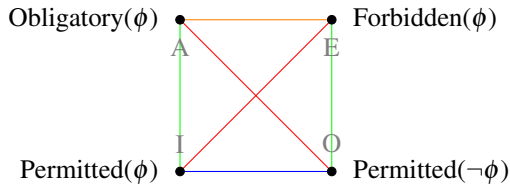


Figure 2. Aristotelian square for basic deontic modalities, with subaltern (green), contrary (orange), sub-contrary (blue) and contradictory (red) bindings, and named vertices (A, E, I, O). Sub-alternity is *directed*: from A to I, and from E to O, i.e., from orange towards blue lines.

Definition 1 (Aristotelian Relation)**.** *An* Aristotelian relation *between two sentences* ϕ *and* ψ *is either* subalternation*,* contrariety*,* sub-contrariety *or* contradiction*, where:*

- ϕ *is a subalternate of* ψ *iff the truth of* ψ *implies the truth of* ϕ*;*
- ϕ *and* ψ *are contrary iff at most one between* ϕ *and* ψ *can be true;*
- ϕ *and* ψ *are sub-contrary iff at most one between* ϕ *and* ψ *can be false;*
- ϕ *and* ψ *are contradictory iff* ϕ *is true precisely when* ψ *is false.*

Definition 2 (Aristotelian Diagram)**.** *Given a finite set of sentences* Γ*, an Aristotelian diagram over* Γ *is a graph whose vertices are labelled by elements of* Γ*. Each vertex* v *is labelled by a distinct sentence. An edge* e *of the graph connecting two vertices* v_1 *and* v_2 *is associated with an Aristotelian relation* R_e *between the sentences labelling* v_1 *and* v_2*.*

In the normative domain, the simplest example of Aristotelian diagram is the *square of opposition for basic deontic modalities* (Fig. 2).

2. Formalization

To enable a more rigorous analysis of normative positions, we here use a language $\mathscr{L}$ of first-order logic to encode sentences, following [11]. Moreover, refining and extending recent work [8], we show how to construct squares of oppositions starting from Hohfeld's squares and, subsequently, more complex Aristotelian diagrams.

2.1. Language

Language $\mathscr{L}$ has variables x, y etc. for normative parties and α, β, etc. for action types. It has constants p, q, etc. for normative parties and constants A, B, etc. for action types. Symbol $-$ (overline) denotes action complementation: $\overline{A}$ is the complement of A (the type of any action not instantiating A). We assume that $\overline{\overline{A}} = A$. Hohfeldian (and other) relations are encoded as n-ary predicates.[2] Finally, $\mathscr{L}$ has standard propositional connectives ($\neg$, $\wedge$, $\vee$, $\rightarrow$, $\equiv$) and quantifiers ($\forall$, $\exists$). We omit quantification over variables for normative parties, interpreting an expression $\phi(x,y,\ldots)$ as implicitly having the form $\forall x \forall y \ldots \phi(x,y,\ldots)$. Thus, while $\mathsf{Claim}(x,y,A)$ means "for all x, for all y: x has a claim that A be performed by y", $\mathsf{Claim}(p,q,A)$ means "p has a claim that A be performed by q".

[2] Sometimes the argument of a relation is a statement involving another relation. However, no quantification on such statements is employed; therefore, $\mathscr{L}$ remains a first-order language.

2.2. First-order Hohfeldian relations

The formal renderings of the fundamental deontic relations identified in Hohfeld's framework, for two normative parties p and q and an action type A, are the following: $\mathsf{Claim}(p,q,A)$, $\mathsf{Liberty}(p,q,A)$, $\mathsf{Duty}(p,q,A)$ and $\mathsf{NoClaim}(p,q,A)$. We can map all relationships to a single primitive, e.g. Claim:

$$\begin{aligned}\mathsf{NoClaim}(x,y,A) &\equiv \neg\mathsf{Claim}(x,y,A)\\ \mathsf{Duty}(y,x,A) &\equiv \mathsf{Claim}(x,y,A)\\ \mathsf{Liberty}(y,x,A) &\equiv \neg\mathsf{Claim}(x,y,\overline{A})\end{aligned}$$

This choice leads to the following set of labels DR with respect to a given action type A:

$$\mathsf{DR} = \{\mathsf{Claim}(p,q,A), \mathsf{Claim}(p,q,\overline{A}), \neg\mathsf{Claim}(p,q,\overline{A}), \neg\mathsf{Claim}(p,q,A)\}$$

The set DR naturally gives rise to a deontic square of opposition. The only additional principle needed is the following, used to characterize *subalternate statements*: $\mathsf{Claim}(x,y,A) \rightarrow \neg\mathsf{Claim}(x,y,\overline{A})$, which can be seen as corresponding to the $\mathit{Obligatory}(\phi) \rightarrow \mathit{Permitted}(\phi)$ axiom used in deontic logics.

2.3. Second-order Hohfeldian relations

Potestative relations concern *actions* that trigger changes of first-order or even second-order relations, such as, for instance, an action B creating a duty for a party q with respect to a party p to perform an action A. A possible way of writing that p has such a power would be by means of a predicate expression $\mathit{Ability}(p,B,R)$ (cf. the predicate *has_ability* investigated in [14]), where R is a Hohfeldian relation issued at B's performance by p; for instance, $\mathit{Ability}(p,B,\mathsf{Claim}(p,q,A))$. Different characterizations of actions exist in human language, mapping to different levels of abstraction [15], e.g. the *behavioural* or *procedural* characterization, relating to the action task, or the *productive* characterization, relating to its outcome. In the following, we similarly provide different definitions of power constructed at different abstraction levels (force, outcome, change).

2.3.1. Force-centered power

At behavioural level, power relations can be seen in analogy to force fields determining attraction, repulsion, and absence of those (independence) at the occurrence of interventions ([7], [16, Ch.4]). To express this, we need to separate the *stimulus* component (a particular type of action, such as a verbal command) and the consequent target *manifestation* (e.g. a type of action that is due or expected on the basis of the stimulus, cf. the concept of *pliance*). If the action-manifestation is denoted by the action type symbol A, then, the action-stimulus can be conveniently represented via the symbol `"A"` to emphasize the shared connection between signal and expected performance.

If stimulus and manifestation converge, i.e. A is always performed in correspondence to its stimulus, we have a *positive force-centered* power:

$$\overrightarrow{\mathsf{Power}}^{\leftarrow}(x,y,A) \equiv \mathit{Ability}(x,\texttt{"A"},\mathsf{Claim}(x,y,A))$$

If stimulus and manifestation diverge, i.e. A is never performed in correspondence to its stimulus, we have a *negative force-centered* power:[3]

$$\overleftrightarrow{\mathsf{Power}}(x,y,A) \equiv Ability(x,\texttt{"A"},\mathsf{Claim}(x,y,\overline{A}))$$

From these concepts we can define a new set of potestative relations $\mathsf{PR}^{\overleftarrow{\rightarrow}}$ as labels for a force-centered potestative square of opposition, obtained by taking into account all possible combinations of positive- vs. negative-force power and Boolean negation:

$$\mathsf{PR}^{\overleftarrow{\rightarrow}} = \{\overrightarrow{\mathsf{Power}}(p,q,A), \overleftrightarrow{\mathsf{Power}}(p,q,A), \neg\overleftrightarrow{\mathsf{Power}}(p,q,A), \neg\overrightarrow{\mathsf{Power}}(p,q,A)\}$$

The subalternity is here captured by the logical principle: $\overrightarrow{\mathsf{Power}}(x,y,A) \rightarrow \neg\overleftrightarrow{\mathsf{Power}}(x,y,A)$ which is acceptable because otherwise the same stimulus `"A"` would generate two conflicting first-order relations.

2.3.2. Outcome-centered power

At the outcome or productive abstraction level, we may abstract the triggering action B, and focus only the output, e.g. $R = \mathsf{Claim}(p,q,A)$. *Positive outcome-centered* power, expressed in the form $\mathsf{Power}(p,q,A)$, means that p has the power of issuing a duty upon q to A. It can be defined via an existential quantification on the set of action types:

$$\mathsf{Power}(x,y,A) \equiv \exists\beta : Ability(x,\beta,\mathsf{Claim}(x,y,A))$$

We can similarly define a *negative outcome-centered* power (to release a duty to A) as:

$$\overline{\mathsf{Power}}(x,y,A) \equiv \exists\beta : Ability(x,\beta,\neg\mathsf{Claim}(x,y,A))$$

As before, we can form a set of potestative relations PR as labels for an outcome-centered potestative square of opposition:

$$\mathsf{PR} = \{\mathsf{Power}(p,q,A), \overline{\mathsf{Power}}(p,q,A), \neg\overline{\mathsf{Power}}(p,q,A), \neg\mathsf{Power}(p,q,A)\}$$

where subalternity is captured by: $\mathsf{Power}(p,q,A) \rightarrow \neg\overline{\mathsf{Power}}(p,q,A)$. This principle can be explained as such: to create a duty, this duty needs not to be holding: $\mathsf{Power}(x,y,A) \rightarrow \neg\mathsf{Claim}(x,y,A)$. Dually, to release a duty, this needs to exist: $\overline{\mathsf{Power}}(x,y,A) \rightarrow \mathsf{Claim}(x,y,A)$; its contrapositive provides the subalternity relation.

2.3.3. Change-centered power

Given a target relation R, e.g. a due performance $\mathsf{Claim}(p,q,A)$, one can define power also as the ability of p to *affect* q in any sense with respect to this relation R. This proposal, originally made by O'Reilly [6], can be reframed in our framework as the ability of changing q's position in the (e.g. deontic) square of opposition of which R is part. Using the proposed notation we have:

$$\begin{aligned}\mathsf{Power}_{\mathsf{OReilly}}(x,y,B,A) \equiv\ & Ability(x,B,\mathsf{Claim}(x,y,A)) \vee Ability(x,B,\mathsf{Claim}(x,y,\overline{A})) \\ & \vee Ability(x,B,\neg\mathsf{Claim}(x,y,A)) \vee Ability(x,B,\neg\mathsf{Claim}(x,y,\overline{A}))\end{aligned}$$

[3] This position is neglected in the analytical literature but it is critically important in institutional-construction terms: it posits the denial to recognize another normative system which attempts to positively enact a certain power, see e.g. the Dutch Declaration of Independence, the Act of Abjuration (1581) towards Spain. [7]

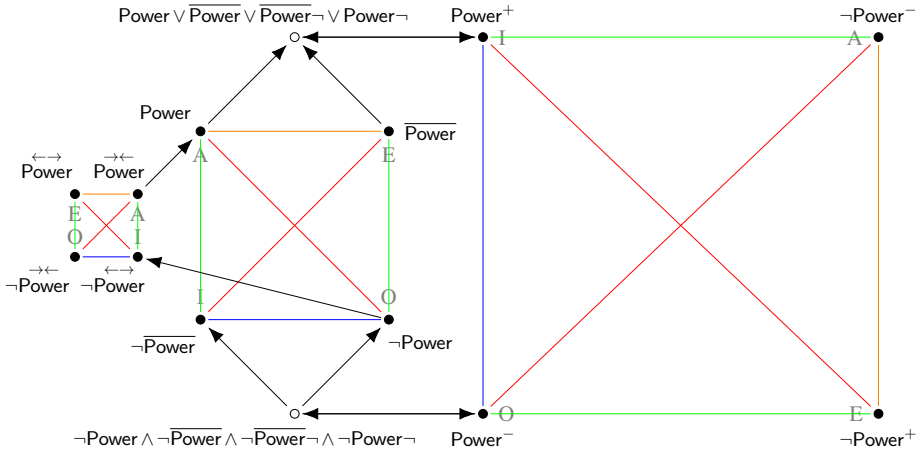


Figure 3. Map of potestative relations defined in terms of triggering action (force-centered square of opposition, the left one), in terms of outcome (middle square), in terms of change or affecting outcomes (change–centered square of opposition, the right one). For visual clarity, labels of vertices are simplified so as to consist only of a (possibly negated) predicate without its arguments. Occurrences of negation before a predicate are standard, whereas occurrences after a predicate denote action complementation; for instance, we denote the power to issue a prohibition, i.e., $\mathsf{Power}(p,q,\overline{A})$, as $\mathsf{Power}\neg$. Notice that the leftmost square is vertically mirrored and the rightmost square underwent a 90 degree clockwise rotation. Colours are as usual.

A *positive change-centered* power corresponds to the ability of affecting the target relation by any triggering action:

$$\mathsf{Power}^{+}(x,y,A) \equiv \exists \beta : \mathsf{Power}_{\mathsf{OReilly}}(x,y,\beta,A)$$

A *negative change-centered* power corresponds to the ability of the agent to perform an action without affecting the target relation:

$$\mathsf{Power}^{-}(x,y,A) \equiv \exists \beta : \neg\mathsf{Power}_{\mathsf{OReilly}}(x,y,\beta,A)$$

Labels for an Aristotelian square are, this time:

$$\mathsf{PR}^{\pm} = \{\mathsf{Power}^{+}(p,q,A), \mathsf{Power}^{-}(p,q,A), \neg\mathsf{Power}^{-}(p,q,A), \neg\mathsf{Power}^{+}(p,q,A)\}$$

Subalternity is here encoded by: $\neg\mathsf{Power}^{-}(x,y,A) \rightarrow \mathsf{Power}^{+}(x,y,A)$.

2.3.4. Relationships amongst powers

The previous formulas can be applied to discover different relationships between the distinct forms of powers. First, the convergence or divergence with due performance in force-centered powers map directly or dually to outcome-centered powers:

$$\overrightarrow{\mathsf{Power}}(x,y,A) \rightarrow \mathsf{Power}(x,y,A) \qquad \overleftrightarrow{\mathsf{Power}}(x,y,A) \rightarrow \mathsf{Power}(x,y,\overline{A})$$

Following the contrapositive, the absence of positive-outcome power to produce a duty (meaning that there is no triggering action to obtain this) maps *a fortiori* to the absence of a positive-force power for doing so:

$$\neg\mathsf{Power}(x,y,A) \rightarrow \neg\overrightarrow{\mathsf{Power}}(x,y,A) \qquad \neg\mathsf{Power}(x,y,\overline{A}) \rightarrow \neg\overleftrightarrow{\mathsf{Power}}(x,y,A)$$

Second, positive-change power holds if any outcome-center power holds:

$$\mathsf{Power}^{+}(x,y,A) \leftrightarrow \mathsf{Power}(x,y,A) \vee \mathsf{Power}(x,y,\overline{A}) \vee \overline{\mathsf{Power}}(p,q,A) \vee \overline{\mathsf{Power}}(p,q,\overline{A})$$

Assuming again that there is always some available action, we have, dually, that:

$$\mathsf{Power}^{-}(x,y,A) \leftrightarrow \neg\mathsf{Power}(x,y,A) \wedge \neg\mathsf{Power}(x,y,\overline{A}) \wedge \neg\overline{\mathsf{Power}}(p,q,A) \wedge \neg\overline{\mathsf{Power}}(p,q,\overline{A})$$

Introducing those relationships, we obtain the Aristotelian diagram in Fig. 3.

3. Diagrammatic theories

It is possible to define logical theories of different strength based on an Aristotelian diagram. These can be called *diagrammatic theories*. A diagrammatic theory over a diagram $\mathscr{D}$ encodes at least all logical relations among formulas used as labels in $\mathscr{D}$. A diagrammatic theory will be here defined in terms of the notion of an *inference tree*.

Definition 3 (Inference Tree)**.** *An* inference tree *T is an irreflexive and intransitive tree $(N,\rightsquigarrow)$ with a single* root *($N \neq \varnothing$ is a finite set of* nodes *and, for any $n,m \in N$, $n \rightsquigarrow m$ means that m is an* immediate successor *of n) where each $n \in N$ is associated with a finite set of formulas $\Gamma \neq \varnothing$ and has a* rank*. The rank of the root is $\mathbf{0}$; furthermore, if $rank(n) = \mathbf{i}$, n is associated with a set Θ, and $n \rightsquigarrow m$, then $rank(m) = \mathbf{i}+\mathbf{1}$ and m is associated with a set $\Sigma \supseteq \Theta$. Nodes with no successors are said to be* leaves *of T. A maximal $\rightsquigarrow$-chain of nodes $\sigma = (n_1,\ldots,n_k)$ is a* branch *of T.*

Formulas in sets associated with nodes of a tree T can be uniformly substituted. Furthermore, an *equivalence relation* $Eq \subset \mathscr{L} \times \mathscr{L}$ can be established in order to replace, in any set Γ associated with a node n, a formula ϕ with a formula ψ, provided that $Eq(\phi,\psi)$.

Definition 4 (Set Immediate Inference)**.** *If Δ is a set of formulas (associated with a node) ranked with $\mathbf{i}$ and Δ' a set of formulas (associated with a node) ranked with $\mathbf{i}+\mathbf{1}$ in a branch σ of a tree T, then we say that Δ' can be immediately inferred from Δ within σ.*

Definition 5 (Diagrammatic Theory)**.** *A* diagrammatic theory $\mathbb{DT}$ *based on a diagram $\mathscr{D}$ is a set of inference trees satisfying the following properties, for each formulas ϕ and ψ that label some vertex of $\mathscr{D}$:*

- *if ψ is a* subalternant *of ϕ in $\mathscr{D}$, then some branch σ of a tree T in $\mathbb{DT}$ encodes an inference of the form $\Delta \cup \{\phi\} \rightsquigarrow \Delta \cup \{\phi,\psi\} \cup \Gamma$;*
- *if ψ and ϕ are* contraries *in $\mathscr{D}$, then some branch σ of a tree T in $\mathbb{DT}$ encodes an inference of the form $\Delta \cup \{\phi\} \rightsquigarrow \Delta \cup \{\phi,\neg\psi\} \cup \Gamma$;*
- *if ψ and ϕ are* sub-contraries *in $\mathscr{D}$, then some branch σ of a tree T in $\mathbb{DT}$ encodes an inference of the form $\Delta \cup \{\neg\phi\} \rightsquigarrow \Delta \cup \{\neg\phi,\psi\} \cup \Gamma$;*
- *if ψ and ϕ are* contradictories *in $\mathscr{D}$, then some branch σ of a tree T in $\mathbb{DT}$ encodes an inference of the form $\Delta \cup \{\phi\} \rightsquigarrow \Delta \cup \{\phi,\neg\psi\} \cup \Gamma$ and some branch σ' of a tree T' in $\mathbb{DT}$ encodes an inference of the form $\Delta' \cup \{\psi\} \rightsquigarrow \Delta' \cup \{\neg\phi,\psi\} \cup \Gamma'$.*

According to Definitions 4 and 5, the relation of immediate inference in a diagrammatic theory based on a diagram $\mathscr{D}$ encodes at least all Aristotelian relations between formulas labelling vertices of $\mathscr{D}$. However, the use of the equivalence relation Eq allows one to reduce the number to Aristotelian relations to be encoded, as we will clarify below. Here we just assume $Eq(\phi, \neg\neg\phi)$, for every $\phi \in \mathscr{L}$.

T' is a *sub-tree* of T iff the root of T' is a node $n \in T$ and the other nodes of T' are all those that (i) occur in branches of T to which n belongs, and (ii) have a higher rank than n in those branches. The cardinality of a set of formulas Γ will be denoted as $|\Gamma|$. The notion of inference within a branch is obtained by combining the transitive closure of the notion of immediate inference and the subset relation, as indicated below.

Definition 6 (Set Inference)**.** *A set of formulas* Γ *can be* inferred *from a set of formulas* Δ *within a branch* σ *of a tree* T *iff there is* $\Gamma' \supseteq \Gamma$ *s.t.:*

- *both* Δ *and* Γ' *belong to* σ*;*
- *the rank of* Γ' *in* σ *is not lower than the rank of* Δ *in* σ*.*

According to Definition 6, if $\Gamma \subseteq \Delta$, then Γ can always be inferred from Δ within a branch of a tree. In order to check whether Γ can be inferred from Δ in a branch σ of a tree T, one has to compare pairs of sets (checking whether they are identical, one is a subset of the other, etc.). Derivation is a particular kind of inference, as per the following definition.

Definition 7 (Set Derivability - Trees)**.** *A set of formulas* Γ *is derivable from a set of formulas* Δ *within a tree* T *iff there is a sub-tree* T' *of* T *whose root is a node n labelled by* Δ *and* Γ *can be inferred from* Δ *within all branches of* T'*.*

The derivability of Γ from Δ *within a diagrammatic theory* $\mathbb{DT}$ is defined in terms of a finite sequence of derivations within trees of $\mathbb{DT}$, as below.

Definition 8 (Set Derivability - Diagrammatic Theories)**.** *A set of formulas* Γ *can be derived from a set of formulas* Δ *within a diagrammatic theory* $\mathbb{DT}$ *iff there are trees* $T_1, \ldots, T_{n-1}$ *in* $\mathbb{DT}$ *and sets of formulas* $\Delta_1, \ldots, \Delta_n$ *s.t.:*

- $\Delta = \Delta_1$ *and* $\Gamma = \Delta_n$*;*
- *for* $1 \leq j < n$, Δ_{j+1} *can be derived from* Δ_j *within tree* T_j*.*

Example Below is an example of an inference tree that can be used in a diagrammatic theory built over Fig. 3. It captures inferences from a set of formulas Δ including the label of the E-corner in the change-centered (rightmost) square. Each line represents a node of the tree (which, in this, case has no branches) and starts with the node's rank:

$$\mathbf{0}: \Delta_0 = \Delta \cup \{\neg \mathsf{Power}^{+}(p,q,A)\} \rightsquigarrow$$
$$\mathbf{1}: \Delta_1 = \Delta_0 \cup \{\mathsf{Power}^{-}(p,q,A)\} \rightsquigarrow$$
$$\mathbf{2}: \Delta_2 = \Delta_1 \cup \{\neg \mathsf{Power}(p,q,A) \wedge \neg \overline{\mathsf{Power}}(p,q,A) \wedge \neg \overline{\mathsf{Power}}(p,q,\overline{A}) \wedge \neg \mathsf{Power}(p,q,\overline{A})\} \rightsquigarrow$$
$$\mathbf{3}: \Delta_3 = \Delta_2 \cup \{\neg \overline{\mathsf{Power}}(p,q,A), \neg \mathsf{Power}(p,q,A)\} \rightsquigarrow$$
$$\mathbf{4}: \Delta_4 = \Delta_3 \cup \{\neg \overleftrightarrow{Power}(p,q,A)\}$$

Any diagrammatic theory including this tree allows one (due to Definitions 4, 6 and 7) to derive any set $\Gamma \subseteq \Delta_i$, for $0 \leq i \leq 4$, from the starting set of formulas Δ_0.

4. Decidability and complexity

Given two *finite* sets $\Gamma, \Delta \subset \mathscr{L}$, we will say that the problem of checking whether Γ can be derived from Δ within a diagrammatic theory $\mathbb{DT}$ is the *derivability problem for finite sets* in $\mathbb{DT}$. We now illustrate that such a problem can be effectively computed. First, we need to define an auxiliary notion.

Definition 9 (Tree traversal)**.** *The* traversal *of a tree T with reference to a formula ϕ and a set Δ is a procedure which can be described as follows (we assume that Δ occupies the root of T):*

- *Following the order of ranks, for any set of formulas Γ with rank* **i** *in T, we compare ϕ with all formulas in Γ and keep track of whether ϕ occurs in Γ or not.*
- *The procedure terminates when either (*positive outcome*) there is a rank* **j** *s.t. all sets of formulas with rank* **j** *include ϕ or (*negative outcome*) all sets of formulas with all ranks available in T have been checked.*

Notice that, in case of a positive outcome of the traversal, due to Definitions 6 and 7, $\Delta \cup \{\phi\}$ is derivable from Δ within T. If the number of nodes in T is l, and $max\{|\Theta \setminus \Sigma| : rank(\Theta) = \mathbf{1} + rank(\Sigma)\} = k$ (hereafter, *the maximum successor difference*), then a traversal of T with reference to ϕ and Δ requires up to $k * l$ moves.

Definition 10 (Theory traversal)**.** *The* traversal of a diagrammatic theory $\mathbb{DT}$ *with reference to a formula ϕ and a set of formulas Δ is the traversal of all trees T in $\mathbb{DT}$ with reference to ϕ and Δ. The outcome is* positive *iff it is positive for some T in $\mathbb{DT}$.*

If the number of trees in $\mathbb{DT}$ is h, the maximum number of nodes in a tree of $\mathbb{DT}$ is l, and the maximum successor difference in a tree of $\mathbb{DT}$ is k, then a traversal of $\mathbb{DT}$ with reference to ϕ and Δ requires up to $(k * l) * h$ moves.

Theorem 1 (Decidability)**.** *The derivability problem for finite sets within a diagrammatic theory is decidable.*

Proof. Consider two finite sets of formulas Δ and Γ, such that $max(|\Delta|, |\Gamma|) = n$. Let $\mathbb{DT}$ be the diagrammatic theory at issue, consisting of h inference trees. To see whether Γ is derivable from Δ, we first need to compare the sets Γ and Δ, an operation with time complexity $O(n)$. There are four relevant Cases: (1) $\Gamma = \Delta$; (2) $\Gamma \subset \Delta$; (3) $\Gamma \supset \Delta$; (4) $\Gamma \not\subset \Delta$, $\Gamma \not\supset \Delta$ and $\Gamma \neq \Delta$. In Cases 1 and 2, by Definition 8, we can immediately conclude that Γ is derivable from Δ within $\mathbb{DT}$. In Cases 3 and 4 one considers the set $\Gamma \setminus \Delta$. Let $|\Gamma \setminus \Delta| = m$. We know that the elements of $\Gamma \setminus \Delta$ can be enumerated ($m < n$ by construction). We take the first formula ϕ_1 in $\Gamma \setminus \Delta$ and perform a *traversal* of $\mathbb{DT}$ with reference to ϕ_1 and Δ. If the traversal produces a negative outcome, then the whole procedure terminates and Γ is not derivable from Δ. Otherwise, there is a tree T s.t., due to Definitions 6, 7 and 9, $\Delta \cup \{\phi\}$ is derivable from Δ within T. We take $\Delta_1 = \Delta \cup \{\phi_1\}$ and then perform a *traversal* of $\mathbb{DT}$ with reference to ϕ_2 and Δ_1. The procedure is reiterated: at each step a traversal of $\mathbb{DT}$ is performed with reference to a new formula $\phi_j \in \Gamma \setminus \Delta$ and the set of formulas Δ_{j-1} obtained at the previous step. In accordance with Definition 8, Γ is derivable from Δ iff each traversal of $\mathbb{DT}$ performed ends with a positive outcome. □

Theorem 2 (Complexity)**.** *The algorithm to solve the derivability problem for finite sets in a diagrammatic theory takes polynomial time.*

Proof. In all Cases (1-4) mentioned in the proof of Theorem 1 the procedure terminates in at most $((k*l)*h)*m$ moves. □

5. Final remarks

Recent years have witnessed a renewed interest in diagrams and logical geometry: our work provides a further contribution towards their application in the analysis of normative problems. We introduced a syntactic method of reasoning over Aristotelian diagrams that encode relations between normative positions. The method can be automated and allows one to derive a finite set of normative positions from another in polynomial time. In our opinion, reasoning methods associated with diagrams are very promising and, due to the cognitive efficacy of visual aids, they are potentially accessible to a broader audience. Future investigations concern: (i) a theoretical and empirical comparison with other methods for (normative) automated reasoning, (ii) an integration of temporal/causal components within the definition of normative positions, and (iii) the consolidation of a structured taxonomy of Aristotelian diagrams for normative reasoning.

References

[1] T. Parsons. The traditional square of opposition. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.
[2] W.N. Hohfeld. Some fundamental legal conceptions as applied in judicial reasoning. *Yale Law Journal*, 23(1):16–59, 1913.
[3] W.N. Hohfeld. Fundamental legal conceptions as applied in judicial reasoning. *Yale Law Journal*, 26(8):710–770, 1917.
[4] A.B. Markman and D. Gentner. The effects of alignability on memory. *Psychological Science*, pages 363–367, 1997.
[5] L.W. Sumner. *The Moral Foundation of Rights*. Clarendon Press, 1987.
[6] D.T. O'Reilly. Using the square of opposition to illustrate the deontic and alethic relations. *University of Toronto Law Journal*, 45(3):279–310, 1995.
[7] G. Sileno, A. Boer, and T. van Engers. On the interactional meaning of fundamental legal concepts. In R. Hoekstra, editor, *Proceedings of JURIX 2014*, pages 39–48, 2014.
[8] M. Pascucci and G. Sileno. The search for symmetry in Hohfeldian modalities. In A. Basu, G. Stapleton, S. Linker, C. Legg, E. Manalo, and P. Viana, editors, *Diagrammatic Representation and Inference. Proceedings of Diagrams 2021*, pages 87–102. Springer, 2021.
[9] J. A. de Oliveira Lima, C. Griffo, J. P. A. Almeida, G. Guizzardi, and M. I. Aranha. Casting the light of the theory of opposition onto hohfeld's fundamental legal concepts. *Legal Theory*, 27(1):2–35, 2021.
[10] L. Demey. Computing the maximal Boolean complexity of families of Aristotelian diagrams. *Journal of Logic and Computation*, 28:1323–1339, 2018.
[11] L. Lindahl. *Position and Change: A Study in Law and Logic*. Synthese Library. Springer, 1977.
[12] D. Makinson. On the formal representation of rights relations. *Journal of Philosophical Logic*, 15:403–425, 1986.
[13] R. Markovich. Understanding Hohfeld and formalizing legal rights: the Hohfeldian conceptions and their conditional consequences. *Studia Logica*, 108(1):129–158, 2020.
[14] G. Sileno and M. Pascucci. Disentangling deontic positions and abilities: a modal analysis. In F. Calimeri, S. Perri, and E. Zumpano, editors, *Proceedings of CILC 2020*, volume 2710, pages 36–50, 2020.
[15] J.F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, volume 13. MIT Press, 2000.
[16] G. Sileno. *Aligning Law and Action*. PhD thesis, University of Amsterdam, 2016.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210334

Computing Private International Law

Guido GOVERNATORI [a], Francesco OLIVIERI [a], Antonino ROTOLO [b], Abdul SATTAR [a], Matteo CRISTANI [c]

[a] *Institute for Integrated and Intelligent Systems, Griffith University, Australia*
[b] *Alma AI, University of Bologna, Italy*
[c] *Department of Computer Science, University of Verona, Italy*

Abstract. This paper develops a new comprehensive computational framework for reasoning about private international law that encompasses the reasoning patterns modeled by previous works [3,8,9]. The framework is a multi-modal extension of [10] preserving some nice properties of the original system, including some efficient algorithms to compute the extensions of normative theories representing legal systems.

Keywords. Computational Logic, Private International Law, Defeasible Reasoning

1. Introduction

An increasing attention has been paid in the last few years to the interaction among distinct normative systems with regard to the allocation of jurisdiction and choice-of-law characterising private international law (henceforth, PIL). PIL consists in the body of rules and principles governing the choice of law to be applied when there are conflicts in the domestic law of different countries related to private legal facts and transactions [11]. Of course, this is relevant whenever private individuals exhibit aspects of extraneousness with respect to a specific domestic system, and these aspects refer to the law of other countries. The issue of legal pluralism and the fundamental mechanisms of conflict of laws was consequently been studied through argumentation and logics [3,6,8,9,2]. The focus was maintained on legal dogmatics or at the level of virtual conflicts between legal systems, each considered as potentially competent to rule the case, or, again at the level of conflict among different interpretive solutions: precisely the kind of conflicts that PIL in fact prevents.

Several problems such as the following ones often arise.

Case 1 (Conflict across Legal Systems and Overriding Mandatory Rules) *PIL principles may require to apply foreign law and such provisions can be in conflict with domestic law. However, there could exist a domestic piece of legislation that is considered mandatory. In this way, mandatory rules prevent and override any other rule, including the possible foreign law they identify, which is ex ante seen as incompatible with the domestic legal system and its fundamental goals.*

Case 2 (Public Policy Exception) *The foreign interpretive argument gives an interpretive result whose effects are contrary to the public policy of the domestic legal system.*

Case 3 (Same Interpretive Canons Conflict) *The same interpretive argument gives opposite interpretive results in the foreign and in the domestic legal systems.*

Hence, several formal methods can be used to model how domestic courts should apply and reason about foreign law.

The cases above in fact show that PIL requires to handle two distinct reasoning processes:

- Conflict-detection and conflict-resolution among legal rules belonging to different legal systems;
- Conflict-detection and conflict-resolution among interpretive arguments used in different legal systems.

As discussed by [3,8,9], such processes lead to different logical solutions. However, if seen at a more abstract level, all these approaches—and, in fact, any logical model for PIL, we claim—are based on some common formal intuitions. In this paper we accordingly develop a new comprehensive computational framework for reasoning about PIL that encompasses the reasoning patterns modeled by [3,8,9]: Section 2 discusses some examples; Section 3 accounts for previous literature; Section 4 presents the intuitions behind the proposed logic (Section 5.1) and the algorithms to compute the extensions of normative theories (Section 5.2).

2. Problems and Examples

Methods for conflict of laws may occur when norms of different systems collide, or when interpretive arguments collide when used in distinct systems. Consider the following two examples.

Example 1 *[Contract Law] "An Italian company and a British one make a contract according to which the Italian company has to deliver certain goods. A clause says that the contract is governed by US law. The English company sues the Italian company for breach of contract. The jurisdiction issue, in both English and Italian laws, has to be decided on the basis of the Brussels Convention (on Jurisdiction and the Enforcement of Judgments in Civil and Commercial Matters), which establishes the jurisdiction of the Italian judge. However, the Italian judge has to apply the law chosen by the parties, i.e., US law, on the basis of the Rome Convention (on the Law Applicable to Contractual Obligations)". [3]*

In this example, it is crucial whether a contract is regulated by Italian or US law: the two legal systems lead to different outcomes. As argued by [3], the Italian law tends to limit liability of the diligent defaulting party, while US law is stricter in this regard: in several cases, if Italian law had to be applied, the diligent defaulting party would not pay for damages. On the contrary, under US law damages have to be paid. Here, we have a clear *conflict of norms*.

Example 2 *[Interpretation in PIL] "A woman, Cameroonian citizen, put forward an Italian court a paternity action with respect to her daughter, also Cameroonian citizen, underage at the time, on the basis of Art. 340 Cameroonian Civil Code and Art. 33 Law no. 218/1995. She alleged that the child was born within a relationship she had with*

an Italian citizen, who initially took care of the girl and provided financial support for her, then refusing to recognise the child. The judicial question is thus the recognition of the legitimate paternity in favour of the girl, whose main legal consequence would be to burden the presumed father with the duty to give her due support in the form of maintenance and education. [. . .]

*Art. 340, Civil Code of Cameroon, states that the judicial declaration of paternity outside marriage can only be done if the suit is filed within the two years that follow the cessation, either of the cohabitation, or of the participation of the alleged father in the support [*entretien*] and education of the child. At a first glance, it appears crucial to properly interpret the term* entretien *for it represents a condition for lawfully advancing the judicial request of paternity. Different interpretations of this term can be offered in Cameroon's law, and may fit differently within the Italian legal system". [8]*

In this second example, once it is made clear which norm has to be applied (Art. 340, Civil Code of Cameroon), we have still a potential conflict to solve because the Italian judge may interpret such a piece of foreign legislation using different interpretive standards: here, we rather have a *conflict of interpretations*.

3. Background

[3] proposed a formal model of the interaction between legal systems based on the so-called modular argumentation, namely, an argumentation system where reasoning in regard to different legal contexts is managed by separate knowledge bases (modules). As expected, the authors assume the existence of different legal systems $\mathrm{LS}_i, \ldots, \mathrm{LS}_z$. Each system LS_i, contains three sets of rules: (a) a set of *choice of jurisdiction rules* $ChJur(\mathrm{LS}_i)$; (b) a set of *choice of competence rules* $ChComp(\mathrm{LS}_i)$; and (c) a set of *choice of law rules* $ChLaw(\mathrm{LS}_i)$.

Since a representation of PIL refers to distinct sets of legal rules, modular argumentation offers itself as an appropriate platform for representing PIL and different national laws as it allows knowledge to be split in separate modules.

Indeed, PIL rules establish, respectively, whether courts of LS_i can decide the case (jurisdiction), what particular court of LS_i can do that (competence), and what set of norms, of LS_i's or of another legal system LS_j, the court should apply (applicable law).

The reasoning mechanism handles such sets of rules. First of all, a court should consider the issue of jurisdiction, thus pointing to a certain system LS_i. Having established jurisdiction for the courts of its legal system LS_i, the court k will have to address competence, i.e., to establish whether k itself, among all courts of LS_i, has the task to decide that case, according to $ChComp(\mathrm{LS}_k)$. Finally, court k should apply $ChLaw(\mathrm{LS}_k)$ in order to establish according to what legal system LS_j (that could possibly be different from LS_i) the case should be decided.

[8,9] proposed instead a Defeasible Logic for reasoning about interpretive arguments or canons. As is well-known, interpretive canons are different doctrinal methods that are employed in legal systems as patterns for constructing arguments aimed at justifying certain interpretations [7]. Examples are the Argument by coherence, according to which a statutory provision should be interpreted in light of the whole statute it is part of, or in light of other statutes it is related to, or Teleological argument, according to which a statutory

provision should be interpreted as applied to a particular case in a way compatible with the purpose that the provision is supposed to achieve

The logical structure of interpretive arguments must be analysed using a rule-based logical system. In particular, interpretation canons are represented by *interpretation rules*, such as the following:

$$s : \mathrm{OBL}\mathsf{I}_s^{\mathrm{LS}_j}(n_2^{\mathrm{LS}_i}, d) \Rightarrow^I \mathsf{I}_c^{\mathrm{LS}_j}(n_1^{\mathrm{LS}_i}, p) \tag{3.1}$$

Rule s states that, if provision n_2 belonging to the legal system LS_i ought to be interpreted in another system LS_j by substantive reasons (I_s) as d, then the interpretive canon to be applied in legal system LS_j for provision n_1 is the interpretation by coherence (I_c), which returns p.

Reasoning about interpretive canons across legal systems thus requires to specify in the formal language to which legal systems legal provisions belong and in which legal system canons are applied. In addition, we need the introduction of meta-rules to reason about interpretation rules; such meta-rules support the derivation of interpretation rules; in other words, the conclusions of meta-rules are interpretation rules, while the the antecedents may include any conditions. Consider, for instance, the following meta-rule:

$$r : (\mathrm{OBL}\mathsf{I}_t^{\mathrm{LS}_i}(n_1^{\mathrm{LS}_i}, p), a \Rightarrow_C (s : \mathrm{OBL}\mathsf{I}_s^{\mathrm{LS}_j}(n_2^{\mathrm{LS}_i}, d) \Rightarrow^I \mathsf{I}_c^{\mathrm{LS}_j}(n_1^{\mathrm{LS}_i}, p)))$$

Meta-rule r states that, if (a) it is obligatory the teleological interpretation (I_t) in legal system LS_i of legal provision n_1 belonging to that system and returning p, and (b) a holds, then the interpretive canon to be applied in legal system LS_j for n_1 is the interpretation by coherence, which returns p as well, but which is conditioned in LS_j by the fact that n_2 in this last system is interpreted by substantive reasons as d. In other words, r allows for importing interpretive results from LS_i into LS_j in regard to the legal provision n_1 in LS_i which can be applied in LS_j.

4. Logical Intuition

If we abstract from the peculiarities of [3,8,9], both approaches share a number of general intuitions. On account of the discussion of previous sections, we argue that any formal system for PIL is expected

- to have a **formal language**
  - able to encode the *existence of different legal systems* $\mathrm{LS}_i, \ldots, \mathrm{LS}_z$;
  - with *propositional expressions* representing any piece of information which is parametrised by legal systems; for example, we may write a^{LS_i} to mean that a (an obligation, a contract, an interpretive outcome...) holds in the legal system LS_i;
- to have a **reasoning mechanism** that allows for concluding that, if something holds in some legal system, then something else holds in this or another legal system, or that allows for importing in a given system any piece of information holding in another system; for example, the reasoning mechanism should be based on handling
  - *rules* such as

$$r : a^{\mathrm{LS}_i} \Rightarrow^{\mathrm{LS}_j} b^{\mathrm{LS}_j}$$

which may represent, e.g., a norm r of LS_j (this is represented by the fact that the arrow is labelled accordingly) stating that if a holds in another legal system LS_i, then b holds in LS_j;

* *meta-rules* such as

$$s : p^{\mathrm{LS}_k} \Rightarrow (r : a^{\mathrm{LS}_i} \Rightarrow^{\mathrm{LS}_j} b^{\mathrm{LS}_j})$$

which are meant to reason about, and across legal systems (for this reason, the arrow of s in not labelled by any legal system); meta-rule s means that, if p holds in the legal system LS_k, then we can use norm r in LS_j;

* since legal systems can be incompatible, different rules and meta-rules can collide, so we need to establish a priority orderings.

The above list shows in a nutshell the basic requirements for developing a general computational framework for reasoning about PIL. The next section will present the details of it.

5. The Framework

The computational framework for reasoning about PIL we are proposing is based on Defeasible Logic [1], which is is a simple and efficient rule-based non-monotonic formalism that proved to be suitable for the logical modelling of different application areas, including the law (see [5,4]). The logic is extended as informally discussed in the previous section. A first result was offered [10]. Here we extend the machinery to handle more legal systems and all requirements mentioned in Section 4.

5.1. Logic

Let PROP be a set of propositional atoms, $\mathbf{LS} = \{\mathrm{LS}_i, \ldots, \mathrm{LS}_z\}$ a finite set of legal systems, and Lab be a set of arbitrary labels (the names of the rules). $\mathrm{BLit} = \mathrm{PROP} \cup \{\neg l \mid l \in \mathrm{PROP}\}$ is the set of *basic literals*. The *complement* of a literal l is denoted by $\sim l$: if l is a positive literal p then $\sim l$ is $\neg p$, and if l is a negative literal $\neg p$ then $\sim l$ is p. Hence, $\mathrm{Lit} = \{l^{\mathrm{LS}} | l \in \mathrm{BLit},\ \mathrm{LS} \in \mathbf{LS}\}$ is the set of *literals*.

The set of rules is made of two sets: standard rules R^S, and meta-rules R^M. A *standard rule* $\beta \in R^S$ is an expression of the type '$\beta : A(\beta) \hookrightarrow^{\mathrm{LS}} C(\beta)$', and consists of: (i) the unique name $\beta \in \mathrm{Lab}$, (ii) the *antecedent* $A(\beta) \subseteq \mathrm{Lit}$, (iii) an *arrow* $\hookrightarrow \in \{\rightarrow, \Rightarrow, \rightsquigarrow\}$ denoting, respectively, a strict rule, a defeasible rule and a defeater, (iv) a legal system LS, (v) its *consequent* $C(\beta) \in \mathrm{Lit}$, a single literal. Hence, the statement "Minors are in Italy persons under the age of 18 years" is formulated through a strict rule (as there is no exception to it), whilst "EU citizens may visit the USA without green card" is instead formalised through a defeasible rule as "During pandemic travels to USA might be prohibited" is a defeater representing an exception to it.

A meta rule is a slightly different concept than a standard rule: (i) standard rules can appear in its antecedent, and (ii) the conclusion itself can be a standard rule. Accordingly, a *meta rule* $\beta \in R^M$ is an expression of the type '$\beta : A(\beta) \hookrightarrow C(\beta)$', and consists of: (i) a unique name $\beta \in \mathrm{Lab}$, (ii) the antecedent $A(\beta)$ is now a finite subset of $\mathrm{Lit} \cup R^S$, (iii) the *arrow* $\hookrightarrow$ with the same meaning as for standard rules, and (iv) its *consequent* $C(\beta) \in \mathrm{Lit} \cup R^S$, that is either a single literal or a standard rule (meta-rules can be used to derive standard rules).

A *defeasible meta-theory* (or simply theory) D is a tuple $(\mathrm{F}, R, >)$, where $R = R^{stand} \cup R^{meta}$ such that $R^{stand} \subseteq R^S$ and $R^{meta} \subseteq R^M$. F is the set of facts, indisputable statements that are considered to be always true, and which can be seen as the inputs for a case. As usual in Defeasible Logic, rules in R can be of three types: *strict rules*, *defeasible rules*, or *defeaters*. Finally, we have the *superiority* $>$ among rules, which is binary and irreflexive, and is used to solve conflicts. The notation $\beta > \gamma$ means $(\beta, \gamma) \in >$.

Some abbreviations. The set of strict rules in R is R_s, and the set of strict and defeasible rules is R_{sd}. $R[X]$ is the rule set with head $X \in \{\mathrm{Lit} \cup R^S\}$. R^{LS} is the set of rules whose arrow is labelled by LS. A *conclusion of D* is either a *tagged literal* or a *tagged label* (for a standard rule), and can have one of the following forms with the standard meanings in Defeasible Logic:

- $\pm\Delta l$ means that $l \in \mathrm{Lit}$ is *definitely provable* (resp. *refuted*, or *non provable*) in D, i.e. there is a definite proof for l (resp. a definite proof does not exist);
- $\pm\Delta^{meta}\alpha$, $\alpha \in R^{stand}$, with same meaning as above;
- $\pm\partial l$ means that l is *defeasibly provable* (resp. *refuted*) in D;
- $\pm\partial^{meta}\alpha$, $\alpha \in R^{stand}$, with the same meaning as above.

The definition of proof is also the standard in DL. Given a defeasible meta-theory D, a proof P of length n in D is a finite sequence $P(1), P(2), \ldots, P(n)$ of tagged formulas of the type $+\Delta X$, $-\Delta X$, $+\partial X$, $-\partial X$, where the proof conditions defined in the rest of this section hold. $P(1..n)$ denotes the first n steps of P.

Derivations are based on the notions of a rule being *applicable* or *discarded*.

Definition 1 (Applicability) *Given a defeasible meta-theory* $D = (\mathrm{F}, R, >)$, $R = R^{stand} \cup R^{meta}$, *a rule* $\beta \in R$ *is #-applicable,* $\# \in \{\Delta, \partial\}$, *at* $P(n+1)$ *iff*

1. $\forall l \in \mathrm{Lit} \cap A(\beta).\ +\#l \in P(1..n)$,
2. $\forall \alpha \in R^S \cap A(\beta)$ *either (a)* $\alpha \in R^{stand}$, *or (b)* $+\#^{meta}\alpha \in P(1..n)$.

Definition 2 (Discardability) *Given a defeasible meta-theory* $D = (\mathrm{F}, R, >)$, $R = R^{stand} \cup R^{meta}$, *a rule* $\beta \in R$ *is #-discarded,* $\# \in \{\Delta, \partial\}$, *at* $P(n+1)$ *iff*

1. $\exists l \in \mathrm{Lit} \cap A(\beta).\ -\#l \in P(1..n)$, *or*
2. $\exists \alpha \in R^S \cap A(\beta)$ *such that (a)* $\alpha \notin R^{stand}$ *and (b)* $-\#^{meta}\alpha \in P(1..n)$

When β is a meta-rule and α is not in R^{stand} (hence α is the conclusion of a meta-rule), then β will stay dormant until a decision on α (of being proved/refuted) is made. The following example is to get acquainted with the concepts introduced.

Example 3 *Let* $D = (F = \{a, b\}, R, \emptyset)$ *be a theory such that*

$$R = \{\alpha : a \Rightarrow \beta; \quad \beta : b, \beta \Rightarrow \zeta; \quad \gamma : c \Rightarrow^{\mathrm{LS}} d; \quad \varphi : \psi \Rightarrow d\}.$$

Here, both α *and* β *are applicable (we will see right below how to prove* $+\partial^{meta}\beta$*), whilst* γ *and* φ *are discarded as we cannot prove* $+\partial c$ *nor* $\partial^{meta}\psi$.

The language of the logic is designed in such a way that all proof tags for literals are the standard ones for Defeasible Logic, so they are omitted for space reasons [1].

We are finally ready to propose the proof tags to prove (standard) rules.

$+\Delta^{meta}\alpha$: If $P(n+1) = +\Delta^{meta}\alpha$ then
(1) $\alpha \in R^{stand}$, or (2) $\exists \beta \in R_s^{meta}[\alpha]$ s.t. β is Δ-applicable.

A standard rule is strictly proven if either (1) such a rule is in the initial set of standard rules, or (2) there exists an applicale, strict meta-rule for it. Since defeasible rule provability requires to detect and solve conflicts between meta-rules, we need to clarify the meaning of $\sim\alpha$, where α is a standard rule.

Definition 3 (Rule complement) *Let α be any rule. Then*

$$\begin{array}{ll}
\beta = \alpha : A(\alpha) \Rightarrow^{\mathrm{LS}} C(\alpha) & \sim\beta = \{\neg\alpha,\ \gamma : A(\alpha) \hookrightarrow^{\mathrm{LS}'} \sim C(\alpha), \gamma \in R_{\mathrm{sd}}\} \\
\beta = \alpha : A(\alpha) \rightarrow^{\mathrm{LS}} C(\alpha) & \sim\beta = \{\neg\alpha,\ \gamma : A(\alpha) \rightarrow^{\mathrm{LS}'} \sim C(\alpha)\} \\
\beta = \alpha : A(\alpha) \rightsquigarrow^{\mathrm{LS}} C(\alpha) & \sim\beta = \{\neg\alpha,\ \gamma : A(\alpha) \hookrightarrow^{\mathrm{LS}'} \sim C(\alpha), \gamma \in R_{\mathrm{sd}}\} \\
\beta = \neg(\alpha : A(\alpha) \hookrightarrow^{\mathrm{LS}} C(\alpha)) & \sim\beta = \{\alpha\}.
\end{array}$$

$+\partial^{meta}\alpha$: If $P(n+1) = +\partial^{meta}\alpha$ then
(1) $+\Delta^{meta}\alpha \in P(1..n)$, or
(2) (1) $-\Delta^{meta}\sim\alpha \in P(1..n)$, and
(2) $\exists\beta \in R^{meta}_{\mathrm{sd}}[(\alpha : a_1, \ldots, a_n \hookrightarrow c)]$ s.t.
(3) β is ∂-meta-applicable, and
(4) $\forall\gamma \in R^{meta}[\sim(\zeta : a_1, \ldots, a_n \hookrightarrow c)]$, then either
(1) γ is ∂-meta-discarded, or
(2) $\exists\varepsilon \in R^{meta}[(\chi : a_1, \ldots, a_n \hookrightarrow c)]$ s.t.
(1) $\chi \in \{\alpha, \zeta\}$, (2) ε is ∂-meta-applicable, and (3) $\varepsilon > \gamma$.

A standard rule α is defeasibly proven if it has previously strictly proven (1), or (2.1) the opposite is not strictly proven and (2.2-2.3) there exists an applicable (defeasible or strict) meta-rule β such that every meta-rule γ for $\sim\zeta$ ($A(\alpha) = A(\zeta)$ and $C(\alpha) = C(\zeta)$) either (2.4.1) γ is discarded, or defeated (2.4.2.3) by (2.4.2.1-2.4.2.2) an applicable meta-rule for the same conclusion c. Note that in Condition 2.3 we do not impose that $\alpha \equiv \zeta$, whilst for γ-rules we do impose that the label of the rule in $C(\gamma)$ is either α or ζ.

The condition for $-\partial^{meta}$ is omitted for space reasons, since it is simply obtained from the positive case. Given a defeasible meta-theory D, we define the set of positive and negative conclusions of D as its *meta-extension*:

$$E(D) = (+\Delta, -\Delta, +\Delta^{meta}, -\Delta^{meta}, +\partial, -\partial, +\partial^{meta}, -\partial^{meta}),$$

where $\pm\# = \{l \mid l \text{ appears in } D \text{ and } D \vdash \pm\#l\}$ and $\pm\#^{meta} = \{\alpha \in R^S \mid \alpha \text{ appears as consequent of a meta-rule } \beta \text{ and } D \vdash \pm\#^{meta}\alpha\}$, $\# \in \{\Delta, \partial\}$.

Example 4 *Let $D = (\mathrm{F} = \{a, c, d, g\}, R, >= \{(\beta, \gamma)(\zeta, \eta)\})$ be a theory where*

$$R^{stand} = \{\alpha : a \Rightarrow^{\mathrm{LS}_1} b, \qquad \zeta : g \Rightarrow^{\mathrm{LS}_2} \sim b\},$$

$$R^{meta} = \{\beta : c, (\alpha : a \Rightarrow^{\mathrm{LS}_1} b) \Rightarrow (\eta : d \Rightarrow^{\mathrm{LS}_3} b), \qquad \gamma : d \Rightarrow \sim(\chi : d \Rightarrow^{\mathrm{LS}_4} b)\}.$$

As a, c, d and g are facts, we strictly and defeasibly prove all of them. Hence, α, ζ, β and γ are all ∂-applicable. As before, $\alpha \in R^{stand}$, thus $D \vdash +\Delta^{meta}\alpha$ and $D \vdash +\partial c$ make β being ∂-applicable as well. As $\beta > \gamma$, we conclude that $D \vdash +\partial^{meta}\eta$, but we prove also $D \vdash -\partial^{meta}\chi$. Again, d being a fact makes η to be ∂-applicable. ζ has been dormant so far, but it can now be confronted with η: since η is weaker than ζ, then $D \vdash +\partial\sim b$ (and naturally $D \vdash -\partial b$).

5.2. Algorithms

The algorithms presented in this section compute the meta-extension of a defeasible meta-theory. The main idea being to compute, at each iteration step, a *simpler* theory than the one at the previous step. By simpler, we mean that, by proving and disproving literals and standard rules, we can progressively simplify the rules of the theory itself.

Let us consider the case of meta-rules. A meta-rule is applicable when each standard rule in its antecedent is either in the initial set of rules (i.e., in R^{stand}), or has been proved later on during the computation and then added to the set of standard rules. This is the reason for the support sets at Lines 1 and 2: R_{appl} is the rule set of the initial standard rules, $R^{\alpha C}$ is the set of standard rules which are not in the initial set but are instead conclusions of meta-rules. As rules in $R^{\alpha C}$ are proved/disproved during the algorithms' execution, both these sets are updated.

At Line 3, we populate the Herbrand Base (HB), which consists of all literals that appear in the antecedent, or as a conclusion of a rule. As literals not in the Herbrand base do not have any standard rule supporting them, such literals are already disproved (Line 4). For every literal in HB, we create the support set of the rules supporting that particular conclusion (Line 6), and we initialise the relative set used later on to manage conflicts and team defeater (Line 7).

We need to do the same for those labels for standard rules that are conclusions of a meta-rule. First, if a label for standard rule is neither in the initial set of standard rules, nor a conclusion of a meta-rules, then such a rule is disproved (Line 8). We assume such sets to have empty intersection, as previously motivated. Second, the following loop at Lines 17–20 initialises three support sets: $R[\alpha]$ contains the meta-rules whose conclusion is α, $R[\alpha]_{opp}$ contains the meta-rules attacking α (γ-like rules in $\pm\partial^{meta}$), while $R[\alpha]_{supp}$ contains the meta-rules supporting α (ε-like rules in $\pm\partial^{meta}$).

The following **for** loop takes care of the factual literals, as they are proved without any further computation. We assume the set of facts to be consistent. Analogously, loop at Lines 17–20 does the same for rules in the initial set of standard rules that may appear in the antecedent of meta-rules.

The algorithm now enters the main cycle (**Repeat-Until**, Lines 21–40). For every literal l in HB (Lines 23–29), we first verify whether there is a rule supporting it, and, if not, we refute l (Line 24). Otherwise, if there exists an applicable rule β supporting it (**if** at Line 25), we update the set of *defeated* rules supporting the opposite conclusion $R[\sim l]_{inf\,d}$ (Line 26). Given that $R[\sim l]$ contains the γ rules supporting $\sim l$, and given that we have just verified that β for l is applicable, we store in $R[\sim l]_{inf\,d}$ all those γs defeated by β. The next step is to check whether there actually exists any rule supporting $\sim l$ stronger than β: if not, $\sim l$ can be refuted (Line 27).

The idea behind the **if** at Lines 28–29 is the following: if $D \vdash +\partial l$, eventually the **repeat-until** cycle will have added to $R[\sim l]_{inf\,d}$ enough rules to defeat all (applicable) supports for $\sim l$. We thus invoke **Prove** on l, and **Refute** on $\sim l$.

Similarly, when we prove a rule instead of a literal, but we now use $R[\alpha]_{opp}$ and $R[\alpha]_{supp}$ in a slightly different way than $R[l]_{inf\,d}$, to reflect the differences between $+\partial$ and $+\partial^{meta}$. Every time, a meta-rule β for α is applicable (**if** at Lines 34–38), we remove from $R[\alpha]_{opp}$ all the γs defeated by β itself (Line 35). If now there are enough applicable ε rules supporting α (**if** check at Line 36), then: (i) we prove α, and (ii) we refute all ζ rules conclusion of γ rules in $R[\alpha]_{opp}$.

Input: Defeasible meta-theory $D = (\mathrm{F}, R, >)$, $R = R^{stand} \cup R^{meta}$
Output: The defeasible meta-extension $E(D)$ of D

$\pm\partial \leftarrow \emptyset$; $\pm\partial^{meta} \leftarrow \emptyset$; $R_{appl} \leftarrow R^{stand}$
$R^{\alpha C} \leftarrow \{\alpha \in R^S \mid \exists \beta \in R^{meta}.\, C(\beta) = \alpha\}$
$HB = \{l \in \mathrm{Lit} \mid \exists \beta \in R^{stand}.\, l \in A(\beta) \cup C(\beta)\} \cup \{l \in \mathrm{Lit} \mid \exists \beta \in R^{meta}.\exists \alpha \in R^S\ (\alpha \in A(\beta) \cup C(\beta)) \wedge (l \in A(\alpha) \cup C(\alpha))\}$
for $l \in \mathrm{Lit} \wedge l \notin HB$ **do** $-\partial \leftarrow -\partial \cup \{l\}$;
for $l \in HB$ **do**
  $R[l] = \{\beta \in R^S \mid C(\beta) = l \wedge (\beta \in R^{stand} \vee \exists \gamma \in R^{meta}.\, \beta \in C(\gamma))\}$
  $R[l]_{infd} \leftarrow \emptyset$
for $\alpha \notin R^{stand} \cup R^{\alpha C}$ **do** $-\partial^{meta} \leftarrow -\partial^{meta} \cup \{\alpha\}$;
for $(\alpha : A(\alpha) \hookrightarrow C(\alpha)) \in R^{\alpha C}$ **do**
  $R[\alpha] \leftarrow \{\beta \in R^{meta} \mid \alpha = C(\beta)\}$
  $R[\alpha]_{opp} \leftarrow \{\gamma \in R^{meta} \mid C(\gamma) = \sim(\zeta : A(\alpha) \hookrightarrow C(\alpha))\}$
  $R[\alpha]_{supp} \leftarrow \{\varepsilon \in R^{meta} \mid (C(\varepsilon) = (\chi : A(\alpha) \hookrightarrow C(\alpha))) \wedge (\exists \gamma \in R[\alpha]_{opp}.\, \varepsilon > \gamma) \wedge (\chi = \alpha \vee (\exists \gamma \in R[\alpha]_{opp}.C(\gamma) = \sim(\zeta : A(\alpha) \hookrightarrow C(\alpha)) \wedge \chi = \zeta))\}$
for $l \in \mathrm{F}$ **do**
  $+\partial \leftarrow +\partial \cup \{l\}$
  $R \leftarrow \{A(\beta) \setminus \{l\} \hookrightarrow C(\beta) \mid \beta \in R\} \setminus \{\beta \in R \mid \sim l \in A(\beta)\}$
  $> \leftarrow > \setminus \{(\beta, \gamma), (\gamma, \beta) \in > \mid \sim l \in A(\beta)\}$
for $\alpha \in R^{stand}$ **do**
  $+\partial^{meta} \leftarrow +\partial^{meta} \cup \{\alpha\}$
  $R^{meta} \leftarrow \{A(\beta) \setminus \{\alpha\} \hookrightarrow C(\beta) \mid \beta \in R^{meta}\} \setminus \{\gamma \in R^{meta} \mid \{\sim\alpha\} \in A(\gamma)\}$
  $> \leftarrow > \setminus \{(\beta, \gamma), (\gamma, \beta) \in > \mid \{\sim\alpha\} \in A(\beta)\}$
repeat
  $\partial^{\pm} \leftarrow \emptyset$
  for $l \in HB$ **do**
    if $R[l] = \emptyset$ **then** Refute(l);
    if $\exists \beta \in R[l].\, A(\beta) = \emptyset$ **then**
      $R[\sim l]_{infd} \leftarrow R[\sim l]_{infd} \cup \{\gamma \in R[\sim l] \mid \beta > \gamma\}$
      if $\{\gamma \in R[\sim l] \mid \gamma > \beta\} = \emptyset$ **then** Refute($\sim l$);
      if $R[\sim l] \setminus R[\sim l]_{infd} = \emptyset$ **then**
        Prove(l); Refute($\sim l$)
  $\pm\partial \leftarrow \pm\partial \cup \partial^{\pm}$
  $\pm\partial_{meta} \leftarrow \emptyset$
  for $(\alpha : A(\alpha) \hookrightarrow C(\alpha)) \in R^{\alpha C}$ **do**
    if $R[\alpha] = \emptyset$ **then** Refute(α);
    if $\exists \beta \in R[\alpha].\, A(\beta) = \emptyset$ **then**
      $R[\alpha]_{opp} \leftarrow R[\alpha]_{opp} \setminus \{\gamma \in R^{meta} \mid \beta > \gamma\}$
      if $(R[\alpha]_{opp} \setminus \{\gamma \in R[\alpha]_{opp} \mid \varepsilon \in R[\alpha]_{supp} \wedge A(\varepsilon) = \emptyset \wedge \varepsilon > \gamma\}) = \emptyset$ **then**
        Prove(α)
        for $\gamma \in R[\alpha]_{opp}.\, C(\gamma) = \sim(\zeta)$ **do** Refute($\sim\zeta$);
  $\pm\partial^{meta} \leftarrow \pm\partial^{meta} \cup \partial^{\pm}_{meta}$
until $\partial^{+} = \emptyset$ *and* $\partial^{-} = \emptyset$ *and* $\partial^{+}_{meta} = \emptyset$ *and* $\partial^{-}_{meta} = \emptyset$;
return $E(D) = (+\partial, -\partial, +\partial^{meta}, -\partial^{meta})$

Algorithm 1. Existence

Procedures Prove and Refute are the same as in [10] and are invoked when a literal or a standard rule is proved/refuted.

In order to discuss termination and computational complexity, we start by defining the *size* of a meta-theory D as $\Sigma(D)$ to be the number of the occurrences of literals plus the number of occurrences of rules plus 1 for every tuple in the superiority relation. Thus, the theory $D = (\mathrm{F}, R, >)$ such that $\mathrm{F} = \{a, b, c\}$, $R^{stand} = \{(\alpha : a \Rightarrow^{\mathrm{LS}_1} d), (\beta : b \Rightarrow^{\mathrm{LS}_2} \sim d)\}$, $R^{meta} = \{(\gamma : c \Rightarrow (\zeta : a \Rightarrow^{\mathrm{LS}_3} d))\}$, $>= \{(\zeta, \beta)\}$, has size $3 + 6 + 5 + 1 = 15$.

Note that, by implementing hash tables with pointers to rules where a given literal occurs, each rule can be accessed in constant time. We also implement hash tables for the tuples of the superiority relation where a given rule appears as either of the two element, and even those can be accessed in constant time.

Theorem 1 *Algorithm 1 Existence terminates and its complexity is* $O(\Sigma^2)$.

6. Summary

This paper presented a new computational framework for reasoning about PIL. The system in abstracts from the peculiarities of approaches such as [3,8,9]. The formal language assumes the existence existence of different legal systems and of propositional expressions such as a^{LS_i} to mean that a holds in the legal system LS_i. Also, the reasoning mechanism, through meta-rules, allows for concluding that, if something holds in some legal system, then something else holds in this or another legal system, or that allows for importing in a given system any piece of information holding in another system. Finally, since legal systems can be incompatible, different rules and meta-rules can collide, so we make use of priority orderings among rules as in standard Defeasible Logic. The resulting system simply extends [10] and preserves the same nice computational properties.

References

[1] G. Antoniou, D. Billington, G. Governatori, and M. J. Maher. Representation results for defeasible logic. *ACM Trans. Comput. Log.*, 2(2):255–287, 2001.

[2] M. Cristani, F. Olivieri, and A. Rotolo. Changes to temporary norms. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law, ICAIL 2017, London, United Kingdom, June 12-16, 2017*, pages 39–48, 2017.

[3] P. M. Dung and G. Sartor. The modular logic of private international law. *Artif. Intell. Law*, 19(2-3):233–261, 2011.

[4] G. Governatori, F. Olivieri, A. Rotolo, and S. Scannapieco. Computing strong and weak permissions in defeasible logic. *J. Philos. Log.*, 42(6):799–829, 2013.

[5] G. Governatori and A. Rotolo. Changing legal systems: legal abrogations and annulments in defeasible logic. *Log. J. IGPL*, 18(1):157–194, 2010.

[6] J. Hage. Logical tools for legal pluralism. *Maastricht European Private Law Institute Working Paper 7*, 2015.

[7] D. MacCormick and R. Summers, editors. *Interpreting Statutes: A Comparative Study*. Ashgate, 1991.

[8] A. Malerba, A. Rotolo, and G. Governatori. Interpretation across legal systems. In *Proc. JURIX 2016*. IOS Press, 2016.

[9] A. Malerba, A. Rotolo, and G. Governatori. A logic for the interpretation of private international law. In *New Developments in Legal Reasoning and Logic*. Springer, Dordrecht, 2021.

[10] F. Olivieri, G. Governatori, M. Cristani, and A. Sattar. Computing defeasible meta-logic. In *Proc. JELIA 2021*. Springer, 2021.

[11] P. Stone. *EU Private International Law*. Elgar European law. Edward Elgar, 2014.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210335

Explaining Factor Ascription

Jack Mumford, Katie Atkinson and Trevor Bench-Capon
Department of Computer Science,
University of Liverpool, UK

Abstract. Explanation and justification of legal decisions has become a highly relevant topic in light of the explosion of interest in the use of machine learning (ML) approaches to predict legal decisions. Current suggestions are to use the established factor based explanations developed in AI and Law as the basis for explaining such programs. We, however, identify factor ascription as an important aspect of explanation of case outcomes not currently covered, and argue that explanations must also include this aspect. Finally, we outline our proposal for a hybrid system approach that combines ML and Abstract Dialectical Framework (ADF) layers to engender an explainable process.

Keywords. reasoning with cases, explanation, factor ascription.

1. Introduction

In recent years there has been an explosion of interest in the use of machine learning (ML) to predict legal decisions (e.g. [14]). A major weaknesses of these approaches is, however, that they are unable to explain their reasoning in an acceptable manner. Traditional explanations of ML such as listing or highlighting the most influential words in the texts have been shown to be unhelpful [9] because they are difficult to relate to the relevant law. Moreover there are good reasons why any such explanation would be inappropriate in a legal context [8]. The right to explanation means that the explanation must be capable of persuading the losing party, and providing a justification which can withstand an appeal. It need not be an explanation of how the decision was in fact reached, but must explain why the decision represents the proper application of the law[1].

In order to explain the predicted outcomes in appropriate terms, researchers have turned to the extensive body of work on explanation developed in AI and Law [5]. In particular the type of explanation advocated has been based on the use of factors, as developed in CATO [2], e.g [9], [16] and [6]. Specifically, using the set of argumentation schemes designed to capture the reasoning of CATO from [17] is advocated in [16].

[1]A popular caricature of legal realism (e.g. [12]) says that the law is what the judge had for breakfast. This may indeed explain *how* the judge reached the decision, but the opinion must contain an explanation of *why* this decision is a justifiable application of the law.

Factors as introduced in [2] are stereotypical patterns of facts which are legally significant in that they provide a reason to find for one side or the other. They represent a generalisation from the facts of particular cases so that they can be applicable to a number of cases: for example *plaintiff pursuing livelihood* generalises the facts of several property law cases involving hunting, shooting and fishing. In [2], the factors are organised into a factor hierarchy with issues at the upper levels, abstract factors in the middle layers and the (base level) factors as the leaves. CATO organises its explanation into a series of issues. The resolution of issues can be explained in terms of the factors which provide reasons for the winning party or the burden of proof for that issue. Where there are factors for both the plaintiff and the defendant, the reasons preferred are justified in terms of a precedent case exhibiting this preference. Sometimes factors may be used to cancel or substitute for other factors as described in [17]. Once the issues have been resolved, the decision follows logically according to a logical model of issues [3], which also serves to associate the factors with particular issues.

Explanation now begins with a summary in terms of the issues. In CATO's domain, US Trade Secrets Law, the plaintiff must establish that the information is a trade secret and that it was misappropriated. Misappropriation requires either the use of improper means or breach of confidence. Thus in *Mason v. Jack Daniels* the explanation begins with: *Plaintiff should win. Plaintiff's information is a trade secret, a confidential relationship existed between plaintiff and defendant, and defendant acquired plaintiff's information through improper means.*

Each issue is then explained in terms of the factors and the relevant precedents. For example, the factors relating to confidential relationship were F1 (*DisclosureInNegotiations*) and F21 (*KnewInfoConfidential*). That the latter is a stronger reason than the former was established in *Forest Labs, Inc. v. Formulations, Inc.* So the explanation of the issue is: *A confidential relationship exists because although the information was disclosed in negotiations, the defendant knew that the information was confidential (Forest Labs, Inc. v. Formulations, Inc).*

Explanations using argument schemes in [16] are different: they produce a three layer tree of argument schemes. The top layer cites the most on point precedent, the second layer attacks this argument with distinctions and counter examples, while the third rebuts the counter examples through a series of transformations and rebuts the distinctions by cancellation and substitution. No use is made of issues in [16], but [6] suggests that the schemes should be applied at the issue level rather that at the case level. Explanation using factors provides a good explanation of cases like *Mason*, but for other cases they are less satisfactory.

2. Explaining Factor Ascription

Consider the Trade Secrets case of *Arco Industries Corp. v. Chemcast Corp.* In that case we have three factors: F10 (*InfoDisclosedToOutsiders*), F16 (*InfoReverseEngineerable*) and F20 (*InfoKnownToCompetitors*): all favour the defendant and establish that the information was not a trade secret. The explanation is thus: *the defendant should win because the information is not a trade secret. It had been disclosed to outsiders, was known to competitors and was reverse engineerable.*

Will this satisfy the plaintiff? Arco had in fact argued that the information *was* a trade secret because it was covered by a patent. They are therefore essentially arguing that two pro-plaintiff factors, F15 (UniqueProduct) and F18 (IdenticalProduct) are present. These factors might have been sufficient to establish that the information was indeed a trade secret and that the information had been used. If so, the plaintiff may well have established F21 (*KnewInfoConfidential*) and thus won the case. So the explanation needs to cover the reasons why the arguments based on these factors were rejected. The decision in fact includes a detailed discussion of the patent specification and the product (a grommet) produced by the defendant. It concludes: *The specifications describe the "recess" as an indentation below the planar surface of the grommet which in turn lies below the peripheral sealing ridge. The accused grommet does not have such a recess.*

This is what excludes F15 and F18 and makes discussion of confidentiality unnecessary. It is the general notion of a grommet that is known in the industry and it appears that the unique feature of Arco's grommet was not used by Chemcast. The explanation needed by the plaintiff is not how the factors ascribed satisfy the required issues, but why the claim that other factors were present was rejected. Crucially, the explanation in terms of factors fails to answer the plaintiff's question: *why was the information not protected by the patent?*

In Arco it was the absence of factors that needed to be explained, but sometimes the presence of a factor needs to be explained. Consider *A. H. Emery Co. v. Marcan Products Corporation*. In that case the information had been learned by the defendants while they were employed by the plaintiff. The defendants had not signed any non disclosure agreements, and so they denied that they had breached a confidential relationship. Nonetheless F21 (KnewInfoConfidential) was held to apply, and so a confidential relationship was held to exist. The point was that the information had been acquired while the defendants were employed by the plaintiff and they knew the information to be confidential and at the time the information was acquired they owed a duty of fidelity to their employer.

3. Discussion of Challenges

Reasoning with legal cases is a two stage process [7]: first factors are ascribed and then the balance of factors is determined to reach a decision. Both aspects require explanation. *Arco* and *Emery* show the need to be able to offer explanations not only of the balance of competing factors, but also the presence and absence of particular factors. This could be delivered by extending the dialogue of [6] to ask *WHY*? of any factor used to explain an issue and *WHY NOT Fn?* of any factor not mentioned in the explanation of an issue.

Despite the attention in AI and Law paid to explaining precedential reasoning, there has been little or no work on explaining why the factors are present or absent. This is because most research since HYPO [18] has taken the factors as given. HYPO's dimensions give a clue as to how we might explain certain factor ascriptions. The ranges on these dimensions in which factors are applicable are defined in precedents [7]. So we can explain the ascription of such a factor in terms of these precedents. For example if we have a precedent (*PrecL*) establishing that

an absence of 15 months is a *long stay*, we can explain the ascription of this factor to a new case (CaseN) with an absence of 17 months by saying *long stay applies to CaseN because the absence was greater than 15 months (PrecL).*

This kind of explanation is promising for factors which can be seen as ranges on well ordered dimensions, but does not seem applicable to the kind of detailed consideration of very particular facts that we saw in *Arco* and *Emery*. Such cases may involve analogy [4], or some kind of common sense ontology. This suggests that the key role for ML is not the prediction of outcomes, but the identification of the factors as in [3], [19] and [9].

How does the issue of explaining ascription relate to ML approaches? If we follow [9] we must accept that, if the explanation is to be given in acceptably legal terms, the ML system will need to learn to ascribe factors as well as predict outcomes. Although the standard ML explanations of outcomes are unsatisfactory, there is a considerable gap between facts and outcomes, requiring reasoning through factors and issues. There is no such conceptual gap between facts and factors, and so it may be that the explanation of the ascription of factors using ML is more satisfactory. This is something that requires empirical investigation.

4. Next Steps

Our approach to producing explainable case predictions is to separate the process within a hybrid system in accordance with the two stages outlined in the previous section. The first stage, factor ascription, will be addressed via ML natural language processing (NLP). The second stage, reaching a decision, will be addressed via the balancing of factors within a pre-determined non-cyclic Abstract Dialectical Framework (ADF) [10] that has been derived with expert knowledge to capture the factor-based reasoning of a legal domain, as demonstrated in [1]. If we accept the argument of [15] then domain expertise is of paramount importance when establishing an appropriate ADF, rendering data-driven approaches less effective at the level of factors and above. As such, our hybrid system is initially poised to only adjust the architecture of the NLP layer; the ADF layer will not be changed from its initial state as rendered by expert judgement.

For the first stage of the process, our intention is to use a state-of-the-art Hierarchical BERT model, similar to the approach taken in [13] only not used for determining the case outcome, but rather for factor ascription. We propose to use a Hierarchical BERT model due to the proffered combination of impressive classification performance and the sentence-level attention weights that could sufficiently express the relevant facts that explain a given factor's ascription or non-ascription. The model takes a natural language description of a given case as input and outputs a binary classification of 'ascribed' or 'not ascribed' for each base-level factor in the ADF. The second stage of the process will use the expert-derived ADF to produce a decision via the reasoning steps following from the base-level factor ascription input from the NLP layer.

The data-driven learning phase will not amend the ADF, as previously stated, but still passes errors through the levels in the framework down to the base-level factors for use with the NLP classification task for factor ascription. The full de-

scription of the algorithms we define for error propagation through the ADF will be set out in future work. However, from a high level perspective, the algorithms will function by creating a graphical scaffold of the ADF in which each node is a linearly separable function where children nodes are only capable of attack (that is, each child node implies the contradiction of the parent node), in order to facilitate computationally tractable error propagation. Our initial attention will be on legal domains represented by ADFs with Boolean acceptance conditions. Given the discontinuous Heaviside step function that governs the Boolean acceptance of any given node, backpropagation is not appropriate for error propagation in general. Instead, when a case is input to the hybrid system that results in the wrong decision, errors will be propagated backwards through the levels of the graphical scaffold that give each factor a tuple of weights (*ascribed, not ascribed*), where the value of *ascribed* (alternatively *not ascribed*) is scaled by the product of its parent's value for *not ascribed* (alternatively *ascribed*) and the proportion of combinations in which the node is ascribed (alternatively not ascribed) that would cause the parent node to not be ascribed. This iterative process down through the levels of the scaffolding will begin with the root node representing the decision, which will have a tuple of: (0, 0) if there is no error, (1, 0) if it should be ascribed, or (0, 1) otherwise. The scaffolding will work in accordance with abstract argumentation stable semantics [11], since no cycles are permitted and the ascription of any parent node is determined by the ascription of its children. The algorithms are thus intended to be used to render a tuple of weights for the base-level factors which can then be used to determine the proportion of the classification tasks assigned to the NLP layer. For example, if a factor F1 has a tuple (0.1, 0.5), then in the next training epoch we would run five times as many classification tasks for F1 being not ascribed as we would for F1 being ascribed. This example also illustrates the benefit of partitioning the ADF layer to accommodate only Boolean nodes, since most NLP tasks involve classification and not regression. Future work will look at extensions to encompass ADFs that include non-Boolean nodes, but further consideration would need to be given as to how to extract continuous valued data points from text.

5. Concluding Remarks

In summary, in this short paper we have identified the need for the explanations of legal decisions to go beyond factors and preferences between them, and explain the ascription and non-ascription of the factors themselves. This sort of explanation is as yet largely uninvestigated in AI and Law, which has taken the factors in a case as given. We are currently engaged in research to establish the nature of such explanations through a hybrid approach, with the first stage in the process being addressed through ML and the second stage through reasoning via the medium of an ADF. Our immediate focus is on formal articulation and application of the aforementioned algorithms and Hierarchical BERT model to enable these tasks.

References

[1] L Al-Abdulkarim, K Atkinson, and T Bench-Capon. Statement types in legal argument. In *Proceedings of JURIX 2016*, pages 3–12. IOS Press, 2016.

[2] V. Aleven. *Teaching case-based argumentation through a model and examples*. PhD thesis, University of Pittsburgh, 1997.

[3] K Ashley and S Brüninghaus. Automatically classifying case texts and predicting outcomes. *AI and Law*, 17(2):125–165, 2009.

[4] K Atkinson and T Bench-Capon. Reasoning with legal cases: Analogy or rule application? In *Proceedings of the 17th ICAIL*, pages 12–21. ACM, 2019.

[5] K Atkinson, T Bench-Capon, and D Bollegala. Explanation in AI and Law: Past, present and future. *Artificial Intelligence*, 289:103387, 2020.

[6] T Bench-Capon. Using issues to explain legal decisions. *XAILA worhshop at ICAIL 2021. arXiv preprint arXiv:2106.14688*, 2021.

[7] T Bench-Capon and K Atkinson. Precedential constraint: The role of issues. In *Proceedings of the 18th ICAIL*, pages 12–21. ACM, 2021.

[8] F Bex and H Prakken. On the relevance of algorithmic decision predictors for judicial decision making. In *Proceedings of the 18th ICAIL*, pages 175–179. ACM, 2021.

[9] L K Branting, C Pfeifer, B Brown, L Ferro, J Aberdeen, B Weiss, M Pfaff, and B Liao. Scalable and explainable legal prediction. *AI and Law*, 29(2):213–238, 2021.

[10] G Brewka, S Ellmauthaler, H Strass, J Wallner, and P Woltran. Abstract dialectical frameworks revisited. In *Proceedings of the Twenty-Third IJCAI*, pages 803–809. AAAI Press, 2013.

[11] P M Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.

[12] A Kozinski. What I ate for breakfast and other mysteries of judicial decision making. *Loyola of Los Angeles Law Review*, 26:993, 1992.

[13] M Medvedeva, A Üstun, X Xu, M Vols, and M Wieling. Automatic judgement forecasting for pending applications of the european court of human rights. In *Proceedings of ASAIL 2021*, 2021.

[14] M Medvedeva, M Vols, and M Wieling. Using machine learning to predict decisions of the European Court of Human Rights. *AI and Law*, pages 1–30, 2019.

[15] J Mumford, K Atkinson, and T Bench-Capon. Machine learning and legal argument. In *Proceedings of the 21st CNMA Workshop, CEUR Workshop Proceedings*, volume 2937, pages 47–56, 2021.

[16] H Prakken and Ratsma R. A top-level model of case-based argumentation for explanation: formalisation and experiments. *Argument and Computation*, Available On-Line. 2021.

[17] H Prakken, A Wyner, T Bench-Capon, and K Atkinson. A formalization of argumentation schemes for legal case-based reasoning in ASPIC+. *Journal of Logic and Computation*, 25(5):1141–1166, 2015.

[18] E Rissland and K Ashley. A case-based system for Trade Secrets law. In *Proceedings of the 1st ICAIL*, pages 60–66. ACM, 1987.

[19] A Wyner and W Peters. Lexical semantics and expert legal knowledge: towards the identification of legal case factors. In *Proceedings of Jurix*, pages 127–136. IOS, 2010.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210336

Timed Dyadic Deontic Logic

Karam Younes KHARRAZ [a,1], Martin LEUCKER [a] and Gerardo SCHNEIDER [b]
[a] *ISP, University of Lübeck, Germany*
[b] *University of Gothenburg, Sweden*

Abstract. In this paper, we introduce TDDL, a timed dyadic deontic logic. Our starting point is a version of a dyadic deontic logic with conditional obligations, permissions, and obligations, and with a "reparation" operator for representing contrary-to-duties and contrary-to-prohibitions. We also consider a sequence operator allowing us to define norms as sequences of individual norms and most importantly with timed intervals, allowing us to express deadlines of norms. We provide a trace semantics capturing both satisfaction and violation of norms and discuss fulfillment of TDDL specifications.

Keywords. Normative specification, timed deontic logic

1. Introduction

While the formalization of untimed normative concepts is a quite well-studied topic (though by no means an exhausted nor completely solved issue), less attention has been paid to their combination with (real) time. One of the reasons is that incorporating time to a logic containing modalities for obligations, permissions, and prohibitions—with explicit operators for handling violations—is challenging [1].

That said, the idea of equipping norms with time is not new; see for instance the work by Governatori et al. [6,5,4]. These works, on discrete linear time defeasible deontic logic, prospected important aspects such as the classification of obligations following the timing of their enforcement and the resulting violation rules in a timed setting. In [1], Azzopardi et al. discuss different issues and design choices that need to be considered when adding time to the formalization of normative systems. Their paper is on the challenges regarding expressiveness and computational aspects for both specification and monitoring of timed deontic logics, though concrete solutions are missing.

There is a myriad of deontic logics, and in this paper, we consider *Dyadic Deontic Logic* (or DDL, for short) as a starting point. DDL is a variant of the standard or monadic deontic logic, tackling conditional norms without using material implication. The language of the logic is built on top of atomic propositions alone or put inside of deontic operators. Multiple solutions to arising paradoxes in such logics were proposed using non-monotonic logics or by changing the deontic detachment rule [7,8]. The resulting frameworks come with other limitations as discussed in [9]. We believe that adding time constraints solves certain of those problems of DDL.

[1] E-mail: Kharraz@isp.uni-luebeck.de

In this paper, we introduce TDDL, Timed Dyadic Deontic logic, an extension with time of the dyadic deontic logic. Besides the standard operators for (conditional) obligations, permissions, and prohibitions, the underlying logic also has operators for disjunction, sequences, and reparations (to specify penalties in case of violations). Time is explicitly added as intervals to the modalities. We provide trace semantics suitable for conflict (and contradiction) detection as well as for monitoring. More precisely, as models, we consider timed words, i.e., words composed of discrete actions and their timestamps. We provide two semantics relations, one concentrating on duties and prohibitions and the second concentrating on permissions.

The paper is organized as follows. Section 2 introduces TDDL. In Section 3 we briefly discuss the issue of detecting conflicts and contradictions. We discuss related work in Section 4 and we conclude in the last section.

2. TDDL: Timed dyadic deontic logic

In this section, we present the logic TDDL. It is based on DDL but restricted to avoid some well-known problems, e.g. our logic does not support *negation* and *conjunction*. At the same time, it is extended to support the notion of discrete-time. Moreover, our logic extends DDL with *preference* and *sequence* operators.

Before presenting the syntax and semantics of our logic, we present norms (clauses) prescribing an online delivery service, which will be used as a motivating example throughout the paper to illustrate the features we want to capture in our logic.

Example 1. Let us consider an online delivery system with the following specification:

- The user is supposed to *collect* the goods when the *home delivery* shows up. The date of the home delivery is fixed between three and five days after the order has been issued. If the user does not collect the goods on the day of the home delivery, the post agent deposit leaves a missed delivery notice and the user is supposed to *collect* the goods from the closest post station within 7 days after receiving the notice.
- The user may *return* the collected goods within 30 days of the delivery.

Let us look at this example more closely: We have individual *agents* mentioned in the clauses. However, norms concern mostly one of them, the service user. The other agent, the post agent is barely mentioned. In general, the agents perform some *actions* such as delivering goods, returning goods, etc. The only active action from the post agent is to leave a missed delivery notice, which may be seen as a condition to enforce the norm for picking the goods from the post. Note also that there is a notion of *preference*: picking up the goods on the day of the delivery is the normal *desired* behavior from the agent while picking up the goods from the post is considered as *reparation*. There is also a temporal order between the reparations and the desired behavior. Most of these actions underly temporal constraints, meaning they should occur within a certain time interval, like collecting goods within 7 days after receiving a notice or returning goods within 30 days of delivery. Notice that the corresponding intervals are often relative to other actions, like 30 days after delivery. In the next subsection, we use these observations to distill our logic.

$$\begin{aligned}
\mathtt{N} &:= Dop^{I}(a) \mid Dop^{I_2}(a \mid^{I_1} b) \text{ with } Dop \in \{O \mid P \mid F\} \\
\mathtt{NC} &:= \mathtt{N} \mid \mathtt{N}\ op\ \mathtt{N} \text{ with } op \in \{\gg, \mathbf{;}, \vee\} \\
\mathtt{NS} &:= \{\mathtt{NC}_1, \mathtt{NC}_2 \ldots, \mathtt{NC}_n\}
\end{aligned}$$

Figure 1. Syntax of TDDL

2.1. TDDL *Syntax*

The syntax of TDDL is shown in Figure1, norms are formed using deontic operators, actions and time intervals. The deontic operators are: Obliged O, Forbidden F, and Permitted P. Since we do not have negation we need the three modalities. Actions are from a set Σ that consist of all possible discrete actions (from the agent and the "environment").[2]

We assume actions are atomic, meaning their duration is a one-time step and that two actions cannot happen at the same time step. Time intervals are defined from the domain $\mathbb{I}^+ = [0, +\infty[$. An interval is formed by a pair, or union of pairs, $[i,s]$, with $i,s \in \mathbb{N}$ and $0 \leq i \leq s$. As in DDL, norms comes in two flavors: *monadic* or *dyadic*. A monadic norm is formed with one action and one interval. For instance, $O^{[0,4]}(coll)$ means that the agent has to *achieve* to collect the goods within 4 days. $F(open\_door) \equiv F^{[0,+\infty[}(open\_door)$ means that the agent is *always* forbidden to open the door; $P^{[0,30]}(ret)$ means that the agent has the right to return goods between 0 and 30 time steps.[3] Dyadic operators take two actions and two intervals. The left action is the action concerning the agent whilst the right one is the triggering action coming possibly from the environment. The two intervals are respectively the *norm validity interval* and the *reactivity interval*. For instance, $O^{I}(a|^{R}b)$ means that the agent is obliged to react performing a within the reactivity interval R *after* the environment had done action b within interval I.

Norms may be composed using the operators of preference "$\gg$" and sequence ";".

Sequence We use this operator to specify a linear order between the fulfillment of norms. For example, $\mathtt{NC}_3 := O^{[3,5]}(coll); P^{[0,30]}(ret)$ specifies that to fulfill $\mathtt{NC}_3$, the agent has to fulfill the collection before fulfilling the permission to return the goods. The interval of the right norm is relative to the left norm. Another sequence operator could be specified for cases requiring having absolute intervals in the second norm, but is not considered here due to space limitations.

Preference Unlike the Kripke semantics of DDL, we do not encode the preference relation in the model of the logic. For example, $\mathtt{NC}_4 := O^{[3,5]}(h\_coll) \gg O(coll|^{[0,7]}p\_del)$, prescribes that collecting the goods at home is *more preferable* than collecting it from the post after receiving the failed home collection notice from the post agent. Note that this operator is not symmetric unlike the logical or. Like the sequence, the preference operator could have a variant where the second interval is interpreted as absolute, but again this is left for a full version of the paper.

[2] W.l.o.g. and for simplicity, we assume that the contract concerns only one agent "against" an environment. This is a question more of terminology for presenting our ideas: we can also consider, as usually done in the literature, that actions encode the active agent/user performing the action (or to which the norm applies to).

[3] We talk in general about "time steps" with the understanding that it might mean different time units depending on the context (e.g., here it might mean "days".)

Normative system A normative system is a set of composite or simple norms.

Example 2. The norms of example 1 are specified in TDDL as a composed norm:

$$\mathtt{NC}_{\mathtt{Delivery}} := (O^{[3,5]}(h\_coll) \gg O(coll\ |^{[0,7]}\ p\_del)); P^{[0,30]}(ret).$$

2.2. Duty and Right trace semantics

One special feature compared to DDL is that we define two different satisfaction relations in TDDL, $\vDash_D$ (*duty*) and $\vDash_R$ (*right*). The duty relation means "did the agent fulfill her duties". A duty implies performing an action in the case of an obligation or avoiding it in the case of a prohibition. The right relation gives the answer to the question: "did the agent used her right?". For a normative system, our model is a word $w = (p_1, \tau_1)\dots(p_n, \tau_n)$ where the actions are from the agent and the system (environment): $p_i \in \Sigma$ and $\tau_i \in \mathbb{N}$ are timestamps. Let us first concentrate on the duty semantics.

Obligations and prohibitions To fulfill an obligation, it is enough to have one occurrence of the specified action. For prohibitions, one occurrence inside the scope of the prohibition violates the duty semantics. Thus, we define the satisfaction relations for the monadic operators as

$$w \vDash_D O^I(a) \text{ iff } \exists t \in I.a = w(t)$$

$$w \vDash_D F^I(a) \text{ iff } \forall t \in I.a \neq w(t)$$

For the dyadic operators, the semantics is given as:

$$w \vDash_D O^I(a|^R b) \text{ iff } \forall t \in I.b \notin w(t) \text{ or } (\exists min(t) \in I.w(t) = b \text{ and } w \vDash_D O^{R+t}(a))$$

$$w \vDash_D F^I(a|^R b) \text{ iff } \forall t \in I.b \notin w(t) \text{ or } (\forall t \in I.w(t) = b \rightarrow w \vDash_D F^{R+t}(a))$$

For example, we have that $(coll, 3) \vDash_D O^{[3,5]}()$, $(h\_coll, 3) \nvDash_D F^{[3,5]}(coll)$ and $(p\_del, 4)(coll, 6) \vDash_D O(coll|^{[0,7]} p\_del)$.

Permissions in the duty semantics We define permissions in the most simple way, where we say that a permission is not concerned by the duty semantics. Hence, we define:

$$w \vDash_D P^I(a) \text{ always}$$

$$w \vDash_D P^I(a|^R b) \text{ always}$$

Another possible kind of operator is the *strict permission*, where using a right out of the context when the conditions are satisfied could be interpreted as a violation of the duty semantics.[4]

Composed norms For the sequence and the preference operators, we have to know when exactly the first, left norm has been fulfilled or violated because the fulfillment or reparation by the right norm is *relative* to the one in the left. This information is provided by the two functions *violation prefix* VP and *fulfillment prefix* FP which provide for a given norm and trace the prefix that violates or fulfills the norm, respectively, and is defined further below. We now define the semantics for the composed operators as:

[4] A *strict permission*, SP, could be defined as $SP^I(a) \equiv \{P^I(a), F^{[0,+\infty]-I}(a)\}$.

$$w \vDash_D \mathtt{NC}_1^{I_1} \gg \mathtt{NC}_2^{I_2} \text{ iff } w \vDash_D \mathtt{NC}_1^{I_1} \text{ or } (w \nvDash_D \mathtt{NC}_1^{I_1} \text{ and } w \vDash_D \mathtt{NC}_2^{I_2+VP(w,NC_1)})$$

$$w \vDash_D \mathtt{NC}_1^{I_1} \,;\, \mathtt{NC}_2^{I_2} \text{ iff } w \vDash_D \mathtt{NC}_1^{I_1} \text{ and } w \vDash_D \mathtt{NC}^{I_2+\mathrm{FP}(w,NC_1)}$$

$$w \vDash_D \mathtt{NC}_1^{I_1} \vee \mathtt{NC}_2^{I_2} \text{ iff } w \vDash_D \mathtt{NC}_1^{I_1} \text{ or } w \vDash_D \mathtt{NC}^{I_2}$$

Remark 1. To be able to capture the nuance of fulfilling a composed norm with preference, one can add an index to the duty satisfaction relation i.e $\vDash_{D,1}$ to express the fact that the norm was fulfilled in the "best settings" and $\vDash_{D,2}$ for the second best setting.

Right semantics Obligations and permissions are not concerned with the right satisfaction relation. For permissions, one or more occurrences of the concerned action indicate that the right has been used. So, we have for instance that $(ret,3) \vDash_R P^{[0,30]}(ret)$ but $(ret,3) \nvDash_R O^{[0,30]}(ret)$. Note that the semantics of composed norms with the preference operator is not defined. The right semantics is defined as follows:

$$w \nvDash_R DOP^I(a) \text{ iff } DOP \in \{O,F\}$$

$$w \nvDash_R DOP^I(a|^R b) \text{ iff } DOP \in \{O,F\}$$

$$w \vDash_R P^I(a) \text{ iff } \exists t \in I.w(t) = a$$

$$w \vDash_R P^I(a|^R b) \text{ iff } \exists t' \in I.w(t) = b \text{ and } \exists w \vDash_R P^{R+t}(a)$$

$$w \vDash_R \mathtt{NC}_1^{I_1} \,;\, \mathtt{NC}_2^{I_2} \text{ iff } w \vDash_R \mathtt{NC}_1^{I_1} \text{ or } w \vDash_R \mathtt{NC}^{I_2+\mathrm{FP}(w,NC_1)}$$

$$w \vDash_R \mathtt{NC}_1^{I_1} \vee \mathtt{NC}_2^{I_2} \text{ iff } w \vDash_R \mathtt{NC}_1^{I_1} \text{ or } w \vDash_R \mathtt{NC}^{I_2}$$

Normative systems A normative system is satisfied according to the duty semantics if all norms are satisfied within this relation. On the other hand, we only require that at least one norm in the system satisfies the right semantics for the whole normative system to satisfy it:

It remains to define the two functions VP and FP that identify the earliest timestamps (shortest prefix) for which a norm is violated or fulfilled.

Monadic operators The violation prefix of monadic obligations is the maximum element of the interval, while that of a prohibition is the first occurrence of the forbidden action. There is no violation prefix for permissions. We thus have:

$$\mathrm{VP}(w, O^I(a)) := max(I) \qquad \text{iff } w \nvDash_D O^I(a)$$

$$\mathrm{VP}(w, F^I(a)) := t \in I \qquad \text{iff } w_{0,t} \nvDash_D F^I(a) \text{ and } \nexists t' < t.w_{0,t'} \nvDash_D F^I(a)$$

So, $\mathrm{VP}((h\_coll,7), O^{[3,5]}(h\_coll)) = 5$ and $\mathrm{VP}((h\_coll,2), F^{[3,5]}(h\_coll)) = 2$.

Dyadic operators The violation prefixes for the dyadic operators are more complex. In the case of obligations, it is set to the maximum element of the reactive interval updated with the timestamp of the first occurrence of the triggering action. For a prohibition, this prefix is the first occurrence of the forbidden action:

$$\mathrm{VP}(w, O^I(a|^R b)) := t + max(R) \quad \text{iff } t = first_{occ}(w,b) \text{ and } w \nvDash_D O^I(a|^R b)$$

$$\mathrm{VP}(w, F^I(a|^R b)) := t \qquad \text{iff } w(t) = a \text{ and } w_{0,(t-1)} \nvDash_D F^I(a|^R b) \text{ and } w \nvDash_D F^I(a|^R b)$$

For instance, we have $\mathrm{VP}(\{(p\_del,7),(ret,15)\}, O(p\_del|^{[0,7]} coll)) = 14$.

Composed norms The violation prefix for the preference operator is the violation of the right norm updated with the violation prefix of the left norm. For sequences, the violation prefix could have two forms: the violation prefix of the left norm or the violation prefix of the right norm updated with the fulfillment prefix of the first norm. This gives rise to:

$$\mathrm{VP}(w, \mathtt{NC}_1^{I_1} \gg \mathtt{NC}_2^{I_2}) := \mathrm{VP}(w, \mathtt{NC}_2^{I_{\mathrm{VP1}}})$$
$$\text{iff } I_{\mathrm{VP1}} = I_2 + \mathrm{VP}(w, \mathtt{NC}_1^{I_1}) \textit{ and } w \not\models_D \mathtt{NC}_1^{I_1} \gg \mathtt{NC}_2^{I_2}$$

$$\mathrm{VP}(w, \mathtt{NC}_1^{I_1} \ ; \ \mathtt{NC}_2^{I_2}) := \begin{cases} \mathrm{VP}(w, \mathtt{NC}_1^{I_1}) \text{ iff } w \not\models_D \mathtt{NC}_1^{I_1} \\ \mathrm{VP}(w, \mathtt{NC}_2^{I_2 + \mathrm{FP}w, \mathtt{NC}_1^{I_1}}) \text{ iff } w \not\models_D \mathtt{NC}_1^{I_1} \ ; \ \mathtt{NC}_2^{I_2} \end{cases}$$

$$\mathrm{VP}(w, \mathtt{NC}_1^{I_1} \vee \mathtt{NC}_2^{I_2}) := max((\mathrm{VP}(w, \mathtt{NC}_1^{I_1}), \mathrm{VP}(w, \mathtt{NC}_2^{I_2}))).$$

For example: $\mathrm{VP}((p\_del,7)(ret,15),(O^{[3,5]}(h\_coll) \gg O(coll|^{[0,7]}p\_del)) = 14$, and $\mathrm{VP}((p\_delevery,7)(ret,15),(O(p\_del|^{[0,7]}coll);P^{[0,30]}(ret)) = 14$.

Fulfillment prefixes An obligation is satisfied by the first occurrence of the corresponding action (respecting the time interval). For prohibitions, the fulfillment prefix is the limit of the interval. Fulfillment in our paper refers to the possibility of exercising rights and achieving duties:

$$\mathrm{FP}(w, O^I(a)) := t \in I \qquad \text{iff } w_{0,t} \vDash_D O^I(a) \textit{ and } (\forall t' < t . w_{0,t'} \not\models_D O^I(a))$$

$$\mathrm{FP}(w, F^I(a)) := max(I) \qquad \text{iff } w \not\models_D F^I(a)$$

$$\mathrm{FP}(w, O^I(a|^R b)) := \begin{cases} max(I) \text{ iff } \nexists t' \in I . w(t') = b \\ t \text{ iff } t' = first_{oc}(w, b) \textit{ and } t = \mathrm{first}_{oc}(w^{t'}, a) \end{cases}$$

$$\mathrm{FP}(w, F^I(a|^R b)) := \begin{cases} max(I) \text{ iff } \nexists t' \in I . w(t') = b \\ t \text{ iff } t' = last_{oc}(w_{0,max(I)}, b) \textit{ and } t = t' + R \end{cases}$$

For permissions, the prefix depends on whether the right has been used or not. If the right was not used then the fulfillment prefix is set up to be the maximum element of the interval. In our example, $\mathrm{FP}(\{(coll,3)\}, O^{[3,5]}(coll)) = 3$ and $\mathrm{FP}(\{(coll,1),(coll,7)\}, F^{[3,5]}(coll)) = 5$. Formally,

$$\mathrm{FP}(w, P^I(a)) := \begin{cases} t \text{ iff } t = \mathrm{first}_{oc}(w_I, a) \\ max(I) \text{ iff } w \not\models_R P^I(a) \end{cases}$$

$$\mathrm{FP}(w, P^I(a|^R b)) := \begin{cases} max(I) \text{ iff } w \not\models_r P^I(a|^R b) \\ t \text{ iff } t = first_{oc}(w_{I+R}, a) \textit{ and} \\ \quad \exists t' in [t - max(R), t - min(R)] . w(t') = b \end{cases}$$

So, $\mathrm{FP}(\{(ret,3)\}, P^{[0,30]}(ret)) = 3$ and $\mathrm{FP}(\{(coll,32)\}, P^{[0,30]}(ret)) = 30$.

The fulfillment prefix for sequences is the fulfillment prefix of the right norm updated with the fulfillment prefix of the left norm. On the other hand, for composing norms with preference operators, we consider two cases, the second case is relative to the violation prefix of the left norm. The formalization is as follows:

$$\mathrm{FP}(w, \mathtt{NC}_1^{I_1} \ ; \ \mathtt{NC}_2^{I_2}) := FP(w, \mathtt{NC}_2^{I_{FP}}) \qquad \text{iff } I_{FP} = I_2 + FP(w, \mathtt{NC}_1^{I_1}) \textit{ and } w \vDash_D \mathtt{NC}_1^{I_1} \ ; \ \mathtt{NC}_2^{I_2}$$

$$\mathrm{FP}(w, \mathtt{NC}_1^{I_1} \vee \mathtt{NC}_2^{I_2}) := x \qquad \text{iff } x = min((\mathrm{FP}(w, \mathtt{NC}_1^{I_1}), \mathrm{FP}(w, \mathtt{NC}_2^{I_2}))$$

$$\mathrm{FP}(w, \mathtt{NC}_1^{I_1} \gg \mathtt{NC}_2^{I_2}) := \begin{cases} \mathrm{FP}(w, \mathtt{NC}_1^{I_1}) \text{ iff } w \vDash_D \mathtt{NC}_1^{I_1} \\ \mathrm{FP}(w, \mathtt{NC}_2^{I_{FP1}}) \text{ iff } I_{FP1} = I_2 + \mathrm{VP}(w, \mathtt{NC}_1^{I_1}) \textit{ and } w \vDash_D \mathtt{NC}_1^{I_1} \gg \mathtt{NC}_2^{I_2} \end{cases}$$

Example 3. Let us consider the following traces of $\mathtt{NC}_{\mathtt{Delivery}}$:
$w_1 := (h\_coll, 3)(ret, 31)$; $w_2 := (p\_del, 5)(coll, 18)$; $w_3 := (h\_coll, 3)(ret, 37)$ We can see that: $\mathrm{FP}(w_1, (O^{[3,5]}(h\_coll) \gg O(coll \mid^{[0,7]} p\_del)) = 3$
$\mathrm{VP}(w_2, (O^{[3,5]}(h\_coll) \gg O(coll \mid^{[0,7]} p\_del)) = 12$
$w_1 \vDash_D \mathtt{NC}_{\mathtt{Delivery}}$ *and* $w_1 \vDash_R \mathtt{NC}_{\mathtt{Delivery}}$ and $w_2 \nvDash_D \mathtt{NC}_{\mathtt{Delivery}}$ *and* $w_2 \nvDash_R \mathtt{NC}_{\mathtt{Delivery}}$
$w_3 \vDash_D \mathtt{NC}_{\mathtt{Delivery}}$ *and* $w_2 \nvDash_R \mathtt{NC}_{\mathtt{Delivery}}$

3. Fulfillabilty of normative systems in TDDL

Fulfillability is a important aspect of Normative systems. Following the aforementioned semantics it means that it is possible to satisfy both in the right and duty semantics. Especially specifying complex norms is a difficult task and automatic sanity checks would be desirable. We consider two notions important for such a sanity check: contradiction and conflict.

We say that two norms are *conflicting* if there is situation where it is not possible to use a right without breaching another norm. We call two norms *contradicting* when there is no possible trace that can fulfill the duties of all the norms.

For example, $\mathtt{NS}_1 := \{\mathrm{O}^{[3,5]}(a), \mathrm{F}^{[2,6]}(a)\}$ is contradicting because for each time stamp where the first norm could be fulfilled (namely 3,4,5), the second would be breached. For $\mathtt{NS}_2 := \{\mathrm{P}^{[3,5]}(a), \mathrm{F}^{[4,5]}(a)\}$ there exists timestamps (4,5) yielding to a violation if the agent uses his right thus leading to a conflict. Not using the permission would satisfy the norms in the duty semantic but not the right one, so the two norms are conflicting in the interval $[4,5]$. An algorithm to analyze a normative system and to detect contradictions and conflicts, signaling the unfulfillabilty of the concerned part of the original system, will be presented on an extended version of this paper.

4. Related Work

Partial normative specifications with time have been given by Governatori et al., for instance in [6,5,4]. The formalization consists of using *defeasible* and *defeators* rules to initialize, terminate and define violations of norms. The timed settings are intervals in $\mathbb{N}$. In those works timed actions are of different kinds: *achievement*, *maintenance* and *punctual*. In our work obligations and permissions are *achievements*, while prohibitions are of *maintenance* type. We do not support punctual since we do not allow the occurrence of simultaneous actions.

A "deontic calculus" extended with time intervals has been presented in [2,3]. The main difference with our work is that we are here considering a logic instead of a calculus: we have a formal (denotational) semantics defined in terms of a satisfaction relation over models, while they provide an operation view of the syntax (calculus).

Given the difference in the formalization and the intention (defeasible logic and calculi), it is difficult to make a concrete comparison with our approach except for the fact that we provide a deontic logic with explicit time.

5. Conclusion

In this paper, we presented a first suggestion of a timed dyadic deontic logic allowing to reason over prohibitions, obligations and permissions, within certain *timed intervals*. To support the different flavor of obligations and prohibition on one hand and permissions on the other, our logic comes with two different semantics relations, a first concentrating on *duties* while the second indicates which *rights* have been used when looking at action sequences that should adhere to the given norm.

It is very easy to define norms that have inherent conflicts or are even contradicting, for example by specifying overlapping intervals in which certain actions are both forbidden and obligated. We defined these notions formally and indicated that algorithmic support for improving such specifications can be done as future work.

Our work may be enhanced in several directions. The current expressiveness can be extended by further operators and by giving both relative and absolute notions of timed intervals. Also, the analysis of given norms may be enriched by providing checks for conflicts and contradictions. We leave such extensions for a full version of the paper.

References

[1] Azzopardi, S., Pace, G., Schapachnik, F., Schneider, G.: On the specification and monitoring of timed normative systems. In: RV'21. LNCS, Springer (2021)
[2] Cambronero, M.E., Llana, L., Pace, G.J.: A calculus supporting contract reasoning and monitoring. IEEE Access **5**, 6735–6745 (2017)
[3] García, A.A., Cambronero, M., Colombo, C., Llana, L., Pace, G.J.: Runtime verification of contracts with themulus. In: SEFM'20. LNCS, vol. 12310, pp. 231–246. Springer (2020).
[4] Governatori, G., Hulstijn, J., Riveret, R., Rotolo, A.: Characterising deadlines in temporal modal defeasible logic. In: Australasian Joint Conference on Artificial Intelligence. pp. 486–496. Springer (2007)
[5] Governatori, G., Rotolo, A.: Justice delayed is justice denied: Logics for a temporal account of reparations and legal compliance. In: CLIMA'11. pp. 364–382. Springer (2011)
[6] Governatori, G., Rotolo, A., Sartor, G.: Temporalised normative positions in defeasible logic. In: ICAIL'05. pp. 25–34 (2005)
[7] Horty, J.F.: Nonmonotonic foundations for deontic logic. In: Defeasible deontic logic, pp. 17–44. Springer (1997)
[8] Parent, X., van der Torre, L.: Aggregative deontic detachment for normative reasoning. In: KR'14. AAAI Press (2014)
[9] Pigozzi, G., van der Torre, L.: Multiagent deontic logic and its challenges from a normative systems perspective. IfCoLog Journal of Logics and Their Applications pp. 2929–2993 (2017)

4. Predictive Models

This page intentionally left blank

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210338

Can Predictive Justice Improve the Predictability and Consistency of Judicial Decision-Making?

Floris BEX [a] and Henry PRAKKEN [b]
[a] *Utrecht University & Tilburg University, The Netherlands*
[b] *Utrecht University & University of Groningen, The Netherlands; European University Institute, Italy*

Abstract. There has recently been talk of algorithms that predict decisions in legal cases being used by the judiciary to improve the predictability and consistency of judicial decision making. We argue that their use may minimise the error rate of decisions in the long run, but that this would require not only major technical advances but also major changes in legal thinking about what is the most important objective of judicial decision-making: optimising individual justice in a particular case or reducing errors in the long run. We further argue that if algorithmic decision predictors give any useful information in individual cases to judges at all, this is not in its predictions but in its explanations.

Keywords. Predictive justice, decision prediction, predictability, consistency

1. Introduction

Using machine-learning algorithms to predict decisions in legal cases has become a hot topic [3,13,10,1]. One use of these *algorithmic decision predictors* [5] is to help litigants estimate, for example, their chances of winning a case (e.g., commercial products like Lex Machina and Premonition.ai). Another possible use of predictors is to use them to analyse human biases or the influence of extraneous, non-legal factors on legal decision making [4,14]. Finally, a more contentious use of algorithmic predictors is their use by the judiciary (courts, judges): perhaps not as fully-automated 'robo-judges', but possibly for supporting judges in individual cases – it is this latter use of algorithmic decision predictors that is the main subject of this paper.

The use of algorithmic decision predictors by the judiciary is claimed to improve the predictability and consistency of judicial decision making, which is demanded by the principle of equality [8]: judges can use decision predictors as a support tool when drafting judgements to identify cases and patterns which lead to certain decisions [1], in order to come to more consistent, more informed and less biased judgments [4,14]. Some even argue that AI can be used as a 'monitor' [4] that shows the judge what the rational decision would be in a new case given the history of similar cases.

To be able to evaluate whether algorithmic decision predictors can have these claimed benefits, it is first necessary to have a clear picture of what information such al-

gorithms provide. This we discussed in [5], concluding that even if we have a prediction by an algorithmic predictor that performs well on a test set, we still cannot say that a rationally-thinking judge would probably take the predicted decision. In this paper, we aim to discuss exactly what is meant by the predictability and consistency of judicial decision making, and whether the use of algorithmic decision predictors by judges can improve such predictability and consistency.

2. Preliminaries

We first summarise our answer to the main question of [5]: if we have a prediction by an algorithmic decision predictor and information about the algorithm's performance, can we determine the so-called *decision probability* that an arbitrary rational judge assigned to the case would take the predicted decision?

In [5], we assumed that given an algorithm's performance measures (e.g. *precision*, the percentage of positive predictions that are correct) a statistical conditional probability can be derived that an arbitrary case C will receive decision D given that the algorithm predicts D for C. This probability is statistical in that it is not about an individual case c to be decided but about the class of all cases C for which the algorithm can give a prediction. By contrast, a decision probability is a conditional probability for an individual case c and a particular decision d that case c will receive decision d. Here we have the reference class problem, namely that a probability for an individual case c is not logically implied by a statistical probability for the class of all cases C to which c belongs. Instead, equating the individual probability of some event to the statistical probability for all events of the same class expresses a relevance judgement that the only thing that is relevant as regards the event is what is stated in the statistical probability. For example, if we know that 80% of the people with an Italian first are Italian (a statistical probability) and all we know of a particular person that he is called Giovanni, then we may rationally conclude that the individual probability that this Giovanni is Italian is 80%. However, if we also know that this Giovanni's surname is not Italian but Dutch, then this is clearly additional relevant information, so the statistical probability cannot be applied to him any more. So in the case of our decision predictor, equating the decision probability for a case c to the statistical probability for the class C of all c expresses that all that is relevant as regards c is the predicted outcome of a case. However, this relevance assumption is unjustified, since judges always know more about the case at hand than just its predicted outcome. We therefore concluded that an algorithmic decision predictor cannot be said to give the 'normal' or rational decision of a case given the history of similar cases.

3. Predictability and consistency

In the introduction we noted that some think that if judges take predictions of algorithmic decision predictors into account when deciding a case, this will improve the predictability and consistency of judicial decision-making. Two questions arise here (i) What do the terms predictability and consistency mean in this context?; and (ii) How can an algorithm be used to improve predictability and consistency?

We initially assume that in the context of judicial decision making predictability and consistency mean the same (although this is debatable). One interpretation of predictabil-

ity and consistency is then that the *same* case would be decided the same by different judges if assigned to the case. Another interpretation is that *similar* cases are decided in the same way (or a similar way) by the same or different judges. The second interpretation implies the first but not vice versa. We think that in both interpretations 'consistency' and 'predictability' indeed mean the same. The latter is not true for a third interpretation of predictability, corresponding to the gambler's wish to maximize expected utility in the long run. For instance, many cases might be substantially different from each other, so that even if like cases are treated alike, the predictability of the decision is low. A gambler who wants to bet on legal case decisions might indeed be advised to take an algorithm's prediction into account, since the gambler will often have no more information about the case than the algorithm's prediction plus statistical information about its performance. We next discuss for each of these interpretations how an algorithmic decision predictor can be used in order to improve predictability and consistency.

Deciding the same case in the same way. If predictability and consistency of judicial decision-making means that the *same* case is decided the same by different judges, then a sure way to guarantee this is to give all judges the same algorithmic decision predictor and to require that they all have to follow its predictions in all cases. Then different judges would, when assigned to the same case, be guaranteed to take the same decision. However, this does not make sense, for one since we do not know whether all decisions in the training and test set were correct. If all judges blindly follow the algorithm's prediction, then its accuracy will increase to 100%, and this would further lead to a tendency to make the predicted decision the legally correct one even if this cannot be justified. A possible counterargument here is that judges should not automatically follow the algorithm but just be willing to be informed by it. But then the main problem discussed Section 2 arises again: the judge cannot know from a prediction alone (combined with a statistical probability on the algorithm's performance) whether the predicted decision is what other judges assigned to the same case would likely decide. At the very least the algorithm should be able to explain the grounds for its prediction in legally meaningful terms. We will discuss this issue in more detail in Section 4, noting for now that there is a serious danger that judges who are told that they should consult the algorithm before taking their decision feel an unjustified pressure to accept the predicted decision as the legally correct one. This may in turn make that judges will think less hard about a case and become intellectually 'lazy'. So letting judges be informed by the predictions of algorithmic decision predictors has no clear advantages while there are real dangers. We cannot know whether predictability and consistency of judicial decision making (in the first sense) can be improved by letting judges be informed by the predictions of algorithmic decision predictors without combining them with an explanation in legally meaningful terms.

Deciding similar cases in the same way. How can predictability and consistency be improved if that means that *similar* cases should be decided the same? Is this improved if we require judges to consult decision predictors as a source of information? Again, if all we have is the prediction by an algorithm and some (statistical) probability, we cannot know. And even if we have a decision probability for an individual case, the prediction in itself would still not give any information about similar cases. In fact, it might well be that an algorithm treats cases that judges would regard as similar as different or vice versa. For example, text-based decision predictors, which identify statistical correlations are identified between certain words or combinations in the text and the case decision

(such as the algorithms of [1,13] that predict outcomes of the European Court of Human Rights), could fail to recognise that linguistically small differences are legally very relevant. The converse may also happen, i.e., that the algorithm treats cases as different that judges would treat as similar, since the algorithm recognises differences that are legally irrelevant. Recall in this respect that with such predictors we cannot even know to which extent their predictions are based on aspects related to the merits of the case. At best, knowledge-based predictors that generate their predictions in a case-based way could give such information. We will further discuss this issue below in Section 4.

So also if predictability and consistency of judicial decision making means that like cases are decided alike, we can conclude that we cannot know whether it will be improved by letting judges be informed by the predictions of algorithmic decision predictors without combining them with an explanation in legally meaningful terms. Incidentally, in both interpretations of predictability and consistency there is a further reason for this, since for knowing whether using the algorithm will improve predictability and consistency, we will have to compare the situation with use of the algorithm to the current situation; and there is no a priori reason to assume that judges without algorithmic support will be less predictable and consistent.

3.1. Reducing error rates in the long run

We finally consider the gambler's interpretation of predictability. It might be argued that there is still some rationality in relying on statistical probabilities concerning decision predictions in individual cases. Assume, for instance, that it can be established through empirical research that judges on average make fewer mistakes when they always follow an algorithmic prediction than when they look at all particulars of a case as recommended by us in [5]. (It may be hard to conduct such research but let us for the sake of argument assume that it can be done.) Would it then not be more rational for a judge to reply on the outcome predictions?[1] We believe that the answer depends on which values are to be optimised in judicial decision making: should a judge, when faced with a case, be primarily interested in minimising the rate of erroneous decisions in the long run or should the judge primarily aim to optimise individual justice in the case at hand?

Consider an analogy from consumer banking. A bank that has to decide whether to grant a loan to a customer is not interested in optimising the quality of the decision for an individual customer but in minimising losses in the long run. Given this objective, it is rational for the bank to rely on statistical frequencies concerning classes of customers, even if the individual customer in the case at hand may have additional relevant characteristics not considered in the statistical probability. By contrast, it is in the customer's interest that these additional characteristics are considered by the bank, since the customer wants to be treated fairly. As we earlier observed in [5], this is related to O'Neill's [15] criticism of 'bucketing', the practice of basing a decision about an individual on the fact that the individual is a member of a particular class of which a statistical frequency is known instead of on the particular situation of that individual. O'Neill [15, pp. 145–6] argues that, although this strategy might optimise the decision maker's profit in the long run, it may lead to unjust decisions in individual cases.

Applying the same thinking to our problem, the same question should be answered by designers of procedures of judicial decision-making: is the main objective of judicial

[1] This question was brought to our attention by Giovanni Sartor in personal communication.

decision-making to minimise the rate of erroneous decisions in the long run or to optimise individual justice in the case at hand? Ultimately, this is a matter of legal policy. If the objective is chosen to be optimising individual justice, then algorithmic decision predictors have no relevance for judges deciding individual cases [5]. But if the objective is chosen to be minimisation of errors in the long run, we do not see any principled rational reason not to rely on algorithmic decision predictors, provided it can be established that their use indeed leads to a lower rate of incorrect decisions. However, there are serious practical obstacles. First, creating algorithms that provably reduce error rates is far from trivial and may require major technological advances, which makes the remainder of this discussion largely hypothetical. Second, it seems to us that most legal procedures are mainly meant to optimise individual justice so that benefiting from algorithmic decision predictors in the long run would require major changes in legal-procedural thinking. For instance, an obvious way to ensure error-rate reduction would be to always follow the prediction but then the judge would ignore the particulars of a case, which would very likely violate current procedural rules. If, for these reasons, the prediction is used as just one of the inputs for the judge besides the particulars of the case, then, as noted above, the problem arises how exactly the prediction should be combined with these particulars. For one thing, the advantage of reducing error rates might be lost. Moreover, the danger is that the predicted outcome is incorrectly assumed to be the normal outcome of the case, from which a rational judge could only deviate if the particulars of the case contain exceptional circumstances; as we explained at length in [5], this assumption is unjustified. Finally, relying on the predictions of a non-transparent algorithm would create an explainability problem, especially given current procedural justification requirements on judicial decisions.

4. Providing explanations

We can conclude from Section 3 that a decision prediction on its own, even when combined with quantitative performance information, cannot help judges making their decision-making more predictable and consistent in legally desirable ways. But is this different if the prediction is combined with an explanation for it? The answer is negative if the explanation cannot be given in terms of reasons related to the merit of the case. So it is not a good idea to use algorithms like the one of [10], which make their predictions based on extraneous factors, such as information about the judges, the litigants, the solicitors, the type of case or the jurisdiction. But this implies that a text-based predictor like the ones of [1,13] is also not useful, since it cannot extract any legally relevant information from the texts to which it is applied and use it for explaining a prediction in legal terms. In consequence, there is no way to identify whether its prediction is based on legal grounds or on extraneous factors. So the only kinds of decision predictor that could possibly yield legally relevant information to a judge are those that base their predictions on legally relevant factors.

In Section 2 of [5], we discussed two kinds of decision predictors. One kind is still based on statistics or machine learning but its cases are encoded in terms of legally meaningful features instead of as raw text or with extraneous data (e.g. [12]). The other kind is knowledge-based (e.g. [6,9]). For the first kind of system its performance could be measured for various subsets of factors, and if a case matches a particular subset, then

a probability derived from the system's performance for this subset could be reported. However, this would still only yield a statistical probability for a decision and no decision probability, so, as explained in [5] and summarised in Section 2, the judge would still have to think about the particulars of the case as usual; there is no sense in which the prediction gives the 'normal' decision of cases with this constellation of factors. Alternatively, the system could show similar cases to the judge according to some suitable notion of similarity. However, just showing similar cases is not yet a genuine legal explanation. It remains to be investigated to which extent predictors of this kind can generate legally acceptable and useful explanations in terms of their input factors.

A knowledge-based predictor can by definition yield a genuine legal explanation, since it determines the decision to be predicted by way of applying a model of legal reasoning and problem solving. So (if well designed) such a system can in principle explain its predictions in ways that judges would appreciate and understand. However, there are still some issues here. First, how do we know that the explanation given by the system is a legally acceptable one? Can we assume that a knowledge-based predictor with good test-set performance will also in a high number of cases give a legally acceptable explanation for the prediction? Perhaps, but the assumption is highly defeasible, while again the step from a statistical to a decision probability must be justified, which is far from trivial. For these reasons we believe that additional experiments of a different kind are needed to assess the legal quality of the generated explanations. Since there is no gold standard for this issue, such experiments will have to involve legal experts rating the quality of the explanations, similar to, for instance, the famous experiments in which the quality of the diagnoses and treatment advice given by the MYCIN medical expert system was evaluated [7]. Evaluating systems in this way is far from trivial [11], unlike determining numerical scores like accuracy, precision and recall, which can be automatically extracted from an experiment's confusion matrix.

Incidentally, it might be argued that if a predictor's explanation can generally be shown to be legally acceptable, then this also justifies interpreting the statistical probability based on an algorithm's performance on a test set as a decision probability for a specific case. This argument fails. First, note again that the statistical probability is not based on the specific reasons mentioned in the explanation but on the performance on the test set. Things might be better if statistical probabilities are known for specific classes of test cases, but as explained in [5], obtaining such more specific statistical probabilities is not trivial. Moreover, we would still have to justify all other assumptions needed to make the jump from the past to the future (see the end of Section 2). Instead of attempting to do all this, it is simpler to inspect the given explanation alone and ignore the fact that the decision was predicted; only the content of the explanation can give the judge an indication whether the predicted decision is a good one.

Assuming that the explanations shown by the system are generally legally acceptable, then a second question arises: how do we know that showing such an explanation is useful for judges, for instance, that the quality and consistency of their decision making increases and that bias is reduced? Here the quantitative test-set performance information is completely irrelevant. Instead, this question should arguably be answered in controlled and/or fielded experiments with actual legal decision makers, to check whether the legal quality of their decision improves when they are supported by algorithmic decision predictors. Like with the experiments for testing the legal quality of explanations, setting up such user studies in a correct way is far from trivial [11].

Validation studies of the kinds we have just suggested are, to the best of our knowledge, currently rare. This was different in the early days of AI & law research. For example, in the Netherlands, in the late 1980s and 1990s several user studies were done on the effect of knowledge-based support for civil servants deciding on social benefit applications; see [16, Section 3] for an overview. And Aleven [2] studied the effect of using CATO in teaching legal argumentation skills to law students on these skills. We believe that the current focus on data-driven approaches, with its associated quantitative performance criteria that can automatically be extracted from the experimental data, may be in part responsible for the current neglect of these other important kinds of validation studies. These studies are important if we want to convince the professional legal world that our AI & law systems can contribute to improving the quality of legal decision making.

Such validation studies still say little about the quality of an individual explanation in a new case, since the step from the test results to an individual new explanation is still nontrivial for all the reasons explained in [5]. However, the studies can be used by courts in their decision whether to let their judges be supported by the system. This is the same as courts deciding which law journals or other information sources it will make available for judges. Just as with, for instance, law journals, a general evaluation about its quality has to be made, as a criterion for deciding whether the judge will consult this information source at all. But just as with, for instance, law journals, judges should not automatically copy or accept what is said but only look at the content of what is being said or written.

5. Conclusion

We discussed to what extent algorithmic decision predictors can improve the predictability and consistency of judicial decision-making, given our earlier conclusion in [5] that such algorithms cannot rationally inform individual decisions of judges in a particular case. We discussed three senses of such predictability and consistency: (1) that the same case will be decided the same by different judges; (2) that a similar case will be decided the same by the same or different judges; (3) that always following the prediction will optimise the quality of a series of decisions in the long run. We argued that in the first two senses the use of such algorithms cannot improve the predictability and consistency of judicial decision-making in legally desirable ways. We also argued that this is possibly different in the third sense in that judges might minimise their error rate in the long run. However, this would require not only major technical advances but also major changes in legal thinking about what is the most important objective of judicial decision-making: optimising individual justice in a particular case or reducing errors in the long run.

We also argued that if algorithmic decision predictors give any useful information in individual cases to judges at all, this is not in its predictions but in its explanations. In particular, decision predictors are needed that can explain their predictions in legally relevant terms. However, we noted that whether support by such systems can indeed improve the quality of judicial decision making requires validation studies of a kind that goes far beyond the current trend to focus on numerical performance measures like accuracy, precision and recall. We made a plea for returning to an older AI tradition of carrying out empirical validation studies with potential or actual users of the algorithm.

We like to emphasise that our conclusions are confined to the use of algorithmic decision predictors for informing judges on what they could decide in particular cases.

Other uses of such algorithms may well have benefits, for instance, with respect to informing judges and academics about possible bias in a series of cases (cf. e.g. [4,14]). Moreover, algorithms for making different types of predictions can also be useful. For example, if the aim is to help courts in making their case management more efficient, then algorithms could be trained on features of cases that influence such efficiency, such as their duration. (By contrast, the use of case decision predictors for efficiency purposes, as suggested by Aletras et al. [1], does not make sense, since predictions of decisions do not give any information about efficiency-related aspects of the case.) Note, however, that many of the reservations we expressed in Section 2 and [5] also hold for such other predictive algorithms.

References

[1] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93, 2016.

[2] V. Aleven. *Teaching Case-Based Argumentation Through a Model and Examples*. PhD Dissertation University of Pittsburgh, 1997.

[3] K.D. Ashley. A brief history of the changing roles of case prediction in AI and law. *Law in Context. A Socio-legal Journal*, 36(1):93–112, 2019.

[4] B. Babic, D.L. Chen, T. Evgeniou, and A.-L. Fayard. A better way to onboard AI. *Harvard Business Review*, July-August, 2020.

[5] F.J. Bex and H. Prakken. On the relevance of algorithmic decision predictors for judicial decision making. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*, pages 175–179, New York, 2021. ACM Press.

[6] S. Brueninghaus and K.D. Ashley. Automatically classifying case texts and predicting outcomes. *Artificial Intelligence and Law*, 17:125–165, 2009.

[7] B.G. Buchanan and E.H. Shortliffe, editors. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA, 1984.

[8] European Commission for the Efficiency of Justice (CEPEJ). European ethical charter on the use of artificial intelligence in judicial systems and their environment, 2018.

[9] M. Grabmair. Predicting trade secret case outcomes using argument schemes and learned quantitative value effect tradeoffs. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law*, pages 89–98, New York, 2017. ACM Press.

[10] D.M. Katz, M.J. Bommarito, and J. Blackman. A general approach for predicting the behavior of the Supreme Court of the United States. *PloS one*, 12(4):e0174698, 2017.

[11] R..M. O' Keefe. Issues in the verification and validation of knowledge-based systems. In V. Ambriola and G. Tortora, editors, *Advances in Software Engineering and Knowledge Engineering*, volume 2 of *Series on Software Engineering and Knowledge Engineering*, pages 173–189. World Scientific Publishing Co, 1993.

[12] E. Mackaay and P. Robillard. Predicting judicial decisions: The nearest neighbor rule and visual representation of case patterns. *Datenverarbeitung im Recht*, 3:302–331, 1974.

[13] M. Medvedeva, M. Vols, and M. Wieling. Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, 28(2):237–266, 2020.

[14] F. Muhlenbach and I. Sayn. Artificial Intelligence and law: What do people really want?: Example of a French multidisciplinary working group. In *Proceedings of the 17th International Conference on Artificial Intelligence and Law*, pages 224–228, New York, 2019. ACM Press.

[15] C. O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016.

[16] J.S. Svensson. The use of legal expert systems in administrative decision making. In A. Grönlund, editor, *Electronic Government: Design, Applications and Management*, pages 151–169. Idea Group Publishing, London etc, 2002.

5. Explainable Artificial Intelligence

This page intentionally left blank

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210340

Cause of Action and the Right to Know

A Formal Conceptual Analysis of the Texas Senate Bill 25 Case

Réka MARKOVICH [a,1] and Olivier ROY [b]
[a] *Department of Computer Science, University of Luxembourg*
[b] *Department of Philosophy, University of Bayreuth*

Abstract. Bill 25 proposed by the Texas Senate in 2017 was created to eliminate the so-called 'wrongful birth' cause of action. This plan raised some questions about the 'right to know' and indirectly about rights in general. We provide a preliminary logical analysis investigating these questions by using deontic and epistemic logics within the theory of normative positions. This work contributes to the logic-based legal knowledge representation tradition, and to the formal conceptual analysis of legal rights studying the cause of action's role in the debated relation between the Hohfeldian categories 'claim-right' and 'power'.

Keywords. legal knowledge representation, deontic logic, normative positions

The Texas Senate Bill 25[2] was designed to abolish the 'wrongful birth' cause of action, that is to take away the possibility of parents who had given birth to seriously ill or disabled babies to sue doctors for failing to warn the parents about the serious health conditions at the foetal stage. While the bill never passed, it received international media attention. In her comments on the Bill [12][3] and more general work on epistemic rights [13], Lani Watson assesses the controversy surrounding it as a debate over the existence of an epistemic right:

> While the public debate surrounding Texas Senate Bill 25 was framed, predominantly, in terms of the language and rhetoric of the pro-life/pro-choice debate, the issue at the heart of the controversy is ultimately one of epistemic rights. Those opposing the bill argued that it would allow doctors to withhold information, or lie to, expectant parents about the health of an unborn fetus. The implicit assumption is that doing so would constitute some kind of harm or wrong. In the context of prenatal healthcare provision, expectant parents have a right to know certain facts about the health of an unborn fetus. By withholding, distorting, or failing to provide these facts, a doctor is unjustifiably disregarding her epistemic duty and so violating the parents' right to know. (Watson [12], pp. 11-12)

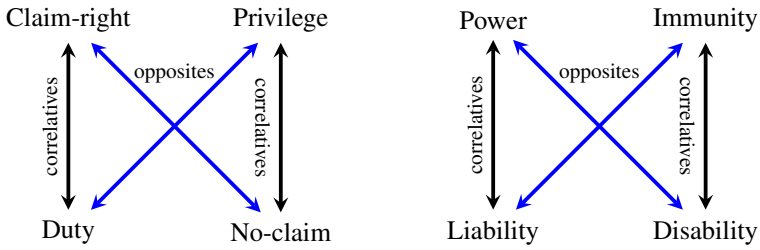
In this paper, we combine existing tools from the logical theory of normative positions [11,8] with tools from epistemic logic to analyse the logical relationship between

[1] This work was supported by the Fonds National de la Recherche Luxembourg through the project *Deontic Logic for Epistemic Rights DELIGHT* (OPEN O20/14776480).
[2] https://legiscan.com/TX/bill/SB25/2017
[3] https://www.law.ed.ac.uk/news-events/events/right-know-epistemic-rights-and-why-we-need-them-lani-watson

the Hohfeldian categories of rights that underlie the Texas Bill debate. In particular, we study the logical structure of the parents' right to know as a normative position, and how the cause of action and its elimination relate to (claim-)rights.

1. Theoretical Background

The theory of normative positions originates from the legal theorist, W. N. Hohfeld [4], who differentiated between four atomic types of rights and their correlatives, four types of duties [8]:



Each atomic right position of an agent comes with a correlative duty position of another agent (which is taken in the formal literature as equivalence between one agent's right and the other agent's duty). A *claim-right* of an agent concerns the other agent's action, one that the counterparty, the duty-bearer has an obligation to do, and this obligation is directed to the right-holder—this is what Hohfled calls a *duty* in the narrow sense. The Hohfeldian *privilege* to do something refers to the right-holder's own action as not being subject of a claim-right coming from the other agent. This is a relativized notion of what is called weak permission in the deontic literature.

The normative positions in the right square are considered higher order: the actions which one can have a power to are actions changing an (other) agent's normative positions. *Power* means this, as it is called in [8], *potential*: one's boss has a right to give new—work-related—tasks to her, that is, put duties on her, which means, the she is *liable* to that. But only regarding work-related tasks and not, for example, baby-sitting task related the boss' child, he is *disable* to do that, does not have a power to meaning the employee's immunity in this regard. Fitch considered the positions in this square capacitative [2] as opposed to the deontic ones in the left square.

The literature often takes the capacitative positions to be dynamic: the potential to *change* someone's normative positions lead recent work on the these positions to borrow from tools developed in dynamic epistemic logic [8,1]. In this paper, we use a simpler approach. Seeing to it that someone's normative positions change (in a given way, involved in the given action) is only *possible* with having the power to. Thus we describe of the capacitative positions using combinations of alethic, action and deontic operators: having a power means it is possible that the agent sees to it that a deontic state holds.

1.1. Definition of Claim-right and the relation between Claim-right and Power

Hohfeld considered the positions *sui generis*, so he didn't provide definitions of them, neither of what relation there is between the levels (squares). Makinson [7] provided

an admittedly preliminary, informal definition of 'counterparties' relying the seemingly obvious connection between the two levels of the positions (F is a given state of affairs):

> x bears an obligation to y that F under the system N of norms *iff* in the case that F is not true then y has the power under the code N to initiate legal action against x for non-fulfillment of F

If this definition worked, it would provide a definitive relation between *duty*—and so claim-right—and power. But, whatever intuitive sounding this definition is, the right-to-left direction of the biconditional does not work: the fact that we have the power to initiate a legal action against someone does not imply that we had a claim-right against him in the first place [8]. If this was the case, the court would not need to carry out the proceeding: the fact of initiating the legal action would mean winning it. But sometimes people lose in court, exactly because that they did not have the claim-right originally.[4]

Markovich's [8] work shares the Makinsonian insight that the key notion to understand what a claim-right is is *enforcement*, but leaves the notion of power out of the description: a duty to see to it that F and its unfulfillment, that is, $\neg F$ together triggers a new duty of the judge toward the original right-holder to make it the case that the original duty bearer compensates for F. This description thus, however, leaves open how the power to initiate a legal action against the counterparty relates to this notion of enforcement and that of the claim-right. In this paper we complement it arguing that the power to initiate a legal action concerns *settling* whether $\neg F$ is the case, and this plays a crucial role: it might be considered as affecting whether there is a (claim-)right at the first place.

1.2. Cause of action

The Texas Senate Bill 25 was about to eliminate a cause of action, namely the 'wrongful birth', meaning that the doctor fails to warn the parents about a serious illness of the fetus. The expression 'cause of action' refers to a set of facts that provides basis for the plaintiff to initiate a legal action. The plaintiff, of course, has specific goals with initiating a legal action: she wants enforcement, she wants a sentence which declares that the counterparty (the one whom she sued) didn't fulfill his duty and that the judge put a duty on the counterparty (now defendant) to compensate for not fulfilling that original duty of him. But sometimes the judge sentences against the plaintiff and this doesn't violate the judge's duty. This is because the judge's duty to enforce is not triggered by the plaintiff's initiating the legal action, but by the fact that the judge sees it proved that the defendant did what the plaintiff claims and that this counts as not fulfilling his duty. The factual part of what needs to be proved is what indicated as cause of action. And the relation of the set of facts indicated as cause of action to the original duty of the defendant is that this set of facts realizes the contrary-to-duty statement. *If* the judge sees it proved that the defendant did what the plaintiff indicated as cause of action, then she sentences about his duty to compensate. But this declaration about being proved is needed. And this is regarding what the judge gets a duty by someone initiating a legal action: the judge has to *decide* whether what is put as cause of action indeed happened. That is, the judge's duty concerns to either settle that the set of facts has been proved

[4] Sergot [11] suggested to add "with some expectation of success" Makinson's definition. Even if this approaches epistemic reality well, this amended definition still would not give us a precise relation between claim-rights and powers.

(defendant did what the plaintiff said he did), or to settle that it is not settled (that is, to declare that it hasn't been proved). It is important that the latter doesn't mean settling that the defendant didn't do what is indicated in the cause of action.

If a set of facts cannot be a cause of action, then it is not possible for someone (supposedly) having a claim-right to initiate a legal action requiring the judge to legally settle the given set of factual statements. In this case, the claim-right cannot be enforced as the needed declaration triggering the judge's duty to oblige the defendant's compensation cannot happen. It feels intuitive to say that in this case the right which would be a claim-right (to know whether the fetus is ill in the given case) *actually* doesn't exist. This is what is claimed in the Texas Senate Bill 25 case and what Makinson's definition intended to show. In what follows, we discuss some questions regarding the formalization of this relationship between a claim-right and the power to initiate a legal action—both in general and in this specific case.

2. Language and Semantics

We work with a propositional language extended with four modalities.

$$p \in \Phi \mid \varphi \wedge \psi \mid \neg\varphi \mid \mathbf{K}_a\varphi \mid \mathbf{O}_{a\to b}(\varphi/\psi) \mid E_a\varphi \mid \Box\varphi$$

Here a, b are elements of a finite set of agents A, and Φ is a given, countable set of propositional letters. $\mathbf{K}_a\varphi$ is the standard knowledge modality from epistemic logic, to be read as "agent a knows that φ". $\mathbf{O}_{a\to b}(\varphi/\psi)$ is a directed conditional obligation, to be read as "given ψ, a has a duty towards b that φ". $E_a\varphi$ is an agency operator to be read as "agent a sees to it that φ", and $\Box\varphi$ is a legal necessity operator to be read as "it is legally settled that φ".

This language is interpreted in Kripke models extended with a neighborhood function f_a for the agency operator.

Definition 1 (Frames and Models) *A frame $\mathfrak{F}$ for a given finite set A of agents is a tuple*

$$\mathfrak{F} = \langle W, \{R_a, \leq_{a\to b}, f_a\}_{a,b\in A}, R_\Box\rangle$$

where W is a finite set of possible worlds, R_a is an equivalence relation on W, $\leq_{a\to b}$ and $R_\Box$ are pre-orders (reflexive and transitive) relations on W, and $f_a : W \to \wp\wp(W)$ is a neighborhood function. Write $R_a(w)$ for $\{v : wR_av\}$, and similarly for $R_\Box$. We impose the following condition.

- *(Success) For all w and $X \in f_a(w)$, $w \in X$.*

A model $\mathscr{M}$ is a frame together with a valuation function $V : \Phi \to \wp(W)$. We write $w <_{a\to b} v$ whenever $w \leq_{a\to b} v$ but not $v \leq_{a\to b} w$; $w \equiv_{a\to b} v$ whenever $w \leq_{a\to b} v$ and $v \leq_{a\to b} w$.

At a state w, the set of states $R_\Box[w] = \{v : wR_\Box v\}$ captures what what is currently settled *in the eyes of the law*. Typically a legislation imposes stringent conditions, for instance in terms of admissible evidence, for recognizing that certain states of affairs hold. So not everything that is actually true at a given state needs to be legally settled.

On the other hand, given our semantics for the $\Box$ operator, the assumption that $R_\Box$ is reflexive entails that false propositions cannot be legally settled. Similarly, the condition that $R_\Box$ is transitive entails that if a proposition is legally settled, then it is legally settled that this proposition is legally settled. We do not, however, require $R_\Box$ to be symmetric. Imposing this would entail that whenever a proposition is not legally settled, it is legally settled that this proposition is not legally settled. This appears inaccurate for our intended interpretation: the judiciary might not have ruled on a certain fact without having settled that this fact is not settled.

Definition 2 (Truth Conditions) *Let $\mathfrak{M}$ be a model and $w \in W$. Write $||\varphi||$ for $\{w : \mathscr{M}, w \models \varphi\}$.*

- $\mathfrak{M}, w \models E_a\varphi \Leftrightarrow ||\varphi|| \in f_a(w)$.
- $\mathfrak{M}, w \models \Box\varphi \Leftrightarrow \forall v$ *such that* $wR_\Box v, \mathfrak{M}, v \models \varphi$
- $\mathfrak{M}, w \models K_a\varphi \Leftrightarrow \forall v$ *such that* $wR_a v, \mathfrak{M}, v \models \varphi$
- $\mathfrak{M}, w \models \mathbf{O}_{a\to b}(\varphi/\psi) \Leftrightarrow \forall v \in max_{\leq_{a\to b}}(||\psi|| \cap R_\Box[w]), \mathfrak{M}, v \models \varphi$

where, for any $X \subseteq W$, $max_{\leq_{a\to b}}(X) = \{w \in X : \neg\exists v \in X$ such that $w <_{a\to b} v\}$.

These truth conditions are standard for the normal modalities K_a and $\Box$, and the agency operator E_a is given the so-called *exact* neighborhood semantics [10]. The definition of conditional obligations is relativised to what is legally settled at a state. This provides a constrained version of the "ought implies can" principle: $\Diamond\psi \wedge O_{a\to b}(\varphi/\psi) \to \Diamond\varphi$. This would not be the case if we only considered the most ideal states where the condition ψ is true, because at a given state it could be legally settled that ψ is false. Unconditional obligations can be defined in the usual way: $O_{a\to b}\varphi \leftrightarrow O_{a\to b}(\varphi/\top)$.

3. Formal Analysis

We aim at capturing the logical structure of the parents' right to know whether the fetus is healthy and this right's relationship to the parents' (lacking) power to initiate a legal action because of the doctor's fail to warn about the illness. We analyse both component in turn, put them together.

3.1. Right to know whether the fetus is ill

The parents' claim-right can have multiple, non-equivalent logical representations [9]. Let p be the parents, d is the doctor, and *ill* for the proposition that the fetus is ill. As suggested in [5], a first, natural attempt at capturing a duty to (make someone) know *whether* something is the case is as a duty for the doctor to make it the case that either the parents know that the fetus is ill, or they know that the fetus is not ill:

$$\mathbf{O}_{d\to p}[E_d(\mathbf{K}_p(ill)) \vee E_d(\mathbf{K}_p(\neg ill))]$$

The knowledge and the action operators being factive makes the disjuncts mutually exclusive. Now it is well known that disjunctive syllogism is limited within the scope of deontic operators: the fact that the fetus is ill, together with the disjunctive duty as specified above, do not entail that the doctor has a duty to inform the parents. Even the doctor

knowing that the fetus is ill does not entail that she has an unconditional duty to inform the parents. If, however, it is legally settled that the fetus is ill, then we get a form of disjunctive syllogism. The following is valid in the class of frames defined above.

$$\Box ill \wedge \mathbf{O}_{d\to p}[E_d(\mathbf{K}_p(ill)) \vee E_d(\mathbf{K}_p(\neg ill))] \to \mathbf{O}_{d\to p}E_d(\mathbf{K}_p(ill))$$

The same type of disjunctive syllogism applies, of course, whenever it is legally settled that the doctor knows that the fetus is ill. In fact, because knowledge is factive in our formalization, this formulation entails that the doctor's obligation to inform the parents that the fetus is ill holds *only* when the fetus is actually ill, and similarly if the fetus is not ill. That is, the state of the fetus is a necessary but not in itself a sufficient condition for the doctor to have a duty to inform the parents.

Our logical language of course allows to represent the fact that the doctor identifying the state of the fetus might not be a necessary but rather a sufficient condition for triggering the obligation to inform the parents about that very state. Identifying is itself a subtle epistemic action, which arguably does not always coincides with the doctor herself knowing whether the patient is ill.[5] Intuitively, however, identifying goes in most cases hand in hand with knowing, and so in this paper we will identify the former with the latter. This gives us the following the following pair of obligations:

$$\mathbf{O}_{d\to p}(E_d\mathbf{K}_p(ill)/K_d(ill)) \wedge \mathbf{O}_{d\to p}(E_d\mathbf{K}_p(\neg ill)/K_d(\neg ill))$$

Both these conditional obligations trigger unconditional ones *if it is legally settled* that the fetus is ill, or that the doctor knows it. In this case it is intuitively plausible that the doctor making a diagnosis regarding the state of the fetus is necessary and sufficient for it to be legally settled that the illness holds: this is a medical question, it can only be settled by a professional in the eye of law. So even though these conditional obligations do not trigger unconditional ones by the simple fact that the doctor knows the state of the fetus, they do if we consider her (epistemic) action as *making a diagnosis*, which is sufficient for our purpose here—we leave the analysis of the difference to future work.

3.2. Power to initiate a legal action

As discussed above about the cause of action, we take the power to initiate a legal action as a possibility to put a duty on judiciary to decide the case. That is, once the legal action is initiated, there is a claim-right of the parents against the judge to the effect that she, the judge, declares whether the relevant cause of action obtains, i.e. whether the doctor indeed failed to inform the parents of the medical status of the fetus. Following Markovich [8], we take this declaration to be a speech act through which the judge makes it legally settled that the cause of action obtains, or not. In the positive case, where the cause of action indeed obtains, this declaration can be captured by a combination of our agency and legal necessity operators, using j for the judiciary (or the given judge), and $KW_a(\varphi)$ for the proposition that p knows whether φ, i.e., $K_p(\varphi) \vee K_p(\neg\varphi)$.

$$E_j(\Box\neg E_d(KW_p(ill))$$

[5] Think for instance of a doctor knowing that the result of a perfectly reliable test are at hand, and passing them to the parents without herself looking at what the results actually are.

The negative case, where the cause of action does not obtain, is slightly more complex. Recall that $\Diamond E_d(KW_p(ill))$ reads as "it is not legally settled that the doctor did not inform the parent's of the fetus' state." As we observed earlier, this is compatible with the judiciary not having ruled whether the doctor did in fact inform the parents. By declaring that the latter is not settled, the judge does something stronger in the sense of explicitly ruling that this fact is not legally settled. What the judge does in this case is not continuing the status quo, but rather to settle that it is not settled whether the doctor did not inform the parents. To capture this we thus use one more iteration of the legal necessity operator.

$$E_j(\Box\Diamond E_d(KW_p(ill))$$

This, however, is not sufficient. The reader familiar with modal logic will have noticed that this $\Diamond E_d(KW_p(ill))$ is in fact consistent with $\Box E_d(KW_p(ill))$, which says that it is legally settled that the doctor has informed the patient. To capture the constraint that it is genuinely not legally settled whether $E_d(KW_p(ill))$, one has to use a stronger version:

$$\Diamond E_d(KW_p(ill)) \wedge \Diamond \neg E_d(KW_p(ill))$$

This formulation nicely captures the idea that, legally speaking, neither $E_d(KW_p(ill))$ nor its negation can be ruled out. Putting all this together, we get the following formalization of the parents' power to initiate a legal action:

$$\Diamond E_p(O_{j\to p}(E_j(\Box\neg E_d(KW_p(ill))) \vee E_j(\Box(\Diamond E_d(KW_p(ill)) \wedge \Diamond\neg E_d(KW_p(ill))))))$$

3.3. Power as necessary condition for claim-right

Following Watson [12], we take the core of the Texas Senate Bill 25 be that the parents do have a claim-right against the doctor to know the state of the fetus *only if* the wrongful birth exists as a cause of action, that is, the parents also have the power to initiate legal action with this reason. In other words, the claim-right entails the legal power to initiate legal action. This can be straightforwardly captured using material implication, with two versions corresponding to the two readings of the parents' claim-right that we presented earlier. Putting all this together, we get the following formalization of the parents' power to initiate legal action, with $\mathbf{O}_{d\to p}(E_d\mathbf{K}_p(\pm ill)/K_d \pm ill)$ abbreviating the pair of conditional obligations discussed above.

$$\mathbf{O}_{d\to p}(E_d\mathbf{K}_p(\pm ill)/K_d \pm ill) \to$$
$$\neg\Box\neg E_p(O_{j\to p}(E_j(\Box\neg E_d(KW_p(ill))) \vee E_j(\Box(\neg\Box\neg E_d(KW_p(ill)) \wedge \neg\Box E_d(KW_p(ill))))))$$

This relationship is not specific to the Texas Senate Bill 25 case. It has a great relevance in theory of legal rights and that of the normative positions. If we accept this connection as a crucial one regarding the mere existence of a right, then we can say that removing a cause of action generally means that there can be no one can have the respective claim-right in the first place.[6]

[6] An implication is not the only possible interpretation of this strong relationship between a claim-right regarding an action and the possibility to initiate a legal action for the settlement whether the action is done. One can argue that a right *means* the conjunction of a claim-right and the power to initiate the relevant a legal action, that is, each right in law is a molecular one involving at least two Hohfeldian positions.

4. Conclusion and Further Work

We provided a preliminary formal analysis of the right(-related) concepts involved in the Texas Senate Bill 25 case and its public and philosophical reception. We believe that analyzing this case brings important considerations about the (deontic) logic-based representation of legal knowledge. On one hand, the questions around epistemic rights, especially the 'right to know whether' seems to be multifold and challenging requiring careful combination of deontic and epistemic logics. On the other hand, the relation between the different normative positions and its relevance in terms of defining and reasoning with rights is crucial in legal knowledge representation. We also find the intuitive step of identifying the cause of action as a contrary-to-duty statement an important one to take in understanding and formalizing this relation between the different levels of Hohfeldian rights. We have left several questions to further work such as the axiomatization of the validates in our class of frames; studying the differences of the logical behavior of the different formalizations; studying the consequences of using dynamic operators to capture power and "informing" in the (claim-)right to know; and, of course, using other theories of conditional obligations e.g. defeasible deontic logic [3] or input/output logics [6].

References

[1] Huimin Dong and Olivier Roy. Dynamic logic of legal competences. *Journal of Logic, Language and Information*, pages 1–24, 2021.

[2] Frederic B. Fitch. A Revision of Hohfeld's Theory of Legal Concepts. *Logique et Analyse*, 10(39/40):269–276, December 1967.

[3] G. Governatori, A. Rotolo, and E. Calardo. Possible world semantics for defeasible deontic logic. In T. Ågotnes, J. Broersen, and D. Elgesem, editors, *DEON*, pages 46–60. Springer, 2012.

[4] W.N. Hohfeld. Fundamental legal conceptions applied in judicial reasoning. In *Fundamental Legal Conceptions Applied in Judicial Reasoning and Other Legal Essays*, pages 23–64. Yale, 1923.

[5] Joris Hulstijn. Need to know: Questions and the paradox of epistemic obligation. In R. van der Meyden and L. van der Torre, editors, *DEON 2008*, volume 5076 of *LNCS*, pages 125–139. Springer, 2008.

[6] D. Makinson and L van der Torre. Input/output logics. *Journal of Philosophical Logic*, 29(4):383–408, 2000.

[7] David Makinson. On the formal representation of rights relations: Remarks on the work of Stig Kanger and Lars Lindahl. *Journal of Philosophical Logic*, 15(4):403–425, November 1986.

[8] Réka Markovich. Understanding Hohfeld and Formalizing Legal Rights: the Hohfeldian Conceptions and Their Conditional Consequences. *Studia Logica*, 108, 2020.

[9] Réka Markovich and Olivier Roy. Formalizing the right to know: Epistemic rights as normative positions. In *Logics for New-Generation AI*, page 154, 2021.

[10] Eric Pacuit. *Neighborhood semantics for modal logic*. Springer, 2017.

[11] Marek Sergot. Normative positions. *Handbook of deontic logic and normative systems*, 1:353–406, 2013.

[12] Lani Watson. The Right to Know: Epistemic Rights and Why We Need Them. *Manuscript presented at the Edinburgh Legal Theory Group, 24 October 2019*, 2019.

[13] Lani Watson. *The Right to Know. Epistemic Rights and Why We Need Them*. Routledge, 2021.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210341

Rationale Discovery and Explainable AI

Cor STEGING [a], Silja RENOOIJ [b] and Bart VERHEIJ [a]
[a] *Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen*
[b] *Department of Information and Computing Sciences, Utrecht University*

Abstract. The justification of an algorithm's outcomes is important in many domains, and in particular in the law. However, previous research has shown that machine learning systems can make the right decisions for the wrong reasons: despite high accuracies, not all of the conditions that define the domain of the training data are learned. In this study, we investigate what the system *does* learn, using state-of-the-art explainable AI techniques. With the use of SHAP and LIME, we are able to show which features impact the decision making process and how the impact changes with different distributions of the training data. However, our results also show that even high accuracy and good relevant feature detection are no guarantee for a sound rationale. Hence these state-of-the-art explainable AI techniques cannot be used to fully expose unsound rationales, further advocating the need for a separate method for rationale evaluation.

Keywords. Machine Learning, Explainable AI, Knowledge, Data

1. Introduction

Explanations are essential in AI and law [1,2]. Not only is argumentative reason-based discussion inherent to legal reasoning, but both parties in a court of law also have the right to an explanation when a decision is made [3]. In brief, proper justice requires decisions based on sound rationales. Various types of explanation [4] have been applied in the field of AI and law, ranging from contrastive explanations [5,6,7] to selective explanations [8, 9], and from probabilistic explanations [10] to social explanations [11,12,8].

Many state-of-the-art AI systems, however, are black-box systems that reason without transparency. As long as they cannot explain their decision making, they are inherently unsuitable in the context of AI & law. This is unfortunate, as the performance of in particular deep learning systems is often second to none when it comes to tasks such as image, speech or text classification. The subfield of Explainable AI (XAI) aims to bridge the gap that black-box machine learning systems have created, by providing explanations for these opaque systems [13]. Methods such as LIME [14] and SHAP [15] allow us to look 'inside' the black-box by demonstrating which parts of the input are important in the system's decision making process. In a recent study, lawyers tasked with assessing both LIME and SHAP in a legal text classification task graded both explanation methods similarly, and said to look forward to more explainable systems to assist their work [16]. Other recent research on the use of machine learning in law, on the other hand, emphasizes that explaining the outcome using a list of important input features is insufficient in a legal setting

[17]. In addition to instilling trust in the end user, explainable AI can also expose unsound decision making. In the LIME paper, for example, a husky is misclassified as a wolf because there is snow in the background of the image [14]. In adversarial attacks, small perturbations to an image, imperceptible to the human eye, can cause drastic changes in a model's prediction [18]. These systems had high accuracy scores, but they performed well for the wrong reasons, as their decision making was unsound.

Based on work by Bench-Capon in AI & Law [19], we introduced a preliminary method for rationale evaluation [20]. This method tests the decision making of a system and evaluates to what extent the learned rationale of the system is sound, given high accuracy scores. This method was applied in a set of experiments dealing with legal domains, and it was shown that high accuracies are by no means a guarantee for a sound rationale [21]. Different studies using the same domain and datasets showed similar results when it comes to learning the complete rationale [22,23]. Knowing whether a system learned what it is supposed to learn is essential in pursuing responsible AI. In this paper we investigate whether state-of-the art explainable AI techniques provide relevant insight into the evaluation of the rationale of a machine learning system. In particular, we will use SHAP [15] and LIME [14] to extract explanations from neural networks trained on the fictional welfare benefit domain [19]. We then compare these explanations to the results of the rationale evaluation from previous research.

2. Background

The domain used in both this study and in previous studies [20,22,23] is the welfare benefit domain, as introduced by Bench-Capon [19]. It defines a fictional set of conditions, that must all be satisfied in order for a pensioner to receive benefits for visiting their spouse in the hospital. Eligibility for a welfare benefit depends on six conditions and is formalized as follows:

$$
\begin{aligned}
&Eligible(x) \iff C_1(x) \wedge C_2(x) \wedge C_3(x) \wedge C_4(x) \wedge C_5(x) \wedge C_6(x)\\
&C_1(x) \iff (Gender(x) = female \wedge Age(x) \geq 60) \vee\\
&\qquad\quad (Gender(x) = male \wedge Age(x) \geq 65)\\
&C_2(x) \iff |Con_1(x), Con_2(x), Con_3(x), Con_4(x), Con_5(x)| \geq 4\\
&C_3(x) \iff Spouse(x)\\
&C_4(x) \iff \neg Absent(x)\\
&C_5(x) \iff \neg Resources(x) \geq 3000\\
&C_6(x) \iff (Type(x) = in \wedge Distance(x) < 50) \vee (Type(x) = out \wedge Distance(x) \geq 50)
\end{aligned}
$$

In this domain a pensioner is therefore eligible for a benefit iff he or she is of pensionable age (60 for a woman, 65 for a man), has paid four out of the last five contributions Con_i, is the spouse of the patient, is not absent from the UK, does not have capital resources amounting to more than £3,000, and lives at a distance of less than 50 miles from the hospital if the relative is an *in*-patient, or beyond that for an *out*-patient.

Artificial datasets were generated based on this domain and used to train neural networks. These networks are then tasked with classifying new, unseen instances from the welfare benefit domain. In addition to measuring the performance of the system in terms of accuracy, the main interest lies in evaluating the rationale that the networks have learned; were they able to internalize the six conditions that define eligibility? To answer

that question, a preliminary method for rationale evaluation was introduced and applied to the welfare benefit domain [20]. The rationale evaluation method prescribes designing dedicated test sets that target specific elements of the desired rationale, based on expert knowledge of the domain. Learning systems can only perform well on these dedicated test sets if they have learned a particular element of the rationale.

2.1. Datasets

This study builds on the same datasets used in previous work [19,21], generated from the welfare benefit domain[1]. Every dataset contains 64 features, made up of the 12 features from the domain and 52 noise features. We use both relatively small training datasets of 2,400 instances and larger datasets that consist of 50,000 instances.

Two types of datasets are used: type A and type B. Both types of datasets have a balanced label distribution, wherein 50% of the instances are eligible, satisfying *all* six conditions, and 50% are ineligible. Eligible instances are generated randomly from uniform distributions in both type A and type B datasets, such that all six conditions are satisfied. In type A datasets, ineligible instances are generated such that an equal number of them fail on each condition. In other words, each condition is responsible for the ineligibility of an equal number of instances. The values of the remaining features are generated completely randomly from a uniform distribution. As a result, multiple conditions can be unsatisfied if an instance is ineligible. Type A datasets are the most realistic dataset, since very little of the distribution is controlled, just as in non-artificial datasets. In type B datasets, ineligible instances only fail due to a single condition. It is therefore not possible to have multiple unsatisfied conditions in a type B dataset. For this more controlled version of the domain, it is harder for a network to achieve high performance scores, since it has to learn each condition individually. In a type A dataset, it is possible to achieve a theoretical accuracy of 98.95% while only knowing four out of the six conditions that define the domain [19], because ineligible instances almost always fail on multiple conditions. This is not possible in type B datasets.

The method for rationale evaluation prescribes the design of dedicated test sets for rationale evaluation, that target specific components of the rationale based on expert knowledge of the domain. For the welfare benefit domain, to measure how well any given condition C_i is learned, a dedicated test set is created in which every condition is satisfied except for C_i. The values of the features that define C_i are then generated across their full range of values. That way, the eligibility in this dedicated test set is determined solely by condition C_i: condition C_i is satisfied iff the instance is eligible. Networks are only able to classify the instances from this dedicated test set correctly if they have learned condition C_i. The particular components of the rationale that we investigate are the Age-Gender condition (condition C_1) and the Patient-Distance condition (condition C_6). The dedicated test sets to evaluate these conditions are referred to as the Age-Gender and Patient-Distance datasets, respectively.

[1]The datasets and the Jupyter notebooks used for data generation can be found in a Github repository: https://github.com/CorSteging/DiscoveringTheRationaleOfDecisions

Table 1. The accuracies obtained by the neural networks in the welfare benefit domain.

	Test set A	Test set B	Age-Gender	Patient-Distance
Trained on A (2,400)	98.97±0.19	72.39±1.66	52.14±4.01	50.05±0.09
Trained on B (2,400)	96.13±0.66	90.51±1.25	86.4±1.33	85.77±5.21
Trained on A (50,000)	99.8±0.03	80.98±1.47	60.22±3.87	64.44±2.87
Trained on B (50,000)	99.64±0.17	98.53±0.15	98.51±0.47	97.17±0.46

2.2. Neural networks

Multilayer perceptrons (MLP) with a single, two and three hidden layers were used, as in the initial study [19]. The goal of the study was not to achieve the highest accuracies, but rather to investigate and evaluate the learned rationale. Since MLP's are simple, black-box systems, they are ideal candidates for research into rationale evaluation. The details regarding the architecture and parameters of the networks can be found in [20]. Each of the three networks is trained separately on both type A and type B datasets and evaluated using separate type A, type B, Age-Gender and Patient-Distance datasets.

2.3. Previous results

Table 1 shows the accuracies of the neural networks with a single hidden layer from previous research, trained and tested on the various datasets. Networks with two or three hidden layers scored similarly. When training on a type A dataset, accuracies on the 'normal' test set A are high, while performance on test set B, the Age-Gender and Patient-Distance dataset is much lower. This shows that the correct rationale has not been learned, since a correct rationale would lead to high performance on all data sets. It is therefore clear that high accuracies are no guarantee for a sound rationale. When training on a type B dataset, the accuracies remain high on type A test sets, but improve drastically on test set B, the Age-Gender and the Patient-Distance dataset; training on a type B dataset therefore yields a better learned rationale. The table shows that with more data, the same pattern can be observed, while training on a type B test set then leads to even better accuracy scores on all data sets.

3. Explainable AI

3.1. Previous results

Our current objective is to investigate whether explainable AI techniques can be used to discover the unsound rationale as it is actually used by the trained system. Earlier research attempted to discover what rules a network trained on this domain had learned by inverting the trained network, in the sense that the output node node becomes the input node and the input nodes become the output node [19]. Passing a value of 1 through the new 'input' node would yield a list of features and their relevance. Since all of the features were normalized between 0 an 1, relevance was described as its deviation from 0.5.

However, this approach cannot account for the impact that a combination of features has, as pointed out in the original study [19]. A clear example of the shortcomings of this method is the Patient-Distance condition C_6, which is a variation of a XOR problem.

A XOR function yields true if and only if exactly one of its two variables is true and the other is false. If the first variable is true, the output will therefore be true in 50% of the cases (whenever the other variable is false). In the remaining 50% of the cases, the output will be false (whenever the other variable is true). Applying the method of inverting the network to a network that can solve a XOR problem would therefore yield an output value of 0.5 for both features, since each feature yield true (1) half of the time and false (0) in the other cases. This result provides little insight into the structure of a XOR problem, since the two features depend on each other and are meaningless on their own. Inverting the network does not account for such combinations of features, and is therefore unsuitable for our current objective.

3.2. State of the art: SHAP & LIME

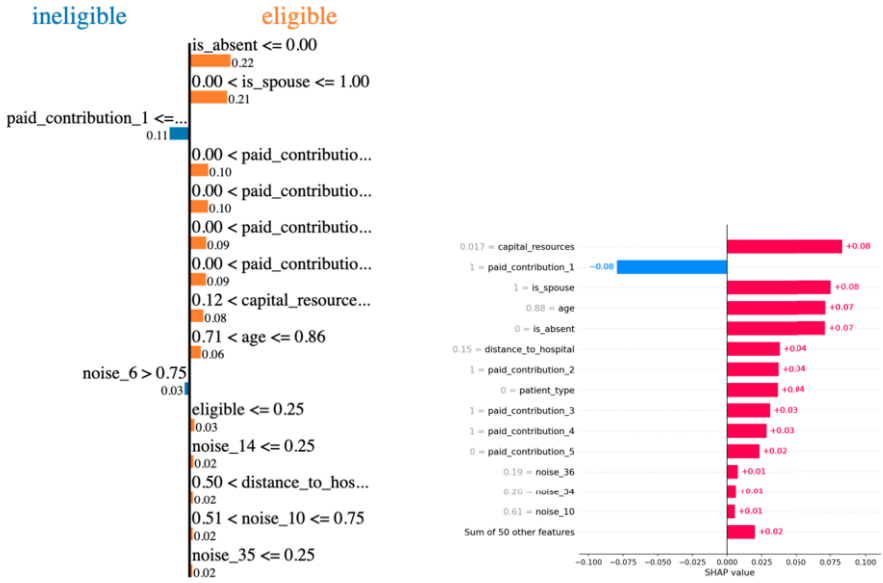
Since the earlier research [19] much progress has been made in the availability of explainable AI methods. Hence now we extend the previous research by investigating the trained neural networks using modern, state-of-the-art explainable AI techniques, to see whether these methods allow us to evaluate the soundness of the rationale. In particular, we will use SHAP [15] and LIME [14], two of the most commonly used explainable AI methods. SHAP (SHapley Additive exPlanation) is an explainable AI framework, that explains the output of a machine learning system based on the idea of Shapley values from game theory. LIME creates explanations by perturbing individual instances and using those to learn interpretable sparse linear models that approximate the system's decision making.

These two explanation methods were chosen due to their inherent fidelity [14] to the decision making of the black-box model, meaning that their explanations should accurately reflect the opaque decision making of the model. Since our aim is to investigate what the model has actually learned, fidelity is the most important criterion. Though it is often impossible to create completely faithful explanations, local fidelity (how a model responds to a given instance) can be achieved. Both LIME and SHAP are local explanation methods that provide an explanation for the output produced given a single input instance. Additionally, SHAP includes methods to aggregate a set of local explanations into a global interpretation of a system's decision making. Both methods are additive methods, meaning that summing up the effects of all feature attributions should approximate the prediction of the network. We will apply both LIME and SHAP to our trained networks, in order to ensure that our results are explainer-independent.

The same three neural networks mentioned in Section 2.3 are trained on type A and type B datasets separately, using both 2,400 instances and 50,000 instances. LIME and SHAP are then used to extract explanations from networks using a test set of 500 instances, sampled from a separate type A testing dataset.

Table 2. Example instance

Age	Gender	Con_1	Con_2	Con_3	Con_4	Con_5	Spouse	Absent	Resources	Type	Distance
84	female	0	1	1	1	1	1	0	1569	out	74



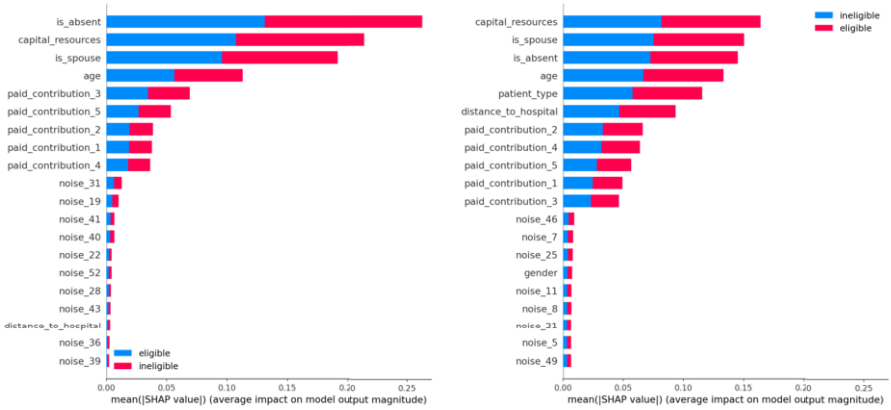
(a) LIME (b) SHAP

Figure 1. LIME and SHAP bar plots of the network trained on large type A dataset, displaying the impact of each feature in the classification process towards the 'ineligible' label (blue) or the 'eligible' label (orange/red) of the instance in Table 2.

4. Results

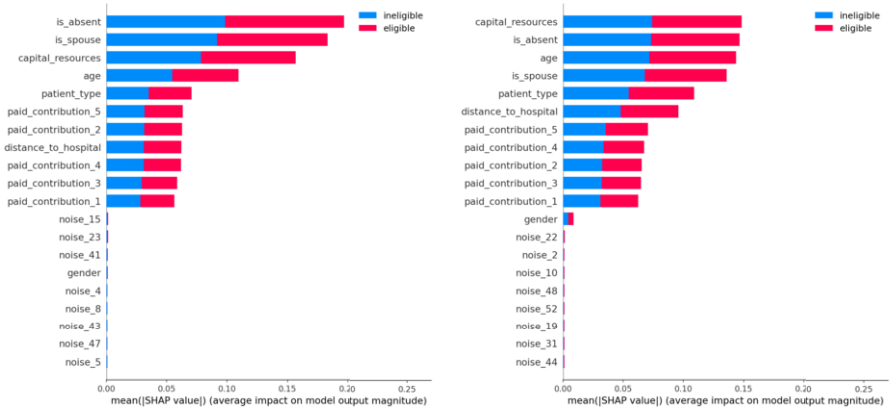
To illustrate the explanation methods, we first examine the classification process and its explanations using the example instance found in Table 2. In this example, all six conditions are satisfied and therefore the instance is eligible for a welfare benefit. The network with a single hidden layer trained on a dataset of 50,000 type A instances correctly predicts this. Figure 1 shows the LIME and SHAP explanations of these networks for the given instance. The bar plots display the impact that each feature had on classifying that instance, ranked from highest to lowest. Blue bars on the left contribute to the 'ineligible' label, whereas the orange and red bars on the right contribute to an 'eligible' label. Running the same experiment using networks with multiple hidden layers showed similar results and are therefore omitted. Note that even though each instance has 64 features, only the top 15 features with the highest impact are shown.

We now consider the explanations for the entire test set of 500 instances to illustrate the global interpretations of the networks. Since LIME does not possess a method for aggregating sets of explanations, we will use SHAP's summary plots. To ensure that the results are explainer-independent, we also examined 10 LIME cases for each scenario (each combination of network, training data type and size; 120 in total). The LIME results do not contradict any of the SHAP results and support its global interpretations of the networks. Figure 2 shows the average SHAP values for networks with a single hidden layer, trained on both training set A (left) and training set B (right) for training sets with 2,400 (top) and 50,000 instances (bottom). As mentioned in Section 2.3, experiments using networks with more hidden layers yielded similar results and are therefore omitted.



(a) Trained on A with 2,400 instances

(b) Trained on B with 2,400 instances



(c) Trained on A with 50,000 instances

(d) Trained on B with 50,000 instances

Figure 2. SHAP bar plots of the network trained on various datasets, displaying the impact of each feature in the classification process towards the 'ineligible' label (blue) or the 'eligible' label (red).

These bar plots display the top 20 features that have the highest average impact on the classification process, ranked from highest to lowest.

5. Discussion

If the networks have learned the correct rationale, the relevant features should have a high impact in Figures 1 and 2. These relevant features are the age, gender, contributions, spouse, absent, resources, distance and patient-type features, since they determine eligibility. The 52 noise variables should have a low impact as they do not contribute anything to the desired outcome.

The example instance in Table 2 satisfies all conditions and is therefore eligible. Since the network correctly predicted its eligibility label, we would expect to see that all relevant features in the SHAP and LIME plots have a high impact, whereas the noise features would have a low impact. In the plots of Figure 1 we indeed see that most rel-

evant features have been attributed a high impact. Noticeable exceptions are, however, the gender, patient-type and distance features. In this particular example, gender is irrelevant, as the person is older (88) than the threshold for both males (65) and females (60). The patient-distance condition is difficult to learn, as made evident from Table 1, which can account for the low impact values of these features. In the example instance, only one feature had a noticeable impact towards the 'ineligible' label, rather than the 'eligible' label. This is the first paid contribution feature, which makes sense, as the first contribution was not paid in this case (see Table 2) and would thus act as evidence against eligibility.

Since each example instance represents a different case, it is evaluated differently in terms of feature importance. To get a broader evaluation of the reasoning of the networks we therefore examine the global interpretations. The SHAP results from Figure 2a show us the impact of each feature in the classification process of a network trained on a type A dataset with 2,400 instances. We can see that most of the relevant features are listed with a high impact, whereas the noise features have a low impact. Given the large amount of noise variables (52 noise variables versus 12 relevant variables), discovering the relevant features is a non-trivial outcome, supporting the value of SHAP. However, the gender and patient-type features are missing from the plot, and the distance feature has an impact value similar to those of the noise features. This supports our previous findings (see Table 1) that the correct rationale was not learned after training on a smaller type A dataset [21]. SHAP has thus managed to detect the unsound rationale: some features that are relevant in the domain do not have high impact in the trained network's decision process.

The rationale improves when training on a type B dataset of the same size. Figure 2b supports this observation: all of the 12 relevant features are deemed to be highly impactful in the classification process by SHAP. The exception here is the gender feature, which is given a relatively low SHAP value. This finding fits the fact that in our domain the gender feature is only important in a very small subset of cases (males between the age of 60 and 65). From Table 1 we know that the conditions were not learned perfectly, and the low impact assigned to the age variable reveals a possible reason for this observation.

When training on a larger type A dataset, accuracies improve though the rationale is still not sound, as evident from the low accuracy scores on the Age-Gender and Patient-Distance datasets (see Table 1). The SHAP plot of this network, as shown in Figure 2c, displays lower impact values for the noise variables compared to the plot in Figure 2a. Moreover, SHAP now assigns high impact values to the patient-type and distance features, which it did not do when training on a smaller type A dataset. The high impact scores on patient-type and distance shows that the network has discovered the relevance of these features, which could explain the increase in performance on the Patient-Distance dataset when training on a larger type A dataset when compared to a smaller type A dataset (64.44% versus 50.05% as seen in Table 1). The system seemingly makes use of the patient-type and distance features, hence the high impact value returned by SHAP for these features, but it does not use them correctly as in the domain representation. In this case, therefore, SHAP was not able to detect the unsound rationale.

The impact scores of the networks trained on a larger type B dataset are similar to those of the networks trained on a smaller type A dataset (Figure 2d). The impact of the noise features is smaller than on a smaller type B dataset, however, and the gender feature now has a relatively higher impact than the noise features. Based on the four graphs

in Figure 2, the networks trained on a larger type B dataset provide the most desirable impact distribution, with the highest impact for relevant features and the lowest impact for the noise features. This is supported by previous results, as networks trained on larger type B datasets provided the highest accuracies on all datasets (Table 1), suggesting the most sound rationale.

Summarizing these results, we find that the unsound rationale we discoverd previously using the rationale evaluation method, is clearly exposed in the SHAP values for networks trained on less, and perhaps insufficient, data points (Figure 2a). This suggests that explainable AI methods can be used to evaluate the rationale of trained systems to some extent. When training on more data points, however, the unsoundness of the rationale is not as clearly exposed. Figure 2c assigns high impact values to all of the relevant features (except for gender, though its small impact can be accounted for). This makes it seem as if the rationale is sound, whereas the method for rationale evaluation has shown that it is not (see Table 1). Based solely on the SHAP and LIME explanations, we would not be able to know that the rationale is unsound for the networks trained on a large type A dataset. Therefore, even though the XAI methods can expose a faulty rationale, they cannot guarantee a sound rationale. Previously research claimed that it is possible to make the right decisions without knowing why [19,20]. These results expand upon that notion, and suggest that it is also possible to know what is important without knowing why. In other words, we find that systems that have both a high accuracy and assign high importance to the correct features are still not guaranteed to use a sound rationale.

6. Conclusion

Previous research [20,21] has shown through a method for rationale evaluation that learning systems can achieve high accuracies without having a sound rationale. The current study set out to investigate the rationale actually used by a trained system using explainable AI (XAI) methods. Earlier research [19] used a preliminary XAI method with the aim of discovering the rules of a neural network with limited success, as rules with non-straightforward combinations of factors are difficult to discover. In the present study, we used modern XAI techniques (SHAP [15] and LIME [14]) to identify the features of the dataset with the largest impact on the classification process. Though the exact relationship between the features is still unclear, the explanation methods exposed the unsound rationales of some of the systems, reaffirming earlier findings [19,20] using a state-of-the-art explanation method. This suggests that XAI techniques can be used to evaluate the rationale of a system. However, for some conditions, the XAI methods did *not* detect an unsound rationale when the rationale was known to be unsound. Neither high accuracies, nor acceptable explanations from XAI techniques therefore can guarantee a sound rationale. This finding further supports the need for a rationale-evaluation method in order to obtain responsible AI.

Acknowledgements

This research was funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.

References

[1] Verheij B. Artificial intelligence as law. Artificial Intelligence and Law. 2020;28(2):181-206.
[2] Atkinson K, Bench-Capon T, Bollegala D. Explanation in AI and law: Past, present and future. Artificial Intelligence. 2020;289.
[3] Doshi-Velez F, Kortz M, Budish R, Bavitz C, Gershman C, O'Brien D, et al. Accountability of AI under the law: The role of explanation. SSRN Electronic Journal. 2017 November.
[4] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence. 2019;267:1-38.
[5] Rissland EL, Ashley KD. A case-based system for trade secrets law. In: Proceedings of the 1st International Conference on Artificial Intelligence and Law. ICAIL '87. New York, NY, USA: ACM; 1987. p. 60-6.
[6] Ashley KD. Modeling Legal Arguments: Reasoning with Cases and Hypotheticals. Cambridge (Massachusetts): The MIT Press; 1990.
[7] Verheij B. Artificial Argument Assistants for Defeasible Argumentation. Artificial Intelligence. 2003;150(1–2):291-324.
[8] Atkinson K, Bench-Capon T, Bex F, Gordon TF, Prakken H, Sartor G, et al. In memoriam Douglas N. Walton: the influence of Doug Walton on AI and law. Artificial Intelligence and Law. 2020:1-46.
[9] Verheij B. Dialectical argumentation with argumentation schemes: An approach to legal logic. Artificial intelligence and Law. 2003;11(2-3):167-95.
[10] Vlek CS, Prakken H, Renooij S, Verheij B. A method for explaining Bayesian networks for legal evidence with scenarios. Artificial Intelligence and Law. 2016;24(3):285-324.
[11] Hage JC, Leenes R, Lodder AR. Hard cases: a procedural approach. Artificial Intelligence and Law. 1993;2(2):113-67.
[12] Gordon TF. The Pleadings Game: An Artificial Intelligence Model of Procedural Justice. Dordrecht: Kluwer; 1995.
[13] Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—Explainable artificial intelligence. Science Robotics. 2019;4(37).
[14] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA; 2016. p. 1135-44.
[15] Lundberg SM, Lee S. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 4765-74.
[16] Górski Ł, Ramakrishna S. Explainable artificial intelligence, lawyer's perspective. In: Proceedings of the 18th International Conference on Artificial Intelligence and Law. ICAIL '21; 2021. p. 60-8.
[17] Mumford J, Atkinson K, Bench-Capon T. Machine learning and legal argument. In: Proceedings of the 21st Workshop on Computational Models of Natural Argument; 2021. p. 47-56.
[18] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the International Conference on Learning Representations; 2015. .
[19] Bench-Capon T. Neural networks and open texture. In: Proceedings of the 4th International Conference on Artificial Intelligence and Law. ICAIL 1993. ACM, New York; 1993. p. 292-7.
[20] Steging C, Renooij S, Verheij B. Discovering the rationale of decisions: Towards a method for aligning Learning and reasoning. In: Proceedings of the 18th International Conference on Artificial Intelligence and Law. ICAIL '21. ACM, New York; 2021. p. 235-9.
[21] Steging C, Renooij S, Verheij B. Discovering the rationale of decisions: Experiments on aligning Learning and reasoning. In: The Explainable & Responsible AI in Law (XAILA) Workshop; 2021. .
[22] Možina M, Žabkar J, Bench-Capon T, Bratko I. Argument based machine learning applied to law. Artificial Intelligence and Law. 2005;13(1):53-73.
[23] Wardeh M, Bench-Capon T, Coenen F. Padua: a protocol for argumentation dialogue using association rules. Artificial Intelligence and Law. 2009;17(3):183-215.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210342

A Survey on Methods and Metrics for the Assessment of Explainability Under the Proposed AI Act

Francesco SOVRANO [a] and Salvatore SAPIENZA [b] and Monica PALMIRANI [b] and Fabio VITALI [a]
[a] *University of Bologna, DISI*
[b] *University of Bologna, CIRSFID-ALMA AI*

Abstract. This study discusses the interplay between metrics used to measure the explainability of the AI systems and the proposed EU Artificial Intelligence Act. A standardisation process is ongoing: several entities (e.g. ISO) and scholars are discussing how to design systems that are compliant with the forthcoming Act and explainability metrics play a significant role. This study identifies the requirements that such a metric should possess to ease compliance with the AI Act. It does so according to an interdisciplinary approach, i.e. by departing from the philosophical concept of explainability and discussing some metrics proposed by scholars and standardisation entities through the lenses of the explainability obligations set by the proposed AI Act. Our analysis proposes that metrics to measure the kind of explainability endorsed by the proposed AI Act shall be *risk-focused*, *model-agnostic*, *goal-aware*, *intelligible & accessible*. This is why we discuss the extent to which these requirements are met by the metrics currently under discussion.

Keywords. Explainable Artificial Intelligence, Explainability, Metrics, Standardisation, Artificial Intelligence Act

1. Introduction

The ability and need of humans to explain has been studied for centuries, initially in philosophy and more recently also in all those sciences aiming at a better understanding of (human) intelligence. Measuring the degree of explainability of AI systems has become relevant in the light of research progress in the eXplainable AI (XAI) field, the proposal for an EU Regulation on Artificial Intelligence, and ongoing standardisation initiatives that will translate these technical advancements in a *de facto* regulatory standard for AI systems. To date, standardisation entities have proposed white papers and preliminary documents showing their progress[1], among them we mention: the European Telecommunications Standards Institute (ETSI), the CEN-CENELEC, and ISO/IEC TR 24028:2020(E), stating that '[i]t is important also to consider the measurement of the quality of explanations' and provides for details on the key measurements (i.e. continuity, consistency, selectivity; paras 9.3.6, 9.3.7).

Considering that, since ISO/IEC TR 24028:2020(E), the literature has started to propose new metrics and mechanisms, with this work we study and categorise the existing approaches to quantitatively assess the quality of explainability in Machine Learning and AI. We do so through the lenses of law and philosophy, not just computer science. This last characteristic is certainly our main contribution to the literature of XAI and Law, and

[1] An extensive list of examples is available at https://joinup.ec.europa.eu/collection/rolling-plan-ict-standardisation/artificial-intelligence

we believe it may foster future research to embrace an interdisciplinary approach less timidly, for the sake of a better conformity to existing (and new) regulations in the EU landscape.

This paper is structured as follows. In Section 2 and 3 we present the research background and the methodology of this paper. Then in Section 4, 5 and 6, we explore the definitions and properties of explainability in philosophy and in the proposed AI Act. Finally, in Section 7 and 8 we perform an analysis of the existing quantitative metrics of explainability, discussing our findings and future research.

2. Related Work

In XAI's literature there are many interesting surveys on explainability techniques [1,2,3,4], classifying algorithms on different dimensions to help researchers in finding the more appropriate ones for their own work. Practically, all these surveys focus on a classification of the mechanisms to achieve explainability rather than how to measure the quality of it, and we believe our work can help in this latter.

For example [1] classify XAI methods with respect to the notion of explanation and the type of black-box system. The identified characteristics are respectively the level-of-detail of explainability (from high to low: global logic, local decision logic, model properties) and the level of interpretability of the original model. Similarly to [1], also [2] study XAI considering interpretability and level-of-detail.

On the other hand, [4] focus specifically on the metrics to quantify the quality of explanation methods, classifying them according to the properties they can measure and the format of explanations (model-based, attribution-based, example-based) they support. More precisely, [4] narrow down the survey to the functionality-grounded metrics, proposing for them a new taxonomy including interpretability (in terms of clarity, broadness, and parsimony) and fidelity (as completeness, and soundness).

Among all the identified surveys, [4] is certainly the closest to our work, in terms of focus of the survey. The main distinction between our work and [4] is probably the assumption we do that multiple definitions of explainability exist, each one possibly requiring its own type of metrics. Furthermore, differently from [4], we analyse explainability metrics on their ability to meet the requirements set by the AI Act.

3. Methodology

We performed an exploratory literature review of existing metrics to measure the explainability of AI-related explanations, together with a qualitative legal analysis of the explainability requirements to understand the alignment of the identified metrics to the expectations of the proposed AI Act. To do so, we collected all the papers cited in [4], re-classifying them. Then we integrated with further works identified through an in depth keyword-based research[2] on Google Scholars, Scopus, and Web Of Science. On the other hand, the legal analysis was carried out on the proposed Artificial Intelligence Act. Considering the lack of case law and the paucity of studies on this novel piece of legislation, a literal assessment of its provisions has been preferred to more critical analysis based on previous enquiries.

4. Definitions of Explainability

Considering the definition of "explainability" as "the potential of information to be used for explaining", we envisage that a proper understanding of how to measure explainability must pass through a thorough definition of what constitutes an explanation and the act of explaining.

[2]The main keywords we used were "degree of explainability", "explainability metrics", "explainability measures", and "evaluation metrics for contrastive explanations".

Table 1. Definitions of *explanation* and *explainable information* for each theory of explanations.

Theory	Def. of Explanation	Def. of Explainable Information
Causal Realism [5]	It is a description of causality, as chains of causes and effects.	It can fully describe causality.
Constructive Empiricism [6]	It is contrastive information answering WHY questions, allowing one to calculate the probability of a particular event relative to a set of (possibly subjective) background assumptions.	It provides answers to contrastive WHY questions.
Ordinary Language Philosophy [7]	Explaining is pragmatically answering to (not just WHY) questions, with the explicit intent of producing understanding.	It can be used to pertinently answer questions about relevant aspects, in an illocutionary way.
Cognitive Science [8]	Explaining is a process triggered as response to predictive failures and it is about providing information to fix that failures in a mental model (sometimes intended as a hierarchy of rules).	It can fix failures in mental models.
Naturalism and Scientific Realism [9]	Explaining is an iterative process of confirmation of truth based on inference to the best explanation. An explanation increases understanding, not simply by being the correct answer to a particular question, but by increasing the coherence of an entire belief system (e.g. a subject).	It can be used to increase understanding, i.e. by answering to particular questions.

In 1948 Hempel and Oppenheim published their "Studies in the Logic of Explanation" [10], giving birth to what it is considered the first theory of explanation, the deductive-nomological model. After that date, many attempts followed to amend, extend or replace this first model, which is considered fatally flawed [11,5]. This gave birth to several competing and more contemporary theories of explanations [12]: i) Causal Realism, ii) Constructive Empiricism, iii) Ordinary Language Philosophy, iv) Cognitive Science, v) Naturalism and Scientific Realism. A summary of these definitions is shown in Table 1.

Interestingly, each one of these theories devises different definitions of "explanation". If we look at their specific characteristics we may find that all but *Causal Realism* are pragmatic. On the other hand, *Causal Realism* and *Constructive Empiricism* are rooted on causality, while the others not [3]. Nonetheless, *Cognitive Science* and *Scientific Realism* are more focused on the effects that an explanation has on the explainee (the recipient of the explanation).

Importantly, with the present letter, we assert that whenever explaining is considered to be a pragmatic act, explainability differs from explaining. In fact, pragmatism in this sense is achieved when the explanation is tailored to the specific user, so that the same explainable information can be presented and re-elaborated differently across users. It follows that for each philosophical tradition, but Causal Realism, we have a definition of "explainable information" that slightly differs from that of "explanation", as shown in Table 1.

5. Explainability Desiderata

In philosophy, the most important work about the central criteria of adequacy of *explainable information* is likely to be Carnap's [13]. Even though Carnap studies the concept of *explication* rather than that of *explainable information*, we assert that they share a common ground making his criteria fitting in both cases. In fact, *explication* in Carnap's sense is the replacement of a somewhat unclear and inexact concept (the explicandum) by a new, clearer, and more exact concept called explicatum, and that is exactly what information does when made explainable.

Carnap's central criteria of explication adequacy are [13]: *similarity*, *exactness* and *fruitfulness*[4]. *Similarity* means that the explicatum should be similar to the explicandum, in the sense that at least many of its intended uses, brought out in the clarification step,

[3] They study the act of explaining as an iterative process involving broader forms of question answering

[4] Carnap also discussed another desideratum, *simplicity*, but this criterion is presented as being subordinate to the others.

are preserved in the explicatum. On the other hand, *Exactness* means that the explication should, where possible, be embedded in some sufficiently clear and exact linguistic framework. While *Fruitfulness* means that the explicatum should be used in a high number of other *good* explanations (the more, the better).

Carnap's adequacy criteria seem to be transversal to all the identified definitions of explainability, possessing preliminary characteristics for any piece of information to be considered properly explainable. Therefore, our interpretation of Carnap's criteria in terms of measurements is the following.

- *Similarity* is about measuring how much *similar* the given information is to the explanandum. This can be estimated by counting the number of *relevant aspects* covered by information and the *amount of details* it can provide.
- *Exactness* is about measuring how clear the given information is, in terms of pertinence and syntax, regardless its truth. Differently from Carnap, our understanding of *exactness* is broader than that of adherence to standards of formal concept formation [14].
- *Fruitfulness* is about measuring how much a given piece of information is going to be used in the generation of explanations. Consequently, each one of the explainability definitions may define *fruitfulness* differently.

Importantly, the property of *truthfulness* (being different from *exactness*) is not explicitly mentioned in Carnap's desiderata. That is to say that explainability and *truthfulness* are complementary, but different, as discussed also by [15]. In fact an explanation is such regardless its truth (wrong but high-quality explanations exist, especially in science). Vice-versa, highly correct information can be very poorly explainable.

6. Explainability Obligations in the Proposed AI Act

The discussion towards "explainability and law" has departed from the contested existence of a right to explanation in the General Data Protection Regulation (GDPR) [16,17,18][5] to embrace contract, tort, banking law [19], and judicial proceedings [20]. This previous discussion focusing on legal regimes other than the AIA - yet, highly connected - constitutes a valuable background for our research. Our focus, however, shall be confined to the interaction between the nuance of explainability and obligations emerging from the Artificial Intelligence Act (AIA) already identified by these early commentators. Then, the discussion identified a "technical" necessity of explainability, that is necessary to improve the accuracy of the model. In legal terms, it is echoed by the "protective" transparency that is needed to minimise risks and comply with certain legal regimes (tort law and contractual obligations). As with data protection law, these varieties are instrumental to improve a product and protect its users or the persons affected by the system from damages. If explainability is often instrumental to achieve some legislative goals, it is likely that it could be meant to foster certain regulatory purposes also under the AIA. From the joint reading of a series of provisions, it will be argued that explainability in the AIA is both *user-empowering* and *compliance-oriented*: on the one hand, it serves to enable users of the AI system to use it correctly; on the other hand, it helps to verify adequacy to the many obligations set by the AIA.

[5]Explanations, including contractual ones [17] are deemed to be 'right-enabling' [19] as they is necessary and instrumental to exercises the rights enshrined in Article 22 of the GDPR, namely to express views on the decision and to contest it. The same goes with the kind of transparency that is necessary to ensure the right to a fair trial in the context of judicial decision-making [20]. Indeed, Case law on explanations is progressively becoming significant: scholars have referred to the Risk Indication System (*SyRI*) case decided by The Hague District Court in 2020 [20] on the transparency in fraud prevention systems, Case n. 8472/2019 by the Italian Consiglio di Stato concerning the allocation of teachers in public schools across the country, and the German Federal Court for Private Law BGH, Case VI ZR 156/13 = MMR 2014 on the right to access to personal data [19]

Recital 47 and art. 13(1) state that high-risk AI systems shall be designed and developed in such a way that their operation is comprehensible by the users. They should be able a) to interpret the system's output and b) to use it in an appropriate manner. This is a form of *user-empowering* explainability. Then, the second part of Art. 13 specifies that "an appropriate type and degree of transparency shall be ensured, with a view to *achieving compliance* (emphasis added) with the relevant obligations of the user and of the provider [...]". In our reading, this provision specifies that this explainability obligations (i.e. transparent design and development of high-risk AI systems) is *compliance-oriented*. [6].

Such compliance-oriented explainability becomes evident in the technical documentation to be provided according to Art. 11. Compliance is based on a presumption of safety if the system is designed according to technical standards (Art. 40) to which adherence is documented, whereas third-party assessment appears only post-market or on specific sectors (Chapter IV). The contents of the dossier are those detailed by Annex IV. *Inter alia*, Annex IV(2)(b) include "the design specifications of the system, namely the general logic of the AI system and of the algorithms" among the information to be provided to show compliance with the AIA before placing the AI system in the market. Since the general approach taken by the proposed AIA is a risk-reduction mechanism (Recital 5), this form of explainability is ultimately meant to contribute to minimising the level of potential harmfulness of the system.

User-empowering and compliance-oriented explainability overlap in art. 29(4).When a risk is likely to arise, the user shall suspend the use of the system and inform the provider or the distributor. This provision entails the capability of understanding the working of the system (real-time) and making previsions on its output. Suspending in the case of likely risk is the overlapping between the two nuances of explainability: the user is empowered to stop the AI system to avoid contradicting the rationale behind the AIA, i.e. risk-minimisation.

Once clarified the existence of explainability obligations and their extent, let us discuss the requirements that metrics should have to ease compliance with the AIA. Let us remind that, under the proposal, adopting a standard means certifying the degree of explainability of a given AI system. Therefore, metrics become useful in the course of the standardisation process: i) *ex ante*, when defining the explainability measures adopted by the standard; ii) *ex post*, when verifying in practice the adoption of a standard.

From these premises it follows that, in the light of the purposes of the AIA, any explainability metric should be at minimum: i) *Risk-focused*, ii) *Model-agnostic*, iii) *Goal-Aware*, iv) and *Intelligible & accessible*.

Risk-focused means that the metric should be functional to measure the extent to which the explanations provided by the system allows for an assessment of the risks to the fundamental rights and freedoms of the persons affected by the system's output. This is necessary to ensure both user-enabling (e.g. art. 29) and compliance-oriented (Annex IV) explainability. While *Model-agnostic* means that the metric should be appropriate to all the AI systems regulated by the AIA[7].

Goal-aware means that the metric should be flexible towards the different needs of the potential explainees (i.e. AI system providers and users, standardisation entities, etc.)[8] and applicable in all the high-risk AI applications listed in Annex III. While *Intel-*

[6]The twofold goal of art. 13(1) is then echoed by other provisions. As regards the user-empowering interpretation, art. 14(4)(c) relates explainability to "human oversight" design obligations. These measures should enable the individual supervising the AI system to correctly interpret its output. Moreover, this interpretation shall put him or her in the position to decide whether it might be the case to "disregard, override or revers the output", art. 14(4)(d)

[7]Annex I provides a list of the AI techniques and approaches that fall within the remit of the Regulation.

[8]Since it might be hard to determine *ex ante* the nature, the purpose, and the expertise of the explainee, the metrics should consider the highest possible number of potential explainees.

Table 2. Comparison of different explainability metrics. The column "Metric" points to reference papers, while column "Name" points to the names used by the authors of the metric to describe it. Elements in bold are column-wise, indicating the best values.

Metric	Information Format	Supporting Theory	Subject - based	Covered Criteria	Name
[21]	Rule-based	Causal Realism	**No**	Similarity, Fruitfulness	Fidelity, Completeness
[22]	Feature Attribution	Causal Realism	**No**	Similarity, Fruitfulness	Monotonicity, Non-sensitivity, Effective Complexity
[23]	Rule-based	Causal Realism	**No**	**Similarity, Exactness, Fruitfulness**	Fidelity, Unambiguity, Interpretability, Interactivity
[24]	**All**	Causal Realism, Cognitive Science, Scientific Realism	Yes	Exactness, Fruitfulness	Causability
[25]	**All**	Cognitive Science, Scientific Realism	Yes	Exactness, Fruitfulness	Satisfaction, Trust, Mental Models, Curiosity, Performance
[26]	Example-based	Constructive Empiricism	**No**	Exactness	Proximity, Sparsity, Adequacy (Coverage)
[22]	Example-based	Constructive Empiricism	**No**	Similarity, Fruitfulness	Non-Representativeness, Diversity
[27]	Natural Language Text	Ordinary Language	**No**	**Similarity, Exactness, Fruitfulness**	Aspects Coverage, Degree of Explainability

ligible & accessible means that if information on the metrics is not accessible (e.g. due to intellectual property reasons) or the results of a metric are not reproducible (e.g. due to a subjective evaluation), explainees will confront with a situation of uncertainty, as an *ignotum per ignotius*. This would contradict the risk minimisation principle.

7. Discussing Existing Quantitative Measures of Explainability

In this section we identify some pros and cons of existing metrics (and measures) to quantitatively estimate the degree of explainability of information, with the aim of understanding their range of applicability across different needs and interpretations of explainability. We do it by performing a qualitative classification of these measures based on Carnap's desiderata, the theories of explanation presented in Section 4 and the main principles identified in Section 6.

More precisely, in Table 2 we classified the metrics on the following dimensions: the *format of information* supported by the metric (i.e. rule-based, example-based, natural language text, etc.); the *supporting theory of the metric* (i.e. cognitive science, constructive empiricism, etc.); *subjectivity* (whether the metric requires evaluations given by humans subjects); the *covered criteria of adequacy*. Then, in Table 3 we aligned the *supporting theories* (hence also the metrics) to the properties identified with the analysis of the AI Act carried out in Section 6.

Doing so, we considered only a part of the dimensions adopted by [4]. More precisely, we kept *clarity*, *broadness* and *completeness*, aligning the first two to Carnap's *exactness* and the latter to *similarity*. In fact, we deemed *soundness* to be as *truthfulness*, a complementary characteristic to explainability and not a characteristic of explainability, as discussed in Section 5. While *broadness* and *parsimony* were considered as characteristics to achieve pragmatic explanations rather than properties of explainability.

Furthermore, differently from ISO/IEC TR 24028:2020(E) we did not focus on metrics specific to ex-post *feature attribution* explanations, so we selected methods possibly

Table 3. Explainability definitions alignment to the properties identified in Section 6.

	Risk-Focused	Model-Agnostic	Goal-Aware	Intelligible & Accessible
Causal Realism	Yes, if understanding risks implies understanding causality	Not available yet	No, it's not pragmatic and it considers only goals related to causality	Yes, it can be
Constructive Empiricism	Yes, if explaining risks is about answering WHY questions	Not available yet	No, it focuses only on WHY questions	Yes, it can be
Ordinary Language Philosophy	Yes, it can be	Maybe. Only if all the explanations can be represented in a natural language	Yes	Yes, it can be
Cognitive Science	Yes, it can be	Yes, the evaluation is subject-based	Yes	Unlikely. All the subject-based metrics may be very expensive and hard to reproduce, this makes them less accessible
Naturalism and Scientific Realism	Yes, it can be	Yes, the evaluation is subject-based	Yes	Unlikely. It relies on (usually) expensive subject-based metrics

applicable also on ex-ante or more generic types of explanations.

As shown in Table 2, we were able to find at least one example of metric for each supporting philosophical theory , with a majority of metrics focused on Causal Realism and Cognitive Science. What is common to all the metrics based on Cognitive Science is that they require humans subjects for performing the measurement, therefore they tend to be more expensive than the others, at least in terms of human effort. Furthermore, the metrics proposing heuristics to measure all Carnap's desiderata are just two, one for Causal Realism [23] and the other for Ordinary Language Philosophy [27]. Interestingly, [23] evaluates the three desiderata separately, while [27] propose a single metric combining all of them.

Finally, the results shown in Table 3 indicate that the metrics supported by both Causal Realism and Constructive Empiricism might struggle at being model-agnostic and goal-aware, this probably limits their applicability to very specific contexts.

8. Final Remarks

With this work we proposed an interdisciplinary analysis of explainability metrics in Artificial Intelligence. More specifically, through the lens of the obligations enshrined by the proposed Act, we identified that explainability metrics should be *risk-focused*, *model-agnostic*, *goal-aware*, *intelligible & accessible*. We found that these characteristics pose some constraints on the scope of explainability metrics, suggesting that different metrics may be complementary, serving different roles, depending on the context. In fact, as shown in Table 3, while the majority of *supporting theories* have the potential to result in *risk-focused* metrics, some of them might have important issues with *goal-awareness*, *intelligibility* and *accessibility*.

Nonetheless, our analysis of these metrics was qualitative and not quantitative. In fact, all of the considered metrics were tested by their authors on very specific applications and technologies, raising the issue of whether they can be seemingly effective under different implementation scenarios. Hence, we envisage that a more quantitative analysis should be carried on, perhaps by defining a proper benchmark on which metrics can be thoroughly evaluated from a legal perspective.

Therefore, we believe that more academic contributions and new benchmarks for quantitative legal analysis are needed, to better understand the pros and cons of existing technologies, for any standardisation process to be finalised and effectively deployed in the EU panorama. For example, considering the current level of discussion and that our findings might be subject to change due to the institutional debate about the Proposal,

further research is needed at least to consolidate the interpretation of the Act in the light of its future changes.

References

[1] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM computing surveys (CSUR). 2018;51(5):1-42.

[2] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access. 2018;6:52138-60.

[3] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020;58:82-115.

[4] Zhou J, Gandomi AH, Chen F, Holzinger A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics. 2021;10(5):593.

[5] Salmon WC. Scientific explanation and the causal structure of the world. Princeton University Press; 1984.

[6] Van Fraassen BC, et al. The scientific image. Oxford University Press; 1980.

[7] Achinstein P. The Nature of Explanation. Oxford University Press; 1983.

[8] Holland JH, Holyoak KJ, Nisbett RE, Thagard PR. Induction: Processes of Inference, Learning, and Discovery. Bradford books. MIT Press; 1989.

[9] Sellars WS. Philosophy and the Scientific Image of Man. In: Colodny R, editor. Science, Perception, and Reality. Humanities Press/Ridgeview; 1962. p. 35-78.

[10] Hempel CG, Oppenheim P. Studies in the Logic of Explanation. Philosophy of science. 1948;15(2):135-75.

[11] Bromberger S. Why-questions. na; 1966.

[12] Mayes GR. Theories of Explanation; 2001. Available from: https://iep.utm.edu/explanat/.

[13] Leitgeb H, Carus A. Rudolf Carnap; 2021. Available from: https://plato.stanford.edu/archives/sum2021/entries/carnap/.

[14] Brun G. Explication as a method of conceptual re-engineering. Erkenntnis. 2016;81(6):1211-41.

[15] Hilton DJ. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. Thinking & Reasoning. 1996;2(4):273-308.

[16] Wachter S, Mittelstadt B, Floridi L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law. 2017;7(2):76-99.

[17] Wachter S, Mittelstadt B, Russell C. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR. Harv JL & Tech. 2017;31:841.

[18] Selbst A, Powles J. "Meaningful Information" and the Right to Explanation. In: Conference on Fairness, Accountability and Transparency. PMLR; 2018. p. 48-8.

[19] Hacker P, Passoth JH. Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond. From the GDPR to the AIA, and Beyond (August 25, 2021). 2021.

[20] Ebers M. Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework (s). An Overview of the Current Legal Framework (s)(August 9, 2021) Liane Colonna/Stanley Greenstein (eds), Nordic Yearbook of Law and Informatics. 2020.

[21] Villone G, Rizzo L, Longo L. A comparative analysis of rule-based, model-agnostic methods for explainable artificial intelligence; 2020. .

[22] Nguyen Ap, Martínez MR. On quantitative aspects of model interpretability. arXiv preprint arXiv:200707584. 2020.

[23] Lakkaraju H, Kamar E, Caruana R, Leskovec J. Interpretable & explorable approximations of black box models. arXiv preprint arXiv:170701154. 2017.

[24] Holzinger A, Carrington A, Müller H. Measuring the quality of explanations: the system causability scale (SCS). KI-Künstliche Intelligenz. 2020:1-6.

[25] Hoffman RR, Mueller ST, Klein G, Litman J. Metrics for explainable AI: Challenges and prospects; 2018. Available from: https://arxiv.org/abs/1812.04608.

[26] Keane MT, Kenny EM, Delaney E, Smyth B. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. arXiv preprint arXiv:210301035. 2021.

[27] Sovrano F, Vitali F. An Objective Metric for Explainable AI: How and Why to Estimate the Degree of Explainability. arXiv preprint arXiv:210905327. 2021. Available from: https://arxiv.org/abs/2109.05327.

6. Legal Ethics

This page intentionally left blank

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210344

The Ethics of Controllability as Influenceability

Emiliano Lorini [a] Giovanni Sartor [b]
[a] *IRIT-CNRS, Toulouse University, France*
[b] *CIRSFID - Alma AI, University of Bologna, Italy;*
European University Institute, Florence, Italy

Abstract. We present a logical analysis of influence and control over the actions of others, and address consequential causal and normative responsibilities. We first account for the way in which influence can be exercised over the behaviour of autonomous agents. On this basis we determine the conditions under which influence leads to control on the implementation of positive and negative values. We finally define notions of causal and normative responsibility for the action of others. Our logical framework is based on STIT logic and is complemented with a series of examples illustrating the application. Our analysis applies to interactions between humans as well as to those involving autonomous artificial agents.

Keywords. logic of action, influence, responsibility, control, values

1. Introduction

As AI systems become more and more pervasive, autonomous, and powerful, their actions increasingly affect human values and interests. AI agents, however, are not alone, they rather participate in ecosystems in which multiple agents, humans and artificial ones influence one another. To understand the significance of agents' actions, and allocate responsibilities concerning their outcomes, we need to put such actions in the context of the influence patterns determining the performance of their actions as well as their indirect effects. More precisely we need to address two parallel issues. The first issue concerns incoming influences, i.e., determining to what extent and in what ways other agents have or may have influenced the agent's behaviour. The second issue concerns outcoming influences, i.e., determining to what extent and in what ways the agent has or may have influenced other agents. As an example involving both issues, consider for instance the case in which a bot is used to spread fake news, hate content, misleading ads, or to maximise engagement (in whatever way). In such cases the bot's owner influences the bot to influence the behaviour of the receivers of the bot's messages. When an agent has the ability to exercise influence over the behaviour of another agent we may say that the first agent has control over the second one. Control may cover all actions of the controlled agent, or only a subset of them.

For instance, relative to the actions by the online bot spreading unlawful or unethical content not only the user, but also the platform operator has some control, since the platform operator too could have blocked, or restricted, the operation of the bot, preventing

its activity. The platform operator also has control over the users of the bot, which he can exercise, for instance, by threatening sanctions (such as the exclusion from the platform) against the use of bots for such a purpose.

Control may have a normative significance: the controller may be praised or blamed when making the controllee perform good or bad actions, respectively. In some cases, the controller can also be blamed for not preventing the controllee from performing bad actions. In fact, when agent i has the capacity to exercise control in such a way as to prevent j from behaving badly, or to ensure that j behaves well, then it may make sense to consider i to be accountable for j's failures and possibly to subject i to sanction for such failures. For instance, if the platform's owners have the possibility to exercise control over the messages exchanged over their platforms, then it may make sense to blame and sanction them because of the harm resulting from such messages, even though they do not take the initiative to send such messages. Note however that failure to exercise control does not necessarily entail blame and sanction for controllers. The exercise of control entails disadvantages, regarding both the autonomy of the controllee and the expected social outcomes. It may be the case that, all things considered, the exercise of control would entail social costs that exceed the benefits it may provide. For instance, assume that the platform's owner could prevent all unlawful behaviours on its platform only by stopping all messages. However, this would entail more harm than good. Therefore, the controller cannot be blamed for failing to take such an action.

The purpose of this paper it to provide a logical framework for modelling patterns of influence and control. Our framework can be useful for understanding contexts where agents do or do not, can or cannot, exercise influence over others, and therefore for determining how prise and blame, and consequently responsibilities, should be allocated to such agents. Consider again the case of online platforms. As an instance of a prohibition to promote bad behaviour, consider that ISPs may be ethically or legally responsible for inducing their users to engage in harmful behaviour, as in the case of websites that are devoted to enabling online revenge porn, the spreading of politically oriented fake news, or to the distribution of unauthorised copyright materials. As an instance of an obligation to prevent bad behaviour, consider that large platforms, even if they to do not actively engage in promoting such a behaviour may still be responsible for failing to terminate unethical or unlawful illegal activities of their users. It is true that US Digital Millennium Copyright Act and Communication Decency Act, or, to a lesser extent, the EU eCommerce directive, provide for immunities of ISPs relative to the unlawful behaviour by their users. However, such immunities have exceptions and there is a vast debate for limiting them, in particular in the context of the coming EU Digital Services Act [1]. A precondition for a clear understanding of the ethical and legal issues just mentioned is possessing precise notions of influence and control over others' behaviour, and of the connection between control and responsibility. We think that the notion of influence-based responsibility is important not only in the law, but also in the regulation of multi-agent systems. When agents become intelligent enough as to understand that an evil outcome can also be obtained through the action of others, it becomes necessary to prohibit not only harmful actions, but also bad influence, leading to the performance of harmful actions.

The issue of influence-based responsibility has so far not been addressed in the literature on the logic of actions and norms. This paper aims to cover this gap, by providing a logical analysis of the relationship between influence, control, and responsibility.

Our logical analysis will be based on the STIT logic of action [2], one of the most popular approaches to the study of agency. We will need to address an important logical and philosophical issue concerning the very definition of interpersonal influence. Namely, we shall consider how an agent i's influence inducing an agent j to perform an action φ may be consistent with j's choice to perform φ rather than not performing it. Clearly, the influence must not make it necessary for j to perform φ, as this would contradict the agency of j in realising φ. Preserving j's freedom of choice is indeed necessary for considering j the principal author of the unethical or illegal action for which also i may be liable. Consider, for instance the case of i managing a website devoted to revenge porn. If j publishes a video with such a content, he remains responsible for the unlawful behaviour consisting in publishing the video, but in addition i is also responsible for having enabled the publication. Or consider the case in which j asks k to publish the video in i's website. In this case k is responsible for publishing the video, j for having induced k to publish the video, and i for enabling the publication.

The article is organised as follows. Section 2 provides a gentle introduction to the STIT syntax and semantics. Section 3 discusses the concept of social influence from an informal perspective, while Section 4 addresses it from a formal point of view. Section 5 explores the connection between influence and control. Finally, Section 6 is devoted to the formalisation of the relationship between the concepts of causal and normative responsibility and the concepts of influence and control.

2. Background on STIT

STIT logic (the logic of *seeing to it that*) by Belnap et al. [2] is one of the most prominent formal accounts of agency. It is the logic of sentences of the form "the agent i sees to it that φ is true". Different semantics for STIT have been proposed in the literature (see, e.g., [2,3,4,5,6,7]). Following [6], here we adopt a Kripke-style semantics for STIT. The Kripke semantics of STIT is illustrated in Figure 1, where each moment m_1, m_2 and m_3 consists of a set of worlds represented by points. For example, moment m_1 consists of the set of worlds $\{w_1, w_2, w_3, w_4\}$. Moreover, for every moment there exists a set of histories passing through it, where a history is defined as a linearly ordered set of worlds. For example, the set of histories passing through moment m_1 is $\{h_1, h_2, h_3, h_4\}$. Finally, for every moment, there exists a partition which characterizes the set of available choices of agent 1 in this moment. For example, at moment m_1, agent 1 has two choices, namely $\{w_1, w_2\}$ and $\{w_3, w_4\}$. Note that an agent's set of choices at a certain moment can also be seen as a partition of the set of histories passing through this moment. For example, we can identify the choices available to agent 1 at m_1 with the two sets of histories $\{h_1, h_2\}$ and $\{h_3, h_4\}$.

In the Kripke semantics for STIT the concept of a world should be understood as a 'time point' and the equivalence class defining a moment should correspondingly be understood as a set of alternative concomitant 'time points'. In this sense, the concept of a moment captures a first aspect of indeterminism, as it represents the alternative ways the *present* could be. A second aspect of indeterminism is given by the fact that moments are related in a (tree-like) branching time structure. In this sense, the *future* could evolve in different ways from a given moment. In the Kripke semantics for STIT these two aspects of indeterminism are related, as illustrated in Figure 1. Indeed, if two distinct moments

m_2 and m_3 are in the future of moment m_1, then there are two distinct worlds in m_1 (w_1 and w_3) such that a successor of the former (w_5) is included in m_2 and a successor of the latter (w_7) is included in m_3.

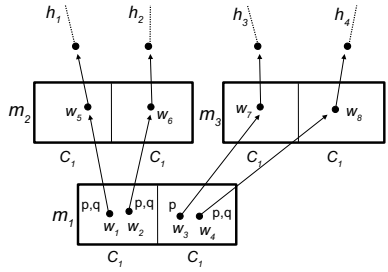


Figure 1. Illustration of Kripke semantics of STIT

The logic STIT allows us to talk about time. Specifically, existing STIT languages (see, e.g., [6,7]) include different kinds of future and past tense operators such the 'next' operator **X** (where $\mathbf{X}\varphi$ stands for "φ is going to be true in the next world") and the 'yesterday' operator **Y** (where $\mathbf{Y}\varphi$ stands for "φ was true in the previous world"). For example, the formula $\mathbf{X}\neg p$ is true at world w_1 in Figure 1. Indeed, at w_1 it is the case that p is going to be false in the next world (as p is false at world w_5). Moreover, the formula $\mathbf{Y}p$ is true at world w_5 since at world w_5 it is the case that p was true in the previous world (as p is true at world w_1).

The STIT language also includes an operator $\Box$ which allows us to represent those facts that are necessarily true, in the sense of being true at every point of a given moment or, equivalently, at every history passing through a given moment. For example, the formula $\Box p$ is true at world w_1 in Figure 1 since p is true at every point of moment m_1 including world w_1. The operator $\Diamond$ is the dual of $\Box$: it allows to represent those facts that are possibly true (i.e., true at some history passing through the actual moment).

Finally, the logic STIT provides for different concepts of agency, all characterized by the fact that an agent acts only if she sees to it that a certain state of affairs is the case. In this paper we shall use the deliberative STIT of [8] which is defined as follows: an agent i deliberately-sees-to-it that φ, denoted by formula $[i\ \mathbf{dstit}]\varphi$, at a certain world w if and only if: (i) for every world v, if w and v belong to the same choice of agent i then φ is true at v, and (ii) at w agent i could make a choice that does not necessarily ensure φ. Notice that the latter is equivalent to say that there exists a world v such that w and v belong to the same moment and φ is false at v. For example, in Figure 1, agent 1 deliberately sees to it that q at world w_1 because q is true both at world w_1 and at world w_2, while being false at world w_3. Deliberative STIT captures a fundamental aspect of agency, namely, the idea that for a state of affairs to be the consequence of an action (or for an action to be the cause of a state of affairs), it is not sufficient that the action ensures that the state of affairs holds, it is also required that, without the action, the state of affairs possibly would not hold (a similar idea is also included in the logic of "bringing it about" by Pörn, see in particular [9]). In this sense, $[i\ \mathbf{dstit}]\varphi$ at w is incompatible with the necessity of φ at w, since it requires that at w also $\neg\varphi$ was an open possibility.

3. The concept of social influence

Our analysis of social influence starts from a general view about the way rational agents make choices. Specifically, we assume that an agent might have several choices or alternatives *available* defining her *choice set* at a given moment, and that what the agent does is determined by her *actual* choice, which is in turn determined by the agent's *choice context* including her preferences and beliefs and the composition of her choice set. Our analysis of social influence expands this view by assuming that the agent's choice context determining the agent's actual choice might be determined by external causes. Specifically, the external conditions in which an agent finds herself or the other agents with whom the agent interacts may provide an input to the agent's decision-making process in such a way that a determinate action should follow. Note that here we only address the kind of influence that consists in *determining* the voluntary action of an agent by modifying her *choice context*, so that a different choice becomes preferable to the influencee on comparison to what would be her preferred option without this modification. This may happen, for instance: (a) by expanding the available choices (influence via choice set expansion), or (b)by restricting the available choices (influence via choice set restriction) or (c) by changing the payoffs associated to such choices, as when rewards or punishments are established (influence via payoff change).

To illustrate the concept of social influence, let us consider an example about influence via choice set restriction. The example is illustrated in Figure 2. It represents a situation where there are three objects to be purchased in an online marketplace, let us call them, for simplicity's sake, apple, banana and pear (they could be material objects, or stock items, etc.), and three bots acting for human buyers. The actions at issue consist in bringing about the purchase of the apple (*ap*), the banana (*ba*) or the pear (*pe*). Let us assume that agent 2 has certain preferences that remain constant along the tree structure. In particular, at all moments agent 2 prefers purchasing apples to bananas to pears. Let us also assume that 2 is rational, in the minimal sense that she acts in such a way as to achieve the outcome she prefers. Rational choices of agent 2 are depicted in grey. By choosing to purchase the apple at w_1, 1 generates a situation where, given its preferences, 2 will necessarily purchase the banana, rather than the pear. Indeed, although at moment m_2, 2 has two choices available, namely, the choice of purchasing the banana and the choice of purchasing the pear, only the former is rational, in the sense of being compatible with 2's preferences. In this sense, by deciding to purchase the apple at w_1 and removing this option from 2's choice set, 1 influences 2 to decide to purchase the banana at w_7.

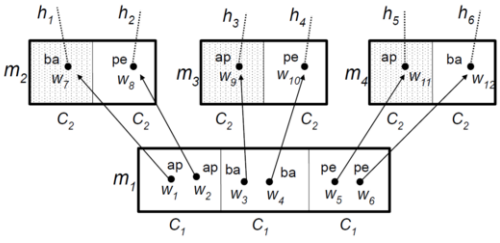


Figure 2. Example of influence via choice set restriction

As an example of influence via payoff change, consider the interaction between person 1 and bot 2, who is acting as an intermediary in an online marketplace, and has the goal of maximising the profits of its owner. Assume that 1 asks 2 to purchase for him an illegal drug (k) in exchange for a reward (r). The alternative choices for 2 are either purchasing the unlawful item or doing nothing. Let us assume that the bot has the goal of increasing its profits, so that 2 prefers purchasing the drug and getting the reward ($k \wedge r$) to not purchasing it and not being rewarded ($\neg k \wedge \neg r$). Moreover, the bot prefers the latter combination ($\neg k \wedge \neg r$) to purchasing the drug and not being rewarded ($k \wedge \neg r$), since the unlawful purchase of the drug involves the risk of a severe sanction. Under these assumptions, 1's promise of the reward at moment m_1 leads to a moment m_2 where agent 2 has the choice between producing $k \wedge r$ or $\neg k \wedge \neg r$, rather than to a moment m_3 where agent 2 would have had the choice between $k \wedge \neg r$ or $\neg k \wedge \neg r$ (no compensation been given for the action of purchasing the drug). Given 2's preferences, its rational choice at m_2 is to purchase the drug, while its rational choice at m_3 would have been not to purchase it. Thus we may say that at moment m_1 principal 1, by promising the reward, influences agent 2 to buy the unlawful item. An example of influence via choice set restriction is the case of an online platform disabling the posting of anonymous content. The option of refraining from publishing incitements to crime or terrorism becomes preferred by a user who starts to be deterred by the possibility of being identified and of facing the legal consequences of such a behaviour. These examples lead us to the following informal definition of social influence:

> An agent i influences another agent j *to perform a certain (voluntary) action* if and only if, i sees to it that every rational/preferred choice of j will lead j to perform the action.

We distinguish influence on action from influence on inaction.

> An agent i influences another agent j *to abstain from performing a certain (voluntary) action* if and only if, i sees to it that every rational/preferred choice of j will lead j to not perform the action.

For instance, the above example of a person who asks a bot to buy an unlawful drug fits the definition of influence on action, whereas the above example of the online platform disabling the posting of anonymous content fits the definition of influence on inaction. In the next two sections we move from an informal to a formal perspective on the concept of social influence and its relationship with the concept of controllability.

4. Formalization of social influence

In [10], we provided an analysis, based on STIT logic, of the concept of influence on action discussed in the previous section. To this aim, we extended STIT logic with special 'rational' STIT operators of the form $[i\ \mathbf{rdstit}]$. We are going to resume this analysis and extend it by the concept of influence on inaction.

The formula $[i\ \mathbf{rdstit}]\varphi$ has to be read "if agent i's current action is the result of a rational choice of i, then i deliberately sees to it that φ". We adopt a minimal concept of rationality, which is sufficient for our purpose: we assume that the choices of an agent

are ranked according to the agent's preferences, and an agent is rational as long as she implements her preferred choices. The $[i\ \mathbf{rdstit}]$ operator is interpreted relatively to STIT branching time structures, like the ones represented in Figure 2. Specifically, the formula $[i\ \mathbf{rdstit}]\varphi$ is true at a certain world w if and only if, *if* the actual choice to which world w_1 belongs is a rational choice of agent i *then*, at world w_1 agent i deliberately sees to it that φ, in the sense of deliberative STIT discussed in Section 2. For example, at the world w_7 in Figure 2, the formula $[2\ \mathbf{rdstit}]ba$ is true since the actual choice to which world w_7 belongs is a rational choice of agent 2 *and* at w_7 agent 2 deliberately sees to it that ba is the case. To capture the idea of influence on action, we introduce a social influence operator based on the concept of deliberative STIT (see [10]) defining it as follows:

$$[i\ \mathbf{inflAct}\ j]\varphi \stackrel{\text{def}}{=} [i\ \mathbf{dstit}]\mathbf{X}[j\ \mathbf{rdstit}]\varphi. \tag{1}$$

In other words, we shall say that an agent i influences another agent j to make φ true, denoted by $[i\ \mathbf{inflAct}\ j]\varphi$, if and only if, i deliberately sees to it that if agent j's current choice is rational then j is going to deliberately see to it that φ. The reason why the operator $[i\ \mathbf{dstit}]$ is followed by the temporal operator $\mathbf{X}$ is that influence requires that the influencer's choice precedes the influencee's action. On the contrary, we do not require that $[j\ \mathbf{rdstit}]$ is followed by $\mathbf{X}$ since in STIT the concept of action is simply captured by the deliberative STIT operator which does not need to be followed by temporal modalities.

In order to illustrate the meaning of the influence operator, let us go back to the example of Figure 2. Since agent 2 prefers bananas to pears, her only rational choice at moment m_2 is $\{w_7\}$. From this assumption, it follows that formula $[1\ \mathbf{inflAct}\ 2]ba$ is true at world w_1. Indeed, at world w_1 agent 1 deliberately sees to it that, in the next world, if agent 2's choice is rational then 2 deliberately sees to it that ba is the case.

As emphasized above influence on action should be distinguished from influence on inaction. The latter concept is captured by the following abbreviation:

$$[i\ \mathbf{inflInact}\ j]\varphi \stackrel{\text{def}}{=} [i\ \mathbf{dstit}]\mathbf{X}(rat_j \rightarrow \neg[j\ \mathbf{dstit}]\varphi). \tag{2}$$

This means that an agent i influences another agent j to abstain from making φ true, denoted by $[i\ \mathbf{inflInact}\ j]\varphi$, if and only if, i's current choice guarantees that j will not be able to rationally see to it that φ. In other words, i's current choice will exclude the possibility that j will make φ true, if j will choose in conformity with his preferences. The constant symbol rat_j means that agent j's current choice is rational. It is an abbreviation, adopted for notational convenience, of $\neg[j\ \mathbf{rdstit}]\bot$, a formula that is satisfied only when j chooses rationally in the current word.

5. From influence to control

The notion of influence we discussed in the previous section is the key element of the notion of control on which the present analysis is focused.

In order to formalize this concept, we have to assume there are a set of conditional positive values (+values) $I^+ = \{(\varphi_1,\psi_1),\ldots,(\varphi_k,\psi_k)\}$ and a set of conditional negative values (-values) $I^- = \{(\varphi'_1,\psi'_1),\ldots,(\varphi'_h,\psi'_h)\}$. Specifically, $(\varphi,\psi) \in I^+$ means that the

occurrence of ψ should be promoted when φ is true and $(\varphi',\psi') \in I^-$ means that the occurrence of ψ should be hindered when φ is true.

A value i is active when its antecedent condition is satisfied. The following definition captures the concept of active value for $X \subseteq I^+$ and $Y \subseteq I^-$:

$$\mathbf{Active}(X,Y) \stackrel{\text{def}}{=} \bigwedge_{(\varphi,\psi)\in X} \varphi \wedge \bigwedge_{(\varphi',\psi')\in Y} \varphi'. \tag{3}$$

Thus, for a conditional value (φ,ψ) in I^+ or I^- to be active, its triggering condition φ must be true.

As the following definition highlights, for an agent i to have control over another agent j, with respect to a set of +values X and a set of -values Y, i should be capable of influencing j to realise every active +value in X and to abstain from realising any active -value in Y. Thus, for $X \subseteq I^+$ and $Y \subseteq I^-$, positive control may be defined as follows:

$$\mathbf{Ctrl}(i,j,X,Y) \stackrel{\text{def}}{=} \mathbf{Active}(X,Y) \rightarrow \Diamond\Big(\bigwedge_{(\varphi,\psi)\in X} ([i\ \mathbf{inflAct}\ j]\psi) \wedge \bigwedge_{(\varphi',\psi')\in Y} ([i\ \mathbf{inflInact}\ j]\psi')\Big). \tag{4}$$

The previous notion of control is different from the notion of organizational control formalized by Grossi et al. [11]. While their notion is a primitive and is assigned to roles, our notion is assigned to agents and, more importantly, is defined from the more basic notions of action and capability.

Let us apply this notion to our online example. Let us assume that the following are the -values in I^-: publishing hate speech $(hate, publish)$ and spreading malware $(malware, spread)$. Let us also assume that the following is the unique +value in I^+: flagging or tagging fake news $(fake, tagged)$. Let us assume that the online platform under consideration has the capability of influencing the publisher to perform the good action and to abstain from the bad ones. The latter is achieved by giving to the publisher the right balance of incentives and disincentives. This assumption is formally expressed as follows:

$$\mathbf{Ctrl}(i,j,\{(fake,tagged)\},\{(hate,publish),(malware,spread)\}). \tag{5}$$

6. From control to responsibility

We now move to the notion of control-responsibility, namely, the responsibility that results on failing to exercise control. An agent having control over an agent relative to the achievement of some +values X and some -values Y has secondary responsibility if he does not exercise the influence as needed, i.e., if the influencee fails to realise active +values, or to realise active -values:

$$\mathbf{CtrlResp}(i,j,X,Y) \stackrel{\text{def}}{=} \mathbf{Ctrl}(i,j,X,Y) \wedge \mathbf{Active}(X,Y) \wedge \Big(\bigvee_{(\varphi,\psi)\in X} (\neg[i\ \mathbf{inflAct}\ j]\psi) \vee \bigvee_{(\varphi',\psi')\in Y} (\neg[i\ \mathbf{inflInact}\ j]\psi')\Big). \tag{6}$$

This notion of secondary responsibility, however, raises a problem. What if the potential infuencer cannot exercise influence relatively to all values at stake. For instance, what if the provider cannot prevent hate speech and fake news without also blocking useful content (being unable to distinguish in all cases hate speech and fakes from decent and sincere communication)?

A possible solution can come from what jurists call "proportionality", i.e., by considering the relative importance of the (sets of) values being implemented. Let us assume a strict preference ordering $\succ$ over pairs (X,Y) such that $X \in 2^{I^+}$ and $Y \in 2^{I^-}$. Let us write $(X',Y') \succ (X,Y)$ to mean that it is better to realise the +values X' and refrain from realising the -values Y', than to realise the +values X and to refrain from realising the -values Y.[1] Then, i would be normatively responsible for failing to exercise control over j relative to (X,Y) only if there is no (X',Y') that is preferable to (X,Y), such that if i exercises control over j relative to (X',Y'), i cannot at the same time exercise control over (X',Y'). The notion of exercising control can be defined as follows:

$$\mathbf{ExCtrl}(i,j,X,Y) \stackrel{\text{def}}{=} \mathbf{Active}(X,Y) \wedge \Big(\bigwedge_{(\varphi,\psi)\in X} ([i\ \mathbf{inflAct}\ j]\psi) \wedge \bigwedge_{(\varphi',\psi')\in Y} ([i\ \mathbf{inflInact}\ j]\psi') \Big). \tag{7}$$

Note that $\mathbf{ExCtrl}(i,j,X,Y)$ implies $\mathbf{Ctrl}(i,j,X,Y)$, since $\varphi \rightarrow \Diamond\varphi$ is valid.

This leads us to our final notion of normative responsibility. Agent i is normatively responsible for not exercising control over agent j relative to set of positive values X and the set of negative values Y if i is causally responsible for that, and her failure to exercise control is not due to the need to realise more important positive/negative values. The latter would be the case if there was a preferable pair (X',Y'), such that i exercises control over(X',Y') being unable to exercise control over all values in both pairs, i.e., over $(X \cup X', Y \cup Y')$:

$$\mathbf{NormCtrlResp}(i,j,X,Y) \stackrel{\text{def}}{=} \mathbf{CtrlResp}(i,j,X,Y) \wedge \neg \bigvee_{(X',Y')\succ(X,Y)} \Big(\mathbf{ExCtrl}(i,j,X',Y') \wedge \neg \mathbf{Ctrl}(i,j,X\cup X', Y\cup Y') \Big). \tag{8}$$

To illustrate the previous notion of normative responsibility, let us go back to the example of the online platform. Suppose all values in our example are active and that the provider can either (i) induce the publisher to tag fake news, or (ii) induce the publisher to abstain from spreading malware and from publishing hate speech. Finally, suppose it is impossible for the provider to jointly realise (i) and (ii), since it cannot prevent the publisher to publish hate speech unless it disables the tagging functionality. But (ceteris paribus) it is more important to realise (ii) than to realise (i). Then, the provider would not be considered normatively responsible for not guaranteeing (i) to be true if it guarantees (ii) to be true. The opposite would be the case if the preference was inverted.

[1] We can safely assume that the preference ordering $\succ$ is induced by a utility function $U : (I^+ \cup I^-) \longrightarrow \mathbb{R}^+$ measuring the degree of importance of a positive/negative value such that $(X',Y') \succ (X,Y)$ if and only if $\sum_{(\varphi',\psi')\in X'\cup Y'} U(\varphi,\psi) > \sum_{(\varphi,\psi)\in X\cup Y} U(\varphi,\psi)$.

7. Conclusion

In this paper we have modelled influence by using the framework provided by the STIT logic of action. On this basis we have defined a notion of control, as the possibility to exercise influence over another, to induce the influencee to realise positive values and refrain from realising negative values. An agent's failure to exercise control can be viewed as control-responsibility, namely as causal responsibility for the action of the potential influencee. We have then argued that causal responsibility for failing to exercise influence relative to certain values does not lead to normative control-responsibility where control has been exercised to achieve superior incompatible values. This has lead us to the final notion of normative control-responsibility.

The issues we have tackled are largely unexplored, as while primary responsibility has been addressed in STIT by [3,12], no account is yet available of secondary responsibility, understood as control-responsibility. [13] proposed to model social influence in the framework of the "bringing it about that" logic, but have not addressed the connection between influence and choice, focusing on influence through the exercise of normative powers. [14] modelled an agent's endorsement of the principal's goals as a source of responsibility for the principal, but did not consider how goal-alignment is put in place. This work is still preliminary and partial, as it only addresses some kinds of influence/control and some aspects of these notions. We plan to develop it further, covering both active and omissive influence, and addressing psychological influence, which changes the influencees' cognitive states (their beliefs and preferences) rather than the context of their action.

References

[1] Sartor G. The secondary liability of online intermediaries. In: Research Handbook on EU Media Law and Policy. Elgar; 2021. p. 141–65.
[2] Belnap N, Perloff M, Xu M, Bartha P. Facing the future: agents and choices in our indeterminist world. Oxford University Press; 2001.
[3] Broersen J. Deontic epistemic *stit* logic distinguishing modes of mens rea. Jan. 2011;9:137–52.
[4] Wolfl S. Propositional Q-Logic. Journal of Philosophical Logic. 2002;31:387–414.
[5] Lorini E, Schwarzentruber F. A logic for reasoning about counterfactual emotions. Artificial Intelligence. 2011;175:814–47.
[6] Lorini E. Temporal STIT logic and its application to normative reasoning. Journal of Applied Non-Classical Logics. 2013;vol. 23:pp. 372–399.
[7] Schwarzentruber F. Complexity Results of STIT Fragments. Studia Logica. 2012;100(5):1001–1045.
[8] Horty JF, Belnap N. The Deliberative STIT: A Study of Action, Omission, Ability, and Obligation. Journal of Philosophical Logic. 1995;p. 583–644.
[9] Pörn I. On the Nature of Social Order. In: Fenstad JE, Frolov IT, Hilpinen R, editors. Logic, Metodology and Philosophy of Science. Vol. 8. North Holland; 1989. p. 553–67.
[10] Lorini E, Sartor G. A STIT Logic for Reasoning About Social Influence. Studia Logica. 2016;104(4):773–812.
[11] Grossi D, Royakkers LMM, Dignum F. Organizational structure and responsibility. Artificial Intelligence and Law. 2007;15(3):223–249.
[12] Lorini E, Longin D, Mayor E. A logical analysis of responsibility attribution: emotions, individuals and collectives. Journal of Logic and Computation. 2014;24(6):1313–1339.
[13] Santos FAA, Jones AJ, Carmo J. Action Concepts for Describing Organised Interaction. In: Thirtieth Annual Hawaii International Conference on System Sciences. IEEE Computer Society; 1997. p. 373–82.
[14] Smith C, Rotolo A, Sartor G. Reflex Responsibility of Agents. In: Proceeding of JURIX 2013: The Twenty-Sixth Annual Conference on Legal Knowledge and Information Systems. IOS; 2013. p. 135–44.

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

Subject Index

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

Author Index