Artif Intell Law
https://doi.org/10.1007/s10506-018-9227-z

Introduction to the special issue on legal text analytics

Jack G. Conrad[1] · L. Karl Branting[2]

1 Introduction

Recent developments in large-scale storage capabilities and related advances in legal data analytics have created unprecedented tools for identifying patterns—such as statistical, semantic, citation-based, and temporal structure—in large legal data repositories. These developments are leading to new insights into both the interconnections among authoritative legal texts and the behavior of courts, legislatures, lawyers, and other legal professionals.

Historically, Artificial Intelligence and Law conferences like the ICAIL (International Conference on Artificial Intelligence and Law) series sponsored by the IAAIL (International Association for Artificial Intelligence and Law)[1] have focused more on argumentation and inference than on empirical and corpora-based approaches to legal analysis and problem solving. However, the rapid growth in new data-centric techniques has opened the door to dramatically new algorithmic approaches for legal problem solving and analysis.

This special issue of the AI and Law journal presents four papers that exemplify these recent trends. Initial versions of all but one of these papers were presented at

[1] https://www.iaail.org (Accessed: 13 April 2018).

✉ Jack G. Conrad
jack.g.conrad@thomsonreuters.com

L. Karl Branting
lbranting@mitre.org

[1] Thomson Reuters, Saint Paul, MN, USA

[2] The MITRE Corporation, McLean, VA, USA

Published online: 02 May 2018

the workshop on “Legal Text, Document, and Corpus Analytics (LTDCA)” held at the University of San Diego School of Law in the summer of 2016.[2] LTDCA was one in a series of recent workshops sponsored by IAAIL-affiliated groups that focus on empirical and machine-learning approaches to legal text analysis. This series includes the ICAIL 2011 workshop on “Applying Human Language Technology to the Law,” the ICAIL 2015 workshop on “Law and Big Data,” and workshops on “Automatic Semantic Analysis of Information in Legal Texts (ASAIL)” at ICAIL 2015 and 2017.[3]

Three of the papers in this issue explore the insights that arise when laws and precedents are viewed as interconnected systems rather than in isolation. “Automated Patent Landscaping,” by Aaron Abood and Dave Feltenberger (Google), presents an approach to understanding patents based on their role in a complex environment of prior patents linked by varying types and degrees of similarity. This approach uses human domain expertise and patent classification schemas to identify prior patents very likely to be related to the current patent (“seeds”) and prior patents very likely to be unrelated (“anti-seeds”). A semi-supervised machine learning model is then employed to prune the resulting high-recall set of candidates down to a high-precision subset. The machine learning component uses several neural network techniques, which are discussed at greater length below.

Two papers, “Bending the Law,” by Greg Leibon (Dartmouth) et al., and “Semantic Edge Labeling over Legal Citation Graphs,” by Ali Sadeghian (University of Florida) et al., take a network perspective on systems of interrelated authoritative legal texts. “Semantic Edge Labeling” addresses a central technical challenge in citation graphs, which are networks of statutory provisions or precedents linked by explicit citations. In Common Law jurisdictions, citations between precedents typically indicate the semantics of the link, e.g., whether the prior precedent is being followed, overturned, distinguished, etc. However, citations in statutory texts often lack any explicit indication of the semantics of the link. Identification of link semantics is often important because, typically, only a subset of edge types is relevant for a given task. For example, analysis of organizational structure might depend solely on a subgraph involving “authority” edges and not at all on “amended by” or “legal basis” edges. When the task is to determine the ramifications of an amendment, by contrast, edges of the latter type might be important and “authority” edges irrelevant. The paper describes the development of a corpus of annotated citations and an experimental evaluation showing that the semantic category of citations can be accurately predicted using a machine learning model trained on the text spans associated with citations.

“Bending the Law” investigates a more expansive network model in which links include not just explicit citations, but also implicit relationships derived from topic models, giving rise to a legal “landscape” similar to that discussed in the Abood and

[2] LTDCA was sponsored by the Center for Computation, Mathematics and Law (CCML) at the University of San Diego School of Law.

[3] ICAIL 2011 was held at the University of Pittsburgh in Pennsylvania, ICAIL 2015 was held at the University of San Diego in California, and ICAIL 2017 was held at King’s College, London.

Feltenberger paper. Random walks on this landscape induce a geometry upon which it is possible to formalize a notion of “curvature” that can account for the tendency of researchers to gravitate away from the neighborhoods of some authoritative sources in the landscape and towards others. This model offers the promise of browsing legal collections by legal relevance rather than simple textual similarity.

The fourth paper reflects the recent revival of interest in the role that neural networks can play in legal AI applications. Neural networks are nothing new to AI and Law: they were cited in work going back to the very first ICAIL conference in Boston in 1987 and in several conferences in the decade that followed. (Bench-Capon et al. 2012) The current application of neural networks to the AI space, known as “deep learning,” differentiates itself from earlier efforts in that numerous layers can be cascaded atop each other to represent different levels of information granularity and to exploit the capabilities of a “long short-term memory” (LSTM). LSTM is a building block for constructing layers within a Recurrent Neural Network (RNN), a kind of neural network that can operate on sequential data and which is suitable for solving sequence labeling tasks. At the most recent ICAIL conference, the number of deep learning papers submitted approached double digits. A number were published in the ICAIL Proceedings, and others appeared in associated workshops. In “Recurrent Neural Network-based Models for Recognizing Requisite and Effectuation Parts in Legal Texts,” Nguyen *et al.* train a RNN to recognize and label two key portions of Japanese legal documents. Their experiments demonstrate superior performance for this identification task relative to alternative approaches and current baselines for both a Japanese and an English-language version of their corpus

One issue that often arises in the context of deep learning applications today is framed in the debate that contrasts explainable AI (XAI),[4] which is transparent and can be used to respond to a “social right to explanation,” and “black box” AI, which relies on complex and typically opaque algorithms that can’t provide explanations for particular decisions. The authors of the deep learning paper in this issue have addressed this topic in two distinct ways. First, they have shown how they can cascade a more traditional and better-known source of input data (Conditional Random Fields) into their RNN model. And second, they have performed a series of error analyses that help us to “look under the hood” and provide their audience with insights into why their trained models make the kinds of decisions (and errors) they do. These are both practical and positive developments, serving to disarm some of the traditional critics of this form of AI.

So, what is one to abstract from the set of research works presented in this special issue? Is it that a key dimension of tomorrow’s AI and Law research will be data-driven and empirical? This outcome will depend on results that are produced by capabilities such as NLP, data mining, and machine learning that are measurable, quantifiable, comparable. Like the work products coming out of hundreds of today’s legal start-up companies, the results must be applicable to and relieve existing pain points among today’s lawyers and other legal practitioners. Beyond our need to simply identify relevant legal materials, we must be able to analyze the underlying

[4] https://www.darpa.mil/program/explainable-artificial-intelligence (Accessed: 13 April 2018).

content and produce predictive models. These are some of the key aspects of today’s AI capabilities in the legal domain—and tomorrow’s successes.

Acknowledgements The guest editors of this special issue wish to thank the numerous contributing reviewers of both accepted and not accepted papers. These reviewers include: Khaled Ammar, Robert Arens, Mohamed Bouadjenek, Roman Boulet, Chris Giannella, Zack Kriegman, Thorne McCarty, Filippo Pompili, Vern Walker, and Ryan Whalen. Without their thorough scrutiny, this special issue would not have come together in the effective manner that it has. We are truly grateful for their efforts. The MITRE Corporation is a not-for-profit Federally Funded Research and Development Center chartered in the public interest. This document is approved for Public Release, Distribution Unlimited, Case Number 18-1276. ©2018 The MITRE Corporation. ALL RIGHTS RESERVED.

References

Bench-Capon T, Araszkiewicz M, Ashley K, Atkinson K, Bex F, Borges F, Bourcier D, Bourgine P, Conrad JG, Francesconi E, Gordon TF, Governatori G, Leidner JL, Lewis DD, Loui RP, McCarty LT, Prakken H, Schilder F, Schweighofer E, Thompson P, Tyrrell A, Verheij B, Walton DN, Wyner AZ (2012) A history of AI and law in 50 papers: 25 years of the international conference on AI and law. Artif Intell Law 20(3):215–319