Proposal for a Full-Day Tutorial on Mining Legal Document Repositories

Kripabandhu Ghosh Sachin Pawar
Girish Keshav Palshikar[*]

Abstract

Legal documents come in a great variety: court case documentation (witness testimonies, evidences, FIR etc.), court judgements, contracts / agreements / memoranda, affidavits, patents, legal statutes and many others. Vast repositories of such legal documents are now available on both on the Internet as well as within enterprise repositories. For instance, the Forum for Information Retrieval (FIRE) corpus [15] contains 30,034 Indian Supreme Court, 1,38,730 Indian High Courts' and 1,83,124 Indian Consumer Courts' judgements. Such legal document corpora continue to grow rapidly; e.g., approx. 100 million cases are filed annually in US state trial courts. Given their complex language and structure, the experience and knowledge of human lawyers are crucial in understanding and using these legal documents in various legal tasks. With rapid advances in NLP and ML, legal text analytics is receiving increasing attention and offers opprtunities for (i) automated understanding, analysis, and knowledge discovery from legal document repositories; and (ii) building practical applications to assist lawyers, to provide legal help to common people, and to reduce the efforts and time required in legal processes. Given the enormous complexities of and workload on legal systems in India, and issues due to legal documents in many languages, the tutorial hopes to interest NLP researchers and students in an exciting application domain of great practical value.

1 Outline of Tutorial Topics

In this tutorial, we will focus on the following topics:

1. **Introduction and Motivation (25 minutes):**
2. **Similarity and retrieval of legal documents (50 minutes):** Efficient retrieval of legal documents is important for obtaining relevant information (stored in electronic textual format) pertinent to a court case or any other legal situation. This is particularly important for *prior case retrieval* for

[*] TCS Research, Tata Consultancy Serviced Limited, Pune, India

an ongoing case. The notion of "similarity" of legal documents is the corner-stone for any legal retrieval engine. Similarity can be captured using various measures for different representations of legal documents. We will discuss about some impactful techniques in this regard, e.g. [9] [10] [14]. For evaluation of a retrieval system, evaluation becomes crucial. Top forums like TREC[1], FIRE[2] etc. designed test-beds for evaluating search systems on legal documents. We will briefly discuss the legal tracks organized in these forums. Building a gold standard is a challenge since it involves expert annotations from legal experts which are expensive to hire. So, intelligent and optimal utilization of legal expertise also becomes crucial. We will also discuss a work in this area [6].

3. **Citations among court cases (30 minutes):** Court cases are known to cite previous cases (or *precedents*) as an integral part of the litigation process. There exist a network between court cases. Efficient harnessing of the citation network can leverage efficient and automatic understanding of cases. Such a system can assist the laywers during an ongoing case to trace important precedents relevant to the current case. We will discuss interesting works in this area, e.g.[11], [23], [5].

4. **Information extraction (IE) from legal documents such as judgements and contracts (60 minutes):** We will discuss how IE techniques [18] can be used to extract many different kinds of generic and domain-specific named entities (e.g., names of lawyers and judges, relevant laws and sections, previous cases), relations (e.g., $lawyer\_for\_defendent$), legal events (e.g., filing of FIR, arrest of accused) and crime events (e.g., location and time of murder) from legal documents [1]. We will discuss an advanced IE application for extracting a visual storyline from crime descriptions in court judgements [20].

5. **Classification of legal documents (15 minutes):** Multi-class (and possibly multi-label) classification of legal documents provide critical insights about the content of these documents. E.g., a narrative of a crime event can be automatically labelled with various articles/sections within the law which it violates. We will cover a few important techniques proposed for document classification in Legal domain [26, 25]. For deeper insight, it is also necessary to classify individual sentences within the legal documents, into multiple semantic classes. Hence, in addition to document classification, we will also cover some key aspects of sentence classification in Legal domain [19, 7].

6. **Legal document summarization (15 minutes):** Legal documents esp. court proceedings are often long and complex. Meticulous reading of the same may often be time consuming. Summarization of legal documents are extremly useful if one wants to get a quick overview of the document

[1] `https://trec-legal.umiacs.umd.edu/`

[2] `https://www.isical.ac.in/~fire/2013/legal.html`

without the need for delving deep into the details. Catchphrases form a good summary of court cases. We will discuss some work on catchphrase detection on legal documents [16] [4]. There exist many document summarization algorithms. However, none seem to produce meaningful summaries from a lawyer's perspectivve. This is because not every aspect of a case e.g. fact, precedent, issue, reason for judgement, judgement etc. appear consistently in a summary [2]. A detailed overview of the state-of-the-art in legal summarization will throw valuable light on the challenges and nuances of the problem.

7. **Mining legal arguments (45 minutes):** Court judgements contain summaries of the legal arguments of both the prosecution and defence lawyers, which are complex and structured pieces involving legal, logical and common-sense reasoning. Given their importance in influencing the court decision, understanding and extracting Legal argument are fast becoming a crucial part of legal text mining. We will cover a few basic techniques for extracting legal arguments from legal documents [21], [17], [3], [24].

8. **Patent analysis (35 minutes):** Patent (or Prior Art) retrieval is a very important research problem. Patent is meant to protect intellectual property (IP) rights of novel scientific inventions. Patent retrieval is widely considered to be a sub-field of Information Retrieval where the search is restricted to patents only and the query being a patent application. Many methods on word-level/phrase-level matching have been proposed in this area [13] [12] [22]. Recent methods have also focussed on deep neural based patent representation and search [8]. We will present the important research contributions in this area. In addition, we will discuss the Intellectual Property (IP) track (CLEF-IP)[3] aimed to foster research in patent retrieval.

9. **Shared tasks and competitions (25 minutes):**

10. **Applications (10 minutes):**

11. **Demo (20 minutes):**

12. **Opportunities for research (30 minutes):**

2 Author Summary

Dr. Kripabandhu Ghosh: completed his Ph.D. in 2016 from ISI, Kolkata. His Ph.D. dissertation was on *Information Retrieval in Legal Domain.* He is currently a researcher at TCS Research, Tata Consultancy Services Ltd., Pune, India. His areas of interest comprise Information Retrieval, Machine Learning, Data Mining, Natural Language Processing etc. with applications in legal

[3] http://ifs.tuwien.ac.at/~clef-ip/

domain, social media text etc. He has co-organized multiple workshops and shared-tasks on data mining of legal documents. He has published his works in the legal domain and other sub-areas at top-tier venues like SIGIR, CIKM, ECIR, etc.

Girish Keshav Palshikar: is an alumnus of IIT, Bombay and IIT, Madras. Since 1992, he is working in the TCS Research, Tata Consultanc Services Ltd., Pune, India, where he is now a Principal Scientist. His areas of research include machine learning, data mining, text mining, and their applications to various domains, including knowledge extraction from domain-specific documents, fraud detection and human resources management. He has more than 120 publications in international journals and conferences. He has given tutorials in 2013, 2014 and 2018 editions of ICON.

Sachin Pawar: works as a Researcher in TCS Research and Innovation. He has received M.Tech. in Computer Science and Engineering from IIT Bombay in 2008. He is currently pursuing his PhD at IIT Bombay under the guidance Prof. Pushpak Bhattacharyya and Girish K. Palshikar (TCS Research). His areas of research include Information Extraction, Text Mining etc. His work is published in leading NLP conferences such as ACL, EACL, IJCNLP, etc. He has given tutorials in 2014 and 2018 editions of ICON.

References

[1] Judith Jeyafreeda Andrew. Automatic extraction of entities and relation from legal documents. In *Proceedings of the Seventh Named Entities Workshop*, pages 1–8, 2018.

[2] Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. A comparative study of summarization algorithms applied to legal case judgments. In *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part I*, pages 413–428, 2019.

[3] Vanessa Wei Feng and Graeme Hirst. Classifying arguments by scheme. In *49th Annual Meeting of the Association for Computational Linguistics*, pages 987–996, 2011.

[4] Filippo Galgani, Paul Compton, and Achim Hoffmann. Towards automatic generation of catchphrases for legal case reports. In *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, CICLing'12, pages 414–425, Berlin, Heidelberg, 2012. Springer-Verlag.

[5] Anton GEIST. The open revolution: Using citation analysis to improve legal text retrieval. In *European Journal of Legal Studies, 2010, 2, 3, The Future of... Law & Technology in the Information Society*, 2010.

[6] Kripabandhu Ghosh and Swapan Kumar Parui. Clustered semi-supervised relevance feedback. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 1723–1726, 2015.

[7] Ingo Glaser, Elena Scepankova, and Florian Matthes. Classifying semantic types of legal sentences: Portability of machine learning models. In *JURIX*, pages 61–70, 2018.

[8] Sebastian Hofstätter, Navid Rekabsaz, Mihai Lupu, Carsten Eickhoff, and Allan Hanbury. Enriching word embeddings for patent retrieval with global context. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval*, pages 810–818, Cham, 2019. Springer International Publishing.

[9] Peter Jackson, Khalid Al-Kofahi, Alex Tyrrell, and Arun Vachher. Information extraction from case law and retrieval of prior cases. *Artif. Intell.*, 150(1-2):239–290, November 2003.

[10] Sushanta Kumar, Polepalli Krishna Reddy, V Balakista Reddy, and Malti Suri. Finding similar legal judgements under common law system. In *Databases in Networked Information Systems: 8th International Workshop, DNIS*, volume 7813, pages 103–116, 03 2013.

[11] Sushanta Kumar, P. Krishna Reddy, V. Balakista Reddy, and Aditya Singh. Similarity analysis of legal judgments. In *Proceedings of the Fourth Annual ACM Bangalore Conference*, COMPUTE '11, pages 17:1–17:4, New York, NY, USA, 2011. ACM.

[12] Parvaz Mahdabi, Linda Andersson, Mostafa Keikha, and Fabio Crestani. Automatic refinement of patent queries using concept importance predictors. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 505–514, New York, NY, USA, 2012. ACM.

[13] Parvaz Mahdabi, Shima Gerani, Jimmy Xiangji Huang, and Fabio Crestani. Leveraging conceptual lexicon: Query disambiguation using proximity information for patent retrieval. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 113–122, New York, NY, USA, 2013. ACM.

[14] Arpan Mandal, Raktim Chaki, Sarbajit Saha, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. Measuring similarity among legal court case documents. In *Proceedings of the 10th Annual ACM India Compute Conference*, Compute '17, pages 1–9, New York, NY, USA, 2017. ACM.

[15] Arpan Mandal, Kripabandhu Ghosh, Arnab Bhattacharya, Arindam Pal, and Saptarshi Ghosh. Overview of the FIRE 2017 irled track: Information retrieval from legal documents. In *Working notes of FIRE 2017 - Forum for*

Information Retrieval Evaluation, Bangalore, India, December 8-10, 2017., pages 63–68, 2017.

[16] Arpan Mandal, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. Automatic catchphrase identification from legal court case documents. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 2187–2190, 2017.

[17] Raquel Mochales and Marie Francine Moens. Argumentation mining. *Artificial Intelligence and Law*, 19:1–22, 2011.

[18] G. K. Palshikar. Techniques for named entity recognition: A survey. In *Bioinformatics: Concepts, Methodologies, Tools, and Applications*, pages 400–426. IGI Global, 2013.

[19] Nitin Ramrakhiyani, Sachin Pawar, and Girish Keshav Palshikar. A system for classification of propositions of the indian supreme court judgements. In *Post-Proceedings of the 4th and 5th Workshops of the Forum for Information Retrieval Evaluation*, page 15. ACM, 2013.

[20] A. Sainani, N. Ramrakhiyani, Preethu R.A., S. Pawar, G.K. Palshikar, and S. Ghaisas. Towards disambiguating contracts for their successful execution - a case from finance domain. In *FinNLP Workshop in 28th Int. Joint Conf. on Artificial Intelligence (IJCAI 2019)*, 2019.

[21] M. Saravanan and B. Ravindran. Identification of rhetorical roles for segmentation and summarization of a legal judgment. *Artificial Intelligence and Law*, 18:45–76, 2010.

[22] Manisha Verma and Vasudeva Varma. Applying key phrase extraction to aid invalidity search. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law*, ICAIL '11, pages 249–255, New York, NY, USA, 2011. ACM.

[23] R. Wagh and D. Anand. Application of citation network analysis for improved similarity index estimation of legal case documents : A study. In *2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, pages 1–5, March 2017.

[24] Vern R. Walker, Dina Foerster, Julia Monica Ponce, and Matthew Rosen. Evidence types, credibility factors, and patterns or soft rules for weighing conflicting evidence: Argument mining in the context of legal rules governing evidence assessment. In *Proceedings of the 5th Workshop on Argument Mining*, pages 68–78, 2018.

[25] Pengfei Wang, Yu Fan, Shuzi Niu, Ze Yang, Yongfeng Zhang, and Jiafeng Guo. Hierarchical matching network for crime classification. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 325–334. ACM, 2019.

[26] Pengfei Wang, Ze Yang, Shuzi Niu, Yongfeng Zhang, Lei Zhang, and ShaoZhang Niu. Modeling dynamic pairwise attention for crime classification over legal articles. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 485–494. ACM, 2018.