Legal Text, Document, and Corpus Analytics (LTDCA 2016) Workshop Report

L. Karl Branting
The MITRE Corporation
McLean, VA, USA
lbranting@mitre.org

Jack G. Conrad
Thomson Reuters
Minneapolis, MN, USA
jack.g.conrad@thomsonreuters.com

29 June 2016

Recent advances in both Human Language Technology (HLT) and techniques for large-scale data analysis have vastly increased capabilities for automated interpretation of legal text, promising improved delivery of legal services for citizens, increased efficiency of court and government agencies, and greater insights into the structure and evolution of law. These advances have coincided with a rapid expansion of interest in automated processing and understanding of legal texts on the part of industry, government agencies, court personnel, and the public.

Growing interest in this field is mirrored in the accelerating pace of meetings on this subject, starting with the 2011 ICAIL workshop on *Applying HLT to the Law* and continuing with *Network Analysis in Law* workshops in 2013, 2014, and 2015, a MITRE/SFI Symposium on *Law and Judicial Systems in the Era of Big Data* in 2014, and the 2015 *ICAIL Workshop on Law and Big Data*.

On June 17, 2016, The University of San Diego (USD) Center for Computation, Mathematics, and Law (CCML) hosted a *Workshop on Legal Text, Document, and Corpus Analytics (LTDCA 2016),* co-chaired by Karl Branting (Program Chair) and USD Law School professor Ted Sichelman (Workshop Chair). The 45 attendees included participants from Silicon Valley firms (including Google, Lex Machina, Casetext, Legal Robot), other large private-sector corporations (including Leidos and Thomson Reuters), non-profits (including the Cornell Legal Information Institute and the Free Law Project), startups such as Aria Acoustics and ClearstoneIP, and Universities, including Dartmouth, Princeton, and the Universities of Florida, Virginia, Warwick, and Nagoya. The program included 10 research papers (4 long presentations, 3 short presentations, and 3 posters), 5 invited talks, and 3 demonstrations.

One of the key features of this workshop that distinguished it from other forums of its kind was the number and caliber of invited speakers, all from industry and all but one from start-ups. In this sense, LTDCA 2016 was more than an academic gathering; it was a practical application-oriented event with all of its invited speakers representing real innovation and potentially disruptive capabilities. These keynotes included:

- Karl Harris from Lex Machina discussing *Using Data-Driven Insights to Set Litigation Strategy*
- Pablo Arrendondo and Ryan Walker from Casetext describing *Harvesting and Leveraging Explanatory Parentheticals*
- Dan Rubins from Legal Robot explaining their approach to *Abstractive Summarization of Legal Texts*
- Leora Morgenstern from Leidos addressing how *Lawyers Know Too Much: Automating Extraction of Executable Logic Programming Rules for Regulatory Text*
- David Lewis from BrainSpace elaborating upon *Information Retrieval in E-Discovery: Progress and Controversy.*

The presentations and papers collectively fell into 3 broad categories: text analytics focused on single documents; corpus analytics that reveal relationships among documents or emergent properties of document collections; and broad perspectives on the legal system.

Starting with document-oriented papers, three Workshop papers addressed document analytics in the context of court filings. In *Language-processing methods for US Court Filings*, Marc Vilain (MITRE) described a range of entity-extraction tasks arising in different genres of Federal court filings and showed how a narrative extractor for a representative legal scenario was built incorporating the output of these entity extractors. Bill Liao and Charles Horowitz (MITRE) described in *An Approach to Identify Stamps on Scanned Court Filings* a technique for identifying stamps on court document so that they can be removed to improve OCR accuracy, benefiting down-stream processes such as classification and entity extraction. *Vocabulary reduction, text excision, and procedural contextual features in judicial document analytics* by Karl Branting (MITRE) addressed the task of detecting document filing errors in judicial databases by comparing the document's apparent type, as determined by text classification, with metadata describing what its filer intended the document to be. Classification accuracy was improved both by filtering terms with low mutual information with the particular category and by adding a meta-classification stage that combines procedural context with the output of the text classification.

Statutory texts were the focus of Leora Morgenstern's (Leidos) presentation, *The Lawyers Know Too Much: Automating Extraction of Executable Logic Programming Rules from Regulatory Text,* which provided a summary and update of her IARPA-sponsored work on the challenges of parsing statutory texts into a standard rule

markup representation. Standard tools of the HLT community were not developed for texts with the unique characteristics of statutory text, such as deeply nested bulleted lists, but development of resources tailored for this genre should make the conversion of statutory text to machine-interpretable rules increasingly feasible.
A different aspect of statutory texts—the meaning of citations to other statutory provisions—was the focus of *Semantic Edge Labeling over Legal Citation Graphs* by Ali Sadeghian (University of Florida) et al. Sadeghian et al. described the development of a corpus of annotated citations and an experimental evaluation showing that the semantic category of citations could be predicted using a machine-learning model trained on the text spans preceding citations.

Disputes over the authorship of anonymous or pseudonymous writings arise in many legal proceedings, such as contested wills or contracts. A system for stylometric determination of authorship described in *Did Aunt Prunella Really Write That Will? A Simple and Understandable Computational Assessment of Authorial Likelihood,* by Patrick Juola (Juola & Associates) uses techniques that Juola applied to several well-publicized authorship disputes.

The important and ubiquitous task of legal document summarization was addressed in two presentations. *Abstractive Summarization of Legal Texts* by Dan Rubins (Legal Robot) described a deep learning-based approach trained on summarization examples. In *Harvesting and Leveraging Explanatory Parentheticals,* Pablo Arredondo & Ryan Walker (Casetext) presented an approach to automatically acquiring such examples from parenthetical descriptions that judicial opinions routinely place after a citation to another case.

Several papers and presentations described analytics intended to provide insights into entire corpora rather than individual documents. *Automated Patent Landscaping,* by Aaron Abood and Dave Feltenberger (Google), addresses the task of identifying all patents related to a given topic. Their approach uses a semi-supervised machine learning model to prune an initial high-recall set of candidates down to a high-precision subset. *Diachronic and Synchronic Analyses of Japanese Statutory Terminology* by Makoto Nakamura & Katsuhiko Toyama (Nagoya University) demonstrated visualizations of the changes in lexical distance between related statutes that can provide insights into the evolution of statutory frameworks. In *Using Data-Driven Insights to Set Litigation Strategy,* Karl Harris (Lex Machina) described a number of analytical approaches for estimating the probability of success of alternative litigation actions offered by their current tools. This work illustrates very clearly both the potential for data-driven analytics to add transparency to the judicial system and the probable strategic advantages of early adopters of such analytics over late adopters. *An Open-Source DB for Empirical Analysis of Judges and Judicial Decisions* by Elliott Ash (Princeton) and Michael Lissner (University of Warwick) introduced a new open-source resource for information about judges that should significantly contribute to judicial transparency. Finally, *Absolute Constructions and the Second Amendment: A Corpus Analysis,* by James Vanden Bosch (Calvin College) shows how analysis of a corpus of historical instances of a specific grammatical construction can provide insight into its use and interpretation in contemporary jurisprudence.

A broader view of the legal system was provided by Thorne McCarty (Rutgers) in his paper *On Semi-Supervised Learning of Legal Semantics,* which summarizes a series of papers proposing a model of legal reasoning that spans abstraction levels from perception through analogical reasoning and intuitionistic logic. *Bending the Law,* by Greg Leibon (Dartmouth) et al. applied concepts from geometry and topology to citations and topic models to characterize the dynamics of US Supreme Court jurisprudence.

Another presentation that took a broader view of an important, currently very active application-based field was David Lewis' talk on *Information Retrieval in E-Discovery: Progress and Controversy*. He identified the contributions the field has made to advance the state of the art while also recounting the special challenges the field presents, including agreement on how to evaluate and compare baseline approaches with newer techniques that foster technology assisted review. Lewis argued that the field needs to comparatively assess the performance of both traditional keyword-based search strategies and supervised learning-based ones.

The workshop also included 3 system demonstrations:

- Alex Lyte et al. (MITRE), *Path Analysis to Evaluate the Impact of Legislation on US Government Agencies*
- Jesse Sukman (ClearstoneIP LLC), *Achieving Accuracy, Speed and Low Cost in Freedom-to-Operate (FTO) and Other Infringement-Based Patent Investigations*
- Jason Summers et al. (Aria Acoustics), *Machine Interface for Contracting Assistance (MICA)*

The full proceedings of the workshop can be found at:

http://law-and-big-data.org/LTDCA_2016_proceedings.pdf

LTDCA 2016's strong program of papers and presentations demonstrated both that the field of legal text, document, and corpus analytics comprises a coherent set of technical issues and that a growing community of researchers and developers from academia, non-profits, and in this case, especially the private sector, share common research goals and methodologies.

The venue planned for LTDCA 2017 is London, concurrent with ICAIL 2017 (http://nms.kcl.ac.uk/icail2017/).