LTDCA-2016

Proceedings of the Workshop on Legal Text, Document, and Corpus Analytics

University of San Diego Law School, San Diego, CA, 17 June 2016
Sponsored by the USD Center for Computation, Mathematics, and Law
http://www.sandiego.edu/law/centers/ccml/

L. Karl Branting (Editor)
Preface
Recent improvements in Human Language Technology (HLT) and in techniques for storage and rapid analysis of large data collections have created new opportunities for automated interpretation of legal text, improved access to statutory and regulatory rules, and greater insights into the structure and evolution of legal systems. These techniques hold promise for the courts, legal practitioners, scholars, and citizens alike. 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.

The LTDCA 2016 workshop focused on research in, and applications of, new techniques for interpretation of legal text, analysis of structured legal documents, improved publication and access to document collections, predictive analysis based on legal text mining, and visualization of legal corpora. These proceedings contain 11 papers and 2 extended abstracts of current research on topics at the intersection of HLT, artificial intelligence, social science, data and network science, and law.

Program Chair
Karl Branting, The MITRE Corporation, USA

Workshop Chair
Ted Sichelman, University of San Diego School of Law

Program Committee
Thomas Bruce, Cornell Law School
Jack G. Conrad, Thomson Reuters, Corp. R&D
William Hamilton, University of Florida Law School
Matt Koehler, The MITRE Corporation
David Lewis, David D. Lewis Consulting LLC
Joonsuk Park, Cornell University
Jana Sukkarieh, NOHA Z, LLC
Daisy Zhe Wang, University of Florida
# TABLE OF CONTENTS

Automated Patent Landscaping ................................................................. 1  
*Aaron Abood and Dave Feltenberger*

An open-source database for empirical analysis of judges and judicial decisions ................. 8  
*Elliott Ash and Michael Lissner*

Absolute Constructions and the Second Amendment: A Corpus Analysis ................................. 13  
*James Vanden Bosch*

Vocabulary reduction, text excision, and procedural contextual features in judicial document analytics .............................................................................................................................................. 30  
*L. Karl Branting*

Did Aunt Prunella Really Write That Will? A Simple and Understandable Computational Assessment of Authorial Likelihood ......................................................... 37  
*Patrick Juola*

Bending the Law ..................................................................................... 42  
*Greg Leibon, Michael Livermore, Reed Harder, Allen Riddell and Daniel Rockmore*

An Approach to Identify Stamps on Scanned Court Filings ....................................................... 51  
*Bill Liao and Charles Horowitz*

Using Path Analysis to Evaluate the Impact of Legislation on U.S. Government Agencies .......... 54  
*Alexander Lyte, Haven Liu, David Slater, Shaun Michel and Sanith Wijesinghe*

On Semi-Supervised Learning of Legal Semantics .................................................................. 57  
*Thorne McCarty*

Diachronic and Synchronic Analyses of Japanese Statutory Terminology ............................ 62  
*Makoto Nakamura and Katsuhiko Toyama*

Semantic Edge Labeling over Legal Citation Graphs ................................................................. 70  
*Ali Sadeghian, Lakshman Sundaram, Daisy Zhe Wang, William F. Hamilton, Karl Branting and Craig Pfeifer*

Achieving Accuracy, Speed and Low-Cost in Freedom-to-Operate (FTO) and Other Infringement-Based Patent Investigations ................................................................. 76  
*Jesse Sukman*

Language-processing methods for US Court Filings ................................................................. 81  
*Marc Vilain*
Automated Patent Landscaping

Aaron Abood
Google, Inc.
aabood@google.com

Dave Feltenberger
Google, Inc.
feltenberger@google.com

ABSTRACT

Patent landscaping is the process of finding patents related to a particular topic. It is important for companies, investors, governments, and academics seeking to gauge innovation and assess risk. However, there is no broadly recognized best approach to landscaping. Frequently, patent landscaping is a bespoke human-driven process that relies heavily on complex queries over bibliographic patent databases.

In this paper, we present Automated Patent Landscaping, an approach that jointly leverages human domain expertise, heuristics based on patent metadata, and machine learning to generate high-quality patent landscapes with minimal effort.

In particular, this paper describes a flexible automated methodology to construct a patent landscape for a topic based on an initial seed set of patents. This approach takes human-selected seed patents that are representative of a topic, such as operating systems, and uses structure inherent in patent data such as references and class codes to "expand" the seed set to a set of "probably-related" patents and anti-seed "probably-unrelated" patents. The expanded set of patents is then pruned with a semi-supervised machine learning model trained on seed and anti-seed patents. This removes patents from the expanded set that are unrelated to the topic and ensures a comprehensive and accurate landscape.

CCS Concepts

- Computing / technology policy ➔ Intellectual Property ➔ Patents

Keywords

Patent landscape; classification; text analytics; semi-supervised machine learning

1. INTRODUCTION

At the height of the smartphone wars, it was rumored that the smartphone was covered by 250,000 patents [13]. Or was it 314,390 [14]? Which is right, and how does one even arrive at such numbers? This also leads to follow on questions such as: who owns the patents? when were the patents filed? where are the inventors from? what fraction of the patents have been litigated?

These questions are often answered by a technique called patent landscaping. Patent landscaping is important to a number of audiences, including (1) companies that desire to assess risk posed by other patent holders and understand their own relative strength, (2) academics and governments that seek to gauge the level of R&D investment and innovation in particular fields [18], and (3) investors looking to value companies and assess risk [3].

Patent landscaping is the exercise of identifying all the patents that are relevant to a topic. This is discussed in detail in the guidelines prepared by the World Intellectual Property Office (WIPO) [18]. It is a challenging task that can involve substantial time and expense. In particular, patent landscaping is made difficult by issues such as: over/under inclusion, needing to fully understand the associated technical/product landscape, scalability/data limitations of commercial tools, and proper search methodology. Building patent landscapes is often seen as more of an art than a science with analysts constructing elaborate queries over bibliographic patent databases and assembling multiple lists of patents for inclusion in the resulting landscape. This can make reproducing and updating landscapes difficult. Further, methodologies may be inconsistent or difficult to explain.

Figure 1 – Expansion levels

This paper describes a semi-supervised machine learning approach for automated patent landscaping. The approach starts with a small human curated seed set of patents narrowly related to the topic of interest. The seed set is expanded by citations (forward and backward) and class codes to identify candidate patents for inclusion in the landscape (Levels 1 and 2 in Figure 1). Patents not included in the expansion are referred to as the anti-seed. A machine learning model is then applied to prune the candidate patents to create a subset of the candidate patents that are relevant to the topic. The model is trained using the seed (positive examples) and a sample of patents from the anti-seed (negative examples). The final landscape is the pruned subset of candidate patents and
the seed set. Alternatively, the model could be applied to the full patent corpus to increase coverage missed by the expansion to Levels 1 and 2.

2. PATENT DATA

A basic understanding of patent data and its traditional application to patent landscaping is useful for understanding the automated approach described in this paper. Those who are familiar with patent data can skip to the next section, AUTOMATED LANDSCAPING.

Patent documents and their associated data are somewhat unique in that they include multiple text segments as well as a significant array of structured metadata. Such metadata includes class codes, citations, inventors, assignees, and family relationships. The following highlights some of the key types of patent data as they relate to patent landscaping. Note that each type has significant limitations when used in patent landscaping.

2.1 Text

Patents contain multiple text segments, such as a title, abstract, and detailed description (body of the patent). Additionally, patents include claims. The claims define the scope of the property right and must be supported by the detailed description.

A common traditional technique for constructing landscapes is to perform boolean keyword searches on some or all of the text segments. However, this can be difficult when words have multiple meanings, when multiple words can describe the same subject, or simply because of spelling variations. Further, constructing such queries can often involve constructing long boolean queries with multi-word phrases and require that the user understand all aspects of the topic.

2.2 Class Codes

Class codes are topical labels applied by a patent office to classify patents. The primary class code regimes are the US class codes (exclusive to the USPTO) and the Cooperative Patent Class (CPC) codes (applied worldwide). Both are large hierarchical taxonomies with thousands of nested labels. Patents are typically assigned to multiple class codes.

A common technique for constructing landscapes is to select all the patents from one or more class codes. However, identifying the desired class codes can be difficult and the underlying patents sometimes deviate from the description of the class codes. Additionally, the codes, either alone or in combination, may not line up well with the topic for which the user wants to construct a landscape. Even if they do, selecting the right combinations of codes requires the user to have a nuanced understanding of the topic.

2.3 Citations

Like scholarly publications, many patents have citations. During prosecution (the process for obtaining a patent), the examiner as well as the applicant may cite publications that are relevant to the patent’s claims in evaluating whether the claims are novel and nonobvious. Frequently, the cited publications are patent publications. A patent publication is a published version of a patent or patent application produced by a patent office. A given patent may publish multiple times in slightly different forms. The most common example of this is the publication of a patent at the application stage and later publishing again as an issued patent.

When building landscapes, analysts sometimes will expand an initial list to include cited patents. However, this can lead to noise as examiners and applicants often cite patents that are not especially relevant.

2.4 Family

One somewhat unique aspect of patents is that they have family relationships. This is often expressed as a family ID. A patent family includes different publications of the same patent/patent application as well as other patents/patent applications that have a priority relationship. One common instance where a priority relationship occurs is where the applicant files the same application in multiple countries (patent rights are country specific). Another is where the applicant files multiple applications in the same country that share a detailed description, but have different claims. Typically, all family members relate to the same technical topic.

It is common to include all family members in a landscape. However, this is typically just a final step to increase coverage for the landscape.

3. AUTOMATED LANDSCAPING

A new automated approach for generating patent landscapes is presented below. This approach greatly mitigates many of the limitations discussed above by leveraging human insights, patent metadata, and machine learning.

This approach takes a human curated set of seed patents and expands it in order to populate a patent landscape. A human curated seed set is a sound starting point as it provides human insight as to the contours of the landscape. The seed set is then expanded using citations and class codes.

The initial results are often over-inclusive, but this is mitigated by pruning out less relevant patents using a machine learning model. The model provides a form of double verification. It is trained using the seed set as positive examples and a random subset of patents not included in the expansion (anti-seed) as negative examples.

The machine learning model can also be applied to the entire patent corpus, not just the expanded patents. This will lead to
greater recall, but will also likely result in a drop in precision as there is no longer double verification outside the expansion. Additionally, this may require significantly more computing resources.

Figure 2 shows a high-level flow of the process. Importantly, the Expansion will be discussed later in the Expansion section.

3.1 Seed Patents

The seed set is the basis for the landscape so its composition is very important. Errors in the seed will propagate throughout the resulting landscape and become magnified. It is important that the seed set be accurate as well as representative. For example, if one wanted to build a landscape for mobile phones, it would be important to include patents relating to screens, antennas, housings, etc.

An advantage to using a seed set is that it minimizes human effort. It frees the user from having to distill search features. No need to identify keywords or class codes. It can also mitigate human review, by allowing the user to simply focus on a narrow subset of the patents to be included in the landscape. Seed sets can often be sourced from pre-existing lists that a user created for another purpose. Further, a user may not need to fully understand all the technical nuances of the topic as much of this is extractable from the seed set. For example, a user building a speech recognition landscape would not need to know about phonemes or hidden Markov models as these common features can be inferred from the patents in the seed set.

Depending on the topic, the size of the seed set may vary. For narrower topics, such as email, a seed set of a few hundred patents may be sufficient. For broader topics, such as processors, a couple thousand patents may be required. Note that these estimates are merely based on informal experimentation taking into account subjective evaluations of patents in the resulting landscapes, the amount of expansion, and the accuracy of the models all in relation to the seed set size. Additional research would be worthwhile here.

3.2 Expansion

To identify candidate patents to include in the landscape, the seed set is expanded based on structured patent metadata. This approach uses class codes and citations, though one could also consider using other types of metadata (e.g., inventors and assignees).

3.2.1 Family Citations

A particularly robust technique for citation expansion is to use bidirectional (i.e., forward and backward) family citations. This captures any patent where that patent, or its family members, cite to or are cited by a patent in the seed set or a family member of a patent in the seed set. This increases the size of the expansion, while sacrificing little accuracy as family members all tend to involve the same technical topic. An efficient way to compute the expansion is to build a family citation graph, where the nodes are families and edges are citation relationships. Then one can simply query the graph for all families within one degree of a family in the seed set.

3.2.2 Class Codes

Class codes are another way to expand the seed set. This can be accomplished by identifying highly relevant class codes and expanding the seed set to include all patents that have those class codes. Highly relevant class codes can be identified by evaluating the distribution of class codes in the seed set relative to the distribution of class codes in all patents. For example, selecting a class code where (1) it occurs in at least 5% of the patents in the seed set, and (2) the ratio of patents in the seed set having the class code is 50 times higher than the ratio of all patents having the class code. This second condition is important because some class codes could be very prevalent in the seed set, but not very relevant. These very prevalent class codes tend to be very general and can span multiple topics (e.g., CPC code G06 “computing; calculating; counting” [7]). Expanding on such class codes could bring in hundreds of thousands of patents, most of which would not be especially relevant.

3.2.3 Combining Approaches

Expansion by citation and expansion by class code are not mutually exclusive and performing both together is beneficial. Expanding by citations is good because the expansion generally reflects all the main technical aspects present in the seed set. However, patent families with few to no citations may not be identified and some citations may be to irrelevant patents. Expanding by class codes is good because all patents have at least one class code and the application of the codes is fairly accurate. However, many technical aspects may not have a specific class code, so expanding on class codes alone will leave gaps. Thus a combination of class codes and citations is preferred.

3.2.4 Running at Scale

We run our automated expansion and machine learning process (discussed in section 3.4) against all US-issued patents since 1980, encompassing approximately 15 million patents and published applications. The data is taken from an internal corpus and is approximately 40 terabytes in size. Essentially three pipelines are run: first, for expansion, second for feature extraction and model training, and thirdly to classify patents. These pipelines are scaled through Map Reduce [4] and an internal implementation of Google’s Cloud Dataflow.

3.3 Types of Expansions

The citation and class code expansion described above can be combined in multiple ways to construct landscapes. The following describes two examples: one with relatively narrow relevance and one with relatively broad relevance. Note that both these approaches use expansion both to identify the anti-seed, but also to pre-filter the results of the final landscape. Both the broad and narrow landscapes use the same sequence of expansions. The specific sequence used here is enumerated in pseudo-code as follows:

Seed Citations =

FamilyOf(ExpandByRef(FamilyOf(SeedSet)))

Codes =

ExpandByCode(HighlyRelevantCodes(
       SeedSet.CPC, FullCorpus.CPC))

Level 1 = Seed Citations U Codes

Level 2 = FamilyOf(ExpandByRef(Level 1))

Anti-Seed = Sample(FullCorpus - Level 2, n)

In step 1, the seed is expanded by family citation. In step 2, the seed is expanded by highly relevant class codes. In step 3, the citation and class code expansions are combined resulting in Level 1 of the landscape. In step 4, Level 1 is expanded by family citation to produce Level 2 of the landscape. Finally, in step 5, the anti-seed is created by sampling patents outside Level 2 (Level 1 and the seed are a subset of Level 2). The anti-seed is used as negative
examples in a later machine learning pruning process described in section 3.4. In this research, we used between 10,000 and 50,000 patents for the anti-seed set. Deeper, more broad topics tended to benefit more from larger anti-seed sets.

3.3.1 Narrow

In this narrow landscape example, only Level 1 is retained. The expanded patents in Level 1 are pruned using the machine learning model to ensure precision. This arrangement is well suited for constructing relatively narrow landscapes. Figure 3 gives the intuition of the narrow landscape. Note that the Level 2 expansion described above is still performed to select the anti-seed set.

![Figure 3 – narrow expansion](image)

3.3.2 Broad

In this broad example, the patents from both Level 1 and Level 2 are retained. The expanded patents from Levels 1 and 2 are pruned using the machine learning model. The broad landscape is shown in Figure 4. This broad landscape example is more useful where a user desires to capture tangential topics as well as the primary topic. For example, in a GPS device landscape, a user may want to capture general touchscreen patents as touchscreens are a popular feature of GPS devices.

![Figure 4 – broad expansion](image)

3.4 Machine Learning Based Pruning

Both the class code and citation expansions will be over-inclusive. The results can be filtered using a supervised machine learning model to exclude non-relevant patents, which is covered in sections below. First, however, we look at some existing machine learning work used to automate patent classification.

3.4.1 Existing Machine Learning Approaches

Most available literature on machine learning based approaches to patent classification focus on fully supervised classification, where both positive and negative training samples are available. Many approaches use the existing hierarchical International Patent Class (IPC) and its replacement (CPC) schemas. These approaches tend to have relatively low accuracy due to the sheer volume of labels, or they focus on a small subset of labels. There are many thousands of classes across the IPC and CPC schemas, which vary greatly in distribution, making for a challenging classification task. For example, C.J. Fall, et. al. [8] focus on a small subset of IPC labels a using bag-of-words approach and achieve between 30-79% accuracy, but conclude that the results are not sufficiently strong to rely on in real-world scenarios. Other interesting approaches incorporate time-based information to increase robustness [12] with accuracy above 80%, or using information retrieval techniques to classify patents by finding similar patents already classified into the IPC and then use those patents’ classes [16]. Additional methods using bag-of-words models have been described as well, e.g. for routing requests to examiners at the European patent office [9] and to assist with patent searching in a manner similar to medical search on PubMed [6], with relatively good results (approaching 90% accuracy), but again against the IPC classes. Most of the existing automated approaches do not consider structured metadata outside of class codes, with some notable exceptions [11].

We are unaware of any techniques that combine 1) semi-supervised techniques to auto-generate portions of the training data, 2) arbitrary labels where the user controls the domain, and 3) running at scale across the entire patent corpus. We were unable to find approaches that do any two of the three, in fact.

3.4.2 Training Data - Seed & Anti-seed

Good training data is essential for any supervised machine learning model. The seed set provides positive examples; however, negative examples are needed as well. One can think the of negative examples as the anti-seed.

An efficient way to create the anti-seed is to sample from patents not included in the expansion. If the expansion is done correctly, only a very small fraction patents not included in the expansion should be relevant to the topic of the landscape. Sampling from these patents provides a large amount of reasonably accurate and representative negative training examples. The best part is that it requires no human effort.

In this research, positive (seed) and negative (anti-seed) data reside at differing poles, with little representative training data that reside in “the middle”. This is by design, since the goal is to prune data that isn’t very similar to the seed set, however, it is also a limitation of the semi-supervised approach since the sampling technique has built-in bias.

3.4.3 Machine Learning Methodologies

Here, one could use a wide variety of supervised machine learning methodology from simple regressions to deep neural networks. The following describes two different methodologies that were applied with good success.

3.4.3.1 Ensemble of Shallow Neural Networks using SVD Embeddings

One methodology that produced good results was an ensemble of shallow neural networks using singular value decomposition (SVD) embedding [10]. This involves constructing SVD embeddings for each patent data type and training multiple neural networks using the embeddings as inputs. The patent data types that proved to be most useful were family citations, text (title plus abstract and claims), class codes, and inventor citations. The outputs of the neural networks are weighted using a logistic
regression model. The results for this approach (below) reflect a single machine implementation with very small SVD embeddings (64 values per data type). Increasing the embedding size should increase accuracy.

Constructing the SVD embeddings is fairly straightforward. For each patent data type construct a sparse matrix of patents by features then apply SVD to compress the sparse feature space into fewer dense features. For family citations the matrix is a binary patent by patent matrix with the positive cells representing citation relationships between patent families. This is very similar to how one would apply SVD embedding for network analysis[17]. For text, use a patent by term frequency matrix. This is essentially a LSA approach [5]. For class codes use a binary patent by class code matrix. For inventor citations use a binary patent by inventor matrix with the positive cells representing a citation relationship between a patent and an inventor. Given the size of the matrices (millions by tens of thousands), a SVD approximation algorithm may be preferred, such as Lanczos bi-diagonalization [1].

The SVD embeddings are then used as inputs to the ensemble of neural networks. A different neural network is trained for each patent data type, as well as a composite neural network that is trained on all of the embeddings concatenated together. The hyperparameters for each neural network were: (1) an input layer that is the same size as the embeddings, (2) a single hidden layer, and (3) an output layer with a single node.

Once all the neural networks are trained, a logistic regression model is used to weigh their outputs. Using the ensemble increased accuracy and helped prevent overfitting making the machine learning more robust. This is especially important for an automated approach that seeks to reduce human effort by avoiding tuning. Interestingly, the weights assigned to each of the patent type specific networks can vary significantly from landscape to landscape. For example, landscapes that involved topics having well defined class codes had higher weights on the class code network. Conversely, landscapes that did not have well-defined class codes had higher weights on citations and examiner citations.

3.4.3.2 Perceptrons using Random Feature Projection Embeddings

A second machine learning approach applied during research was the Perceptron [15] algorithm and embeddings using Random Feature Projection (RFP) [2]. When dealing with data at a large scale, such as the universe of patents, sometimes it’s preferable to sacrifice a small amount of accuracy for large increases in efficiency. In addition to more traditional methods such as TF/IDF, a very fast way to reduce a high-dimensionality bag-of-word feature space -- in this case, hundreds of millions of tokens across tens of millions of instances -- is by using RFP. Random feature projection is very fast and depending on the dimensionality of the resulting feature vector, can provide increased efficiency of machine learning algorithms at little-to-no loss of accuracy.

As with the shallow neural net approach described above, many types of features were used. In particular, the following proved useful:

- n-grams of size 1 and 2 for abstract, description, claims, and references’ titles
- bigrams occurring at least twice across a patent’s description and claims
- the CPC classifications of the patent

Notably absent in this approach, however, was the family citation and LSA of the text. Instead, bag of word text features and CPC classifications were combined into a single matrix. RFP was then applied on the full matrix, reducing the dimensionality to only 5,000 columns. Larger feature vectors, e.g. 10,000 or 15,000 columns, proved to have negligible increase in accuracy while increasing model size and slowing down the training and inference process.

Finally, while many algorithms were experimented with on the RFP data embeddings, including support vectors and logistic regression, the perceptron proved the most accurate on RFP embeddings by a minimum of 5% F1 score across most experimental landscapes.

4. RESULTS

As discussed above, most automated classification research focuses around automating classification into the existing IPC or CPC classification hierarchies [6; 9; 12; 16].

The current research was performed on an internal Google dataset containing all patents since 1980 that are English-language and US-issued, which is approximately 15 million patents and published applications. However, this automated technique should apply to non-US patents as most countries make use of citations published applications. However, this automated technique should further be applied to non-US patents as most countries make use of citations published applications. However, this automated technique should apply to non-US patents as well.

4.1 Analysis Methodology

The machine learning analysis is primarily done by evaluating F1 scores [19]. Reported here are the F1 scores and additional details for three example landscapes. Further, a plot of 60 landscapes illustrating the broad applicability of the methodology.

In addition to the standard machine learning analysis, internal patent domain experts reviewed the resulting landscapes for validity. This was somewhat subjective and involved reviewing representative portions of the results, comparing overlap to traditionally generated landscapes, and analysis of aggregate statistics (e.g., assignee distributions).

4.1.2 Machine Learning Results

Figure 5 shows the landscapes across the top columns, and in subsequent rows are the F1 scores for the machine learning process, the number of patents in the seed set, the number in the narrow expansion, the number in the narrow landscape after pruning, and similarly for the broad landscape.

The F1 score was calculated by training the machine learning models on the seed and anti-seed sets, running k-folds cross-validation (k=10), and retaining the maximum F1 score. Note that the scores reflect the F1 score for the minority class, that is, the patents in the seed set. What this tells us is that, while the training set is imbalanced (as heavily as 1:10 in extreme cases), by calculating the F1 for the minority class we show the classifier isn’t always choosing the majority class to “cheat.” The majority (negative) class also scores strongly, but results are not presented as it is not important to the task of pruning with high precision. We also vary the randomization seed for cross-validation splits between runs to help avoid overfitting.

While the results are very promising, they may overstate the actual accuracy of the model as there are likely to be few borderline positive/negative examples in the training data. In the current context, this is acceptable, as the goal is to prune patents from the expansion that are not clearly related to the seed topic.
The results are in line with internal expert two sets of validation (expansion and machine learning), as detailed here. This also suggests the model is dealing with close positives and negatives fairly well, at least in aggregate. We see similar citation and class code expansions are a reasonably good way to identify potentially relevant patents and strong negatives (anti-seed). This suggests the model is dealing with close positives and negatives fairly well, at least in aggregate. We see similar shapes of expansion and pruning across other landscapes not detailed here.

<table>
<thead>
<tr>
<th>Topic</th>
<th>browser</th>
<th>operating system</th>
<th>machine learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFP F1 Score</td>
<td>.954</td>
<td>.882</td>
<td>.929</td>
</tr>
<tr>
<td>SNN F1 Score</td>
<td>.930</td>
<td>.797</td>
<td>.895</td>
</tr>
<tr>
<td># Patents in Seed Set</td>
<td>1094</td>
<td>1083</td>
<td>1238</td>
</tr>
<tr>
<td># in Level 1 Expansion</td>
<td>72,895</td>
<td>44,779</td>
<td>46,570</td>
</tr>
<tr>
<td># in Narrow Landscape (RFP)</td>
<td>49,363</td>
<td>30,319</td>
<td>28,947</td>
</tr>
<tr>
<td># in Level 2 Expansion</td>
<td>511,173</td>
<td>505,176</td>
<td>527,203</td>
</tr>
<tr>
<td># in Broad Landscape (RFP)</td>
<td>92,362</td>
<td>118,675</td>
<td>86,432</td>
</tr>
</tbody>
</table>

**Figure 5 – seed and expansion counts for three sample landscapes**

As shown in Figure 5, patents identified in the Level 1 expansion are much less likely to be pruned by the machine learning model than patents identified in the Level 2 expansion. For the browser topic 68% (49,363/72,895) of patents identified in Level 1 are retained whereas only 10% (92,362-49,363)/(511,173-72,895) of patents identified first in Level 2 are retained. This, especially in light of the high F1 scores, supports the notion that the citation and class code expansions are a reasonably good way to identify potentially relevant patents and strong negatives (anti-seed). This suggests the model is dealing with close positives and negatives fairly well, at least in aggregate. We see similar shapes of expansion and pruning across other landscapes not detailed here.

**Figure 6 – distribution of F1 scores for some sample patent landscapes produced using RFP & Perceptron approach**

The plot from Figure 6 below shows the distribution of F1 scores across a range of 60 topics in ascending order by F1 score. Details of each topic are omitted, however this demonstrates that the Automated Patent Landscaping technique can be applied to numerous topics with success. Most landscapes achieved an F1 score above .90. There were some with poorer scores, of course. For example, the lowest score was for a landscape about “car infotainment.” In that model, the classifier frequently confused car-related patents that had no infotainment components as being part of the topic. Were this landscape to be used in a real world scenario, more tuning of the model and seed set would likely be necessary.

**4.1.3 Patent Landscape Results**

The resulting patents in both the broad and narrow sets have two sets of validation (expansion and machine learning), as described previously. The results are in line with internal expert evaluations that found the resulting landscapes to be both accurate and reasonably comprehensive using the methodology described above in section 4.1.1.

As demonstrated in the results and expert evaluation, this automated patent landscaping approach provides a repeatable, accurate, and scalable way to generate patent landscapes with minimal human effort. This has significant implications for companies, governments, and investors that benefit from patent landscape analysis, but are discouraged because of cost and effort. Further this approach, especially if widely adopted, makes patent landscapes more compelling to external audiences as they can focus on the results rather than questioning the underlying methodology.

It should be noted that the landscaping approach described here could be applicable to other domains that have similar metadata. Two that come to mind, in particular, are scholarly articles and legal opinions.

**5. REFERENCES**


An open-source database for empirical analysis of judges and judicial decisions

Elliott Ash* and Michael Lissner†

April 29, 2016

Motivation

A growing body of evidence points to a well-functioning legal system as a key component of economic growth and prosperity (e.g. Djankov et al., 2003). The inclusiveness of institutions, for example, is often mediated by access to courts and an impartial judiciary (Acemoglu and Robinson, 2012). In particular, a recent empirical literature in economics suggests that common law systems, where a large number of judges working on a large number of cases gradually construct our body of legal rules, tend to work well in regulating disputes and conflicts (Gennaioli and Shleifer, 2007; La Porta et al., 2008). Judges are tasked with interpreting and developing the law through their written decisions, and therefore are important actors in the regulation of our civil society.

A systematic empirical analysis of the role of judges in common law systems has posed a substantial challenge to social science researchers. First, constructing large-scale statistical datasets from large corpora of legal opinions has been difficult due to restrictions imposed by proprietary legal data sources such as WestLaw and LexisNexis. Second, these legal documents have not been merged with data on institutional variables that may be influencing the written caselaw, nor with data on socioeconomic outcomes that may be determined in part by the legal incentives imposed by written caselaw. We even lack a public data set that systematically and consistently matches judges to

*Post-doctoral Research Fellow, Center for Study of Democratic Politics, Princeton University; Assistant Professor of Economics (on leave until 2017), University of Warwick. Contact: e@elliottash.com.
†Lead Developer and Co-Founder, Free Law Project. Contact: mike@free.law. This research is supported by the National Science Foundation and the John S. and James L. Knight Foundation.
their authored opinions. Third, even if one had these data, there would still be the challenge of analyzing large text corpora in a statistical framework. The technology for organizing and analyzing a large corpus of a judge’s written text, along with associated metadata, is still somewhat experimental, and rapidly changing due to new advances in computational capacity and machine learning algorithms (e.g. Ash, 2016).

This proposal introduces CourtListener’s judicial database, an important step forward in the scientific analysis of the judiciary. CourtListener is a free and open-source repository, search engine, and research platform for analysis of court opinions and other legal documents. The new system is a key set of metadata linked to those opinions: a comprehensive database of judges and other persons involved in the operation of the legal system. We hope that this database, its APIs, and its bulk data will provide a valuable tool for practicing attorneys, social-science researchers, and others that are interested in the functioning of the legal system.

This database will contribute to an active literature on judicial institutions and judge performance. Epstein et al. (2013) is the leading book-length treatment of federal judges. Lim (2013) compares the behavior of elected versus appointed state court trial judges and finds that as a group appointed judges are more homogeneous and tend to make less harsh sentencing decisions. Finally, Ash and MacLeod (2015, 2016) analyze how the behavior of state supreme court judges responds to changes in the institutional work environment. The new database will allow researchers to build on and improve these existing studies.

A New Judges Database

The Free Law Project, along with researchers at Princeton and Columbia, has recently launched a database of thousands of judges working in U.S. courts. The initial version of the database has nearly 8,500 judges from federal and state courts, all of which are available via the APIs, in bulk data, and via a new judicial search interface. The database is aimed to be comprehensive, including as many facts about as many judges as possible.

An important technological feature of this data, and an improvement on existing datasets, is that it lives in a relational database schema, implemented in PostgreSQL using a Django interface. One can think of this platform as a “People Database” rather than a “Judges Database,” since many individuals in the data have not worked as judges.
Figure 1: William Howard Taft's Profile Page

(a) Judicial Positions

JUDGE of SUPREME COURT OF THE UNITED STATES (June 30, 1921 – February 3, 1930)

- Appointed By: Warren G. Harding
- Selected By: Appointment (President) on June 30, 1921
- Confirmation Date: June 30, 1921
- Vote Info: Senate voted by voice vote
- Termination Reason: Resigned

(b) Non-Judicial Positions

Includes a variety of positions such as governor, president, clerk, practitioner, public defender, attorneys general, and professor.

KENT PROFESSOR OF LAW at YALE UNIVERSITY (1913 – 1921)

PRESIDENT OF THE UNITED STATES (1909 – 1913)

(c) Political Affiliations

- Republican (1891)
- Beginning December 16, 1901: Republican (1901 – 1909)

Education History

- Attended Yale University and received a Bachelor's degree in 1878

(c) Non-Judicial Positions

- Richard Rhoads
  Judge of District Court, D. Arizona (August 26, 1912 – March 3, 1913)
- John Moses Cheney
  Judge of District Court, N.D. Florida (August 26, 1912 – March 3, 1913)
- Clinton Woodbury Howard
  Judge of District Court, W.D. Washington (August 26, 1912 – March 3, 1913)
- James Madison Morton Jr.
  Judge of District Court, D. Massachusetts (August 12, 1912 – January 19, 1913)
- Arthur J. Tabb
  Judge of District Court, B.D. Michigan (August 6, 1912 – December 2, 1944)
- Frederic O. Dodge
  Judge of Court of Appeals for the First Circuit (July 22, 1912 – June 20, 1918)
- Joseph Whittaker Thompson
  Judge of District Court, E.D. Pennsylvania (July 16, 1912 – February 3, 1931)
- Edward E. Cushman
  Judge of District Court, W.D. Washington (May 1, 1912 – January 25, 1940)
- John Bayard McPherson
  Judge of Court of Appeals for the Third Circuit (April 3, 1912 – January 20, 1919)

SECRETARY OF WAR (1904 – 1908)

CIVIL GOVERNOR of PHILIPPINE ISLANDS (1901 – 1904)

PRESIDENT of U.S. PHILIPPINE COMMISSION (1900 – 1901)

PROFESSOR AND DEAN at UNIVERSITY OF CINCINNATI (1895 – 1900)
For example, all the U.S. presidents are in the database, and they are systematically linked as appointers to the federal judges they have appointed to office. An interesting historical wrinkle that is easily visualized in this schema is the unique situation of William Howard Taft, who served as a Sixth Circuit Judge, as U.S. President, and as Chief Justice of the U.S. Supreme Court.

Taft’s profile page is illustrated in Figure 1 across three columns. Column a shows the non-career information collection. The data has basic biographical information, such as date/place of birth/death, as well as race, gender, and religion. There is political affiliation information, both from running on the presidential ballot as a Republican, and from being appointed by a Republican president. The data allow for the full education history of a judge, including the school, the degree, the major or concentration, and the year earned.

Column b shows the section of Taft’s profile page associated with his judicial career positions. This section lists the positions at all the courts, a range of information about when and how the judge was appointed, and also how the position ended. This section will also provide statistics about and links to the set of opinions written during these positions.

Finally, Column c shows the profile section associated with non-judicial positions. In Taft’s case, this section is particularly interesting because he was a president who himself appointed many judges to the federal bench. These judges are listed under Taft’s position as President of the United States. In addition, we can see that Taft was a man of many talents: He served in academia, as secretary of war, and even as governor of the Philippines.

The database includes a large set of additional information that can be accessed through the API or through the bulk data service. Besides the full name, we have nicknames and aliases, for example. For each judicial position, we have information on how they were appointed or elected, and the number of votes for and against. We have clerkships, which allows tracing of clerkship networks. We have collected information on ratings that a judge has been given by the American Bar Association. For elected judges, we have records of the reelection events, and whether the judge prevailed or lost reelection. In the future we are open to expanding on the fields for this data.

We have collected all available public datasets and added a large amount of data ourselves. Most of the information on federal judges comes from Federal Judicial Center (FJC), which compiles and maintains a spreadsheet with that information. The data on state court judges have been collected as part of an ongoing project on judicial behavior.
funded by the National Science Foundation (Ash and MacLeod, 2015, 2016). The data on schools and their aliases was compiled by the Department of Education and made available via their Developer Portal.

However, there are many actors in the U.S. legal system and the database is far from complete. For example, we have been unable to find a comprehensive source of data for magistrate judges or federal bankruptcy judges. Moreover, the data on lower court judges in the states is poor. We hope that interested researchers will collaborate with us in contributing more data.

References

Absolute Constructions and the Second Amendment: A Corpus Analysis

I. Introduction

The ongoing conversation regarding the grammar and meaning of the Second Amendment is not likely to be resolved anytime soon, but over the last 20 years there have been excellent contributions to that conversation that have made the context of that 1791 amendment richer, deeper, and more complex. David Yassky’s amicus brief of 1999 did some basic work in that direction, clarifying some points of punctuation, making helpful distinctions regarding the then-current meanings of *keeping* and *bearing arms*, and setting a broad picture of the larger work being done by provisions such as the Second Amendment at the national and local constitutional level. The grammatical structure of the Second Amendment and the semantic implications of that structure received a good deal of attention in the 2008 Supreme Court ruling (see especially Volokh [1998]; and Scalia, Stevens, and the Linguists’ Brief [Baron, Bailey, and Kaplan] in *District of Columbia v. Heller*, 2008). And a year after that ruling, David Thomas Konig performed a significant scholarly service by providing a study of the public meanings of and the political culture surrounding the uses of preambles in constitution-making during the Revolutionary period of American history.

Over the last 25 years, there has also been a rich new body of work done on the absolute construction in English, and this scholarship has direct bearing on the Second Amendment conversation. An absolute construction (AC) introduces the main clause of that amendment, and the meaning and import of that opening absolute construction are among the most controversial matters in the larger contemporary discussion. One of the things that makes that subject so difficult is the relatively small body of ACs used as points of reference and comparison in the discussion of the Second Amendment.

In this study, I will do three things: (1) describe briefly the modern English AC in terms of its grammatical features; (2) describe a variety of ACs in publications from Great Britain and America in the years before and after 1791; and (3) provide some context for examining the semantic implications of such ACs and the main clauses (the matrix clauses) they belong to.

What follows is the text of the Second Amendment as ratified in 1791:

A well regulated Militia, being necessary to the security of a free State, the right of the people to keep and bear Arms, shall not be infringed.

The main or matrix clause of this amendment is *the right of the people to keep and bear arms shall not be infringed*, and the introductory material, the preamble, preceding the matrix clause is the absolute construction—*a well regulated militia being necessary to the security of a free state*—that has caused some confusion and controversy over the years.
II. Absolute Constructions in Modern English

There is a rich body of work describing the absolute construction in English, much of it appearing in the last 25 years, from the 1991 publication of Bernd Kortmann’s *Free Adjuncts and Absolutes in English* to a flurry of recent articles by other linguists in Europe, particularly a group of linguists at KU Leuven (Belgium), on the history and development of the AC in English and other Germanic languages. [Note: the AC is also called a nominative absolute, an absolute phrase (USA), and an absolute clause (UK).]

An AC is made up of an expressed subject and a non-finite verb form (often a present or past participle), with predicate structures similar to and as varied as those of complete English clauses. Here are several examples from the Corpus of Contemporary American English (COCA):

(1) Greece having been “saved” by “Europe,” the ordinary Greek citizen is allowed to withdraw $460 per week for all the purposes of life. (COCA, *National Review* 2015)

(2) When these statistics were split for gender, as shown in Figure 3, notable differences emerged, the most significant being in friend encouragement, with 25% of girls receiving encouragement from their friends compared to only 11% of boys. (COCA, *Journal of Adolescent & Adult Literacy*, 2015)

(3) When these statistics were split for gender, as shown in Figure 3, notable differences emerged, the most significant being in friend encouragement, with 25% of girls receiving encouragement from their friends compared to only 11% of boys. (COCA, *Journal of Adolescent & Adult Literacy*, 2015)

(4) With small-company shares having grown expensive in recent years, funds that focused on the cheaper “value” end of this universe tended to outperform in 2014. (COCA, *Money*, 2015)

(5) With the club having done little to endear itself to its local fan base, attendance is expected to be at a record low for the home opener Sunday when the club plays Seattle—which has been in the Super Bowl the past two seasons. (COCA, *St. Louis Post-Dispatch*, 2015)

(6) Her face a mask of polite civility, she focused on the duke. (COCA, *The Courtesan Duchess*, 2015)

And here are several examples from late-18th-century American English:

(7) This being the apparent design of the constitution, it would be highly improper to magnify the function into an important prerogative, even where no rights of other departments could be affected by it. (Corpus of Historical American English; *Federalist Papers*, 1788)
(8) This was the way by which the psalmist’s soul was elevated in high devotion, his heart enlarged in gratitude, and his lips opened wide in praise to God. (American Thanksgiving sermon, 1766)

(9) This was the way by which the psalmist’s soul was elevated in high devotion, his heart enlarged in gratitude, and his lips opened wide in praise to God. (American Thanksgiving sermon, 1766)

(10) A man going one day along the street, an impudent fellow came swaggering up to him, and thrust between him and the wall. (American Jest Book, 1785)

(11) But he telling them, if they did not get pretty near the Land, the Current would drive them to the Leeward of all the Islands, they left him to steer as he thought best, but intended to throw him overboard before he got to Land. (American piracy trial, 1769)

(12) Searching for the foundations of this proposition, I can find none which may pretend a colour of right or reason, but the defect before developed, that there being no barrier between the legislative, executive, and judiciary departments, the legislature may seize the whole: that having seized it, and possessing a right to fix their own quorum, they may reduce that quorum to one, whom they may call a chairman, speaker, dictator, or by any other name they please. (Jefferson, Notes on the State of Virginia, 1785)

In these twelve examples, some of the main features of the absolute construction are noteworthy:

- In contemporary English, many ACs are introduced by the word “with,” which is not a preposition in such instances but instead an augmentor, used mainly to set off and announce an AC, as in (3), (4), and (5).
- The AC typically has a non-finite verb (that is, a verb form not marked for person, number, and tense), but there are also non-verb versions of the AC, as in (6); in an AC such as this one, the dropped verb form is generally understood to be “being,” and the internal grammar of this AC is Subject–Copula–Subjective Complement (the noun variety, here), with the copula deleted.
- ACs can be placed before the matrix clause (as in (1), (4), (5), (6), (7), (10), and (11)), after the matrix clause (as in (2), (3), (8), and (9)), or somewhere in the middle of the matrix clause (as in (12)).
- ACs can be quite simple grammatically, as in (6), described above, and (10), with a Subject–Intransitive Verb followed by two adverbial units; but they can also be quite complex grammatically, as in (11), with its Subject–Transitive Verb–Indirect Object–Direct Object structure, the Direct Object being a noun clause introduced by an adverb clause.

The AC introducing the Second Amendment (“A well regulated militia, being necessary to the security of a free state”) has the following basic features: (1) it is placed in the initial position in the sentence, appearing before the matrix clause it belongs to and modifies; and (2) its
grammatical structure is as follows: Subject (“militia”)–Copula (“being”)–Subjective Complement, specifically, a Predicate Adjective (“necessary”).

III. The Corpus, the Search, and the Results

Last year, with the assistance of Carrie Ott, one of my undergraduate students at Calvin College, I constructed a corpus made up of British and American works published between twenty to thirty years before 1791 and twenty years after, using materials from five genres—sermons, fiction, political writing, legal writing, and nonfiction prose. To this original collection I added “Book I” of Blackstone’s *Commentaries on the Laws of England* (1765) and the entirety of *The Federalist Papers* (1787/88). This augmented corpus has a total of 39 works from America and 41 from Great Britain; the entire corpus has in it 1,039,862 words, with approximately 35,000 more words in the British part of the corpus than in the American part.

I used the AntConc concordancing software to analyze this corpus, searching mainly for ACs built on the present participle “being,” the participle used in the Second Amendment AC, and for ACs built on the present participle “having,” just in case interesting features turned up with the “having been” form of the non-finite verb. I was not actively searching for ACs featuring other verbs or those displaying the non-verb model, but if such ACs turned up in the neighborhood I was analyzing, I added them to the collection.

I extracted a total of 708 ACs from this corpus, the great majority of them based on “being,” alone or in combination with other verbs (440 ACs in all: 279 British, 161 American), and “having,” alone or in combination with other verbs (61 ACs in all: 27 British, 34 American).

**Placement of the AC:** Of the 708 ACs I extracted, 354 appear in the initial position, that is, immediately before the matrix clause that the AC modifies; 90 appear in the medial position, that is, inside the boundaries of the matrix clause; and 266 appear in the final position, following the matrix clause.

**“Being” as copula (linking verb):** Of the 440 ACs with “being” as the main verb or as part of it, 76 of them feature copulas (linking verbs) that have Predicate Adjectives (PA) as the complement. And of the 76 ACs with a copula and a PA complement, there are 35 ACs that are in the initial position, introducing the matrix clause. These 35 ACs are excellent matches for the Second Amendment AC, which has the same basic features: it is an AC in the initial position, with “being” as the copula and a PA as its complement. There are another 34 ACs with a copula plus Predicate Noun structure instead of a Predicate Adjective; these 34 ACs also appear in the initial position. (See Appendix A, Charts 1 and 2.)

**Distribution of ACs:** The 35 S–Copula-PA ACs (Appendix B) are distributed across the lines of nation, date, and genre: 20 are from American English texts, and 15 are from British; the earliest appears in 1760, the latest in 1793. Although this corpus is relatively small, it’s interesting that all but one of the initial-position ACs with the S–Copula–PA grammatical structure were in print before the ratification of the Second Amendment in 1791. And of the 34 S–Copula–PN ACs
(Appendix C), all but six were in print before the 1791 ratification. The AC in general and this particular kind of AC specifically were in wide distribution during this period of time.

Register of the ACs: The AC was a quite common feature of written English at the time. Across the five genres I assembled, sermons and nonfiction prose had the lowest number of absolutes, and political writing, legal writing, and fiction had many more absolutes. But for a grammatical structure that is typically associated with formal and literary writing, it’s worth noting that 4 of the 35 target absolutes come from a much less formal and literary source, The American Jest Book (1789), a collection of jokes, witticisms, and humorous stories. In fact, I found 77 ACs of all sorts in this collection, all but three of them appearing before the main clause and almost invariably providing the setup for a joke or a story.

IV. The Semantic Implications of Such Absolute Constructions

Absolute Constructions have traditionally been named and defined as “absolute” because the AC almost invariably has no modifying relationship with any particular word in the matrix clause; instead, the AC modifies the entire clause, as if the AC were an adverb clause. ACs can provide many different kinds of information, from cause, condition, concession, or time (referred to as “adverbial absolutes” when ACs perform these functions), to reason, manner, purpose, result, additional or accompanying circumstance, and exemplification or specification. Kortmann, Killie and Swann, Timofeeva, and others have thoroughly described, cataloged, and categorized the semantic relationships of an AC to its matrix clause; the AC is a truly versatile modifier, capable of providing very clear semantic signals or quite ambiguous ones. But there is substantial agreement regarding the semantic relationship of an initial-position absolute with the “being” participle as its verb: such an AC regularly functions as an adverb clause of cause.

According to Timofeeva (2012, p. 231), adverbial absolutes “establish a . . . framework within which the main predication holds”; Fonteyn & van de Pol (2015) observe that “sentence-initial adverbial clauses will express those adverbial relations that serve as frame-setting background information, i.e. temporal and causal relations,” and Kortmann (1991, p. 173) refers to the “apparent tendency of being-adjuncts/absolutes to function as idiomatic constructions of causality.”

There is, of course, nothing automatic about such ACs indicating primarily a causal relationship to the matrix clause, but in the 69 initial-position ACs I’ve referred to, it is by far the most common reading of the relationship of the AC to the matrix clause. In nearly every instance, it makes perfect sense of the entire sentence to turn the AC into an adverb clause of cause by introducing it with “because” and changing the participial “being” into a verb marked appropriately for person, number, and tense.

This is not a study of how to interpret the Second Amendment; that is for legal scholars and courts to determine. However, as Justice Scalia wrote in the majority opinion of District of Columbia v. Heller, “In interpreting this text, we are guided by the principle that ‘[t]he Constitution was written to be understood by the voters; its words and phrases were used in their normal and ordinary as distinguished from technical meaning.’”
In this context, it is important to keep in mind that the Second Amendment is not alone in its grammatical construction, nor does such a construction live mainly in constitutional statements; there are enough similar sentences from that period to provide a framework for understanding the semantic relationship between an initial absolute construction and its matrix clause. Such ACs work this way in legal writing, in nonfiction prose, in political writing, and in the setup for jokes and humorous stories. There is, thus, strong evidence that points to that Second Amendment AC as being put to use for that same reason, to establish “a framework within which the main predication holds,” functioning as an adverb clause of cause—an “idiomatic construction of causality.” Such a reading of the Second Amendment would challenge the central interpretive strategy of the majority opinion of District of Columbia v. Heller, namely, that the preamble of the Second Amendment (the absolute construction) can safely be ignored until after the clear meaning of the main clause has been examined. But such a reading of the Second Amendment would no longer be a reading of those “words and phrases . . . in their normal and ordinary . . . meaning.”

Grammatical analysis alone cannot solve the interpretive issues of the Second Amendment. The current conversation about how best to read that amendment is necessarily complex, involving legal precedent, cultural contexts, and the changing tides of history. But this conversation clearly does also involve a grammatical matter, and the Second Amendment’s relationship to the larger context of the history and meanings of the absolute construction is arguably significant enough to allow it a voice in this conversation as well.

---

**Corpus Tool and Corpora**

AntConc: a freeware corpus analysis toolkit for concordancing and text analysis, by Laurence Anthony. Available online at http://www.laurenceAnthony.net/software/antconc/


Works Cited


Appendix A: Second Amendment Corpus—Genres and Results

<table>
<thead>
<tr>
<th>Genres: Great Britain</th>
<th># of Words</th>
<th>All ACs</th>
<th>Initial ACs, “being,” PA</th>
<th>Genres: America</th>
<th># of Words</th>
<th>All ACs</th>
<th>Initial ACs, “being,” PA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Political</td>
<td>89,622</td>
<td>142</td>
<td>4</td>
<td>Political</td>
<td>268,004</td>
<td>112</td>
<td>10</td>
</tr>
<tr>
<td>Nonfiction</td>
<td>31,601</td>
<td>13</td>
<td>1</td>
<td>Nonfiction</td>
<td>32,559</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Legal</td>
<td>240,109</td>
<td>191</td>
<td>8</td>
<td>Legal</td>
<td>87,481</td>
<td>56</td>
<td>4</td>
</tr>
<tr>
<td>Fiction</td>
<td>85,735</td>
<td>61</td>
<td>1</td>
<td>Fiction</td>
<td>27,484</td>
<td>78</td>
<td>4</td>
</tr>
<tr>
<td>Sermons</td>
<td>90,078</td>
<td>25</td>
<td>1</td>
<td>Sermons</td>
<td>87,189</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Totals</td>
<td>537,145</td>
<td>432</td>
<td>15</td>
<td>Totals</td>
<td>502,717</td>
<td>276</td>
<td>20</td>
</tr>
</tbody>
</table>

Chart 1: All ACs, British and American = 708; Initial ACs with a “being” participle and a PA =35

<table>
<thead>
<tr>
<th>Genres: Great Britain</th>
<th># of Words</th>
<th>All ACs</th>
<th>Initial ACs, “being,” PN</th>
<th>Genres: America</th>
<th># of Words</th>
<th>All ACs</th>
<th>Initial ACs, “being,” PN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Political</td>
<td>89,622</td>
<td>142</td>
<td>9</td>
<td>Political</td>
<td>268,004</td>
<td>112</td>
<td>11</td>
</tr>
<tr>
<td>Nonfiction</td>
<td>31,601</td>
<td>13</td>
<td>0</td>
<td>Nonfiction</td>
<td>32,559</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Legal</td>
<td>240,109</td>
<td>191</td>
<td>9</td>
<td>Legal</td>
<td>87,481</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>Fiction</td>
<td>85,735</td>
<td>61</td>
<td>1</td>
<td>Fiction</td>
<td>27,484</td>
<td>78</td>
<td>1</td>
</tr>
<tr>
<td>Sermons</td>
<td>90,078</td>
<td>25</td>
<td>0</td>
<td>Sermons</td>
<td>87,189</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Totals</td>
<td>537,145</td>
<td>432</td>
<td>19</td>
<td>Totals</td>
<td>502,717</td>
<td>276</td>
<td>15</td>
</tr>
</tbody>
</table>

Chart 2: All ACs, British and American = 708; Initial ACs with a “being” participle and a PN =34
Appendix B

35 Absolute Constructions (Subject–Copula–Predicate Adjective) in the Initial Position, British and American (1760-1793)

A. British Texts (15):

1. but this sum being scarce sufficient to maintain my wife and children, who are five in number, I agreed to read prayers in the afternoon at another church, about four miles from hence; and for this additional duty I receive ten pounds more. (Fiction: The Adventures of Sir Launcelot Greaves, by Tobias Smollett, 1760)

2. But the Cynaethians, who inhabit the most rude and savage parts of Arcadia, having neglected all those helps of which, on that account, they had so much the more occasion; and being, on the contrary, subject to mutual divisions and contests, they are, at length, become so fierce and barbarous, that there is not a city in Greece, where such frequent and enormous crimes are committed, as in that of Cynaetha. (Nonfiction: An essay on musical expression, by Charles Avison, 1775)

3. By this Time, the Day being pretty far spent, the KING having Water brought him by the Earl of Pembroke and his Assistants, washed and rose from Dinner, before the third Course was brought in, and retiring into the Inner Court of Wards, and, being disrobed, went privately to his Barge, and so to Whitehall, where he landed. (Political: Coronation, 1760)

4. The real cause of the present scarcity being, in general, too well known to need any formal discourse to ascertain its origin, it may, perhaps, be thought impertinent in me to undertake that task. (Political: An inquiry into the connection between the present price of provisions, and the size of farms. With remarks on population as affected thereby, by John Arbuthnot, 1773)

5. But if these consequences do not hold, yet, their own calculations being just, the proportion of their numbers to that of the established church is four times, and their relative strength five times, what it was at the revolution. (Political: Test Act, 1790)

6. Therefore let there be supposed to be two bodies, absolutely without weight; whereof the first is projected upwards, with two degrees of velocity, to continue uniformly the same; and let the second be in like manner projected at the same instant, with one degree of celerity: now all consideration of weight, with respect to both, being out of the question: at the end of a certain time, the second will be below the first by a given distance; and at the end of double that time, by double that distance. (Political: The alteration of the constitution of the House of Commons, and the inequality of the land-tax, considered conjointly, by John Brand, 1793)

7. The house not being large enough to contain the people, he preached in an open field. (Sermon: “On The Death of The Rev. Mr. George Whitefield,” by John Wesley, 1770)
8. But this state of dependence being almost forgotten, and ready to be disputed by the Irish nation, it became necessary some years ago to declare how that matter really stood. (Legal: *Commentaries on the Laws of England, Book I*, by William Blackstone, 1765)

9. But this being liable to be attended with either fraud, or at least caprice, in the persons paying; and with either jealousies or mean compliances in such as were competitors for receiving them; it was now ordered by the law of king Edgar[n], that “dentur omnes decimae primariae ecclesiae ad quam parochia pertinet.” (Legal: *Commentaries on the Laws of England, Book I*, by William Blackstone, 1765)

10. They did not assemble without writ, and then make the throne vacant; but the throne being previously vacant by the king's abdication, they assembled without writ, as they must do if they assembled at all. (Legal: *Commentaries on the Laws of England, Book I*, by William Blackstone, 1765)

11. And, his own emoluments being probably considerable, the common council of London endeavoured to erect another post-office in opposition to his, till checked by a resolution of the commons, declaring, that the office of postmaster is and ought to be in the sole power and disposal of the parliament. (Legal: *Commentaries on the Laws of England, Book I*, by William Blackstone, 1765)

12. The sheriff being answerable for the misdemeanors of these bailiffs, they are therefore usually bound in a bond for the due execution of their office, and thence are called bound-bailiffs; which the common people have corrupted into a much more homely appellation. (Legal: *Commentaries on the Laws of England, Book I*, by William Blackstone, 1765)

13. And these canonical disabilities are either grounded upon the express words of the divine law, or are consequences plainly deducible from thence: it therefore being sinful in the persons, who labour under them, to attempt to contract matrimony together, they are properly the object of the ecclesiastical magistrate's coercion; in order to separate the offenders, and inflict penance for the offence, pro salute animarum. (Legal: *Commentaries on the Laws of England, Book I*, by William Blackstone, 1765)

14. But such marriages not being void ab initio, but voidable only by sentence of separation, they are esteemed valid to all civil purposes, unless such separation is actually made during the life of the parties. (Legal: *Commentaries on the Laws of England, Book I*, by William Blackstone, 1765)

15. It cannot be a trustee; for such kind of confidence is foreign to the ends of it's institution: neither can it be compelled to perform such trust, because it cannot be committed to prison[k]; for it's existence being ideal, no man can apprehend or arrest it. (Legal: *Commentaries on the Laws of England, Book I*, by William Blackstone, 1765)
B. American Texts (20):

1. **Hendrick being very solicitous to know what it was**, Sir William informed him, that he had dreamed that he (Hendrick) had made him a present of a particular tract of land (the most valuable on the Mohawk river) of about 5000 acres. (Fiction: *The American Jest Book, Containing a Curious Variety of Jests, Anecdotes, Bon-Mots, Stories, Etc.*; Philadelphia, 1789)

2. A quaker lodging at an inn, the **house being full**, a damning blade came up into his room, and would have hector’d him out; but he told him ’twas his room, and by yea and nay, he should not come there. (Fiction: *American Jest Book*, 1789)

3. **The father being unwilling to part with his mare**, the match was broke off. (Fiction: *American Jest Book*, 1789)

4. **A young lady being sick**, a physician was sent for to feel her pulse; she being very coy, and loth he should touch her naked skin, pulled her smock sleeve over her hand. (Fiction: *American Jest Book*, 1789)

5. **General Stark being indisposed**, colonel Spencer is added as a member. (Legal: “Trial of Major General Arnold,” 1780)

6. A considerable quantity of stores being forwarded to river, and some continental teams coming in, which were not expected, and **the general being desirous to have the waggons sent, to save the property at Eggharbour from falling into the hands of the enemy**, I desired mr. Jordan to go to general Arnold, who wanted him to go to Eggharbour, and that he would pay him, and give him his directions. (Legal: “Trial of Major General Arnold,” 1780)

7. **One of the waggoners being sick**, and Jordan's quitting the service at that time, several of the waggoners came out under another waggon master; but I cannot say whether they came out to complete their former tour of duty or not. (Legal: “Trial of Major General Arnold,” 1780)

8. that on the 8th day of September, being then in sight of said port, **the said captain Underwood, &c being so confined and under their power**, they were discovered, fired upon, and the said sloop seized by the libellant. (Legal: “The Case of the Sloop Active,” by James Holmes, 1788)

9. **The garrison being asleep, (except the centries)** we gave three huzzas which greatly surprized them. (Nonfiction: *A Narrative of Col. Ethan Allen's Observations during his Captivity*, 1779)

10. but **the sun, by this time, being near two hours high**, and the sign failing, I began to conclude myself to be in a premunire, and would have crossed the river neck again, but I knew the enemy would have discovered such an attempt; and as there could not more than one third part of my troops cross at one time, the other two third, would of course fall into their hands. (Nonfiction: *A Narrative of Col. Ethan Allen's Observations during his Captivity*, 1779)
11. But the residue of the earth being double the extent of America, the exact proportion would have been but as 4 to 8. (Political: *Notes on the State of Virginia*, by Thomas Jefferson, 1785)

12. Journals of observations on the quantity of rain, and degree of heat, being lengthy, confused, and too minute to produce general and distinct ideas, I have taken five years observations, to wit, from 1772 to 1777, made in Williamsburgh and its neighbourhood, have reduced them to an average for every month in the year, and stated those averages in the following table, adding an analytical view of the winds during the same period. (Political: *Notes on the State of Virginia*, by Thomas Jefferson, 1785)

13. The bones of infants being soft, they probably decay sooner, which might be the cause so few were found here. (Political: *Notes on the State of Virginia*, by Thomas Jefferson, 1785)

14. In December 1776, our circumstances being much distressed, it was proposed in the house of delegates to create a dictator, invested with every power legislative, executive and judiciary, civil and military, of life and of death, over our persons and over our properties: and in June 1781, again under calamity, the same proposition was repeated, and wanted a few votes only of being passed. (Political: *Notes on the State of Virginia*, by Thomas Jefferson, 1785)

15. Many of the laws which were in force during the monarchy being relative merely to that form of government, or inculcating principles inconsistent with republicanism, the first assembly which met after the establishment of the commonwealth appointed a committee to revise the whole code, to reduce it into proper form and volume, and report it to the assembly. (Political: *Notes on the State of Virginia*, by Thomas Jefferson, 1785)

16. Hence, the number of representatives in the two cases not being in proportion to that of the two constituents, and being proportionally greater in the small republic, it follows that, if the proportion of fit characters be not less in the large than in the small republic, the former will present a greater option, and consequently a greater probability of a fit choice. (Political: *The Federalist Papers*, 1787)

17. It being therefore evident that the supposition of a want of power to require the aid of the POSSE COMITATUS is entirely destitute of color, it will follow, that the conclusion which has been drawn from it, in its application to the authority of the federal government over the militia, is as uncandid as it is illogical. (Political: *The Federalist Papers*, 1787)

18. Its constitutional powers being at once more extensive, and less susceptible of precise limits, it can, with the greater facility, mask, under complicated and indirect measures, the encroachments which it makes on the co-ordinate departments. (Political: *The Federalist Papers*, 1787)

19. On the other side, the executive power being restrained within a narrower compass, and being more simple in its nature, and the judiciary being described by landmarks still less uncertain, projects of usurpation by either of these departments would immediately betray and defeat themselves. (Political: *The Federalist Papers*, 1787)
20. In the constitution of the judiciary department in particular, it might be inexpedient to insist rigorously on the principle: first, because peculiar qualifications being essential in the members, the primary consideration ought to be to select that mode of choice which best secures these qualifications. (Political: *The Federalist Papers*, 1787)
Appendix C

34 Absolute Constructions (S–Copula–PN) in the Initial Position, British and American (1760-1793)

A. British Texts (19):

1. This being her situation, he had long looked upon her as his future property; as such he had indulged his admiration, and as such he had already appropriated her estate, though he had not more vigilantly inspected into her sentiments, than he had guarded his own from a similar scrutiny. (Fiction: Cecilia, 1782)

2. But this being a lie, he was farther given up to the tempter, and first took the highway in order to make up that sum, intending then to buy a coach and horses and be married. (Sermon: Ordinary, 1760)

3. All Things being thus prepared, and it being about ten o’Clock, the Proceeding began from out of the said Hall, into the Palace Yard, through the Gate House and the End of King Street, thence along the Great Sanctuary, and so to the West End of the Abby Church, all upon blue Cloth, which was spread upon the Ground, from the Throne in Westminster Hall, to the great Steps in the Abby Church; by Sir George Carteret, Knight, Vice Chamberlain, appointed by the KING, to be his Almoner for this Day. (Political: Coronation, 1761)

4. The greater part of the inhabitants of Holland being Calvinists, their religion would have been established: it would likewise have been the religion of all the magistrates. (Political: Test Act, 1790)

5. Moreover, the total land-tax being £1,962 ms. its proportional charge on this class is £1,434 ms. and their total payment to all taxes £4,746 ms. exceeding their proportional charge, in the ratio of 14,330 to 10,000. (Political: Alteration of the Third Estate, 1793)

6. The amount of an adequate surplus fund being £3,485 ms. a year, the existing fund, which very little exceeds 1½ million, must be admitted to be totally inadequate. (Political: Alteration of the Third Estate, 1793)

7. This defalcation may be stated in another point of view: thus its proportional payment being £10,000, its effective charge is £5,002 only: and for every sum of £10,000 thus paid, the just charge amounts to £19,989. (Political: Alteration of the Third Estate, 1793)

8. And the land-tax on this part of the kingdom being £652,200, the charge in the pound was 2s. 1 2/4d. very nearly. (Political: Alteration of the Third Estate, 1793)

9. Now the part paid out of the rents of the two districts, is to their whole charge, as that rent to the whole national income: hence the amount of these general taxes being £15,162 ms., their charge upon the provincial proprietors of land is £4,1827 ms. or 3s. 3⅓d. in the pound rent. (Political: Alteration of the Third Estate, 1793)
10. and, as I mentioned before, the expence of the food necessary to bring those young animals to market being as much in four months as would maintain them as a breeding stock for two years, that must undoubtedly have an effect upon all other butcher's meat, and this will as surely affect other provisions, in a certain proportion. (Political: The Proper Size of Farms, 1773)

11. This being my idea, I shall not scruple to point out, as the first necessary step to be taken, the inclosing and parcelling out of all the King's Forests and Chaces, which now lie a disgrace to a Minister who wishes to be thought to listen to the complaints of a nation. (Political: The Proper Size of Farms, 1773)

12. For the end and intent of such laws being only to regulate the behaviour of mankind, as they are members of society, and stand in various relations to each other, they have consequently no business or concern with any but social or relative duties. (Legal: Blackstone, Commentaries on English Law, Book I, 1765)

13. and, as sir Matthew Hale observes, this being the highest and greatest court, over which none other can have jurisdiction in the kingdom, if by any means a misgovernment should any way fall upon it, the subjects of this kingdom are left without all manner of remedy. (Legal: Blackstone, Commentaries on English Law, Book I, 1765)

14. This reason would be unanswerable, if the commons taxed none but themselves: but it is notorious, that a very large share of property is in the possession of the house of lords; that this property is equally taxable, and taxed, as the property of the commons; and therefore the commons not being the sole persons taxed, this cannot be the reason of their having the sole right of raising and modelling the supply. (Legal: Blackstone, Commentaries on English Law, Book I, 1765)

15. For, the victory obtained at Hastings not being a victory over the nation collectively, but only over the person of Harold, the only right that the conqueror could pretend to acquire thereby, was the right to possess the crown of England, not to alter the nature of the government. (Legal: Blackstone, Commentaries on English Law, Book I, 1765)

16. These canonical disabilities, being entirely the province of the ecclesiastical courts, our books are perfectly silent concerning them. (Legal: Blackstone, Commentaries on English Law, Book I, 1765)

17. The main end and design of marriage therefore being to ascertain and fix upon some certain person, to whom the care, the protection, the maintenance, and the education of the children should belong; this end is undoubtedly better answered by legitimating all issue born after wedlock, than by legitimating all issue of the same parties, even born before wedlock, so as wedlock afterwards ensues. (Legal: Blackstone, Commentaries on English Law, Book I, 1765)

18. But, this being a matter of some uncertainty, the law is not exact as to a few days. (Legal: Blackstone, Commentaries on English Law, Book I, 1765)
19. And, in general, **the king being the sole founder of all civil corporations, and the endower the efficient founder of all eleemosynary ones**, the right of visitation of the former results, according to the rule laid down, to the king; and of the latter, to the patron or endower. (Legal: Blackstone, *Commentaries on English Law*, Book I, 1765)

**B. American Texts (15):**

1. **This being so undeniably the case**, it is somewhat wonderful that any persons, and still more surprizing that any who pretend to the least share of reason and goodness, should set themselves to oppose the laudable endeavors of such as have of late attempted to make the singing in our publick assemblies regular; to retrieve the credit of our psalmody, and rescue it from that barbarism and uncouthness, that ridicule and contempt into which it had very unhappily, and almost universally fallen. (Sermon, Lancaster, 1771)

2. **It being winter**, the doctor insisted he should compose some verses on the fifth of November, and repeat them under his window, which accordingly he did. (Non-fiction: *American Jest Book*, 1789)

3. **General Arnold being commanding officer in the state**, he made that request to me, and I conceived it as my duty to oblige him as my superior officer in every thing in my power, which I would not have done to any others officer in the united states, who was then present. (Legal: Arnold Trial, 1779)

4. **It being the opinion of the Court, that the act incorporating the bank is constitutional; and that the power of establishing a branch in the State of Maryland might be properly exercised by the bank itself**, we proceed to inquire Whether the State of Maryland may, without violating the constitution, tax that branch? (Legal: Maryland Trial, 1819)

5. **This being our comprehension of longitude**, that of our latitude, taken between this and Mason and Dixon's line, is 3.13'.42.4" equal to 223.3 miles, supposing a degree of a great circle to be 69 m. 864 f. as computed by Cassini. (Political: Jefferson, *State of Virginia*, 1785)

6. **The Alleghany being the great ridge which divides the waters of the Atlantic from those of the Mississipi**, its summit is doubtless more elevated above the ocean than that of any other mountain. (Political: Jefferson, *State of Virginia*, 1785)

7. **Civil government being the sole object of forming societies**, its administration must be conducted by common consent. (Political: Jefferson, *State of Virginia*, 1785)

8. **There being three parties interested in these several charters**, what passed between the first and second it was thought could not affect the third. (Political: Jefferson, *State of Virginia*, 1785)

9. **The first stage of this education being the schools of the hundreds, wherein the great mass of the people will receive their instruction**, the principal foundations of future order will be laid here. (Political: Jefferson, *State of Virginia*, 1785)
10. The memory is then most susceptible and tenacious of impressions; and the learning of 
languages being chiefly a work of memory, it seems precisely fitted to the powers of this 
period, which is long enough too for acquiring the most useful languages antient and modern. 
(Political: Jefferson, State of Virginia, 1785)

11. Architecture being one of the fine arts, and as such within the department of a 
professor of the college, according to the new arrangement, perhaps a spark may fall on some 
young subjects of natural taste, kindle up their genius, and produce a reformation in this elegant 
and useful art. (Political: Jefferson, State of Virginia, 1785)

12. Heresy, thus circumscribed, being an offence at the common law, our act of assembly of 
October 1777, c. 17. gives cognizance of it to the general court, by declaring, that the jurisdiction 
of that court shall be general in all matters at the common law. (Political: Jefferson, State of 
Virginia, 1785)

13. Power being almost always the rival of power, the general government will at all times 
stand ready to check the usurpations of the state governments, and these will have the same 
disposition towards the general government. (Political: The Federalist Papers, 1787)

14. And this being the case, I perceive at present no impediment to the establishment of an 
appeal from the State courts to the subordinate national tribunals; and many advantages attending 
the power of doing it may be imagined. (Political: The Federalist Papers, 1787)

15. This being the case, let me ask if it is consistent with common-sense to suppose that a 
provision obliging the legislative power to commit the trial of criminal causes to juries, is a 
privation of its right to authorize or permit that mode of trial in other cases? (Political: The 
Federalist Papers, 1787)
Vocabulary Reduction, Text Excision, and Procedural-Context Features in Judicial Document Analytics

L. Karl Branting
The MITRE Corporation
7515 Colshire Drive
McLean, Virginia 22102
lbranting@mitre.org

ABSTRACT
Collections of documents filed in courts are potentially a rich source of information for citizens, attorneys, and courts, but courts typically lack the ability to interpret them automatically. This paper presents technical approaches to two applications of judicial document interpretation: detection of document filing errors; and matching orders with the motions that they rule on. An empirical evaluation identified several techniques that exploit genre-specific aspects of judicial documents to improve performance on these two tasks, including vocabulary reduction to task-specific terms, excision of the portion of documents unlikely to contain relevant text, and optimizing error detection by separating document classification into two stages: classification of the document’s text followed by interpretation of this text classification based on procedural context.

1. INTRODUCTION
The transition from paper to electronic filing in Federal, state, and municipal courts, which began in the late 1990s, has transformed how courts operate and how judges, court staff, attorneys, and the public create, submit, and access court filings. However, despite many advances in judicial access and administration brought about by electronic filing, courts are typically unable to interpret the contents of court filings automatically. Instead, court filings are interpreted only when they are read by an attorney, judge, or court staff member.

Machine interpretation of court filings would open a rich source of information for improving court administration and case management, access to justice, and analysis of the judiciary. However, there are numerous challenges to automating the interpretation of case filings. Courts typically accept documents in the form of PDFs created from scans. Scanned PDFs require optical character recognition (OCR) for text extraction, but this process introduces many errors and does not preserve the document layout, which contains important information about the relationships among text segments in the document. Moreover, the language of court filings is complex and specialized, and the function of a court filing depends not just on its text and format, but also on its procedural context. As a result, successful automation of court filings requires overcoming a combination of technical challenges.

This paper (1) describes the nature of court dockets and databases, (2) describes two classes of representative judicial document analysis tasks: docket error detection; and order/motion matching, and (3) presents technical approaches to each of the tasks together with preliminary empirical evaluations of the effectiveness of each approach.

2. COURT DOCKETS AND DATABASES
A court docket is a register of document-triggered litigation events, where a litigation event consists of either (1) a pleading, motion, or letter from a litigant, (2) an order, judgment, or other action by a judge, or (3) a record of an administrative action (such as notifying an attorney of a filing error) by a member of the court staff. Contemporary electronic docket systems are typified by CM/ECF [4], which was developed by the Administrative Office of US Court (AO) and is used in all Federal Courts. Each docket event in CM/ECF includes both (1) metadata generated at the time of filing, including both case-specific data (e.g., case number, parties, judge) and event-specific data (e.g., the attorney submitting the document, the intended document type) and (2) a text document in PDF format (except for administrative entries). One typical CM/ECF database for a large federal court contains approximately 420,000 cases involving 1,200,000 litigants, attorneys, and judges, roughly 10,900,000 docket entries, and approximately 4,000,000 documents. The experiments described below were performed on a collection of 267,834 documents that were filed consecutively in 2015 in a large federal district court.

3. DOCKET ERROR DETECTION
There are many kinds of docket errors, including defects in a submitted document (e.g., missing signature, sensitive information in an unsealed document, missing case caption) and mismatches between the content of a document and the context of the case (e.g., wrong parties, case number, or judge; mismatch between the document title and the document type asserted by the user). For attorneys, detection of defects at submission time could prevent the embarrassment of submitting a defective document and the inconvenience and delays of refileing. For court staff, automated filing error detection could reduce the auditing staff required for filing errors, a significant drain of resources in many courts. Automating error detection could significantly reduce both of these problems.

This section focuses on four types of docket errors:

- Event-type errors. Specifying the wrong event type for a document, e.g., submitting a Motion for Summary Judgment as a Counterclaim. In the experiments below, there were 20 event types, such as complaint, transfer, notice, order, service, etc.
Main-vs-attachment errors. Filing a document, such as an exhibit, that can only be filed as an attachment to another document, as a main document or filing a document, such as a Memorandum in Support of a Motion for Summary Judgment, that should be filed as a main document, as an attachment.

Show-cause order errors. Only judges are permitted to file show-cause orders; it is an error if an attorney does so.

Letter-motion errors. In some courts, certain routine motions can be filed as letters, but all other filings must have a formal caption. Recognizing these errors requires distinguishing letters from non-letters.

Each of these error-detection tasks requires classifying a document with respect to the corresponding set of categories (event type, main vs. attachment, show-cause order vs. non-show-cause order, and letter vs. non-letter) and evaluating whether the category is consistent with the metadata generated in CM/ECF by the filer’s selections. Event type document classification is particularly challenging both because document types are both numerous (20 in the test dataset) and skewed (roughly power-law frequency distribution in the test set).

3.1 Text Classification

The first set of experiments attempted to identify each of the four docket errors above by classifying document text and determining whether there is a conflict between the apparent text category and the document’s metadata. An initial barrier was that OCR errors greatly expand the apparent vocabulary size, making term-vector representations of documents extremely verbose and leading to very slow training and large models. One approach to reducing this verbosity is to classify documents using a language model (LM)[1, 9], which can be trained incrementally. Language-model classification is relatively fast even if the feature sets include n-grams with large n. The experiment in this paper used the lingpipe1 LMClassifier, which performs joint probability-based classification of token sequences into non-overlapping categories based on language models for each category and a multivariate distribution over categories.

A second approach is to reduce the vocabulary to terms likely to be relevant to the particular domain [7]. A third approach is to excise those portions of documents that contain the least information about the document type. All three approaches were explored in this work.

3.1.1 Vocabulary Reduction and Text Excision

Court filings can be thought of as comprising four distinct sets of terms:

- Procedural words, which describe the intended legal function of the document (e.g., “complaint,” “amended,” “counsel”)
- stop-words (uninformative common words, such as “of” and “the”)
- Words unique to the case, such as names, and words expressing the narrative events giving rise to the case; and

Terms in the first of these sets—procedural words—carry the most information about the type of the document. These words tend to be concentrated around the beginning of legal documents, often in the case caption, and at the end, where distinctive phrases like “so ordered” may occur.

We experimented with several approaches to vocabulary reduction: two ad hoc and domain-specific and one general and domain-independent. The first approach was to eliminate all terms except non-stopwords that occur in the Federal Rules of Civil Procedure [5]. An alternative approach was to remove all terms except for non-stopwords occurring in “event” (i.e., document) descriptions typed by filers when they submit into CM/ECF. The third approach was to select terms based on their mutual information with each particular text categories [2]. The first lexical set, termed FRCP, contains 2658 terms; the second, termed event, consists of 513 terms. Separate mutual-information sets were created for each classification task, reflecting the fact that the information gain from a term depends on the category distribution of the documents.

For example, Figure 1 shows the 10 highest information terms for
Table 1: Thresholds and size of large and small high information-gain term sets.

<table>
<thead>
<tr>
<th></th>
<th>showcause</th>
<th>main_attachment</th>
<th>types</th>
<th>letter</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ig_small</strong></td>
<td>0.01 (135)</td>
<td>0.025 (262)</td>
<td>0.1 (221)</td>
<td>0.0005 (246)</td>
</tr>
<tr>
<td><strong>ig_large</strong></td>
<td>0.0025 (406)</td>
<td>0.0125 (914)</td>
<td>0.05 (689)</td>
<td>0.000001 (390)</td>
</tr>
</tbody>
</table>

three different classification tasks: event-type classification, distinguishing letters from non letters, and show-cause order detection, illustrating that the most informative terms differ widely depending on the classification task.

Figure 2 illustrates the reduction of full document text to just FRCP terms, which typifies the vocabulary-reduction process.

Several approaches to document excision were explored as well. The first was to limit the text to the first $l$ tokens of the document (i.e., excise the remainder of the document). If $l$ is sufficiently large, this is equivalent to including the entire document. A second option is to include the last $l$ tokens of the suffix as well as the prefix. For simplicity, the same $l$ is used for both the prefix and the suffix.

The initial set of experiments using language model (LM) classification evaluated the effect of varying the following parameters:

- Vocabulary reduction: none, FRCP, event, ig_small, ig_large
- Prefix length, $l$
- Whether $l$ tokens of the suffix are included, in addition to the prefix
- The maximum n-gram length, $n$

Two different information-gain thresholds were tested for each classification type, intended to create one small set of very-high information terms (ig_small) and a larger set created using a lower threshold (ig_large). The thresholds and sizes of the large and small high information-gain term sets are set forth in Table 1. The text of each document was obtained by OCR using the open-source program Tesseract [11]. Each text was normalized by removing non-ASCII characters and standardizing case prior to vocabulary reduction, if any.

Figure 3 shows a comparison of four vocabulary alternatives on the four text classification tasks described above. These tests measured mean f-measure in 8-fold cross validation using a 1-gram language model, 50-token prefix length, and no suffix. In the baseline vocabulary set, normalize, non-ASCII characters, numbers, and punctuation are removed and tokens were lower-cased. The results show that classification accuracy using an unreduced vocabulary was significantly lower than the best reduced vocabulary performance for show-cause order detection and type classification. Choice of vocabulary had little effect on accuracy for the letter and main vs. attachment detection tasks. No reduced-vocabulary set consistently outperformed the others, although ig_large (with a lower information-gain threshold) was consistently slightly better than ig_small (with a higher information-gain threshold).

Figure 4 shows the results of running this same set of experiments with more-expressive 4-gram models. Accuracy rose for all test
Figure 5: Classification accuracy as a function of maximum n-gram size (8-fold cross validation using event vocabulary, 50-token prefix length, and no suffix).

Figure 6: Classification accuracy as a function of prefix length (8-fold cross validation using event vocabulary, maximum n-gram length n = 6, and no suffix).

Figure 7: The ratio of classification accuracy using both prefix and suffix to accuracy using prefix only, as a function of prefix and suffix length (8-fold cross validation using event vocabulary, and maximum n-gram length n = 6).

Figure 7 shows the ratio of prefix-only to prefix-plus-suffix as a function of prefix and suffix length. For these classification tasks, the greatest improvement in performance from including the suffix as well as the prefix occurred when the prefix and suffix lengths were quite short (8 tokens), and there was no improvement for l = 64.

Summarizing over the tests, the the highest mean f-measure based on text classification alone and the particular combination of parameters that led to this accuracy for each document category were as follows:

1. Event type: 0.743 (prefix=50, no suffix, max n-gram=4, ig_large vocabulary, 20 categories)
2. Main-vs-attachment: 0.871 (prefix=256, no suffix, max n-gram=6, event vocabulary)
3. Show-cause order: 0.957 (prefix=50, no suffix, max n-gram=5, ig_small vocabulary)
4. Letter-vs-non-letter: 0.889 (prefix=50, no suffix, max n-gram=4, ig_large vocabulary)

3.2 Incorporating Procedural Context Features

The accuracy of event-type detection (f-measure of roughly 0.743 under the best combinations of parameters) is sufficiently low that its utility for many auditing functions may be limited. An analysis of the classification errors produced by the event-type text classification model indicated that a document’s event type depends not just on the text of the document but also on its procedural context. For example, motions and orders are sometimes extremely similar because judges grant a motion by adding and signing an order stamp to the motion. Since stamps and signatures are seldom accurately OCR’d, the motion and order may be indistinguishable by the text alone under these circumstances. However, orders can be issued only by a judge, and judges never file motions, so the two cases can be distinguished by knowing the filer. In addition, attachments have the same event type as the main document in CM/ECF. So, for example, a memorandum of law is ordinarily a main document, but an already-filed memorandum can sometimes be filed as an attachment, in which case its event type is the same as that of the main document. So, determining the event type of a document...
requires knowing both (1) whether it was filed as a main document or as an attachment and (2) the filer.

The first and simplest step to incorporating procedural-contextual information is to assign to each document filed as an attachment the type of the main document to which it is attached. This reduces the event-type classification problem to classifying documents that are filed as main documents. The f-measure for LM classification restricted to main documents was 0.809, as compared to .707 for main documents and attachments combined (event vocabulary, prefix length 25, max n-gram size 6, suffix not included).

Four approaches were explored for adding procedural-contextual information to improve event-type classification for main documents. The first was the highly ad hoc procedure of prepending procedural-context features as unique tokens to the document text to which the LM classifier was applied. Using filer (crt,aty,rep,unk,pty,jud), document page length, the first token of the filer’s document summary description, and the nature of the suit as procedural-contextual features raised the f-measure to 0.847.

To avoid the unnatural procedure of adding context features to the input into a language model, the next two approaches represented documents as collections of term features together with context features. Using the event vocabulary yields 513 term features. Two types of term features were evaluated: term frequency; and inverse first-occurrence-position (IFOP). The rationale for IFOP is that it is typically the first occurrence of a term that indicates its procedural function, e.g., the first time “motion” occurs is likely to be in the case caption or in the introductory sentence of a letter, in either case indicating the purpose of the document. Subsequent occurrences are likely to refer to other documents. Inverse first position weights early occurrences more heavily than later occurrences. Applying the J48 decision tree classifier from WEKA [6] yielded f-measures of 0.847 (term frequency) and 0.870 (IFOP).

The fourth approach was to use the classification predicted by an LM classifier as a feature combined with metadata features (filer, nature of suit, page count), yielding an f-measure of 0.964. These results are summarized in Table 2.

Table 2: Event-type classification results incorporating different term and procedural-contextual features.

<table>
<thead>
<tr>
<th>Method</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM classifier main and attachments</td>
<td>0.707</td>
</tr>
<tr>
<td>LM classifier main only</td>
<td>0.809</td>
</tr>
<tr>
<td>LM + features</td>
<td>0.847</td>
</tr>
<tr>
<td>J48 features + terms-freq</td>
<td>0.846</td>
</tr>
<tr>
<td>J48 features + terms-pos</td>
<td>0.870</td>
</tr>
<tr>
<td>LM output + features</td>
<td>0.964</td>
</tr>
</tbody>
</table>

Figure 8: A fragment of the J48 decision tree for event-type classification using both procedural-contextual features and LM predictions.

4. ORDER/MOTION MATCHING

In many federal courts, docket clerks are responsible for filing orders executed by judges into CM/ECF, a process that requires the clerk to identify all pending motions to which the order responds and to link the order to those motions. This entails reading all pending motions, a tedious task. If the motions corresponding to an order could be identified automatically, docket clerks would be relieved of this laborious task. Even ranking the motions by their likelihood of being resolved by a motion would decrease the burden on docket clerks. Moreover, order/motion matching is a subtask of a more general issue-chaining problem, which consists of identifying the sequence of preceding and subsequent documents relevant to a given document.

A straightforward approach to this task is to treat order/motion matching as an information-retrieval task, under the hypothesis that an order is likely to have a higher degree of similarity to its corresponding motions than to motions that it does not rule on. An obvious approach is to present pending motions to the clerk in rank order of their TF-IDF2-weighted cosine similarity to the order.

The evaluation above showing that vocabulary reduction improves document classification raises the question whether vocabulary reduction might be beneficial for order/motion matching as well. A second question is whether the IDF motion should be trained on an entire corpus of motions and orders or whether acceptable accuracy can be obtained by training just on the order and pending motions.

To evaluate the effectiveness of this approach to order/motion match, a subset of the document set described above was collected consisting of 3,356 groups, each comprising (1) an order, (2) a motion

2Term Frequency/Inverse Document Frequency
that the order rules on (a triggering motion), and (3) a non-empty set of all motions that were pending at the time of the order but not ruled on by the order (non-triggering motions). The mean number of motions per group was 5.87 (i.e., there were on average 4.87 non-triggering motions). For each group, all motions were ranked by similarity to the order under the given metric. The proportion of triggering motions that were ranked first and mean rank of the triggering motion were calculated from each group’s ranking.

These groups were evaluated using three vocabulary reduction approaches: the raw document text (which often contains many OCR errors); normalization, as described above; and event terms. The two alternative TF/IDF training models were applied to each of the three vocabulary reduction approaches, for a total of 6 combinations. For each combination, the mean rank of the triggering motion among all the motions was determined.

Figure 9 shows that the highest accuracy, as measured by the proportion of triggering motions that were ranked first among all pending motions, was achieved by normalizing the text but not by vocabulary reduction. Intuitively, reduction to procedurally relevant terms improves the ability to determine what docket event a document performs, but reduces the ability to discern the similarity between pairs of documents. TF/IDF training on just the order and pending motions (local) is at least as accurate as training over all orders and motions (all). Figure 10 shows the mean rank (zero indexed) of the most similar motion under each of the six conditions. The best (lowest) mean rank was achieved with normalization and local TF/IDF training.

It is not unusual for a single order to rule on multiple pending motions. A more realistic assessment of the utility of pending motion ranking is therefore to determine how many non-triggering motions a clerk would have to consider if the clerk read each motion in rank order until every motion ruled on by the order is found. One way to express this quantity is as mean precision at 100% recall. In the test set described above, using text normalization and local TF/IDF training, mean precision at 100% recall was 0.83, indicating that the number of motions that a clerk would have to read was significantly reduced.

5. RELATED WORK

There is a long history of applying text classification techniques to legal documents dating back at least to the 1970s [3]. Text classification has been recognized as of particular importance for electronic discovery [10]. Little prior work has addressed classification of docket entries other than Nallapati and Manning [8], which achieved an f-measure of 0.8967 in distinguishing Orders to Show Cause from other document types using a hand-engineered feature set. As shown above, we obtained an f-measure of 0.9573 using the reduced vocabulary approach as well as good performance on other classification tasks and scalability. Thus, the approach described in this paper represents a significant advance over prior work.

6. SUMMARY AND FUTURE WORK

Judicial document collections contain a rich trove of potential information, but analyzing these documents presents many challenges. This paper has demonstrated how vocabulary reduction, text excision, and procedural-context features can be used in combination to improve the accuracy of recognizing the nature of legal documents, including whether the document is a main document or an attachment, the document’s event type, and whether the document is a show-cause order. Reduced vocabularies based on domain-specific information—FRCP terms and the document description field of the CM/ECF database—performed with comparable accuracy to reduced vocabularies based on information gain, illustrating that useful reduced term sets can be derived without domain-specific information.

These results demonstrate the feasibility of automating the process of auditing CM/ECF submissions, which is currently a significant drain on court resources. The experiment with order/motion matching demonstrates that while vocabulary reduction may improve accuracy for document classification, it can decrease accuracy for tasks that involve matching based on overall similarity rather than procedural similarity.

Many of the challenges addressed in this work arise from the current inability to reason about the layout of legal documents. For example, many documents have a case caption in which the case title and other standard information fields have standard spatial relationships to one another. We are currently engaged in developing an annotated corpus of court documents for use in training 2-dimensional conditional random fields and other spatially-aware document analysis tools. However, even when these tools are available, in many cases it will remain necessary to reason about the text itself.

No single technology is applicable to all judicial documents, nor is
any approach sufficient for all document analysis tasks. However, each addition to this suite of technologies adds to the capabilities available to the courts, government agencies, and citizens to exploit the deep well of information latent in judicial document corpora.

Acknowledgment
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 16-1541. ©2016 The MITRE Corporation. All rights reserved.

7. REFERENCES
ABSTRACT
Authorship attribution, the determination of the authorship of a disputed document by analyzing the document itself, is an important legal and technical question. By studying aspects of writing style such as word choice, punctuation, and preferred syntax, statistical analysis can produce probabilistic assessments of the likelihood of any specific author. We have developed a program called Envelope that implements a proposed protocol [23] using an ad hoc distractor set, five separate analyses, and Fisher’s exact test on the rank order sum as a data fusion method. We discuss testing results and some suggested extensions and improvements.

CCS Concepts
•Security and privacy → Human and societal aspects of security and privacy; •Computing methodologies → Information extraction; •Applied computing → Investigation techniques; Evidence collection, storage and analysis;

Keywords
authorship attribution, text classification, digital evidence

1. INTRODUCTION
The authorship of documents is a key question in many legal cases, as a skim of many of Agatha Christie’s novels will show. Electronic documents [6, 16, 17] bring their own set of issues, as handwriting cannot be used to validate the documents. Stylometry, the study of individual writing style [15, 31], can. In the case of Ceglia v. Zuckerberg, et al.[27], for example, ownership of a significant part of Facebook depended in part on the validity of an emailed agreement ostensibly between the two parties.

In this paper, we describe the authorship attribution problem, with some examples. We then describe a formal protocol to address this kind of problem and present a software system (Envelope) that implements this protocol in a simple and easy-to-use way.

2. BACKGROUND
2.1 Authorship analysis
Language is among the most individualized activities people engage in. For this reason, much can be learned about a person by looking at his or her writings. An easy example is distinguishing between different regional groups. A Commonwealth English speaker/writer can easily be spotted by her use of “lorry” instead of “truck,” spelling “labor” with a ‘u,’ and less obviously by grammatical constructions such as “could do” or “in hospital.” These insights can be extended to questions of authorship without regard to handwriting. The basic theory of traditional stylistics is fairly simple. As McMenamin [27] describes it,

At any given moment, a writer picks and chooses just those elements of language that will best communicate what he/she wants to say. The writer’s “choice” of available alternate forms is often determined by external conditions and then becomes the unconscious result of habitually using one form instead of another. Individuality in writing style results from a given writer’s own unique set of habitual linguistic choices.[27]

Coulthard’s [8] description is also apt:

The underlying linguistic theory is that all speaker/writers of a given language have their own personal form of that language, technically labeled an idiolect. A speaker/writer’s idiolect will manifest itself in distinctive and cumulatively unique rule-governed choices for encoding meaning linguistically in the written and spoken communications they produce. For example, in the case of vocabulary, every speaker/writer has a very large learned and stored set of words built up over many years. Such sets may differ slightly or considerably from the word sets that all other speaker/writers have similarly built up, in terms both of stored individual items in their passive vocabulary and, more importantly, in terms of their preferences for selecting and then combin-
ing these individual items in the production of texts. [8]

These choices express themselves in a number of ways. In an expert witness report, McMenamin [27] analyzed eleven different and distinct “features” of the writing in both a known (undisputed) email and a disputed email. One feature, for example, hinged on the spelling of the word cannot, and in particular whether it was written as one word (cannot) or as two (can not). Another feature was the use of the single word “sorry” as a sentence opener (as opposed, for example, to “I’m sorry”). Coulthard similarly discussed (among other features) the use of the specific phrase “disgruntled employees.” (Why “disgruntled” and not one of its myriad synonyms?) In both cases, significant differences in these features can be held as evidence of differences in authorship.

2.2 Computational text analysis

Computer-based stylometry applies the same general theory, but with a few major differences. The basic assumption that people make individual choices about language still holds, but instead of ad hoc features selected by examination of the specific documents, the analysts use more general feature sets that apply across the spectrum of problems.[3, 1, 13, 32, 28, 24] Using these feature sets or others [30], the features present in a document are automatically identified, gathered into collections of feature representations (such as vector spaces), and then classified using ordinary classification methods [15, 29, 26, 14, 20] to establish the most likely author.

Binongo [1] provides a clear example of this. In his study of the authorship of the Oz books, he collected the frequencies of the fifty most frequent words in English from the books of undisputed authorship (his feature set). He applied principle component analysis (his classification method) to obtain a data visualization of the stylistic differences, then showed that the disputed 13th book clearly lay in the stylistic space corresponding only to one candidate author. This would clearly be highly relevant evidence if the authorship (perhaps for copyright reasons) were being disputed in court.

From a legal standpoint, there are three key issues with this technology. The first, admissibility, has been addressed in detail elsewhere [5, 8, 23], but is closely tied to the second issue, the scientific validity of the technology itself. Numerous surveys [15, 26, 14, 31] and TREC-style conferences [33] have shown that authorship can be determined with high accuracy (typically 80% or better) using realistically-sized samples. Large-scale studies [20, 34] have confirmed that there are often many different “best practices” that perform well based on different features. This allows for ordinary data fusion techniques such as mixture-of-experts [18, 10] to boost accuracy rates to practical levels.

3. JUOLA’S PROPOSED PROTOCOL

The usefulness of the above technology has been demonstrated in actual disputes. For example, Collins [7] used mixture of experts to validate a newly discovered short story by Edgar Allan Poe, and Juola [2, 21] used a similar method to identify J.K. Rowling as the author of A Cuckoo’s Calling. In a legal context, Juola [22] was able to validate authorship of anonymous newspaper columns in support of an asylum claim in US immigration court.

Partly to address admissibility needs and partly to address the needs of a lawyer, judge, and jury to understand this sort of evidence, Juola [23] proposed a formal protocol for handling and developing this type of authorial evidence. Key elements of this proposed protocol are:

- Suitable data for analysis, including an ad hoc set of distractor authors not believed to be connected to the case,
- A set of independent analysis methods that have been validated to perform well on similar tasks
- A predefined data fusion framework amenable to formal statistical analysis, so that the likelihood of error can be assessed mathematically.

Juola [23] showed how this protocol could be applied several separate authorship disputes. We have implemented this protocol in a SaaS platform, named Envelope, to provide low-cost, high-accuracy resolution of authorship disputes.

4. ENVELOPE: OUR IMPLEMENTATION

Envelope, in its current version, focuses on a specific (and relatively common) type of disputed document, electronic mail [6] written in English. The system is presented with client-supplied copies of the disputed email(s) as well as samples known to be by the purported author. These documents are compared against a set of distractor authors (currently a set of ten gender-balanced authors extracted from the Enron corpus [25]) and rank-ordered for similarity along five human-understandable features that have been shown to work well in large-scale testing [20, 33].

The five measured dimensions are as follows:

- Authorial Vocabulary (Vocabulary overlap) : Words are, of course, what a work is fundamentally all about. A crime novel is usually about a dead body and how people deal with the problem it poses; a romance novel is about a small group of people and their feelings for each other. Even email differs in word choices as discussed above [27, 8]. Authorial vocabulary is also one of the best ways to tell individual writers apart, by looking at the choices they make, not only in the concepts they try to express, but the specific words they use to create their own individual expression. The degree of shared vocabulary is thus a key authorial indicator.

- Expressive Complexity (Word lengths) : One key attribute of authors is, on the one hand, their complexity, and on the other, their readability. A precise author who uses exactly the specific word to every event — “that’s not a car, that’s a Cadillac? that’s not a cat, but a tabby”—will more or less be forced to use rarer words. These rarer words, by their very nature, are longer [35]. A large and complex vocabulary will naturally be reflected in larger words, producing a very distinctive style of writing. By tracking the distribution of word lengths (n.b., not just the average word length, which is known not to perform well), we can assess the expressive complexity of a given author.

- Character n-grams : In addition to comparing words directly, scholarship has shown [29, 32, 28] that com-
parison of character clusters (for example, four adjacent letters, whether as part of a word like “eXAM-Ple” or across two words as in “iN_iTh”). This allows matching of similar but not identical words, such as different forms of the same stem or words with similar affixes, and even preferred combinations of words.

- Function words: One of the most telling and oft-studied aspects of an individual writer is their use of function words [3, 13, 1], the simple, short, common, and almost meaningless words that form a substantial fraction of English writing. (To understand how these words lack meaning, consider that the three most common words in English are “the,” “a/an”, and “of.” What would you offer as a dictionary definition of “the”?) These words are called “function words” because they do not carry meaning of their own, but instead describe the functions of the other words in the sentence in relation to each other. These words thus provide a good indication of the tone of the writing and the specific types of relationships expressed throughout the manuscript.

- Punctuation: Although not necessarily linguistically interesting, and often the choice of editor instead of author, punctuation offers an insight into social conventions that have little effect on the meaning of the text. Because they have little effect, they are often freely variable between authors. For example, an author’s decision to use an Oxford comma, their choice of marking extraneous material (for example, with commas, parentheses, or brackets), the way they split sentences with semicolons, periods, or comma splices, and even whether punctuation is put inside or outside quotation marks, do not change the meaning. In unedited documents (such as email), they therefore provide a strongly topic-independent cue to authorship that is not directly related to the other dimensions. (See [27] for some non-computational examples.)

Along each document, the eleven possible authors (ten implausible distractor authors plus one plausible suspect) are thus ranked from #1 (most similar/likely) to #11. The rank sum of the purported author across all dimensions is calculated and used to fuse the different analyses. For example, if the purported author scored as the most similar author on all five dimensions (the most compelling possible result), the rank sum would be five. The system then uses Fisher’s exact test [9] to determine a likelihood that the specific experimental result could have been obtained by chance.

In more detail, we consider the null hypothesis that the disputed document was not written by the purported author, and that there is, in fact, no relation between them. Under this assumption, the purported author would rank anywhere from #1 to #11 (with equal probability), averaging at the sixth slot. Consistently appearing closer than the sixth slot, then, is evidence of systematic similarity between the two authors across a variety of independent stylometric variables. An unrelated person is unlikely to show this kind of systematic similarity, and hence if the calculated rank sum is small enough, we can reject the null hypothesis at any specific alpha cutoff desired. The system as currently developed uses standard cutoffs: if the p-value is 0.05 or less, we consider this to be “strong indications of common authorship,” while trend-level values (p-value of 0.10 or less) are “indications of common authorship.” “Weak indications” occur at p-values of 0.20 or less. Inconclusive or outright contraindications are handled appropriately.

The system is therefore capable of delivering a sophisticated stylometric analysis quickly, cheaply, and without human intervention (thereby minimizing analyst bias effects).

5. ACCURACY AND VALIDITY

To enhance validity, the system performs a certain amount of data validation. Both the known and disputed documents need to be of sufficient length (currently defined as =< 200 words), and cannot include header information (which can be picked up, for example, by looking for From: lines). Furthermore, the documents must be in English [4] (which we currently approximate by confirming the existence of “the” in the files). Violations of these conditions are documented in the generated report but do not prevent analysis; more sophisticated (and expensive) human-based analysis may be necessary in these circumstances. For example, stylometric analysis technology is known to transfer well between languages [19, 11, 12], but a new distractor corpus would be necessary.

The accuracy of this system has been preliminarily tested on a variety of other email samples drawn from 20 additional authors in the Enron [25] corpus. Out of 375 trials, 179 produced “strong” indications of authorship, and all 179 (100%) were correct. Similarly, “Weak” indications were correct in 21 of 23 cases (91%). Only 2 cases showed just “indications”, and 1 of those 2 (150%) was correct. “Weak” indications were correct in 22 of 23 cases, while the remaining 43 inconclusive cases could not be validated, but showed significant numbers of both same (6) and different (37) author pairs. Thus, as expected, this method does not return an answer in all cases, but when an answer is returned, the accuracy is very high.

6. DISCUSSION AND CONCLUSIONS

The Envelope system thus delivers a high-quality analysis at low cost. Being fully automatic, the analysis is reproducible and is not influenced by analyst bias in any specific case. The probability of error has been confirmed empirically to be low, and the estimated error rate (in the form of the calculated p-value) is realistic. It is easy to extend the current system to additional languages, additional document types, or even additional classification tasks such as author profiling.

Interpreting an Envelope report can be fairly straightforward. In the event of a finding of similarity, this means that the two documents were shown to be highly similar across a very wide range of linguistic and stylistic characteristics. If the author of the email was not Aunt Prunella, it was, at a minimum, someone who used Aunt Prunella’s characteristic vocabulary, her characteristic syntax, her characteristic style of punctuation, and even used the little words (“function words”) in the same way that she did. The computer can characterize the likelihood of this kind of match occurring from a person off the street with high precision and high reliability. As the old joke has it, “if it looks like a duck, walks like a duck, and uses punctuation like a duck...”

There are a number of fairly obvious extensions and possible improvements. Extension to new genres and/or lan-
guages [19, 11, 12] can be as simple as the creation of a new set of distractor documents. It may be possible to improve the accuracy by the incorporation of other analysis methods and feature sets (for example, the distribution of part-of-speech tags), although high-level processing such as POS tagging may limit its use in other languages. We continue our preliminary testing and will be expanding our offerings in terms of genre.

So, who did write Aunt Prunella’s will, or at least her email setting out her last wishes? Who wrote the email ostensibly dividing up ownership of the startup, or revealing confidential business information? Computational analysis, as typified by Envelope, may not be able to provide definitive answers, but the evidence it creates can provide valuable information to help guide investigations or suggest preliminary conclusions. This system provides low-cost advice without the time and cost of human analysis, while retaining high accuracy.

7. REFERENCES

[19] P. Juola. Cross-linguistic transference of authorship attribution, or why english-only prototypes are


Bending the Law

Greg Leibon
Coherent Path, Inc.
and
Department of Mathematics,
Dartmouth College
gleibon@gmail.com

Michael Livermore
School of Law
University of Virginia
Charlottesville, VA 22904
mal5un@virginia.edu

Reed Harder
Thayer School of Engineering
Dartmouth College
Hanover, NH 03755
reedharder@gmail.com

Allen Riddell
Department of Mathematics
Dartmouth College
Hanover, NH 03755
abr@riddell.org

Dan Rockmore†
Department of Mathematics
Department of Computer Science
Dartmouth College
Hanover, NH 03755
rockmore@math.dartmouth.edu

ABSTRACT

Legal reasoning requires identification, through search, of authoritative legal texts (such as statutes, constitutions, or prior judicial decisions) that apply to a given legal question. In this paper we model the concept of the law search as an organizing principle in the evolution of the corpus of legal texts, apply that model to U.S. Supreme Court opinions. We examine the underlying navigable geometric and topological structure of the Supreme Court opinion corpus (the “opinion landscape”) and quantify and study its dynamic evolution. We realize the legal document corpus as a geometric network in which nodes are legal texts connected in a weighted and interleaved fashion according to both semantic similarity and citation connection. This network representation derives from a stylized generative process that models human-executed search via a probabilistic agent that navigates between cases according to these legally relevant features. The network model and (parametrized) probabilistic search behavior give rise to a PageRank-style ranking of the texts – already implemented in a pilot version on a publicly accessible website – that can be compared to search results produced by human researchers. The search model also gives rise to a natural geometry through which we can measure change in the network. This enables us to then measure the ways in which new judicial decisions affect the topography of the network and its future evolution. While we deploy it here on the U.S. Supreme Court opinion corpus, there are obvious extensions to larger bodies of evolving bodies of legal text (or text corpora in general). The model is a proxy for the way in which new opinions influence the search behavior of litigants and judges and thus affect the law. This type of legal search effect is a new legal consequence of research practice that has not been previously identified in jurisprudential thought and has never before been subject to empirical analysis. We quantitatively estimate the extent of this effect and find significant relationships between search-related network structures and propensity of future citation. This finding indicates that influence on search is a pathway through which judicial decisions can affect future legal development.

CCS Concepts

• Applied computing → Law; • Information systems → Digital libraries and archives; Document topic models; Content analysis and feature selection; Search interfaces; Link and co-citation analysis; • Theory of computation → Structured prediction;

Keywords

Topic model, law search, citation networks, multi-networks, PageRank, network curvature

1. INTRODUCTION

Judicial decision making is characterized by the application by courts of authoritative rules to the stylized presentation of disputed claims between competing litigants. These authoritative rules are set forth in legal source materials such as constitutions, statutes, and decisions in prior cases. For a legal source to have bearing on a current dispute, it must be retrievable by the relevant legal actors. The problem of organizing legal texts into a comprehensible whole has been recognized since Justinian I’s Corpus Juris Civilis issued in 529-34. The acute problems of identifying relevant legal sources (i.e., legal precedent) presented by the common law tradition has spurred codification and classification...
efforts that have ranged from Blackstone’s “Commentaries on the Laws of England (1765-69)” to the codification movement in the late nineteenth century [15], to the development and spread of the West American Digest System in the twentieth century [31]. Most recently, the effect of digitization on the evolution of the law, primarily in its impact on legal research, has become a subject of inquiry (see e.g., [21, 3, 4, 18, 25, 19, 10, 14]).

In this paper we consider the textual corpus of legal sources as an evolving, geometrically defined landscape that encompasses regions of the law and that is influenced by the dynamics and feedback of law search. Everything devolves from a model of the process of legal research in which “actors” start from a case or opinion and then build out an understanding of the relevant issues by following citations, searching for cases that cite the initial case of interest, and identifying textually similar cases. This has a natural network formulation, in which legal sources are connected to each other based on citation information and a “topic model” representation of their textual content. Topic models represent texts (embodied as a “bag-of-words”) as mixtures of “topics”, probability distributions over the vocabulary in the corpus (see e.g., [6]). By encoding three kinds of connectivity this becomes a multi-network representation, a combinatorial structure that has proved useful in a number of different contexts, such as biology and economics (e.g., [2, 7, 11]). In this work we bring the multi-network concept to the novel contexts of text-mining and text search, with a specific application to judicial texts.

Distance in this landscape reflects the ease with which a human user of the legal corpus could navigate from one legal source to another, based on the underlying citation network as well as via topical similarity, which in standard resources (e.g., through a commercial database such as Lexis-Nexis) is usually reduced to a keyword search. Our proxy for keyword navigation is a similarity network enabled via a topic modeling of the corpus. The underlying distance (metric) produces well-defined regions (i.e., groups of legal sources) that are relatively close to each other, but relatively distant from other regions. Distance is also a proxy for relevance. When new judicial decisions are issued and incorporated into the legal corpus, they interact with search technology to change the legal sources that will be discovered during the next search. This is a new kind of legal effect that, as far as we know, has never been identified as a theoretical possibility, much less formalized and subjected to an empirical test.

Use of the citation network to measure the influence of judicial opinions is now well-studied (see e.g., [13, 12, 20]), although interesting potential avenues of investigation remain underexplored (see e.g., [30] for a citation network analysis in the context of scientific articles). On the other hand, topic models have only very recently entered legal studies where they have thus far showed promise as a new quantitative analysis framework [29, 27, 16, 24]. Both citation networks and topic modeling are examples of computational approaches to legal studies. Early conversations concerning law and digitization focused on distinction in “context” between digital and physical forms, for example, whether digitization enhanced or reduced comprehension or facilitated or undermined serendipity in conducting searches. In particular, the legal significance of the effects of various search modalities (citation-based, keyword, unstructured text) are only just beginning apparent (see e.g., [26]).

The landscape (metric) structure is based on a natural Markov model derived from the multi-network representation of the legal text corpus. The Markov model in turn gives rise to a natural notion of curvature for the underlying state space of the multi-network. As per the usual interpretation of this geometric notion, the more negative the curvature of a region of the legal landscape, the easier it is to navigate to legal sources outside that region from legal sources that are inside of the region. Curvature may change over time as new legal sources are added to the corpus. An increase in curvature in a given region indicates increasing difficulty in navigating to legal sources outside that region from within. This has the interpretation that the region has become more isolated from the rest of the legal corpus and thus is less relevant to new decisions outside of the region. We refer to this effect as the puddling of a region. The opposite effect wherein curvature decreases is referred to as drainage of a region. Drainage of a region is characterized by ease of navigation from points inside the region to legal sources that are outside the region. Notions of network curvature have only just begun to make their way into applied literature. Some early work has adapted the idea of Ricci curvature to the network setting, mainly for its relation to various isoperimetric inequalities (see e.g., [9, 23]). More recent work approaches the idea from the point of view of optimal transport [28]. This in turn makes strong connections to discrete Markov chains – as does ours – but this other work is quite different from the approach taken herein.

We apply our framework to an analysis of all U.S. Supreme Court cases from 1951 to 2002 and investigate the temporal evolution of the curvature over time. Key to our analysis is controlling for something we call legal momentum which captures that fact that some regions of the law remain active and relevant to new decisions while others become vestigial over time. A variety of social and cultural factors may explain the tendency of areas of the law to become vestigial, but, regardless of the factors involved, it is useful to focus on the active regions. When looking at the active part of the legal corpus, we find that regions of the law that experience puddling are less likely to be relevant to future cases, while regions of the law that experience drainage are more likely to be relevant to future cases.

2. RESULTS

We have a data analytic result wherein we show that the metrics we have developed to determine “impact” of an opinion allow us to predict its ongoing relevance. We postpone the technical details of the construction to the next section. As indicated above, our results depend on a notion of distance derived from a multi-network built on the corpus of Supreme Court opinions. The multi-network is realized via the incorporation of three kinds of edges:

- “Cited by” edges – such an edge from \( a \) to \( b \) represents that opinion \( a \) is cited by opinion \( b \).
- “Cited” edges – such an edge from \( a \) to \( b \) represents that opinion \( a \) cites opinion \( b \) and
- “Similarity edges” – this is a weighted edge between \( a \) and \( b \) (i.e., symmetric in \( a \) and \( b \)) that encodes a kind of textual comparison of the texts that depends
on a topic model representation of the opinions (see Methods – Section 3 – for details).

Using these three kinds of edges we create a Markov chain on the space of opinions, which in turn gives rise to a notion of distance between any two opinions A and B, which we denote as PageDist(a, b). The Markov chain further enables us to construct a notion of (local) curvature on this multi-networked set of opinions. For a state (opinion) a let κ(a) denote the curvature at a. Like the traditional differential geometric notion of local curvature (curvature at a point), it reflects the ease of escape from the neighborhood near by the point: the more negative this value, the easier it is to escape.1

If the degree of difficulty of escape is large, a walk will have a tendency to “get stuck” in the neighborhood of the state. This can be interpreted as an opinion that doesn’t connect usefully beyond its surrounding or nearby opinions. Conversely, a more “fluid” area around an opinion suggests that it engages usefully with the broader opinion landscape. This kind of idea will be key to understanding the impact and momentum of an opinion.

As the network of opinions evolves, a measure of change in the local connectivity of the opinions can be expressed in terms of changing κ. We think of it as measuring how the network is bending. Suppose now that we consider the network at two different time points t0 and t1 with corresponding node (opinion) sets N0 and N1. We want to be a little careful as to how we measure the effect of the introduction of new cases and to that end we define κ(a; N0, N1) to be the curvature of the induced chain obtained by lumping into a single connection all opinions that enter the corpus between times t0 and t1 that connect a pair of opinions. Basically, it allows for the new cases to enable potential “shortcuts” not present in the t0 corpus. We then quantify a change in the induced exploration geometry as

Bending(N1, N0)(a) = κ(a; N0, N1) − κ(a; N0)

where κ(a; N0) is the curvature at a as a point in the multi-network built on N0 (i.e., at time t0). Identifying the network with the timestamp we might also write

Bending(a; t1 > t0) = κ(a; t1 > t0) − κ(a; t0).

Figure 1 shows the distribution of κ(·; 1990) as well as bending relative to 1995 in the Supreme Court opinion corpus (Bending(·; 1995 > 1990)).

Bending is easy to interpret, it indicates whether the induced geometry at a point evolves in such a way that it became easier or more difficult to escape from the point. Regions where it is more difficult to make such transitions we call puddling regions and regions where it is easier are called drainage regions. A precise definition should work with the distribution of Bending values, so we call the subset corresponding to the bottom quartile of Bending(·; t1, t0) the Drainage region (relative to a given date) or Drainage(t1, t0). Similarly, we call the subset corresponding to the top quartile of Bending(·; t1, t0) the Puddling region (relative to a given date) or Puddling(t1, t0).

To make precise the utility of these definitions we first quantify what it means for a case to be “impactful”. For this,

\[ \text{Impact}_{t_2, t_1, t_0, d} = \{ a \in N_{t_0} \mid \text{PageDist}(a, b) < d \text{, for some } b \in N_{t_2} - N_{t_1}. \} \]

Thus, this set (with these parameter values) comprises the “early” opinions a (i.e., those that could serve as precedent) that find themselves close to newly arrived (later) opinions (those issued in the period between t1 and t2). Thus the opinions in Impact_{t_2, t_1, t_0, d} have remained relevant to the new opinions.

The threshold d can be set based on various criteria. A natural way to set it is by taking into account the PageDist distribution. A guiding principle that we often follow sets d according to the percentage of cases that we want to declare as “impactful” over a given initial or baseline period. That is, for fixed time periods t0 < t1, as d increases, so does the fraction of opinions in the corpus at time t0 that are considered impactful. Conversely, as the fraction of cases that will be viewed as impactful grows, this implicitly corresponds to an increased threshold d.

We further define the Initial Impact Probability (IIP) (for t1 > t0 and a given threshold d) as the fraction of opinions present at time t0 that are in Impact_{t_2, t_1, t_0, d} – i.e., those opinions that remain impactful at time t1 according to a threshold d. The goal is to understand how to predict which cases remain impactful as time goes on. Figure 2 shows how IIP varies with the impact on future cases \( P(x \in \text{Impact}_{t_2, t_1, t_0, d} \mid \text{Impact}_{t_1, t_0, t_0, d}) \). Therein we graph

\[ P(x \in \text{Impact}_{t_2, t_1, t_0, d} \mid \text{Impact}_{t_1, t_0, t_0, d}) - \text{IIP} \]

(with t0 = 1990, t1 = 1995, and t2 = 2000) against IIP (recall that as d increases monotonically with IIP, so that we can view both axes as functions of d). This behaves as might be expected, with an increasing percentage of opinions remaining impactful, until such a time as too many initial cases are tossed in, some of which will be opinions that have become vestigial.

Let us now fix d so as to correspond to the maximum IIP setting in Figure 2. With the choice of d set, we now have fixed the parameter by which we identify opinions as impactful. We can now examine how drainage and puddling effects the impact on future cases. This is shown in Figure 3. We see the impact on future cases (the blue line) compared to impact on future cases in the “drainage” and “puddling” regions. Therein we see that indeed, drainage regions (low bending) have roughly a greater than 10% chance of having an impact on future cases than do puddling regions (high bending). That is, the drainage regions that are connecting up the space are more associated to future impact. The caption for Figure 3 contains some detail around the statistical significance of the result.

3. THE MATHEMATICAL FRAMEWORK

3.1 A random walk model for legal research

The geometry we construct for the legal corpus is based on a model of how the legal corpus is utilized as a network – that is, the geometry is derived from a model of the search process. We frame legal search as a process of “local” exploration of the opinion corpus, i.e., modeling the way in which
Figure 1: On the left we see a histogram of the curvature $\kappa(\ast; 1990)$ and on the right we see the bending $Bending(\ast; 1990, 1995)$. This gives a sense of the variation of the curvature over time.

a user of the legal corpus might navigate from opinion to opinion in the process of researching an issue. This kind of navigation is naturally viewed as a Markov chain (see e.g., [17]), formulated as a matrix $T$ of transition probabilities where the entries are indexed by the opinions. For opinions $a$ and $b$ the value of the entry $T(a, b)$ is the probability of “moving to” opinion $b$ from opinion $a$ in an exploration of the legal corpus. More precisely, framing this as a “random walk” in “opinion space” this is the probability of moving at the next step to case $b$, given that you are at case $a$, i.e., the conditional probability

$$T(a, b) = P(b|a),$$

in standard notation.

Our transition probabilities are constructed as a combination of several terms, reflecting a model of navigation of the space of legal opinions. One of the defining features of judicial opinions is their citation of relevant prior legal sources.

Our model of legal search thus makes use of a combination of three basic types of local exploration from an initial opinion $a$: consideration of (1) opinions cited by $a$; (2) opinions that cite $a$, and (3) opinions that are similar to $a$ from two other legal sources, including statutes and constitutions, have other types of internal ordering (such as organization by chapter or article) that may be relevant for law search. For purposes of this analysis, we restrict our application to the body of U.S. Supreme Court decisions and do not incorporate other sources of law. The framework of search that we develop, however, is generalizable to these other legal sources.

a textual point of view. The last of these is to be determined by a notion of similarity based on the use of a topic model. The topics are derived automatically from the overall corpus (see [6] for a friendly explanation of topic modeling). While there are a number of different kinds of topic models, the “latent Dirichlet allocation” (LDA) model (the “Dirichlet” refers to an underlying assumption of a Dirichlet distribution in the model) is perhaps the best known and most widely used [5]. This is the topic model that we use here. A detailed description of topic modeling is beyond the scope of this paper. Suffice to say that a topic model derives a representation of each text in the corpus as a mixture of probability distributions over the vocabulary in the corpus. Each distribution is a “topic”.

As mentioned, the Markov chain (transition matrix) can be written as a linear combination of chains, $T_{\text{cited}}, T_{\text{cited-by}}$, and $T_{\text{sim}}$. Moreover, it is possible that the exploratory mode (i.e. the weights given to the three forms of connection in the network) may vary for a given search. That is,

$$T(a, b) = p_{\text{cited}}(a)T_{\text{cited}}(a, b) +$$

$$p_{\text{cited-by}}(a)T_{\text{cited-by}}(a, b) +$$

$$p_{\text{sim}}(a)T_{\text{sim}}(a, b)$$

(1)

with the proviso that

$$p_{\text{cited}}(a) + p_{\text{cited-by}}(a) + p_{\text{sim}}(a) = 1$$

reflecting that these are all the possible ways in which one
Figure 3: Here the x-axis is the year the case was decided. The blue curve is $P(x \in \text{Impact}_{t_2, t_1, t_0, d} \mid \text{Impact}_{t_0, t_1, t_2, d})$ with $t_0 = \text{date}$, $t_1 = \text{date} + 5$, and $t_2 = \text{date} + 10$ and $d$ fixed by the 20th percentile, as in Figure 2. In black we see the same curve conditioned on Drainage regions, while in red the same curve conditioned on Puddling regions. Notice that indeed, the bending is correlated with long term impact as predicted, and that after the geometry has really “warmed up” (about 1978), we see a fairly stable 10% difference. To confirm that this correlation is statistically significant, let the null hypothesis be that there is nothing but a random difference between the Drainage and Puddling regions. So for a fixed measurement, under the null hypothesis there would be a fifty-fifty chance that we confirm our suspicion (technically, bounded by 50% when allowing for ties). Furthermore, for events that differ by at least 5 years, the $N_{t_2} \setminus N_{t_1}$ populations are distinct, so that the measurements are suitably independent. Thus, we have 6 independent measurements with a perfect track record and can conclude that the $p$-value is less than $\frac{1}{2^6}$ and the correlation significant.
navigates the corpus. (The notation suggests the weights may vary depending on the initial state of the search.)

The transition matrices $T_{\text{cited}}$ and $T_{\text{cited-by}}$, based on the citation network are straightforward to construct. A natural and standard choice is to weight equally all opinions cited by a given opinion, and similarly for all opinions that cite the given opinion. This could be varied in some way, perhaps accounting for some notion of the importance of an opinion. We choose to work with equal weights. We make use of the excellent “Supreme Court Citation Network Data” database created by Fowler and Jeon [1].

The construction of $T_{\text{sim}}$ requires more detailed explanation. We only consider as relevant to a given opinion the “top” topics and similarly for a given topic, only consider as relevant to our exploration those opinions who express it most strongly. More precisely, we fix integer parameters $M_T$ and $M_O$ such that for a given opinion $a$, Topic$_a$ is the set of the $M_T$ most heavily weighted topics in opinion $a$ and for a topic $t$ within Topic$_a$, we let Opinion$_t$ comprise the $M_O$ other opinions in which $t$ was most strongly expressed. Thus for a given opinion $a$ we can create an $M_T \times M_O$ matrix in which the $i, j$ entry is the $j$th most significant opinion in the corpus for the $i$th most significant topic in opinion $a$. If we define $W_{a,b}$ to be the number of times opinion $b$ occurs in this matrix, then $T_{\text{sim}}$ is the random walk produced by normalizing according to these weights.

The Markov chain that we have derived to mimic the search process is a natural generalization of the famous PageRank algorithm [8]. Of interest to us is the geometry that this search model produces. In particular, this kind of Markov-based search produces a metric on the network space that we call PageDist. We call the induced geometry an exploration geometry.

To define PageDist we attach one last parameter $r$ to the random walk of (1): at each step assume a probability $r$ of ending the exploration. Hence, starting at an opinion $a$ the expected time (number of steps it takes) for a search to end at opinion $b$ is

$$R(a, b) = \sum_{k=0}^{\infty} r^k T^{(a,b)}.$$

With this we define the PageDist metric as

$$\text{PageDist}(a, b) = \left\| R(a, \cdot) - R(b, \cdot) \right\|_p$$

where $p$ denotes the $p$-norm. The PageDist metric captures our notion of distance within the landscape. Figure 4 shows the distribution of distances among our corpus of Supreme Court opinions.

Figure 2: Here the $x$-axis is Initial Impact Probability (as a percentage), the function of $d$ that gives the fraction of early opinions with impact on the initial set of new opinions. Recall that as IIP increases, so does $d$. In blue we see we see $P(x \in \text{Impact}_{t_2,t_1,t_0,d} | \text{Impact}_{t_1,t_0,d} - \text{IIP})$ with $t_0 = 1990$, $t_1 = 1995$, and $t_2 = 2000$ (and $d$ a function of IIP). Thus, this is the proportion of early (pre-1990) opinions that continue to have impact in the 1995-2000 period, given that they had impact in the 1990-1995 period, minus the fraction of opinions that initially have impact on opinions written between 1990 and 1995. Thus, we are subtracting out some baseline guess of how many of these early cases you would expect to have impact in this time based on earlier information. This measures how much larger than random the future impact is given recent impact. This is all a function of $d$ or equivalently, IIP. We see that IIP = 20 is roughly an optimal value.

\(^3\)Recall that this notation means $(\sum_x |R(a, x) - R(b, x)|^p)^{1/p}$.
The random walk setting also makes possible a definition of curvature that encodes a level of difficulty for escape from a given point (in the execution of a random walk). If the degree of difficulty is large, a walk will have a tendency to get “stuck” in the neighborhood of the state. This can be interpreted as an opinion that doesn’t connect usefully with its surrounding or nearby opinions. Conversely, a more “fluid” area around an opinion suggests that it engages usefully with the broader opinion landscape. This kind of idea will be key to understanding the impact and momentum of an opinion.

We define curvature as

$$\kappa(a) = \log(\mathcal{R}(a, a) - 1).$$

As the network evolves we measure how local connectivity in terms of changing $\kappa$. We think of it as measuring how the network is bending. Let us make this precise. Given a network $N$ with a transition matrix $P$ reflecting a Markov process on the network, let $S < N$, be node sets. A Markov chain on $N$ induces a chain on $S$ by using the weights

$$W_S(a, b) = P(a, b) + \sum_{k \in N \setminus S, a \neq b} P(a, k)P(k, b),$$

for $a, b \in S$. Note that we are simply lumping together into one term all transitions $a$ to $b$ that go outside of $S$. We form a new transition matrix $P(a, b; S, N)$ normalizing $W_S(a, b)$ so that the weights sum to one at each vertex. We call this the induced local exploration. This induces a corresponding exploration geometry and a curvature $\kappa$ for $S$ relative to $N$ which we denote as $\kappa(a; S, N)$. This is the curvature notion introduced above.

As per the description in Section 2 we consider the network at two different time points $t_0$ and $t_1$ with corresponding node sets $N_0$ and $N_1$. Then we can quantify a change in the induced exploration geometry as

$$\text{Bending}(N_1, N_0)(a) = \kappa(a; N_0, N_1) - \kappa(a; N_0, N_0).$$

Identifying the network with the timestamp we might also write

$$\text{Bending}(a; t_1 > t_0) = \kappa(a; t_1 > t_0) - \kappa(a; t_0).$$

4. CLOSING THOUGHTS

In this paper we introduce a new multi-network framework integrating citation and textual information for encoding relationships between federal judicial opinions. The citation component derives from the underlying citation network of opinions. The textual piece derives from an LDA topic model computed from the text corpus. The network is the reification of a basic model of legal search as would be executed by a prototypical legal researcher (“homo legalus”) looking for cases relevant to some initial case. The notion of search turns into a Markov chain on the network, built as a linear combination of the individual chains on the citation and topic networks. The Markov process produces a notion of distance between opinions which can also be thought of as a proxy for relevance. Along with distance, there is a notion of curvature, and with this an implicit framing of the opinion corpus as a “landscape” which we call “the legal landscape”.

We have implemented a first generation website that will allow users to explore a smallish subset of Supreme Court opinions using this search tool (www.bendingthelaw.org).

The text corpus evolves in the sense that cases enter the corpus regularly and in so doing continually deform the asso-
cated text landscape. Of particular interest are those cases that remain relevant over long periods of time—such cases we call impactful. Some regions of the legal landscape have the property that they serve as nexuses of connection for regions of the landscape. We show that those regions which over time become significantly more negatively curved are such connective areas. With the analogy of flow in mind, we call such areas, regions of “drainage”. Areas which experience a significant increase in curvature we call “puddling regions”. We show that drainage areas are more likely to contain the impactful cases than the puddling regions. We further show that opinions that start off impactful, in the sense of entering the landscape highly relevant to many cases over a short period of time tend to remain impactful, thereby suggesting a property of legal momentum.

There are natural next steps to take with this idea. In one direction we will expand the text corpus to include all Supreme Court and Appellate Court Opinions. We also plan to validate and compare our model by asking users to compare the results of our search algorithm (under a range of parameter choices) with their own usual research approaches. Our newly introduced opinion distance function gives a new variable to explore the relations of opinions to all kinds of social and economic variables. It is also natural to export this model to other court systems that produce English language opinions. In this regard it would be interesting to see the ways in which the “bending” of the courts systems vary, and try to understand what might account for such (possible) variation. Ultimately, it would also be of interest to effect the integration of distinct corpora via this model.

5. ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the Neukom Institute for Computational Science at Dartmouth College.

6. REFERENCES

APPENDIX

A. IMPLEMENTATION.

The ideas presented in this paper form the foundation of new web-based search tool for exploring a space of legal decisions using the exploration geometry introduced in the body of this paper. Specifically, we have built a prototype for a website and user interface that will enable the exploration of an opinion database, that ultimately will encompass all Federal Court and Supreme Court cases. At present it is running on a small subset (SC cases 1950–2001). This prototype can be found at www.bendingthelaw.org.

The current user interface (UI) introduces users to cases in the “vicinity” (in the sense of our exploration geometry) of a pre-identified case specified by the user. The anticipation is that these cases will be strong candidates for precedent-based reasoning. As per (1) the return depends on the database of cases as well as the individual weights assigned to the three-component random walk process encoding the exploration geometry— that is, a choice of weights $p_{\text{cited}}$, $p_{\text{citedby}}$, and $p_{\text{sim}}$. As a first step we allow a choice of weights from $\{0, 1, 2\}$. Recall that the similarity piece of the random walk, $T_{\text{sim}}$ requires that we construct the “topic by opinion” matrix of a given size. We choose that to be 10 – i.e., that for any given topic we consider the 10 opinions that make the most use of it and conversely, for any opinion, we consider the 10 topics that make the strongest contribution to it.

Given an initial query, the UI provides two complementary representations: (1) a ranked list of geometrically closest (in terms of PageDist) cases and (2) a map of the space, centered on a case of origin (the original input). As a “map”, this representation shows not only the relation of cases to the initial query, but also the relations of the closest cases to each other. The associated visual integrates the citation network representation with a 2-d multidimensional scaling visualization of the thirty (including the query) intercase distances. (An arrow from case A to case B means that case A cites case B.) The map is generated by clicking on “View Case Network” (after executing the query). The opinion map produced from the query “329 US 187: Ballard v. United States” is shown in Figure 5.

Figure 5: Here is a snapshot from our alpha version UI for exploring the space of legal opinions. The current UI is built on the database of Supreme Court opinions over the time period 1950–2001. What we see here is the 2-d MDS visualization of the PageDist neighborhood of 30 closest cases to “329 US 187: Ballard v. United States”. Note that the exploration weights have been set to 2 (“cited”), 1 (“cited by”), and 2 (“topic similarity”).
ABSTRACT
Automatic interpretation of court filings would help improve the efficiency and effectiveness of judicial process. The many technical challenges hindering such interpretation include the introduction of noise during Optical Character Recognition (OCR) due to the stamps on scanned court filings. This paper presents an approach that employs an image analysis technique to identify the stamps for removal prior to OCR, thereby improving OCR quality. The results show 98% accuracy for document images with clean stamps.

CCS Concepts
• Applied computing → Law, social and behavioral sciences → Law

Keywords
Court filings; judicial; document image; image processing; stamp; optical character recognition; OCR

1. INTRODUCTION
Since it began in the late 1990s, the transition from paper to electronic filing has transformed how the U.S. federal courts operate. Yet despite many advances in judicial access and administration, the courts still lack the capability to automatically interpret the contents of filings. Instead, court filings are interpreted only when they are reviewed by humans, such as a judge, court staff, or attorney. This results in low-performance quality control and lengthy processing time when human resources are constrained. Machine interpretation of court filings would open a rich source of information for improving court administration, access to justice, and analysis of the judicial process. However, automating the interpretation of court filings poses many technical challenges, such as improving OCR quality, recovering document layout, and understanding the context.

Both native Portable Document Format (PDF) documents and scanned PDFs are permitted in court filings. These scanned PDFs require OCR to extract the text. Figure 1 shows some examples of stamped court filings in scanned PDF format. Inserted by court staff or judge, the stamps on court filings may introduce unwanted text during OCR, thus reducing OCR quality. Further noise can be introduced when stamps interfere with segmentation algorithms included in OCR process. This paper presents an approach to identify the stamps for removal prior to OCR, thus improving subsequent OCR quality and therefore enhancing automatic interpretation of court filings.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 2010 ACM 1-58113-000-0/00/0010 ...$15.00.
DOI: http://dx.doi.org/10.1145/12345.67890

2. METHODOLOGY
2.1 Hypothesis
Shape analysis in images has been an active research area because it has broad applications in many fields, such as medical imaging, intelligence, robotics, and transportation [1, 2, 3 and 4]. In general, the methodologies for shape analysis fall into two categories: contour-based and region-based. The contour-based method uses only the information from the contour, such as perimeter and Fourier transform. The region-based method examines the entire shape region, including area and moment.

A collection of publicly available filings from the U.S. District Court for the Southern District of New York shows four common types of stamps. As displayed in Figures 1 and 2, the stamps are either rectangular or oval and have distinctive enclosed areas compared to other components in the document.
We hypothesize that the stamps can be automatically identified by their significant area. The first step needs to establish the cut-off threshold.

### 2.2 Establishing Threshold

The process for finding the threshold involves the following steps:

1. Select document images with clean stamps
2. Reduce the document size by 4 x 4
3. Segment the document into connected components
4. Dilate connected components
5. Convert connected components into polygons
6. Sort the polygons by area size
7. Remove any interior polygons
8. Plot the histogram
9. Identify a threshold

First, we select a set of document images (scanned PDFs) with clean stamps. “Clean” stamps are those that have fair quality and do not overlap with other text. Second, both height and width of the image are reduced by a factor of 4, because processing the original image (e.g., with a typical scanning resolution of 300 dots per inch) requires impractically significant time and computing resources. A factor of 4 was experimentally determined to reduce processing time sufficiently without degrading performance. Then the document image is segmented into connected components using a graph-based segmentation algorithm [5]. Each connected component is dilated sufficiently to fill in the gap caused by poor scanning quality and is then converted into a polygon. All polygons are sorted by area size. The polygon is removed if it is completely covered by another polygon (e.g., an interior polygon such as the inner space of a letter ‘o’). Finally, a histogram will be plotted and a threshold will be identified. OpenIMAJ®, an open-source software package, is used for image processing [6].

We selected seven documents with eight clean stamps in total (one image had two stamps) to establish the threshold. The documents included all four types of stamps shown in Figure 2. Figure 3 displays the histogram of polygon areas, with the X-axis representing the polygon area size, and the Y-axis representing the frequency. For example, the column at 60 K means that two polygons have an area size between 55 K and 60 K pixels.

It is manually verified that the polygons with area larger than 35 K are all related to stamps. As the histogram shows, the areas of stamps clearly differ strongly from other areas of typical textual components. Therefore, it was reasonable to set the cut-off threshold as 35 K.

### 2.3 Algorithm

Figure 4 describes the steps implemented to identify the stamps. All the polygons will be checked against the threshold (35 K). For each document image as the input, the final outcome can be zero, one or more than one stamps.

![Flow chart of the process to identify stamps.](image)

**Figure 4.** Flow chart of the process to identify stamps.

### 3. RESULTS AND DISCUSSION

We collected additional 56 documents with 59 clean stamps and applied the threshold of 35 K pixels to detect the stamps. As shown in Figure 5, all 59 stamps are correctly identified. However, two non-stamp components are misidentified as stamps. This results in 100% recall and 97% precision for stamp identification.

![Confusion matrix of stamps and non-stamps.](image)

**Figure 5.** Confusion matrix of stamps and non-stamps.
Figure 6 shows some examples of document images in which the stamps were correctly identified. Figure 7 displays one example in which the algorithm misidentified the document’s signature block as a stamp. This occurred because the signature is large and the presence of cursive characters produces an enclosed area larger than the 35 K threshold.

Figure 8 shows some examples of non-clean stamps that overlap adjacent text. In addition to identifying the real stamps, the algorithm misidentified some overlapped words as part of the stamps, resulting in a “tail” shape.

We examined Scale-Invariant Feature Transform (SIFT) matching as an alternative technique to area size, but determined that the stamps lack sufficient features to establish reliable SIFT matches between stamp templates and actual stamps.

4. Future Work
Since rectangles and ovals are symmetric shapes, this property of symmetry can be calculated and used to prevent large and cursive signatures from being misidentified as stamps. To remove the “tail” shape, an algorithm could be developed that identifies two points on the contour where the ratio of walking distance along the contour to the direct distance exceeds a threshold. This direct distance could be used to replace the walking distance as part of the contour. Future research will investigate whether this method is generally applicable to other document sets that contain different sets of stamps. For example, if other courts use different stamps, the threshold needs to be recalibrated. The robustness of the threshold-setting process should be studied as well.

5. ACKNOWLEDGMENTS
This work was funded under contract number USCA16D2019. The MITRE Corporation is a not-for-profit Federally Funded Research and Development Center chartered in the public interest. © 2016 The MITRE Corporation. All rights reserved.

6. REFERENCES


ABSTRACT

Forecasting the potential impact of newly passed legislation on the greater legal system, and the governmental agencies that the legislation regulates, is often difficult due to the multiple cross-referenced layers of legal corpora. However, by constructing a network for the U.S. laws, regulations, and policies based on the citations between them, it may be possible to assess the potential impacts introduced by new legislation. Accordingly, a legal path analysis framework was developed to evaluate the impact of legislation on U.S. government agencies. This framework and its implementation as a database driven web application will enable policy analysts and researchers to trace and visualize the “path” of changes to the United States Code and Code of Federal Regulations resulting from a given new legislation. Additionally, the developed tool will also help to elucidate the interconnected web of regulations and policies created by new legislation, and highlight agencies that are ultimately affected by the changes in the legal system.

Categories and Subject Descriptors
D.3.3 [Programming Languages]: Python, Neo4j, Cypher, JavaScript

General Terms
Graph Theory; Legal Analysis

Keywords
Policy networks, rulesets

1. INTRODUCTION

The U.S. legal system includes several major bodies (or corpora) of legal texts, which, among other things, outline the authorities and responsibilities of federal government agencies. Changes to the scope of the agencies are often initiated by the Congress via the introduction and passage of bills and resolutions. Enacted bills become Public Laws (PL), which can amend the United States Code (USC) [1]. Amendments may create and eliminate agencies, change their funding, or expand or contract their jurisdiction [2]. These amendments are further interpreted by relevant government agencies via the rulemaking process by proposing new regulations that in turn may update the Code of Federal Regulations (CFR) or make changes to the internal policy documents of the agencies. Eventually, agencies act on these new changes. This chain of events may take several years, and may include in-progress revisions and interpretation by Congress and the courts [2].

According to GovTrack [3], since 1973 there have been over 11,000 enacted PLs, averaging 500 a year (although the rate of new laws passed has decreased by half this decade) [4]. Furthermore, by some estimates the USC contains over 47,000 pages of text [5], and the number of words in the USC has grown from about 2.5 million in 1925 to about 43 million in 2011 [6].

Similarly, according to Regulations.gov [7], federal agencies and departments issue nearly 8,000 Federal Regulations (FR) per year. With the CFR being on the same order of magnitude, in terms of number of sections, as the USC, the collective legal corpora are an immense, complex, and constantly changing system.

The legal system can be modeled as a network, with the legal text (e.g., titles, chapters, sections, clauses of USC) and citations within corresponding to the entities (or nodes) and relationships (or edges), respectively. Using such an approach, the interconnected set of Bills, PLs, Statutes, the USC, Rules, and the CFR can be treated as a network, to explore the chain of changes that result from a new legislation.

In network science, this is referred to as path analysis. A path is defined as a sequence of ‘edges’ that connect a sequence of ‘nodes’ [8]. In this context, the nodes are legal texts and the edges are citations. Some nodes are further associated with government agencies, either because the agency created the corresponding legal text or the text governs the operation of the agency. The goal
of this path analysis is to assess, given the paths of nodes and edges linked from a legislative act, the agencies that are associated with the changed document nodes, to infer potential impact of the legislation on the agency.

To perform such analysis, a database driven web application (PolicyNet [9]) was developed to model this network. PolicyNet performs automated download, citation extraction, and ingestion for each of these document corpora from FDSys [10]. Regulations.gov [11], and other legal document sources into a graph database (Neo4J). A web-based application is provided to explore the network by querying of the network based on topic or structure of the documents, as well as perform path analysis.

The proposed demonstration will focus on using the path analysis to explore this vast network of structured documents using this tool. PolicyNet has several features for exploring the network of laws, regulations, and policies, including a graph-based search and a neighborhood view of any node in the network. This demonstration will highlight a new feature, in which users can enter a Public Law and see all the laws, regulations, and policies that have cited the law. This network of documents is organized hierarchically, such that the provenance of changes is represented. This view of the data demonstrates the magnitude of changes that a given law can create, and provide an indication of the agencies and policies that are affected by the law. This tool, backed by a database of all public laws and regulations, provides a generalized method to answer empirical questions about regulatory impact.

2. RELATED WORK

Citation network analysis has been used effectively in judicial analysis to determine the impact of supreme court decisions [12]. Several papers have described network representations of legal code, and found properties similar to many traditionally-studied networks. In particular, Boulet, et al. [13] explored the French environmental code with several traditional network measures and found it to be similar to many small-world networks in every-day life, including a network of German companies and a medieval social network.

In Measuring the complexity of the law: the United States Code, Katz, et al. [14] found several scaling factors within the US Code, indicating it’s similarity to other real-world networks. Further, Katz et al. used the network structure and other measures of knowledge acquisition to propose an empirical framework for evaluating the complexity of the USC, based on its structure, language, and interdependency.

By combining network analysis with legal insights about the classes of citations, Hamdaqa, et al. [15] differentiated between citing and amending citations, and defined a set of graph patterns that may indicate overlaps or conflicts in law documents. Each of these authors have found useful insights about the legal system using citation network analysis. This analysis builds upon their research by leveraging similar network analysis techniques, including edge classification, to tell the story of a law’s journey through the legal system.

3. METHODS

To construct a complete network of the legal system, it is necessary to incorporate all relevant corpora of legal documents. Accordingly, the most updated revision of the USC and CFR have been ingested into the PolicyNet database. Most of the federal legal corpora is available online in machine-readable format, along with document metadata in various levels of specificity. For example, current and some historic Bills, Public Laws, Statues, the USC, Federal Rules, the CFR, and many other official documents are available in bulk from the Government Publishing Office’s FDSys site [16].

To perform the path analysis of a given PL, each amendment to the USC by the PL is obtained by parsing the PL using a citation parsing algorithm. Several open-source citation parsers are available [17], including one developed specifically for the PolicyNet tool [9].

Subsequently, a list of Proposed and Final Rules issued after the enactment of the PL, which have citations to the amended parts of USC, are found using the citation network in the PolicyNet database.

Lastly, the parts of CFR modified by the Proposed and Final Rules are obtained by parsing the rule. Additionally, if agency level policy and regulations (e.g., operating manuals) are available, those that cite the amended and modified sections of the federal laws are also listed. Once all the relevant laws are itemized, the affected agencies can be identified as those that issued the Proposed and Final Rules, or those whose operations are governed and regulated by the identified sections of CFR and USC. The impact to a given agency can be quantified using various metrics, such as the number of sections of the laws affecting the agency that are amended or modified as a result of the PL.

These documents are organized into a hierarchical citation network and displayed in the browser using an open-source network visualization library [18]. Laws and agencies are highlighted, and colors are used to differentiate the types of legal corpora. This network provides an indication of the path of implementation of a given law by agencies.

4. REFERENCES


ABSTRACT
This short discussion paper reviews my prior work (from 1998, 2007 and 2015) and proposes a research strategy to produce a computational summary of a legal case, which can be scaled up to a realistic legal corpus.

1. THE CHALLENGE
In a paper written almost twenty years ago, I advocated the development of an intelligent legal information system based on structured casenotes[14].¹

The general argument was that editorial enhancements to primary legal materials (statutes, regulations, cases, etc.) should not take the form of additional natural language texts (treatises, annotations, practitioner’s guides, etc.), but should take the form of computational structures, “using recent advances in Knowledge Representation (KR) and Natural Language (NL) techniques.” Specifically, for the editorial enhancement of a legal case, I proposed the following definition:

A structured casenote is a computational summary of the procedural history of a case along with the substantive legal conclusions articulated at each stage of the process. It would play the same role in the legal information systems of the 21st century that West Headnotes and Key Numbers have played in the 20th century.

The main body of the paper then explored this proposal by analyzing a recent copyright case. Quality King Distributors, Inc., v. L’Anza Research International, Inc., 118 S.Ct. 1125 (1998), and its three opinions, in the District Court, the Court of Appeals for the Ninth Circuit, and the Supreme Court.

The challenge, of course, was to build a database of structured casenotes at the appropriate scale. I suggested two approaches, which would most likely be used in combination in any practical system: First, we could attempt a semi-automated system, in which a human editor would produce a computational summary of a typical case” (the NL component). The structured casenote could then be elaborated as follows:

For example: We could focus on copyright law. The paper concluded by comparing the Supreme Court’s decision in Quality King v. L’Anza Research with the prior decision in the Ninth Circuit, which turned on a disagreement about the scope of §§106(3), 109(a) and 602(a) of the 1976 Copyright Act. A structured casenote would represent this disagreement in a formal knowledge representation language (the KR component again), and link it back to the procedural history of the case. The paper showed how this would work.

My focus on procedural history was based on the traditional “brief” that students are taught to write in their first year of law school. I explained the idea as follows:

The traditional case brief focuses on the procedural context first: Who is suing whom, and for what? What is the plaintiff’s legal theory? What facts does the plaintiff allege to support this theory? How does the defendant respond? How does the trial court dispose of the case? What is the basis of the appeal? What issues of law are presented to the appellate court? How does the appellate court resolve these issues, and with what justification?

To ask these questions and answer them in a structured casenote, I wrote, we need “a representation of the rules of civil procedure at some reasonable level of abstraction” (the KR component), and we need a computational grammar “with coverage of the procedural expressions that occur in the synopsis of a typical case” (the NL component). The structured casenote could then be elaborated as follows:

Within this procedural framework, we would represent the substantive issues at stake in the decision. This is more complicated, since the territory is so vast, potentially encompassing the entire legal system. We can get started on this project, however, by focusing on particular areas of the law . . .

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

LTIDCA’16, June 17, 2016, San Diego, CA, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM XXX-XX, $15.00.
http://dx.doi.org/xxx.xx.
2. TWO STEPS TOWARD A SOLUTION: ICAIL ’07 AND ICAIL ’15

When I wrote my paper on structured casenotes in 1998, the field of computational linguistics was not advanced enough to handle the complexities of judicial opinions. But the state of the art changed dramatically with the publication of Michael Collins’ thesis in 1999 [5]. At the International Conference on Artificial Intelligence and Law (ICAIL) in 2007, I published a paper that explored the potential impact of this research on the law [16].

There were two main contributions in my paper. First, I showed that a “state-of-the-art statistical parser . . . can handle even the complex syntactic constructions of an appellate court judge” by presenting the output of Collins’ parser applied to the full text of the judicial opinions in 111 federal civil cases, comprising a total of 15,362 sentences. Second, and more important, in my opinion, I showed that “a semantic interpretation of the full text of a judicial opinion can be computed automatically from the output of the parser.” The main technical contribution of my paper was the specification of a quasi-logical form, or QLF, to represent the semantic interpretation of a sentence, and a definite clause grammar, or DCG, to compute the correct quasi-logical form from the output of the parser. The DCG was constructed manually, but with automated assistance, and it stabilized at approximately 700 rules. The QLF was based loosely on my Language for Legal Discourse (LLD) [13], and it was intended to serve as an intermediate step towards the construction of the full representation of a legal case in LLD, thus realizing some of the goals articulated in my paper on structured casenotes.

To see how this might work, take a look at the QLF interpretation of a sentence from 526 U.S. 795 (1999) in Figure 2 of my ICAIL ’07 paper. You should be able to see that this data structure provides a partial answer to the first two questions from the procedural history in a traditional case brief: Who is suing whom, and for what? What is the plaintiff’s legal structure provides a partial answer to the first two questions from the procedural history in a traditional case brief: Who is suing whom, and for what? What is the plaintiff’s legal

How accurate are these syntactic analyses and semantic interpretations? Unfortunately, we do not have the data to answer this question. Collins’ parser was trained on Sections 02-21 of the Wall Street Journal corpus (approximately 40,000 sentences) and tested on Section 23 (2,416 sentences). His overall results for labeled recall and labeled precision were 88.0% and 88.3%, respectively, and his parser performed much better on the core structure of a sentence than it did on the fringes. Qualitatively, our results for judicial opinions seem to be similar, but not quite as good, because we

are outside the genre of the training set. But since we do not have an annotated test set of judicial opinions, we cannot answer this question quantitatively. For the semantic interpretations, the situation is even worse. It is difficult to get people even to agree on what semantic (or logical) representations we should be annotating!

Let’s consider a different strategy. It is not difficult to write a set of hand-coded extraction patterns to map information from the QLF interpretations in Figures 2 and 3 into the format of a structured casenote. (As an experiment, I have written many batches of these extraction patterns, for various components of a structured casenote, all encoded in horn clause logic.) Of course, a manual technique like this will not scale up to a large database. But now we can try to generalize our hand-coded patterns by the unsupervised learning of the legal semantics implicit in a large set of unannotated legal cases. The total system would thus be engaged in a form of semi-supervised learning of legal semantics.

To pursue this strategy further, we need to look at my recent paper at the International Conference on Artificial Intelligence and Law (ICAIL) in 2015 [18]. It may not be apparent why this paper is relevant, however, since it is extremely theoretical. The mathematical foundations of the work are presented in [17], where I developed the theory of differential similarity. There are two mathematical components in the theory of differential similarity:

1. A probabilistic model. This is a diffusion process determined by a potential function, \( U(x) \), and its gradient, \( \nabla U(x) \). The general theory is developed in an arbitrary \( n \)-dimensional Euclidean space, but for a three-dimensional illustration, see Figures 1 and 2 in [18]. It turns out that the invariant probability measure for the diffusion process is proportional to \( e^{U(x)} \), which means that \( \nabla U(x) \) is the gradient of the log of the stationary probability density.

2. A geometric model. This is a Riemannian manifold with a Riemannian metric, \( g_\Omega(x) \), which we interpret as a measure of dissimilarity. Using this dissimilarity metric, we can define a radial coordinate, \( \rho \), and the directional coordinates \( \theta_1, \theta_2, \ldots, \theta_{n-1} \) in our original \( n \)-dimensional space, and then compute an optimal nonlinear \( k \)-dimensional subspace. The radial coordinate is defined to follow the gradient vector, \( \nabla U(x) \), and the directional coordinates are defined to be orthogonal to \( \nabla U(x) \). For a three-dimensional illustration, see Figure 3 in [18].

The two components are connected to each other, because of the role played by \( \nabla U(x) \) in each. Specifically, the dissimilarity will be small in a region in which the probability density is high, and large in a region in which the probability density is low. Thus, when we compute a nonlinear subspace, we are simultaneously minimizing dissimilarity and maximizing probability.

The theory of differential similarity thus fits within the framework of “manifold learning” as described in Section 2.1 of my

3The corpus consisted of 9 Supreme Court cases from May, 1999: 57 Second Circuit cases from May and June, 1999; and 45 Third Circuit cases from May and June, 1999. I have sometimes been asked why these cases, in a paper published in 2007, all came from May and June of 1999. The answer is that I started experimenting with Collins’ parser as soon as it became available, and this was simply the first database that I assembled.

Hence the title of my paper. Section 2.2 explains how the theory of differential similarity can be applied to a hierarchical learning problem. Figure 4 shows an architecture for “deep learning” on the MNIST dataset of handwritten digits, and Figure 5 shows a small example of the output from the first stage in this process. The stages alternate as follows:

(i) Assume that we have computed a low-dimensional non-linear subspace for each of the (two or more) prior stages in the architecture. We use the encoded values given by the radial coordinate systems in these non-linear subspaces to construct a product manifold in the current stage.

(ii) Our input data can be represented in this product manifold using any coordinate system that is convenient for the task at hand, but our probabilistic model (and our statistical estimation procedure) only works in a Euclidean space with Cartesian coordinates. Thus we resample the data in Cartesian coordinates and use our Riemannian dissimilarity metric again to construct a nonlinear submanifold in the current stage.

Notice that we are cycling back and forth between two geometries as we proceed through this architecture: a linear Euclidean geometry and a nonlinear Riemannian geometry.

The main technical contribution of my ICAI’15 paper is in Section 3, where I show how to use the machinery of manifold learning and deep learning to define a semantics for a logical language. The theory of differential similarity is thus extended to encompass a third component:

3. A logical language. The proposed logical language is a categorical logic based on the category of differential manifolds (Man), which is weaker than a logic based on the category of sets (Set) or the category of topological spaces (Top).

For a quick intuitive understanding of what this means, assume that we have replaced the standard semantics of classical logic, based on sets and their elements, with a semantics based on manifolds and their points. The atomic formulas could then be interpreted as prototypical clusters, similar to the geometric objects illustrated in Figures 3 and 5, and the geometric properties of these clusters could be propagated throughout the rest of the language. The same strategy could be applied to the entirety of my Language for Legal Discourse (LLD), as sketched briefly in Section 4 of my ICAI’15 paper. There is more work to be done here, of course, but these are feasible tasks.

Several people have asked me to explain the title of my ICAI’15 paper: *How to Ground a Language for Legal Discourse in a Prototypical Perceptual Semantics*. Why do I call it a “prototypical perceptual semantics”? It is a prototypical semantics because it is based on a representation of prototypical clusters, as just mentioned. Why is it a prototypical perceptual semantics? Notice that my primary illustrations of the probabilistic/geometric model are drawn from the field of image processing, such as the MNIST dataset of handwritten digits. If we can build a logical language on these foundations, we will have a plausible account of how human cognition could be grounded in human perception. Hence the title of my paper.

3. CAN WE LEARN A GROUNDED SEMANTICS WITHOUT A PERCEPTUAL GROUND?

The model that emerges from my ICAI’15 paper is a complex structure, with multiple layers. Each layer is a hybrid drawn from three areas of mathematics: Probability, Geometry, Logic. Moreover, each layer proceeds from probability *through* geometry to logic, according to the theory of differential similarity. Thus, each layer is subject to the following constraints, as viewed from the top down:

- The logic is constrained by the geometry.
- The geometric model is constrained by the probabilistic model, since the Riemannian dissimilarity metric depends on the probability measure.
- The probability measure is constrained by the distribution of sample data in the actual world.

It is the existence of these mutual constraints, I suggest, which makes it possible to learn the semantics of a complex knowledge representation language.

For a legal domain, the concepts at the top level would be legal categories, and the concepts at the intermediate levels would be various common sense categories: space, time, mass, action, permission, obligation, causation, purpose, intention, knowledge, belief, and so on. Indeed, this is basically the design of my Language for Legal Discourse (LLD) [13]. If we posit this edifice as a model of human cognition, there would be additional lower levels for human perception: visual, aural, haptic, etc. Presumably, this is how human beings learn the law! However, our machines cannot learn a legal semantics this way, starting at the perceptual level, because they only have access to legal texts. Is there a way around this problem? In fact, there are two reasons to think that a computational solution might be possible.

First, the theory of differential similarity is not acutely sensitive to the precise details of the representations used at the lower levels. As long as the input data is encoded in a high-dimensional Euclidean space and is approximately correct, according to the true perceptual semantics, the statistical resampling at the higher levels will tend to insure that the Riemannian geometry computed at those levels is also approximately correct.

Second, there is increasing evidence that the semantics of lexical items can be represented, approximately, as a vector in a high-dimensional vector space, using only the information available in the texts. There are now several methods to compute these *word embeddings*, ranging from highly non-linear mappings based on recurrent neural networks [20] to much simpler mappings based on word co-occurrence statistics [19], and they all yield essentially the same results: The semantic representations exhibit a surprising *linear* structure. The standard example is the following approximate equality:

\[ v(\text{"queen"}) \approx v(\text{"king"}) - v(\text{"man"}) + v(\text{"woman"}) \]

where \( v : \text{Words} \rightarrow \mathbb{R}^n \) is the computed mapping from the set of lexical items into a high-dimensional real-valued vector space. Thus, the dimensions of the embedding space
seem to be capturing the latent continuous semantic features of the lexicon.

The research strategy should now be obvious: We initialize our model with a word embedding computed from legal texts, and we try to learn the higher level concepts in a legal domain by applying the theory of differential similarity.

4. RELATED WORK

There are now quite a few commercial applications of Natural Language Processing and Machine Learning in legal domains. The explosive growth of e-Discovery is well-known and well-documented. See, e.g., [2]. The next growth area seems to be Contract Analysis, where the initial commercial entrants include: Counselytics,6 Kira Systems,7 KM Standards,8 RAVN Systems9 and Seal Software.10 Unfortunately, by the nature of their business model, it is difficult to determine precisely which technologies these companies are using, or developing. In the broad subfield of legal research and analysis, the most extensive press coverage has been devoted to ROSS Intelligence, which is applying IBM’s Watson technology to the law.11 However, at the time of this writing, the ROSS team has yet to release a product, or even to publish a functional specification of their system. By contrast, the IBM Watson group has published numerous useful and informative papers about the system that defeated two human opponents playing Jeopardy. See, e.g., [8], [7].

In the recent academic literature on artificial intelligence and law, there are several relevant papers: [11], [3], [25], [27], [22], [9], [10]. The paper by Grabmair, et al. [10], demonstrates that the semantic annotation of a set of vaccine injury cases can improve document retrieval and reranking over the baseline set by WestlawNext. The paper by Gaur, et al. [9], describes an interesting technique, using two operators called “Inverse Lambda” and “Generalization,” to learn the semantics of unknown words in a legal text from syntactically similar words with known meanings. For both projects, the challenge is to scale this work up to a more realistic corpus.

At a much larger scale, there are systems currently under development that construct enormous knowledge bases from information extracted from the Web, and then use these knowledge bases to improve subsequent information retrieval and question answering, also from the Web. Some representative examples include: Freebase [4], YAGO [26], NELL [21] and Knowledge Vault [6]. For a review of the techniques of Statistical Relational Learning that support most of these systems, see [23]. In some respects, our task is easier, since we can work with a focused collection of texts instead of the entire World Wide Web. But in other respects, our task is harder: We are trying to extract a conceptual model from our sources, not just a network of facts. This brings our challenge closer to the DARPA initiative on “Deep Exploration and Filtering of Text (DEFT)”.12

There is a growing literature on word embeddings, including several recent papers that try to explain why this technique works: [12], [24], [1]. The paper by Arora, et al. [1], proposes a “random walk on context spaces” as an explanation, which is interesting because it has some conceptual similarities to my model of “Brownian motion with a drift term” in [17].

5. PROSPECTUS

I have traced my proposed research strategy through three prior papers, written over a period of almost twenty years. The overall strategy calls for a form of semi-supervised learning, in which hand-coded extraction patterns for the information that we would like to represent in a structured casenote would be supplemented by the unsupervised learning of a legal semantics from a corpus of legal texts. This last step relies on the very theoretical model presented in my ICAIL’15 paper, which takes the form of a prototypical perceptual semantics. There is a mismatch in this last step, however, which I propose to overcome by the use of word embeddings. Applying this new technique, we should be able to work with linguistic data from the start.

This is obviously a speculative proposal, but it blends theory and practice in an interesting way. If the theory in my ICAIL’15 paper is correct, we should be able to build a successful practical application. Conversely, if the practical application succeeds, it will lend empirical support to the theory.

6. REFERENCES


http://counselytics.com/
http://kirasystems.com/
http://kmstandards.com/
http://www.ravn.co.uk/
http://www.seal-software.com/
http://www.rossintelligence.com/


Diachronic and Synchronic Analyses of Japanese Statutory Terminology

Case Study of the Gas Business Act and Electricity Business Act

Makoto Nakamura
Japan Legal Information Institute,
Graduate School of Law,
Nagoya University
Furo-Cho, Chikusa-ku, Nagoya
Aichi, 464-8601, Japan
mnakamura@nagoya-u.jp

Katsuhiko Toyama
Information Technology Center / Graduate School of Information Science,
Graduate School of Information Science,
Nagoya University
Furo-Cho, Chikusa-ku, Nagoya
Aichi, 464-8601, Japan
toyama@is.nagoya-u.ac.jp

ABSTRACT

In this paper, we report on our ongoing research on the development of a diachronic legal terminology, which deals with temporal changes in legal terms. Focusing on provisions for definitions in acts, we successfully extracted legal terms, their explanations, and their relations. Thus far, we have collected 27,737 terms and 36,698 relations. Picking up two acts, we analyzed their legal terms from diachronic and synchronic points of view. Our analysis showed that similarities between the two acts can be measured by their legal terms.

Keywords
legal text processing; legal terms; Japanese statutes; terminology

1. INTRODUCTION

In this paper, we report on our ongoing research on the development of a Japanese legal terminology. In general, important terms in statutes are explicitly defined prior to use. As such, we focus on the legal terms defined in a provision, each of which consists of a tuple of a legal term and its explanation. The provision for the definition of terms is typically placed in Article 2 in an act, following the act’s purpose. In other words, the legal terms in a provision are regarded as governing the whole act.

Legal statutes are not only established, but also often amended by changes in social situations. In some cases, legal terms are also revised, added, or deleted depending on the scale of the amendment. Therefore, an amendment to provisions for legal terms implies a drastic change of the entire act. The terminology for legal terms must deal with temporal changes that are dependent on amendment acts.

The study of the diachronic evolution of terminology can be regarded as a set. Such study deals with the narrative of the historical origins of a domain’s terminology and its actual historical evolution to the present state [2]. Applied to the legal domain, this is achievable with an exhaustive number of legal terms from all of the acts in the statutory corpora.

Thus far, we have constructed a diachronic legal terminology that consists of legal terms, their explanations, and their relations. Our approach to automatic extraction is based on a rule-based method [5]. Our purpose in this paper is to analyze our terminology from diachronic and synchronic viewpoints. This thorough study makes it possible to deal with synchronic similarities in legal terms to uncover the hidden relations among acts. For example, the submission of bills is often motivated by such changes in social situations as economic and political issues, wars, and natural disasters, during which multiple acts are simultaneously enacted. Although similar terms are often defined in different acts with explanations that resemble each other, identifying such relations is difficult without knowledge. The natural language processing technique makes it possible to calculate the similarity between terms and explanations. Therefore, a diachronic legal terminology provides legal scholars with a method of analyzing the dynamics of legal changes.

We organized our paper as follows: In Section 2, we introduce the diachronic legal terminology. In Section 3, we explain the target issues. In Section 4, we explain how to extract legal terms. In Section 5, we propose a method for analyzing acts with legal terms and, in Section 6, we evaluate the terminology. We present our conclusion in Section 7.

2. DIACHRONIC CHANGES IN LEGAL TERMS

In this section, we introduce diachronic changes in legal terms and, in Section 2.1, we explain these changes with examples. Section 2.2 shows actual changes in definitions.

2.1 Examples

We cite the Gas Business Act (Act No. 51 of 1954) as an example to explain diachronic changes in legal terms. As of September 2015, this act has been amended 36 times, six of which include the revision of terms and definitions in Article
Figures 1(a) to 1(c) show the diachronic changes in the terms at three time points:

1. At the new enactment, only two terms, “Gas Business” and “Gas Facilities,” were defined in the Gas Business Act (Act No. 51 of 1954), which came into effect as of April 1, 1954 (Figure 1(a)).

2. The term “Gas Business” was changed to “General Gas Utility Business,” becoming a hyponym of the newly defined term “Gas Business” with the newly added term “Community Gas Utility Business,” by the Act on the Partial Amendment of the Gas Business Act (Act No. 18 of 1970), which came into effect as of October 12, 1970. Note that, unlike language changes as a natural phenomenon, the sense of legal terms was forced to change on the enforcement date (Figure 1(b)).

3. Although no enforcement date has been determined yet for the Act on the Partial Amendment, etc., of the Electricity Business Act, etc. (Act No. 47 of 2015)\(^2\), the number of terms defined in the Gas Business Act will be increased to 16 (Figure 1(c)). In the period between (2) and (3), the terms “General Gas Utility Business,” “Community Gas Utility Business,” “Gas Pipeline Service Business,” “Large-Volume Gas Business,” “Large-Volume Supply,” “General Gas Utility,” “Community Gas Utility,” “Gas Pipeline Service Provider,” “Large-Volume Gas Supplier,” and “Wholesale Supply” were defined, but deleted later. In addition, the term “Intra-Area Wheeling Service” was replaced with “Wheeling Service.” These were basically eliminated by social selection.

2.2 Amendment of Legal Terms

Statutes are written in two types of languages: an object language for new enactments and metalinguage, which rewrites the description in object language, for revisions, rearrangements of provisions, and to repeal acts. While the former describes a law itself, the latter shows how to rewrite it with patch expressions. The amendment of a statute is realized by applying the latter to the former. This amendment method is called consolidation.

Figure 2 shows an excerpt from the acts dealing with the changes of the term “Gas Business” from Figure 1(a) to Figure 1(b). The revised act is shown in Figure 3. Note that we referred to the Japan Law Translation Database System\(^3\) for the English translation of these acts. When there is no translation for the acts or act titles in the website, we manually translated them using the database. The original Japanese versions of the acts are shown in Appendix A.

3. TARGET ISSUES


2. This act provides in Supplementary Provisions that Article 2 of this Act shall come into effect as of the date specified by a Cabinet Order that has not been promulgated yet. Article 39-2 will be deleted at the same time by this amendment.


Related statutes are defined as those that directly or indirectly refer to each other. Finding statutes that are directly related to a target statute is not difficult; simply list up all of the statutes that refer to and are referred to by the target statute.

Some related statutes share the same or similar terms with explanations that differ from each other. They are often amended simultaneously. In this case, the amendment statute would play the key role as a bibliographic coupling. In other words, it would be difficult to find such a relationship between statutes without explicit reference. By focusing on legal terms, we may be able to estimate the strength of connections between statutes. The Electricity Business Act (Act No. 170 of 1964) is related to the Gas Business Act (Act No. 51 of 1954).

We see a synchronic similarity in legal terms among related acts. In the same way, we may be able to find hidden related acts that are indirectly related, but not explicitly described. Therefore, we picked up these known two acts to illustrate linguistic analysis of legal terms.

4. LEGAL TERMINOLOGY

4.1 Extraction of Provisions for Definitions from Corpora

Some private companies provide advanced legal databases. The database of Dai-Ichi Hoki Co., Ltd.\(^4\) serves all of the acts and ordinances that are currently in effect, precedents, literature, and so on. It stores not only all 2,501 currently effective acts, but also their old versions at any time point. In addition, the history is listed with each article, from which we can designate any two time points to create a comparison table.

The problem with this database is that old acts are only available for inspection back about 15 years, which is too short to appreciate the dynamics of legal terms. Although we employ this database due to the lack of need for a consolidation process, this is a pilot version.

We searched all the provisions for the definitions of legal terms. Since each article in an act typically has an article caption, we extracted articles whose captions include a particular string denoting definitions with a regular expression (Figure 8(a)). Restricting our target articles to those having a particular article caption, we can extract them with high precision (100\%), and detect deletion of the articles by an amendment statute. The number of acts that include one or more articles for the definition of legal terms is 1,033 out of 2,501. If an act includes a number of articles for the definition of legal terms, we consider them separately. Therefore, we deal with 1,081 articles, 540 of which were revised during the last 15 years.

4.2 Extraction of Legal Terms

What are recognized as legal terms to be collected depends on the purpose [1, 4, 6]. In this paper, we define legal terms as those explicitly defined prior to use in a statute, each of which consists of a tuple of a legal term in the quotations and its explanation. They typically take the following forms:

1. An independent provision

2. An inserted statement in parentheses

\(^1\)http://www.japaneselawtranslation.go.jp/

\(^2\)This act provides in Supplementary Provisions that Article 2 of this Act shall come into effect as of the date specified by a Cabinet Order that has not been promulgated yet. Article 39-2 will be deleted at the same time by this amendment.

\(^3\)https://www.d1-law.com/
(a) Legal terms and relations in the Gas Business Act (Act No. 51 of 1954)

(b) Legal terms and relations as of enforcement of the Act on the Partial Amendment of the Gas Business Act (Act No. 18 of 1970)

(c) Legal terms and relations as of enforcement of the Act on the Partial Amendment, etc. of the Electricity Business Act, etc. (Act No. 47 of 2015)

Figure 1: Dynamics of definitions in the Gas Business Act

For (1), an article often consists of a number of paragraphs, each of which defines a legal term. They are described with boilerplate expressions including a legal term and its explanatory sentence, which can be extracted with a set of regular expression rules (Figure 9(f)). The underlined phrases in the upper box in Figure 2 match one of the rules. Some paragraphs include items for a list of either conditions for the term defined in the paragraph’s body or legal terms. A set of regular expression rules distinguishes them (Figure 8(b)). In the former case, the explanatory sentence includes all of the items. For the latter, a legal term and its definition can be extracted with a simpler set of regular expression rules (Figure 8(c)).

For (2), a defined term appears in parentheses following a phrase as its explanation in the main text. Abbreviations of terms are often defined in parentheses. An example is shown in Figure 3, where the term “Community Gas Utility Business” and the term in parentheses, “Specified Gas Generating Facility,” are defined as follows:

**Term:** Community Gas Utility Business

**Definition:** The business of generating gas at a simplified gas generating facility specified by a Cabinet Order.

We extracted the explanation of the latter, the underlined part in Figure 3, from the beginning of the definition to just before the beginning of the parentheses.

Although the definitions in the parentheses often appear in the main text regardless of the article’s content, we deal with those in the article for the definition of legal terms. This is because legal terms can have a relation that shares the term defined in the parentheses in their explanatory sentences.

We successfully extracted legal terms, their explanations, and their relations. We found 27,737 terms and 36,698 relations. The precision of the relations was 88.0%, which might be improved with additional regular expression rules [5]. In addition, taking repealed statutes into account, they are revised as 27,737 terms and 36,698 relations.

5. CLUSTERING ACTS

We analyzed a set of acts in effect as of a designated date. The following paragraphs explain how to (1) define features, (2) make a word vector, and (3) cluster acts. We repeat this procedure for each year (2001-2015).

Since Japanese words in a sentence are not separated by a space, we need morphological analysis, which separates words and attaches a part of speech tag to each morpheme. We used Mecab [3]. In addition, we registered all of the act titles and legal terms in the dictionary in the morphological analyzer in advance. Therefore, each is dealt with as a single word.

\[ \text{Note that the original statute does not include the underlines, which were added by the author.} \]
We used bayon\(^{a}\), which is a clustering tool for large-scale data sets. We employed repeated bisection clustering. We set an option to ‘-2 1’, which specifies the limit value of cluster partitions.

### 6. EVALUATION

Table 1 shows the chronological table of clusters for the Electricity Business Act and Gas Business Act. The column ‘#Acts’ denotes the numbers of the effective acts that have articles for definitions in that year. The number of clusters denoted by ‘#Clst’ changes due to the clustering option. The column ‘Clst ID’ lists the cluster ID to identify the cluster. The numerals in the column ‘Act IDs in the clusters’ show the cosine similarity between the two acts. Note that the clustering result is not fixed, but contains a random factor. In fact, the clustering ‘Clst ID’ might have no meaning, but we could see the change in the relation between the two acts. They belong to the same cluster for the years 2005, 2006, and 2014, at which the cosine similarity shows high values.

The provisions of the definitions of legal terms in the Gas Business Act were amended in 2004, while the Electricity Business Act was amended three times in 15 years: 2005, 2012, and 2014. Although the amendment of the Electricity Business Act in 2005 followed that of the Gas Business Act in 2004, they were implemented by the same amendment act. Therefore, it would be reasonable that they were amended for the same purpose and in the same way. In fact, they belonged to the same cluster after both were amended. Likewise, the Electricity Business Act’s amendment in 2014 might have been close to the Gas Business Act.

Other acts are amended every year, which affects the val-

\(^{a}\)https://github.com/fujimizu/bayon/
ues of word vectors with tf-idf weight. As a result, the cosine similarity between the two acts changes even when neither is amended.

Since the Act ID 9: the Hot Spring Act (Act No. 125 of 1948) sometimes appears in the same cluster as the Gas Business Act (Act No. 51 of 1954), we investigate the relation between these three acts, adding the Electricity Business Act (No. 170 of 1964). The more similar the acts from the viewpoint of the cosine similarity, the higher the probability that they are in the same cluster. Figure 4 shows the coupling rate for two out of the three acts every year by 10,000 trials, where the indices 1, 2, and 9 denote the Electricity Business Act, Gas Business Act, and Hot Spring Act, respectively.

The blue and red dashed lines denote the years that the articles for definitions in the Gas Business Act and the Electricity Business Act are amended, respectively. The orange line represents the coupling rate between the Electricity Business Act and the Gas Business Act.

Since only the Gas Business Act was drastically amended in 2004, the similarity between the two acts decreased. Instead, it returned in 2005 due to a similar amendment. Figure 5 illustrates the change of legal terms about the service provided by these two acts, which explains what happened in the two years. Although the Act on the Partial Amendment of the Electricity Business Act and the Gas Business Act (Act No. 92 of 2003) amended both acts, the Gas Business Act preceded the Electricity Business Act for one year with respect to the enforcement date. The Electricity Business Act had originally defined “Cross-Area Wheeling Service.” The amendment act added “Intra-Area Wheeling Service” and “Wheeling Service” as a broader concept of these two terms. On the other hand, the Gas Business Act had originally defined “Intra-Area Wheeling Service.” Since the amendment act assimilated the name of the service with the Electricity Business Act, the term “Wheeling Service” was defined not as a broader concept of but as a replacement with the term “Intra-Area Wheeling Service.” As a result, there is no common term between the two acts about the service as of April 1, 2004. This is one of the reasons why the coupling rate suddenly dropped down in this year. As of April 1, 2015, the acts shared a common term again, which got the coupling rate back. Therefore, we could say these two acts are connected, and amendment acts occur close to each other except the case in 2004. As a result, these two acts retain a high degree of similarity in 2012 and, finally, the coupling rate comes to a peak in 2014. We demonstrated the change in relation by the enforcement of amendment acts.

The blue line shows the coupling rate between the Electricity Act and the Hot Spring Act, which have little relation. The green line denotes the coupling rate between the Gas Business Act and the Hot Spring Act, which seems to complement the orange line. We compared the legal terms between these acts. Unfortunately, they seem to have little relation, but we did find the word ‘gas’ in an explanation. This is because the tf-idf value was blown up. Although this does not indicate a hidden relation to the Gas Business Act, we could suggest a method for finding it.

7. CONCLUSION

In this paper, we extracted synchronic similarity in the legal terms among related acts. Related statutes are defined as ones that directly or indirectly refer to each other. Finding directly related statutes of a target is not difficult; simply list all of the statutes that refer to and are referred to by the target. On the other hand, finding indirect relations between statutes without explicit references is difficult. We introduced a clustering method for statutes represented by legal terms at a particular date. Although we picked up two acts, further analysis will reveal relations among legal terms.

8. ACKNOWLEDGMENTS

This research was partly supported by the Japan Society for the Promotion of Science KAKENHI Grant-in-Aid for Scientific Research (S) No. 23220005, (A) No. 26240050, and (C) No. 15K00201.

9. REFERENCES

Table 1: Chronological table of clusters for the (1) Electricity Business Act and (2) Gas Business Act

<table>
<thead>
<tr>
<th>Date</th>
<th>Acts</th>
<th># Clst</th>
<th>ID Act IDs in the clusters of 1&amp;2</th>
<th>Sim</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>769</td>
<td>156</td>
<td>1, 3</td>
<td>0.199</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>139</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>789</td>
<td>162</td>
<td>1, 3, 5</td>
<td>0.201</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>823</td>
<td>172</td>
<td>1, 3</td>
<td>0.200</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4, 118</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>851</td>
<td>176</td>
<td>2, 6, 7</td>
<td>0.182</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>881</td>
<td>179</td>
<td>1, 3, 5, 2, 4, 13</td>
<td>0.208</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>143</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>902</td>
<td>194</td>
<td>1, 3, 2, 5, 4, 6</td>
<td>0.206</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>179</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>919</td>
<td>183</td>
<td>1, 3, 5, 13</td>
<td>0.201</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>948</td>
<td>191</td>
<td>2, 4, 5, 6</td>
<td>0.200</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>53, 172</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>966</td>
<td>189</td>
<td>2, 4, 5, 6, 7</td>
<td>0.198</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>119, 187</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>988</td>
<td>195</td>
<td>2, 4, 5, 6, 7, 15</td>
<td>0.199</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>24, 161</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>999</td>
<td>200</td>
<td>1, 3</td>
<td>0.200</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6, 18, 19, 20</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>1,021</td>
<td>211</td>
<td>1, 3</td>
<td>0.189</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8, 161</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>1,034</td>
<td>202</td>
<td>2, 4, 10</td>
<td>0.190</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>106, 169</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>1,051</td>
<td>213</td>
<td>2, 4, 10</td>
<td>0.213</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>124, 232</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>1,074</td>
<td>217</td>
<td>2, 4, 10</td>
<td>0.210</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>62, 232</td>
<td></td>
</tr>
</tbody>
</table>

**APPENDIX**

A. THE ORIGINAL STATUTES

Figures 6 and 7 show the original version of Figures 2 and 3, respectively. Also, Table 3 shows the original Japanese act titles of Table 2.

B. REGULAR EXPRESSION RULES FOR EXTRACTION AND IDENTIFICATION

Table 2: List of Acts

<table>
<thead>
<tr>
<th>ID</th>
<th>Act Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Electricity Business Act</strong> (Act No. 170 of 1964)</td>
</tr>
<tr>
<td>2</td>
<td><strong>Gas Business Act</strong> (Act No. 51 of 1954)</td>
</tr>
<tr>
<td>3</td>
<td>Act on Promotion of Power-Resources Development Tax (Act No. 79 of 1974)</td>
</tr>
<tr>
<td>4</td>
<td>Heat Supply Business Act (Act No. 88 of 1972)</td>
</tr>
<tr>
<td>5</td>
<td>Industrial Water Supply Business Act (Act No. 84 of 1958)</td>
</tr>
<tr>
<td>6</td>
<td>Act on the Supervision of Construction Project to Install Specified Gas Appliances (Act No. 33 of 1979)</td>
</tr>
<tr>
<td>8</td>
<td>Act on the Prevention of Disasters in Petroleum Industrial Complexes and Other Petroleum Facilities (Act No. 84 of 1975)</td>
</tr>
<tr>
<td>9</td>
<td>Hot Spring Act (Act No. 125 of 1948)</td>
</tr>
<tr>
<td>10</td>
<td>High Pressure Gas Control Act (Act No. 204 of 1951)</td>
</tr>
<tr>
<td>11</td>
<td>Gravel Gathering Act (Act No. 74 of 1968)</td>
</tr>
<tr>
<td>12</td>
<td>Act on Water Washing Coal Business (Act No. 134 of 1958)</td>
</tr>
<tr>
<td>13</td>
<td>Waterworks Act (Act No. 177 of 1957)</td>
</tr>
<tr>
<td>14</td>
<td>Act on Rational Use and Proper Management of Fluorocarbons (Act No. 64 of 2001)</td>
</tr>
<tr>
<td>15</td>
<td>Act on Development of Areas Around Electric Facilities for Electricity Generation (Act No. 78 of 1974)</td>
</tr>
<tr>
<td>16</td>
<td>Act on Special Measures Concerning the Promotion of Establishment Area for the Nuclear Power Facilities etc. (Act No. 148 of 2000)</td>
</tr>
<tr>
<td>17</td>
<td>Specified Multipurpose Dam Act (Act No. 35 of 1957)</td>
</tr>
<tr>
<td>18</td>
<td>Electrician Act (Act No. 139 of 1960)</td>
</tr>
<tr>
<td>19</td>
<td>Act on Regulation of Electrical Contracting Business (Act No. 96 of 1970)</td>
</tr>
<tr>
<td>20</td>
<td>Basic Act on the Advancement of Utilizing Geospatial Information (Act No. 63 of 2007)</td>
</tr>
<tr>
<td>22</td>
<td>Special Measures Act on Support for Independence of Homeless People (Act No. 105 of 2002)</td>
</tr>
<tr>
<td>23</td>
<td>Act on Special Measures Concerning Procurement of Electricity from Renewable Energy Sources by Electricity Utilities (Act No. 108 of 2011)</td>
</tr>
<tr>
<td>24</td>
<td>Act on the Promotion of Use of Nonfossil Energy Sources and Effective Use of Fossil Energy Materials by Energy Suppliers (Act No. 72 of 2009)</td>
</tr>
<tr>
<td>25</td>
<td>Act on Special Measures for Prevention of Mining Pollution by Metal Mining etc. (Act No. 26 of 1973)</td>
</tr>
<tr>
<td>26</td>
<td>Basic Act on the Advancement of Utilizing Biomass (Act No. 52 of 2009)</td>
</tr>
</tbody>
</table>
ガス事業法 (昭和 29 年法律第 51 号)

(定義)
第二条 この法律において、「ガス事業」とは、一般の需用に応じ導管によりガスを供給する事業をいう。

ガス事業法の一部を改正する法律

(昭和 45 年法律第 18 号)

第二条 第一項後「ガス事業」を「一般ガス事業」に、「需用」を「需用」に改め、「供給する事業の下に「(第三項
に規定するガス発生設備においてガスを発生させ、導管に
よりこれを供給するものを除く。)」を加え、(*snip*) 同
条第一項の次に次の五項を加える。

(*snip*)

3 この法律において「簡易ガス事業」とは、一般の需要
に応じ、政策で定める簡易なガス発生設備 (以下「特定ガ
ス発生設備」という。) においてガスを発生させ、導管に
よりこれを供給する事業であって、一の地域内におけるガ
スの供給地点の数が七十以上のものをいう。

(*snip*)

5 この法律において「ガス事業」とは、一般ガス事業及
び簡易ガス事業をいう。

---

Figure 6: The original version of Figure 2: Excerpt from the Gas Business Act (Act No. 51 of 1954) and the Act on Partial Amendment of the Gas Business Act (Act No. 18 of 1970)

The regular expression rules used for extraction and identification are shown in Figures 8 and 9. Note that they are written in Japanese. There is no English translation, because they include fragments of Japanese words.

<table>
<thead>
<tr>
<th>Regular Expression Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) to extract articles for legal term definitions:</td>
</tr>
</tbody>
</table>
| "*\{ ('\*')* )\}*定義 { ' ' } 等及び'
| (b) to identify presence of itemization for definitions: |
| *語句、+当該各号*、 |
| *規定の解釈、次の定義*、 |
| *」とは、次に掲げるものとする。
| (c) to extract definitions from an explanatory sentence: |
| "*と、(.+ )をいう。
| "*とは、(.+ )をいい、.+ (も)とする。
| (d) to identify the legal term as a hypernym of terms in its explanatory sentence: |
| "( ['\*'] )* (又は|若しくは|及び|並びに)+ ['\*'] )* ' |
| "( ['\*'] )* (又は|若しくは|及び|並びに)+ ['\*'] )* ' |
| (e) to identify the legal term as a hyponym of the term in its explanatory sentence: |
| "( ['\*'] )* で (あって)|、" |

Figure 8: Regular expression rules (a)-(e)
Table 3: The original version of Table 2: List of Acts

<table>
<thead>
<tr>
<th>ID</th>
<th>Act Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>電気事業法 (昭和39年法律第170号)</td>
</tr>
<tr>
<td>2</td>
<td>ガス事業法 (昭和29年法律第51号)</td>
</tr>
<tr>
<td>3</td>
<td>電源開発奨励税法 (昭和49年法律第79号)</td>
</tr>
<tr>
<td>4</td>
<td>熱供給事業法 (昭和47年法律第88号)</td>
</tr>
<tr>
<td>5</td>
<td>工業用水道事業法 (昭和33年法律第84号)</td>
</tr>
<tr>
<td>6</td>
<td>特定ガス消費機器の設置工事の監督に関する法律 (昭和54年法律第33号)</td>
</tr>
<tr>
<td>7</td>
<td>濃化石油ガスの保安の確保及び取引の適正化に関する法律 (昭和42年法律第149号)</td>
</tr>
<tr>
<td>8</td>
<td>石油コンピュータ等災害防止法 (昭和50年法律第84号)</td>
</tr>
<tr>
<td>9</td>
<td>温泉法 (昭和23年法律第125号)</td>
</tr>
<tr>
<td>10</td>
<td>高圧ガス取締法 (昭和26年法律第204号)</td>
</tr>
<tr>
<td>11</td>
<td>原子力発電施設等立地地域の再開に関する特別措置法 (平成12年法律第148号)</td>
</tr>
<tr>
<td>12</td>
<td>電気工事土法 (昭和43年法律第74号)</td>
</tr>
<tr>
<td>13</td>
<td>水道法 (昭和32年法律第177号)</td>
</tr>
<tr>
<td>14</td>
<td>フロン類の使用の合理化及び管理の適正化に関する法律 (平成13年法律第64号)</td>
</tr>
<tr>
<td>15</td>
<td>高圧線路施設用地等整備法 (昭和49年法律第78号)</td>
</tr>
<tr>
<td>16</td>
<td>原子力発電施設等立地地域の再開に関する特別措置法 (平成12年法律第148号)</td>
</tr>
<tr>
<td>17</td>
<td>特定多目的ダム法 (昭和32年法律第35号)</td>
</tr>
<tr>
<td>18</td>
<td>電気工事土法 (昭和43年法律第74号)</td>
</tr>
<tr>
<td>19</td>
<td>電気工事の業務の適正化に関する法律 (昭和45年法律第96号)</td>
</tr>
<tr>
<td>20</td>
<td>地理空間情報活用推進基本法 (平成19年法律第63号)</td>
</tr>
<tr>
<td>21</td>
<td>コンテナに関する運送条約及び国際道路運送手帳による担保の下で行う貨物の国際運送に関する運送条約（TIR条約）の実施に伴う関税等の特例に関する法律 (昭和46年法律第65号)</td>
</tr>
<tr>
<td>22</td>
<td>ホームレスの自立の支援等に関する特別措置法 (平成14年法律第105号)</td>
</tr>
<tr>
<td>23</td>
<td>電気事業者による再生可能エネルギー電気の調達に関する特別措置法 (平成23年法律第108号)</td>
</tr>
<tr>
<td>24</td>
<td>エネルギーサイクル事業者による非化石エネルギー源の利用及び化石エネルギー原料の有効利用の促進に関する法律 (平成21年法律第72号)</td>
</tr>
<tr>
<td>25</td>
<td>金属鉱物等鉱物資源特別措置法 (昭和48年法律第26号)</td>
</tr>
<tr>
<td>26</td>
<td>バイオマス活用推進基本法 (平成21年法律第52号)</td>
</tr>
</tbody>
</table>

Figure 9: Regular expression rules (f)
ABSTRACT

Citations, as in when a certain statute is being cited in another statute, differ in meaning, and we aim to annotate each edge with a semantic label that expresses this meaning or purpose. Our efforts involve defining, annotating and automatically assigning each citation edge with a specific semantic label. In this paper we define a gold set of labels that cover a vast majority of citation types that appear in the United States Code (US Code) but still specific enough to meaningfully group each citation. We proposed a Linear-Chain CRF based model to extract the useful features needed to label each citation. The extracted features were then mapped to a vector space using a word embedding technique and we used clustering methods to group the citations to their corresponding labels. This paper analyzes the content and structure of the US Code, but most of the techniques used can be easily generalized to other legal documents. It is worth mentioning that during this process we also collected a human labeled data set of the US Code that can be very useful for future research.

Keywords

US Code; Legal citation graph; Automatic citation analysis; Conditional Random Fields; K-means clustering

1. INTRODUCTION

New regulations and laws in the United States are legislated or existing ones are evolved through a complex legal cycle. This system involves numerous organizations, parties, and individuals. Individual legal rules seldom exist in isolation, but instead typically occur as components of broader statutory, regulatory, and common-law frameworks consisting of numerous interconnected rules, regulations, and rulings. The complexity of these frameworks impedes comprehension and compliance by government agencies, businesses, and citizens and makes amending legislation laborious and error-prone for regulatory agencies and legislative drafters.

Citation networks are a promising recent approach to improving the intelligibility of complex rule frameworks. In a citation network, rules are represented by nodes and citations are represented by edges [16, 20]. Citation networks can often permit a complex regulatory framework to be comprehended at a glance. Techniques for automatically representing and displaying citation networks is an active area of research.

Computer assisted and automatic systems have been and are growing rapidly in every field. The legal domain is also no exception to this trend [11, 18, 7]. Specially there has been extensive research in designing programs and intelligent software that can address the challenging and expensive task of information extraction from general text. Information extraction is of special importance from a legal perspective since almost all the information in this domain is collected in natural human language. This techniques can be utilized to aid in the automation of creating and displaying meaningful citation networks.

An important aspect of citation-network use is that, generally, only a small subgraph is relevant for any particular application or task. Indeed, visualizations of entire citation networks are generally incomprehensible “hairballs.”

The subgraph of a citation network relevant to a particular task depends both on the attributes of the nodes (i.e., rules) and edges (i.e., citations). For example, a subgraph relevant to public health emergencies would include both nodes defining the powers and duties of agents (e.g., doctors, epidemiologists, coroners) and citations indicating the relative authority among these agents. In general, the portion of a statutory framework relevant to a given task consists of the subgraph induced by nodes and edges having a semantic relationship to the task.

While nodes relevant to a given task (e.g., UAV licensing) can typically be found using information-retrieval techniques, such as term-vector or topic similarity, identification of relevant edges, is much less well understood. Various researchers have proposed different taxonomies of edges in citation graphs [9, 13, 5], but there is not yet a consensus on the most useful set of edge types. Moreover, there has been little progress in automatically applying semantic labels to citations edges, which is essential for large-scale citation network visualization and analysis tools.
This paper first reviews the related work in Section 2. Followed by precisely describing our research problem in Section 3 and the proposed automated system to tackle this problem in Section 4. In Section 5, we describe the data set used to evaluate our system as well as the proposed gold standard label set used for labeling the citation graph. And finally we conclude the paper in Section 6 by a summary of the results and a plan for future research on this study.

2. RELATED WORK

There has been various previous research projects addressing the detection, resolution and labeling of citations in the legal domain. But to our knowledge there has not been any prior work on a systematic approach to automatically detecting and labeling of cross references with a detailed semantic label set.

In [9] M. Hamdaqa et al., lay the grounds and propose techniques for analysis of citation networks. One of their key contributions is to review methods of automatically detecting the presence of citation in legal texts. They note that even this simple sounding task alone, is not easy. Although there have been numerous standards and books devoted to proper citation, in many cases the citation text does not follow the correct format and style thus making it hard for automatic extraction of citations from legal documents. They also propose a categorization schema for citations which groups a citation as either an Assertion or an Amendment, which they elaborate in their second paper [10], we will discuss more on this later in this section.

In a more recent work [1], Madedjouma et al., study and investigate the natural language patterns used in cross reference expressions to automatically detect and link a citation to its target. One of their main contributions is in the detection of complicated cross references that are written in natural language. But, unlike us, they do not approach the task of labeling the citations and limit their work on resolving the citation links.

In [13] Maxwell et al., aim to develop a system to help software companies comply with all the regulations. They study the taxonomy of legal-cross references in the acts related to healthcare and financial information systems. They claim to be the first to identify concrete examples of conflicting compliance requirements due to cross-references in legal texts. They analyze different patterns of cross-references that occur in these case studies to obtain seven cross-reference types: constraint, exception, definition, unrelated, incorrect, general, and prioritization and use grounded theory (the discovery of theory from data [8]) to conjecture that this set of labels are generalizable to other legal domains. Their definitions of constraint, exception, definition and prioritization are very similar to our "Limitation", "Exception", "Definition", "Delegation of Authority". While their unrelated label does not apply to general purpose citation labeling and only points out the cross-references that are not related to laws governing software systems. Although we have a more detailed set of labels, we do not have a label that corresponds to incorrect since we do not look at the cited text and thus we are not able to determine if the citation is indeed correctly citing the desired section of the cited.

T.D. Breaux et al., in [5] design and propose "Frame-Based Requirements Analysis Method (FBRAM)". FBRAM is a software which helps to generate context-free markup language. Their system facilitates the creation of a model used to systematically acquire a semi-formal representation of requirements from legal texts. The set of labels used in this work are Exclusion, Fact, Definition, Permission, Obligation, Refraining. Their approach in this paper is quite different from ours, since they group/label the text and requirements in the cited text while we are interested in the bigger picture of why the statute is being cited. We must also note that FBRAM is utterly relying on a human analyst and mainly helps only if an analyst manually annotates the whole regulatory document first while we use artificial intelligence and machine learning methods to label cross-references.

In a sequel to their first paper [9], M. Hamdaqa et al. explore the relationships between the citing and the cited law in [10]. Their work is the closest approach to ours in the sense that they also offer an automated system that classifies each citation based on its semantic role in the context. They give a list of advantages in why one want to explore the relationships among provisions created through citations from one to the other. In short: it is useful in understanding the impact of changes in a law and those depending on it; checking consistencies/conflicts between multiple regulations; eases navigation through laws and their dependencies. They also propose grouping of each edge into Assertions (Definition, Specification, Compliance) and three subtypes of Amendments. They claim that using the verb that is directly related to the citation, one can label the citation into one of the two main groups but do not talk about the possibility of grouping them to the smaller subgroups nor they give numerical evaluations of the accuracy of their approach. In contrast we label each citation into a more refined set and also provide experimental results.

3. PROBLEM STATEMENT

As described in the previous sections, dealing with citations in the legal documents is an important task. In this paper we propose a system that can label each cross-reference according to a predefined set of labels. For the purposes of this paper we only discuss the US Code and its underlying citation graph, but in general our approach can be modified to apply to any other legal citation graph.

A citation graph refers to a graphical representation of all the cross-references in a document to other documents or parts of itself. Nodes, or vertices, in a citation graph are representing the section that is being cited or is citing another section. Edges in this graph are directed and if part of statute A is citing a part in statute B, there is an edge from A to B.

In this paper we introduce an automated semantic labeling of edges in a citation graph. We label/group the edges into a set of predefined labels that classify each edge based on their reason for being cited. For example, in:

subsection (a) of this section shall not apply to that portion of the employee’s accrued benefit to which the requirements of section 409(h) of title 26 apply

The cited statute, section 409(h) of title 26, imposes a limitation to where the obligations of the citing text would apply.

In the next section we will provide a descriptive summary of each part of the overall system.
4. THE AUTOMATED SYSTEM

As we stated in the previous sections, the main focus of this work is to build a system that can automatically label the edges in a citation graph with a predefined set of labels, each of which represents a possible relationship between the citing provision and the cited provision, that is, the purpose for the citation. The first step towards this goal is to be able to automatically detect the presence and span of each citation in the document. We will next describe our citation extraction method.

4.1 Extracting the Citation Text

The first step towards building this system is to be able to identify a citation. Cross-references in the legal domain mostly follow standards and predefined templates. The Bluebook [3] or the newer Citation Manual from US Association of Legal Writing Directors (ALWD) [17] are among the manuals that contain rules for proper citing of legal texts. But as previously mentioned these rules are not always followed.

To extract the citations from a document (e.g., the US Code), we used a complex regex pattern-matching schema that attempts to locate and identify a variety of known formats for citations. The result is the extraction of a number of known corpora types, which then go through an additional processing schema developed to split each extraction - which can potentially include multiple references to the same or different corpora, such as "26 USC sections 1, 2, and 3 ..." or "28 USC 121 and 10 CFR" - into individual elements and then re-combine them according to basic citation rules, so that it would produce the following: "26 USC 1", "26 USC 2", "26 USC 3", "28 USC 121" and "10 CFR" as 5 separate references.

4.2 Feature Extraction

A key idea in this method is our novel feature selection. We find a section of the text related to the citation, the predicate, and use this as the main feature in our classification. The predicate of a citation to be that portion of the text immediately preceding the citation that expresses the citation’s meaning.

During the annotation process along with collecting a labeled set of citations we also asked each annotator to tag the span of the corresponding "predicate", which we will talk about in more details in section 5.3. For the purposes of this work, we define the predicate as:

1. The full span of words, that
2. Directly expresses the relationship of the cited provision to something in the current section, and
3. That would make sense if applied to any other provision, i.e., contains nothing specific to the subject matter of the particular section (e.g., funds, exemption), and
4. That expresses as much of the semantics (meaning and purpose) of the relationship as possible without violating 1-3.

For example, in:

... all provisions excluded from this chapter under Section 42 U.S.C 1879 ...

the word under is not the full possible span that still satisfies (2)-(4), thus violating criterion (1). The phrase provisions excluded from this chapter under includes provisions, which is not a relationship but is instead the thing that the citation applies to, violating criteria (2) and (3). However, excluded from this chapter under satisfies all 4 criteria.

To automatically extract the predicate we designed and trained a linear-chain Conditional Random Field (CRF) on our collected annotated data. The correlated sequential structure of the words in a predicate can be well captured with this type of graphical models, which our experimental results in section 5.4 demonstrate too.

4.3 Classification

One of our main contributions is the automatic process of labeling citations in a legal citation graph. To achieve this goal we utilize an unsupervised learning algorithm to cluster the citations based on a simple word embedding of the extracted predicates.

More precisely we first train a shallow two layered neural network on a large corpus of English text extracted from wikipedia and fine tuned it by another round of training on the whole corpus of US Code. This approach is a well known method for representing words as vectors in a high dimensional space of real numbers first introduced by Tomas Mikolov et al. in [15]. We then use these vectors as the underlying representation of words in the predicate and cluster them using k-means. Subsequently each citation is labeled based on the cluster representing it. More detailed explanation and experimental evaluations are presented in Section 5.
In summary the complete system enables automatic labeling of the citations in a legal document. After the legal document is given to the system input, in step one it detects all the citations present in the document using the methods described in Section 4.1. It then automatically extracts what we call the predicate which contains information about the type of the at hand citation, this step was described in Sections 4.2 and 5.4. In the next step it utilizes machine learning techniques described in Sections 4.3 and 5.5 to assign to the citation, an appropriate label. The final labeled graph is then illustrated using our graphical user interface where each edge type is colored according to its type. Figure 1 shows a diagram of the complete system.

5. EVALUATION

In this section we evaluate the proposed gold standard label set used to capture the purpose of the citations, Annotated Dataset, CRF model and final Clustering Algorithm. We first briefly acquaint the reader with the dataset used, i.e. the US Code.

5.1 The Dataset

The dataset used to demonstrate the use of our system is the US code, which is a consolidation of the general and permanent laws of the United States. There are in total over 130,000 different citations in the US Code. The collection of US Code used was taken from the online repository of Legal Information Institute of Cornell Law School [6]. There are over 29000 distinct citations to statutes in the US Code, Code of Federal Regulations and other sources. These laws cite 26417 distinct US codes with the US law code "42 USC 1395x" being cited the highest.

Next we introduce the set of labels used in the semantic labeling of the citations.

5.2 Designing the Golden Labels

We inspected many random provisions found in the US Code and proposed a primary set of labels that could capture the relations found there in. This labels along with a set of unlabeled citations from the US Code was then annotated by a group of expert human annotators.

After analysing the results, we merged the labels that were too close to be separated and caused confusion; also expanded the labels by adding new labels found to be necessary. Integrating the feedback we got from the first round of annotations we updated the labels. We believe that the purpose of each citation can be effectively captured with this set of 9 labels.

- **Legal Basis**: A relationship between a program/entity and the statute that is its legal basis.
- **Authority**: A relationship under which one party is permitted to direct another party to perform an action.
- **Definition**: A citation that directly defines the subject, brings a definition for a term used in the rule.
- **Example or illustrations**: A citation to a rule that is used to introduce something chosen as a typical case or is defining the subject by illustrating/describing it.
- **Exception**: A link between a rule and a set of circumstances where that rule doesn’t apply.
- **Criterion**: A link from a conclusion to the "standard/criterion", but not how (not the procedure), of reaching that.
- **Limitation**: A relationship between a description and a restriction on that.
- **Procedure**: A link from an activity to a description of how that activity should be performed
- **Amended by/Amendment to**: A relationship between two versions of the rule.

As we discuss in Section 5.3, the final round of annotations by the human experts confirmed the validity of this labels. The result is a label set long enough to cover almost all of the citations and also short enough for practical use.

5.3 Annotation Process

To apply machine learning paradigms for labeling citations and also test the coverage of the gold standard label set described in the earlier section, we need to have a set of data manually annotated. Manual annotations lead to semantic problems similar to ones discussed in [5]. A crowd sourced option like Amazon Mechanical Turk, as mentioned in [2] can be a good medium for the manual annotation process. The problem with crowd sourcing is the absence of critical legal expertise with the annotators, which impacta their judging abilities for a domain specific task. The manual annotators could experience problems like Logical Ambiguity, which would need legal expertise to be resolved. To mitigate these issues associated, the manual annotator group for the project comprised of 7 Graduate law students, with the guidance of a team of legal experts.

The experiment was designed to run in two stages. The first stage comprised of a set of 200 randomly selected citations that were distributed to the annotators. The first set of annotations helped us expand and modify the gold standard to include for the citations that were deemed to be included in newer labels by the manual annotators. The second round then generated the training samples for our machine learning paradigms to label the citation. The second round also validated the gold standard as the manual annotators did not find a need to expand the labels set to accommodate for any citation. The second round of annotations produced 394 labeled citations that served as the basis for the the Clustering algorithms explained in the following sections.

Out of the 394 citations that were manually annotated, only one was found to need a new label not in our label set. This confirms that our label set covers the citations in the US Code very well.

5.4 Predicate Extraction

To find the proper predicate in the context of the citing provision, we used a linear chain CRF. Conditional Random Fields are probabilistic graphical models that very well capture the sequential property present in words in natural language [12]. A detailed description of CRFs can be found in [12, 19].

To create the features, we manually looked at samples of the raw text and considered the properties of the predicate. First, the predicate is almost always preceding the citation. Second, the predicates usually have a certain part of speech (POS) role. For example:
• Preposition-Verb-Preposition The term "commodity broker" means futures ... or commodity options dealer, as defined in section 348d., or

• Preposition Notwithstanding subsections (a), (b), and (c) of this section and paragraph (2) of this subsection, the court shall disallow any claim for reimbursement or . . . , or

• Preposition-Noun-Preposition Without regard to the the Richard B. Russell National School Lunch Act (4 U.S.C. 1751 et seq.), or etc.

Third, specific words such as under, defined, amended, tend to appear more in predicate span.

Trying to keep the features as simple as possible and keeping these properties in mind, we defined the following set of features for each token. Also, we replace the whole span of the target citation text, for which we intend to find the predicate, with a unique character sequence not present in the rest of the corpus, i.e., CIGITE. This will make it easier for the CRF to recognize the the citation and work with it as single word. To mark the span of each predicate we used the standard Begin/In/Out (BIO) encoding for tagging the predicate chunks and the other tokens.

Exact word features We used the exact lowercase token of each word and its neighboring words (before and after it) as three features for each token. We must note that this and other multi-valued categorical variables were binarized for use in the model.

Is digit feature We used a boolean feature to determine if a token is a digit or not.

Part of speech features Based on the lexical tags produced by NLTK [4], each word and its neighboring words were assigned with their corresponding POS tags. In addition to that we used the first two and the last two letters of the tag as additional features for the word and its neighbors. This helps when NLTK produces refined POS, for example NNP and NN might have to be treated the same in detecting the predicates.

Distance to citation features We used 5 boolean features determining the relative position of the word to the target citation. $f_1 = 1$ if the word appears after the citation. $f_2 = 1$ if there are no tokens between the word and the citation. $f_3 = 1$ if there is exactly one token between the word and citation. $f_4 = 1$ if there are more than two words in between. $f_5 = 1$ if there are more than four words in between.

Miscellaneous features Other features used were to determine if the word was at the beginning of a sentence, end of a sentence or if the token is a punctuation.

To evaluate the performance of this model, we applied the system to a dataset of 1000 citations and their corresponding predicates\footnote{This dataset was also obtained during the annotation process, but lacked a semantic label for the citations.}. We performed a 10-fold cross validation and presented the performance results in Table 1.

*Following BIO encoding the beginning of a predicate is tagged with B_PRD, any other word in the predicate span is tagged with L_PRD and any word that is not a part of the predicate is tagged with O.

5.5 Clustering Accuracy

As mentioned before, after extracting each citation’s predicate we used word2vec [15, 14] to represent each word in the predicate as a vector in a 300 dimensional space. To further simplify the clustering, we correspond each predicate with the average of the vectors representing each of the words in that predicate. Although this averaging results in a loss of information, but due to the properties in the embedding method we used most the meaning in the predicate is still preserved.

To cluster the data we use k-means classification and cluster the whole US Code using 15 cluster centers. Note that since we have a relatively large number of labels and there is no guarantee that each form exactly one cluster in the projected space. For this reasons, we use more cluster centers to capture the spread as much as possible. This might slightly over-fit or even decrease the accuracy, but its effects are negligible compared to the relatively large dataset and number of labels.

To evaluate the performance of our clustering algorithm, we use the annotated dataset obtained from the human expert annotators. Each cluster is labeled according to the label of the closest point to the center of the cluster. We present the classification accuracy and the confusion matrix in Table 2.

As mentioned before, after extracting each citation’s predicate we used word2vec [15, 14] to represent each word in the predicate as a vector in a 300 dimensional space. To further simplify the clustering, we correspond each predicate with the average of the vectors representing each of the words in that predicate. Although this averaging results in a loss of information, but due to the properties in the embedding method we used most the meaning in the predicate is still preserved.

To evaluate the performance of our clustering algorithm, we use the annotated dataset obtained from the human expert annotators. Each cluster is labeled according to the label of the closest point to the center of the cluster. We present the classification accuracy and the confusion matrix in Table 2.

Among the 394 citations that were manually annotated 1 needed a label that was not in the gold standard, we denote it with New Label Necessary (NL) in the table.

6. CONCLUSION AND FUTURE WORK

We presented an automated system that automatically determines the purpose behind a citation. This enables lawyers

Table 1: Predicate extraction performance

<table>
<thead>
<tr>
<th></th>
<th>Prec.</th>
<th>Recall</th>
<th>F1</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>B_PRD*</td>
<td>0.91</td>
<td>0.84</td>
<td>0.875</td>
<td>100</td>
</tr>
<tr>
<td>L_PRD</td>
<td>0.93</td>
<td>0.88</td>
<td>0.898</td>
<td>119</td>
</tr>
<tr>
<td>O</td>
<td>0.99</td>
<td>0.99</td>
<td>0.998</td>
<td>6518</td>
</tr>
</tbody>
</table>

Table 2: Empirical Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Am.</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Au.</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cr.</td>
<td>0</td>
<td>0</td>
<td>47</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>29</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>De.</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>44</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Il.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ex.</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Le.</td>
<td>2</td>
<td>0</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>7</td>
<td>98</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Li.</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NL.</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pr.</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
and policymakers better analyse the relation between different laws or users to find the necessary regulations much easier.

Our system has three main part. We first automatically extract the citations from the document, then find an informative expression from the text related to that citation which we call it as the predicate. Using Natural Language Processing (NLP) and machine learning techniques we then label the citation into one of the predefined set of citation types.

Our contributions in this paper are three-fold. We propose a gold standard label set that almost all the citations in the legal domain (specially laws and regulations) can be categorized according it and verified its coverage in manual experiment by a group of experts. We also produced a dataset of 394 annotated citations from the US code that can be used for future research on this topic. Finally we built a fully automated system for semantic labeling of the edges over a legal citation graph.

In future work we plan to have a more in depth analysis of the results from annotation process and the accuracy of a human expert. We further plan to use advanced machine learning techniques to increase the accuracy of our system by using the whole context related to the citation.

7. ACKNOWLEDGEMENTS

The authors would like to thank Vironica I Brown, Roman Diveev, Max Goldstein, Eva L Lauer, Nicholas W Long, Paul J Punzone and Joseph M Ragukonis for their contributions in the annotation process. We would also like to thank Benjamin Grider for his help in designing the graphical user interface for our system.

This work is partially supported by UF CISE Data Science Research Lab, UF Law School and ICAIR program.

8. REFERENCES


Achieving Accuracy, Speed and Low-Cost in Freedom-to-Operate (FTO) and Other Infringement-Based Patent Investigations

Jesse Sukman
ClearstoneIP LLC
www.clearstoneip.com

ABSTRACT

This extended abstract describes a cloud-based system that assists patent analysts and their clients in saving time and money when carrying out freedom-to-operate and other infringement-based patent investigations, while realizing improved accuracy. The system provides an administrative interface for reviewers to memorialize patent analysis as it relates to the technical features of patent claims. A second system interface presents an interactive taxonomy of these technical features to a searcher for filtering a large set of patents. Specifically, this second interface prompts users to select technical features of the presented taxonomy that are not embodied by a product or process undergoing analysis. This eliminatory approach ensures objectivity, and thus fidelity, of the resulting output patent set.

CATEGORIES AND SUBJECT DESCRIPTORS


1. INTRODUCTION

It is well-established that the patent system is intended to promote innovation. Patents convey rights in inventions, permitting their owners to exclude others from practicing the same. Naturally, for the patent system to function effectively, the public must be able to identify, and thereby have the opportunity to avoid, infringing others’ patents without undue cost.

Unfortunately, many believe the cost to carry out infringement-based investigations to be significant. Some believe, in fact, that these costs are sufficiently high as to warrant legislative change. While reasonable minds may differ as to the extent of this cost, the bases for these concerns are valid.

FTO (and other infringement-based investigations) are uniquely difficult research exercises. The unique difficulties of FTO may be summarized as follows:

A. In FTO, relevance of patent results is determined by claim scope, not description. The technical aspects described by a patent’s disclosure are distinct from its claims.

B. Products tell a thousand stories. Products, due to their physical existence, can be described in thousands of ways. Each way could be a basis for infringement. Patentability searching, instead, is more discrete.
C. Missing patents in an FTO search could be dire. Finding relevant patents in an FTO search is no indication whether additional relevant patents exist. An entire technology space must be cleared. In patentability searching, producing a few close results is more acceptable.ii

Experienced patent analysts have long recognized these unique difficulties. However, conventional analytics tools have not so evolved. They apply virtually the same processes to all types of patent searches despite their compelling distinctions. Our technology specifically targets infringement-based patent investigations, presenting an opportunity to achieve high accuracy at low cost.

2. SYSTEM OVERVIEW

As shown below in Figure 1, the current system architecture includes an Annotator-side (or administrative) interface, a Searcher-side interface and a Central Analytics Module. The Central Analytics Module receives patent information from one or more electronic Remote Patent Databases.

```
Figure 1: System Architecture
```

Based on information received from the Remote Patent Database and Records Storage, an Annotator GUI is provided. The Annotator GUI enables an Annotator to create and correlate annotations with patent references thereby generating workspaces. Specifically, the Annotator Interface enables the Annotator to annotate patents on a claim-by-claim basis. Thus, the annotations intendedly signify technical aspects of patent references required for infringement. Also, by enabling claim-by-claim annotation, differences in claim scopes within a single patent are respected and stored.
A Searcher may then interface with the system through a Searcher GUI. The Searcher GUI enables the searcher to access stored records including workspaces to provide a display of annotation-patent correlations. The Searcher GUI further enables the searcher to select annotations intendedly not embodied by a product or process undergoing search. A Deduction Engine operates to determine which patent references are to be included and excluded in an output patent set based on the Searcher’s selections. A Reporting Module enables delivery of detailed patent information to the Searcher GUI for display and/or exportation of detailed patent information in data files readable by other application (e.g., Microsoft® Excel®).

In summation, this architecture provides for quick and accurate paring down of large patent sets to small, relevant collections.

3. OVERVIEW OF THE APPROACH

This system is configured to memorialize patent analysis in a manner that enables expedited infringement-based searching. Accordingly, the system relies on the development of databases of annotations correlated with patent references (i.e., “workspaces”). These workspaces are then used by Searchers to filter relevant patent references from large sets of patents.

a. Workspace Development

For the system to provide benefit, a collection of patent references must first be annotated. To carry this out, the Annotator identifies a list of patent references to be annotated. The system automatically populates an annotation queue with detailed information for each listed patent reference (not shown). The Annotator may then select each patent reference, one-by-one, to associate annotations therewith. Figure 2 below illustrates a screenshot of the Annotator GUI in which an Annotator may associate annotations with a patent reference.

Figure 2: Annotator GUI
On the left-side pane, the GUI presents each independent claim of a selected patent reference. The right-side pane displays annotations of the active workspace in taxonomy form. At any time, additional annotations may be added to this taxonomy at any desired location. Attributes of the annotations may be edited or relocated at any time also. Annotations are selected in the right-side pane and added to the central pane and, thus, correlated with the highlighted independent claim of the patent reference. The workspace is created by annotating all independent claims of a specified collection of patent references.

b. Searching

A Searcher may conduct an FTO search by accessing the Searcher GUI as shown below in Figure 3. The Searcher selects a Workspace and an Initial Set of patents to be reviewed (left-side pane). The system then displays an annotation taxonomy (central pane) based on the selected Initial Set of patents. The right-side pane displays attributes of the resulting set. The Searcher reviews the taxonomy and selected annotations corresponding to technical features not embodied in the product undergoing FTO. In this manner, patents deemed to require such selected features are eliminated from the search results. In a short period of time, a large proportion of patents may be eliminated, leaving a small fraction for manual review. In typical cases, patent sets of 2,000-5,000 may be reduced by 85-95% in 30-60min. The user may then page through patent details of the Remaining Patents directly through the Searcher GUI or by exporting Remaining Patent information to Microsoft® Excel®.
This system is best-suited for handling FTO at various points throughout the development cycle of a product. Using conventional tools to handle FTO, product revisions often require starting from scratch. However, with this system, annotation selections may be saved as a “Product Record” and subsequently loaded and revised. The result is the ability to quickly identify patents that may become an issue solely due to a variation in product.

4. USE CASES

The above Overview exemplifies use of this system for handling FTO analyses. This same platform could be used for portfolio monitoring, i.e. reviewing one’s own patent portfolio for infringement by others. In this case, the Initial Set corresponds to the user’s own patent portfolio and annotations are selected in view of competitor products. As another use case, this platform could assist in patent marking. By, again, setting the Initial Set to one’s own patent portfolio, patents corresponding to one’s own products could be easily identified. Additional uses are also possible, limited only by creativity. For example, this platform may provide an efficient means to identify licensing opportunities, assess patent or portfolio value, analyze competitors’ strategies and perform invalidity and patentability searches.

5. CONCLUSIONS

Ensuring that products are infringement-free is a shared desire of many. However, the cost of this compliance has grown significantly and continues to grow. These costs could be greatly reduced by applying a software platform specifically designed for infringement-based searching. By focusing on the elimination of patents understood to be non-relevant, instead of focusing on descriptive similarity, a large set of patents could be reduced to a fraction thereof, accurately and in little time.

---


ABSTRACT
This paper is concerned with the application of language-based analytics to legal writing, focusing especially on filings to the US Federal Court system. We present analytic applications that require language processing at the level of entities and events, and discuss challenges to the application of current language-processing tools that arise from both the format of court filings, and their specialized language use. The paper describes the development of a substantial gold-standard corpus of court filings that were manually annotated for named and non-named entities. This corpus was used to develop several named-entity recognition systems based on a statistical learning method, conditional random fields, and targeting different aspects of the corpus. We describe these systems and present experimental results and accuracy evaluations. Going beyond entities, we also describe an approach to identifying events in a complex task, and provide an approximate accuracy evaluation with indirect proxy measures.

CCS Concepts
• Information systems → Information retrieval → Information tasks and goals → Information extraction
• Information systems → Information retrieval → Evaluation of retrieval results
• Computing methodologies → Artificial intelligence → Natural language processing → Information extraction
• Computing methodologies → Machine learning

Keywords
Court filings; human language technology; natural language processing; information extraction; named-entity recognition; event or scenario extraction; statistical language processing; machine learning; conditional random fields; corpus linguistics.

1. ELECTRONIC COURT FILINGS
Over the last two decades, the system of courts in the United States has been engaged in a wholesale transition to electronic record-keeping. This process has been very similar to that undergone by the medical community, as doctors and other practitioners have come to adopt electronic medical records. What these two large-scale transition efforts have in common is universal agreement that the change is wholly necessary for future viability, efficiency, and effectiveness. Both, however, also have been slow in coming, and in many instances are still far from enabling some of the most exciting benefits of electronic data.
with violations of these statutes. These details are typically only found in the unstructured text of the filing, so to perform this kind of analysis requires surfacing this information from unstructured text and structuring it so it can be used by a mapping tool.

## 2. LANGUAGE PROCESSING OVERVIEW

Extracting this kind of information out of narrative text is a classic application of named entity recognition and related natural language processing methods. In our work with court filings, we have approached the problem at three different levels.

- Since entities are foundational to many language processing tasks, we developed entity recognition models that we then deployed against three different collections of filings to the US Federal Courts. In this work, we consider entities in the broad sense to include both classical named forms (such as person names) as well as non-names of relevance to the legal domain (such as references to the legal code).
- For one of our data sets, indictments related to financial crimes, we wanted to capture financial transactions, which we modeled as a sentence-wise categorization problem for money and related expressions.
- Lastly, for a data set that originated from medical liability complaints, we developed a scenario extraction component that captured surgery events.

These processes produce the elements that directly inform the analytic tasks that we’ve considered, but they occur in the context of a more complex architecture, which we briefly summarize here.

### 2.1 Processing preliminaries

Anyone with even passing experience with the courts understands that the legal system is deeply based on paper documents. In civil cases, especially at the state level, it is not unusual for court hearings to be headed off at the last minute when litigants settle their claims through a handwritten agreement drafted on a page of yellow legal paper. While most filings in the US Federal Courts are not handwritten, a substantial majority of these documents are submitted to the Clerks of Court as printouts. These paper documents enter the CM-ECF electronic docket system as page scans that get captured as image-file PDF (Adobe’s portable document format). The minority of documents that come in as electronic submissions do so as native PDF, not as plain text.

As a result, some effort is required to render these filings into a format that can be more readily accepted by language-processing algorithms. The ideal target is plain text, or alternatively HTML for the case of table-rich data. Optical character recognition (OCR) is thus required for the scanned filings, and PDF-to-text conversion must be applied to the native PDF files. Both these processes are error-prone.

Regarding OCR, Figure 2 shows the first (cover) page of a scanned court filing that illustrates a bevy of nightmarish challenges. Line-to-line flow is often compromised by unaligned text blocks, line numbers in the margin, and parentheses used as table rules. When text is “kerned” with intra-letter spaces, it will not likely be read as the intended words. Stamps, handwriting, and smudges cause noise in the recognized text. Even scratches on the scanner glass can break up letters and masquerade as a table rule. While modern-day OCR packages are remarkably effective in the presence of modest noise, many court filings present the harder kinds of challenges shown in Figure 2.

![Figure 2: challenges to OCR (defendant names smudged out)](image)

PDF-to-text conversion is not prone to the kind of noise caused by smudges and scratches, but line-flow problems arise nonetheless around non-paragraph text blocks. These affect, in particular the cover pages of court filings and any embedded tables.

### 2.2 Basic language processing steps

Once a filing has been rendered to plain text, processing begins in earnest. We first segment a document into sentences, and then further break the sentences down into tokens. This latter process separates punctuation from actual words, but also attempts to capture the extent of punctuation-rich forms like numbers, URLs, email addresses and the like.

This is followed by two stages that are based on jCarafe, a trainable language processing toolkit that implements conditional random fields (CRFs). The first application of jCarafe assigns part-of-speech labels to tokens, a useful linguistic feature that informs in particular the determination of name boundaries in subsequent processing. Names and non-named entities are identified by a second application of jCarafe; more on this in section 3.

Once entities have been identified, they are further normalized into a more usable form. For example, dates (e.g., “11/17/2004”) are rendered into Oracle-standard format (11-NOV-2007). Money expressions (“$10 million”) are assigned an attribute capturing their actual value (10000000). In the case of person and organization names, we also attempt to partition name mentions into groups that designate the same actual individual or the same organization. For instance, “John Hackman” in “Plaintiff John Hackman” would be grouped with “Hackman” in “Hackman filed a motion”. This is a limited instance of the more general process of coreference [8], which also encompasses mapping pronouns (“his motion”) and description phrases (“the Plaintiff”).

...
The financial crime and medical liability complaints required more sophisticated downstream processing to find the transaction and surgery events relevant to their respective tasks. Details of the surgery event extraction appear below.

### 3. ENTITY DETECTION

Entities are central to the kinds of analytics that concern us here. As mentioned, we consider entities to encompass both names (persons, places, organizations, etc.) and non-names (dates, money, bank accounts, and much more). All fall under the somewhat misnamed rubric of named entity recognition (NER).

Court filings, however, present a number of challenges to conventional named entity recognition tools. Chief among these is the fact that the language of court filings does not follow the conventions of typical written text. Since most NER systems are designed (or trained) to process news reporting, they will typically be flummoxed by the very un-news-like nature of legal writing, resulting in much lower accuracy with the law than with news.

Another key challenge around the language of the law is that the critical content of legal documents includes many kinds of elements that would not be expected in conventional text: references to the legal code, criminal nomenclature, and so forth. To capture the full scope of what is available in court filings therefore requires adapting our methods to the idiosyncratic entities available in legal text.

To address these two related challenges, we ported our in-house statistical language-processing tool, jCarafe, re-configuring it to legal language. jCarafe [17] is a machine-learning toolkit that provides a flexible implementation of conditional random fields (CRFs), a leading approach to named entity recognition, part-of-speech tagging, and other tasks that can be understood as sequence labeling [7,9,11]. CRFs are trained through any of a number of discriminative learning methods that maximize the conditional probability of a label vector (the output), given a vector of input tokens and features that hold true of those tokens. This requires manually annotated gold-standard training data, sometimes augmented with un-annotated same-genre data to support feature engineering based on word clusters or co-occurrence vectors.

Towards this end, we built a large corpus of manually annotated court filings that were marked according to the extended labeling task shown in Table 1, to the left. The task specifies 39 labels, a substantial increase over academic named entity tasks such as CoNLL (3 commonly used labels, from the Conference on Natural Language Learning) [14], ACE (Automated Content Extraction, 7 labels) [5], or MUC (7 labels, the Message Understanding Conferences) [12]. Figure 3 shows a small sample of this markup.

All the data were drawn from filings submitted to the Federal District Courts’ CM-ECF database, which is publically accessible through the PACER Web interface (for Public Access to Court Electronic Records). The corpus was assembled from three distinct data sets: (i) 3,000 indictments related to financial crimes, (ii) 5,000 complaints related to medical malpractice, and (iii) a sample of non-specific randomly-selected filings from the Federal District Courts. All the data reported on here, both training and test components, were annotated by the same senior experienced annotator. We have recently begun further annotation using junior annotators whose work is then reviewed by our senior annotator.

### 3.1 Entity detection in indictments

This first portion of the dataset consists primarily of Grand Jury indictments and prosecutorial briefs, along with a handful of plea

---

**Table 1: repertoire of entity types identified in court filings**

<table>
<thead>
<tr>
<th>Person, 4 subtypes</th>
<th>(i) person names (“John Smith”); (ii) anonymized names (“Informant A”); (iii) titles (“Judge”); (iv) initials (“JGH”)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization, 7 subtypes</td>
<td>(i) businesses (“American Medical”), unless further specified; (ii) anonymized businesses (“Money Transmitter A”); (iii) banks; (iv) courts; (v) law enforcement (“Boston Police”); (vi) criminal organizations (rare); (vii) other organizations (universities, etc.)</td>
</tr>
<tr>
<td>Legal (4 subtypes)</td>
<td>(i) criminal code (“18 USC § 1422(a)”); (ii) crimes (“Narcotics Conspiracy”); (iii) filing type† (“civil complaint”); (iv) case number</td>
</tr>
<tr>
<td>Location, 6 subtypes</td>
<td>(i) named locations (towns, counties, etc.), except as further specified; (ii) street addresses and PO boxes; (iii) states or provinces; (iv) countries; (v) zip codes; (vi) districts (“Southern District NY”)</td>
</tr>
<tr>
<td>Times, 4 subtypes</td>
<td>(i) dates (“Nov. 6 2013”, “11/6/2013”, etc.); (ii) times; (iii) dates of birth (rare; may be redacted); (iv) date ranges (rare)</td>
</tr>
<tr>
<td>Financial, 4 types</td>
<td>(i) money expressions (“$10,000”); (ii) cash-equivalent instruments (bank checks, money orders, etc.); (iii) commodities (gold, etc. – rare); (iv) bank account numbers (rare)</td>
</tr>
<tr>
<td>Identifiers, 7 types</td>
<td>(i) driver’s license; (ii) social security number; (iii) phone number; (iv) automotive (license plate or VIN); (v) bar (legal) license; (vi) email address; (vii) any other ID</td>
</tr>
<tr>
<td>Structure marks 3 types</td>
<td>(i) page header; (ii) page footer; (iii) section number</td>
</tr>
</tbody>
</table>

---

On **MAY 19, 2005,** *Crockett* filed an adversary proceeding against *Travelers* and *Travelers*, seeking roughly $950,000$ from the alleged breach of *Crockett’s* subcontract. *Crockett* asserts as an objection to the *Travelers’ claim that it did not perform any of its work negligently. *Travelers* and *Travelers* have impeded *BASENY* into the adversary proceeding as a third-party defendant.

**Figure 3:** sample entity-marked text, as displayed in the Callisto annotation tool

### 2.3 Task-specific processing

These steps, starting with OCR or PDF-conversion, and running up through normalization, are of common value to all the tasks we have approached with these techniques. The processing we have applied beyond those steps has varied from task to task, however.

In one instance, we were interested in determining the geographic variation of the legal code, as it applied to certain crimes; the map in Figure 1 shows an instance of this work. To support our GIS (mapping) tool, we further processed the normalized system output to capture location blocks (“Springfield, Mass.”) so that the GIS could properly distinguish similarly-named locations from each other. We also attempted to correct common OCR errors that arise with references to legal statutes, e.g. by replacing the numeral “1” with the letter “i”, when the numeral appears in contexts that call for a lower case letter.

---

**Table 2:** repertoire of entity types identified in court filings  
† not marked in indictments and complaints
bargains, judgments, and other court orders. These documents, for the most part, are focused on the crimes allegedly committed by a defendant, along with facts supporting the determination of criminal behavior. From a language-processing perspective, the emphasis is therefore on entities, facts, and events. There are many references to the places and times when the crimes were committed, as well as to the statutes that defendants are said to have violated. Where financial crimes are being alleged, there are also many references to money, financial instruments (checks and money orders), bank, bank accounts, money transmitters, etc. In other words, this is an extremely rich data set in which to explore information extraction.

To train iCarafe to process these indictment filings, we manually annotated 50 files (out of 3,000), which we then divided into 41 training files (148,000 words) and 9 test files (15,000 words). We used a time-tested development process, focusing initially on exploratory annotation to develop the tag set, followed by repeated rounds of corpus production. From the second production round on, we relied on models trained from the previous rounds to pre-annotate the data: since correcting system output is less time-consuming than annotating from scratch, this sped up corpus development substantially.

We trained and evaluated an entity recognition model based on this training-test split, using the venerable MUC scorer to assess accuracy [3]. We found an overall accuracy of F=87; precision (P=93) was substantially higher than recall (R=82), an asymmetry that is typical for conditional random fields.\(^2\) Table 2 shows type-specific scores for several kinds of entities. The highest-performing entities included person names (predicted in part by name gazetteers), stereotypical organization types (courts, districts, government, etc.), references to the legal code (which follow stereotypical patterns), location types, and many entities anchored on numbers (dates, times, money, phone numbers). Entity types that proved more problematic were references to crimes (which vary widely in length), less-frequent number-anchored entities (e.g., bank accounts), and especially business names (which display hard-to-capture variability).

These results are lower than high-water marks for newswire data like MUC-6, where top performers scored in the range of F=93-95 (MUC scorer) [12]. While current event news is known to be easy for NER [16], what makes court filings especially harder are several factors: (i) the task has many more entity types, and with added fine distinctions that are absent from MUC; (ii) corrupted text caused by optical character recognition noise; and (iii) absence of linguistic cues that are common in newswire (e.g., corporate designators like “inc.” or honorifics like “Mr.” that facilitate the recognition of their respective kinds of names).

3.2 Entity detection in medical complaints

This second data set is drawn from consolidated litigation related to manufacturer liability around certain medical devices. While these liability complaints also consist of Federal Court filings, they are very different in both style and substance from the indictments.

In particular, whereas indictments are primarily focused on the facts and events of a defendant’s alleged crimes, the medical liability complaints tend to dispense with these kinds of events in a small number of paragraphs. (The facts are typically that a patient was implanted with a medical device that was later found to have failed.) Instead, these medical liability complaints focus much more on the chain of responsibility that marks the device maker as liable for the failure of the device. Much of the legal writing therefore focuses on ownership chains between device makers and their subsidiaries or distributors, manufacturers’ foreknowledge of the potential for device failure, continued marketing of the devices to doctors in spite of known failures, and so forth.

From a corpus-linguistic perspective, these reports offer much less variability and richness than do the indictments. While the specifics of the incidents (and failures) vary from one complaint to the next, a same cadre of device makers and their subsidiaries appear over and over. Likewise, the lines of argumentation are often repetitive, as chains of ownership are re-established for each of multiple counts in the complaints. Finally, as many of these complaints were prepared by the same law offices, they often bear a formulaic resemblance the one to the other. All of these factors mark the language of the liability complaints as substantially different from the indictments.

This divergence in reporting content was immediately evident when applying the indictments-trained entity model to the complaints. An initial evaluation based on 15,000 words of manually-annotated complaints produced a disappointing F-score of 41 (R=29, P=69). The leading factor behind this poor performance was poor recognition of the two most frequent entities in the test sample: company names (43 per cent of entities, with F=15) and person names (15 per cent of all entities, with F=31). Qualitative review of the errors revealed that a small number of companies (the medical device makers) were consistently failing to get recognized, leading to poor recall for company names (R=10). In addition, both company names and many non-names (e.g., legal terms like “STRICT LIABILITY”) were being reported as persons when they appeared in upper case; because of the partial-credit scoring scheme in the MUC scorer, this caused poor precision for both person names (P=53) and companies (P=31).

Why this race to the bottom with upper-case text? It turns out that prosecutorial indictments tend to report the names of alleged perpetrators in upper case, and the named entity tagger had simply over-fit the training data by excessively weighting case-related features as predictors of person entities. In a learning-based framework such as this one, the proper way to counter this poor fit between model and text is to retrain the model. We were able to use the same label set and annotation guidelines for the complaints as we had for the indictments, so this amounted to simply annotating complaint filings and adjoining the newly

---

2 Recall is the false-negative rate, precision the false-positive rate, and F is the harmonic average of P and F. Formally, say key is the total count of entities in the answer keys, act is the number of entities detected by the system, and cor is the number of correct system entities. Then R=cor/key, P=cor/act, and F is the harmonic mean F=2PR/(P+R).
annotated data to the existing indictments data prior to retraining. We avoided more complex cross-domain adaptation strategies such as those of Daumé III [4] or Sutton and McCallum [13], which are more appropriate when adapting dissimilar label sets.

We annotated a total of 234,000 words of complaint data, with 145,000 designated for training, and the other 89,000 words available for testing and other purposes (such as configuring downstream processing). Retraining the model on the combination of the original 148,000 words of indictment data adjoined to the 145,000 words of complaint data yielded a model with greatly improved accuracy on the complaint data (F=93, an 88% reduction in error from F=41). Somewhat disappointingly, the combination model loses accuracy on the original indictments data, dropping from F=87 to F=85, with the additional errors scattered in ways that don’t demonstrate any particular pattern.

3.3 Entity detection in district filings

Our third data sample consist of an undifferentiated sample of filings to the Federal District Courts. This sample covers both civil proceedings as well as Federal criminal cases, and reflects the full range of documents that become attached to a case in the District Courts. This includes original filings, defendants’ responses to plaintiff claims, motions to amend or dismiss other motions, motions to extend deadlines for discovery, transcripts of depositions and similar affidavits by plaintiffs or informative parties, and so forth. The sample also contains findings of the court: judges’ orders approving or rejecting motions, sentencing orders, dismissal of cases, and more.

These data, on the whole, are rich in all the linguistic aspects of the law. As with the indictments, they make frequent reference to entities, facts, and events, thus serving as a good data set for exercising information extraction methods. In addition, they also contain significant amounts of legal argumentation, especially as regards application of the legal code and reasoning by analogy to legal precedent. Legal precedent, in particular, adds a dimension that contain significant amounts of legal argumentation, especially as regards application of the legal code and reasoning by analogy to legal precedent. Legal precedent, in particular, adds a dimension that

Table 3: Accuracy for select types in District Court filings

<table>
<thead>
<tr>
<th>Words</th>
<th>R</th>
<th>P</th>
<th>F</th>
<th>person R/P/F</th>
<th>code R/P/F</th>
</tr>
</thead>
<tbody>
<tr>
<td>41,000</td>
<td>61</td>
<td>84</td>
<td>70</td>
<td>R=74 P=92 F=82</td>
<td>R=49 P=92 F=63</td>
</tr>
<tr>
<td>63,000</td>
<td>67</td>
<td>87</td>
<td>75</td>
<td>R=83 P=95 F=88</td>
<td>R=67 P=94 F=78</td>
</tr>
<tr>
<td>95,000</td>
<td>72</td>
<td>87</td>
<td>79</td>
<td>R=86 P=91 F=88</td>
<td>R=78 P=93 F=84</td>
</tr>
</tbody>
</table>

Table 4: Effect of adjunction (District filings, 95,000 words)

<table>
<thead>
<tr>
<th>Data configuration</th>
<th>R</th>
<th>P</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>95000</td>
<td>72</td>
<td>87</td>
<td>78.9</td>
</tr>
<tr>
<td>95000 + cmplnts.</td>
<td>74</td>
<td>87</td>
<td>79.7</td>
</tr>
<tr>
<td>95000 + indict.</td>
<td>74</td>
<td>87</td>
<td>80.0</td>
</tr>
<tr>
<td>95000 + indict. + cmplnts.</td>
<td>75</td>
<td>88</td>
<td>80.8</td>
</tr>
</tbody>
</table>

Note in particular that the trend from 41,000 to 63,000 to 95,000 words shows continued improvement with no evidence that extraction accuracy is reaching an asymptotic maximum. This is encouraging, as past experience suggests that the addition of further training data from this segment could readily boost recognition accuracy by another 10-15 points of F-score.

As Table 3 also shows, the improvement in accuracy that results from adding more training data has beneficial effect on recognition of some key entities. In the case of person names, recall goes from R=74 to R=86 (a 30% reduction in error), with no substantive change in precision (a 1-point drop). For references to the legal code, recall improves from R=49 to R=72 (45% error reduction), again with no more than a 1-point drop in precision.

Table 4 addresses the question of how much benefit is gained from adjunction of training data from our other corpus segments. In all cases, we find modest improvement: adjoining the complaints yields a gain of .7 points of F (for a 3.6% reduction in error); adjoining the indictments yields a gain of 1.1 points (5.6% reduction in error); and adjoining both yields a gain of 1.9 points (9.2% reduction in error). While the gains are modest, one encouraging aspect of the adjunctions is that they especially improve the recognition of business names. Conjoining both complaints and indictments to the 95,000-word training set produces a gain of 7 points of F score (15% error reduction), with gains in both recall and precision. This is significant in that business names are well understood in the information extraction community as the hardest kind of entity to recognize correctly.

It is worth noting that system output above F=80, while still having much room for improvement, is often acceptable for applications that look to draw conclusions from aggregate analyses of large data sets. Having precision at P=87, while again leaving room for improvement, also means that individual precision errors are less likely to aggregate in ways that could affect aggregate-level analyses.

3.4 Configuration of the CRF

In the general case, the statistical mathematics around conditional random fields encompasses any graph-like configuration of interacting statistical decisions, with the statistical theory of CRFs handling issues of conditional non-independence that befuddles traditional generative models (such as hidden Markov models). For the special case of linear-chain CRFs, as are used to model word sequences (sentences), the mathematics allows for tractable polynomial-time training of the model. Decoding, the application of the model to new sentences, is through the polynomial Viterbi algorithm.
The key to successful deployment of a CRF-based entity recognition model is the appropriate use of features, as these provide the grist for the statistical decisions behind the CRF. In our case, we used a conventional set of context features that is typical for named entity recognition: various configuration of left and right unigrams (single words), bigrams (word pairs), and trigrams (word triples). Likewise, we used bigram features that captured contextual parts of speech. We also had features that applied to individual tokens: the word itself, its part of speech, its prefixes and suffixes up to length three, as well as a number of regular expression features. The latter capture, for example, capitalization patterns, as well as number patterns associated with telephones, dates, and similar forms like email addresses.

A key component of our feature configuration is the use of word lists [6]. These inform the learning procedure by capturing known a-priori generalizations that might not be inferable from the training data alone. Among these are gazetteers of place names (countries of the world, country adjectives, states of the US, and the like). Each word list is encoded as its own feature: any token that matches an entry in the word list will have that feature flagged as true. This allows the learning procedure to generalize from (say) country names found in the training data to previously unseen country names during decoding, since both are found in the same country name list, and the corresponding feature will be true for both. We also have extensive lists of given names and surnames, which allow for the same kind of generalization from person names found in training instances to non-training names. Similar lists capture portions of business names (e.g., designators like “inc.”), days of the week, months of the year, and so forth.

One particular point regarding gazetteer lists related to companies: we restrict their content to atomic terms such as designators or headwords (“company,” “partnership,” . . . ), not full names. This is largely because names are linguistically productive, meaning they display wide variation that is generally hard to predict. So while some frequently-noted businesses appear repeatedly in our data (Refco Capital Markets Ltd., for instance), most business names in these data may never be repeated outside the scope of a single case. There is little practical value, therefore, in attempting to capture a comprehensive registry of business names, since the likelihood of any one name matching is so rare. Never mind that building such an inventory would be a non-trivial task. A further impediment is that the mathematics of linear-chain CRFs encompasses individual tokens, not sequences of tokens, as would be required to use gazetteers for multi-word names (semi-Markov CRFs allow for this, but at a computational cost; see [9]).

Another use of word lists is to capture long-distance effects in the context of a single filing [15]. This happens, for instance, when individuals who are introduced in predictable context (“John White”, “Mr. White”) are subsequently referred to in non-predictive context (“White has been the subject of . . .”). In the absence of a given name like “John” or an honorific like “Mr.”, the subsequent reference to “White” is unlikely to be identified as a person. To capture these linguistic dependencies between cases that are contextually predicted and those that are not, we use feature displacement on select word lists, in this instance, the list of given names. The way this works is that if a feature is true of a token in some context, that fact is copied to other instances of that token. For example, with “John White”, the given-name list matches one position to the left of the token “White,” which causes the token to have a true value for the feature corresponding to a one-word-to-the-left match for that list. To displace this feature, we mark all other instances of “White” in the document with a second version of the feature: this displaced version captures the fact that the original feature matched elsewhere in the document. This effectively enables contextually predictive information to be displaced to cases in non-predictive contexts, leading to significantly improved entity recognition.

Lastly, training algorithms for a CRF make critical use of optimization as they adjust the belief weights associated with features during training (their parameters). While there has been much recent work on rapid approximate optimization methods for CRFs, for this work we relied on a standard non-approximating stochastic gradient descent, using the L-BFGS method.

3.5 Handling citations
For a final topic around entity recognition, we return to the question of citations, such as the one noted above: “Perlman v. McDonald’s Corp., 396 F.3d 508, 511 (2d Cir. 2005)”. The issue with citations, from an entity-recognition perspective, is that they operate simultaneously at two levels: as conventional named entities (“Perlman”, “McDonald’s Corp.”) and as complex structured entities (“Perlman v. McDonald’s Corp.”). This duality poses a challenge to conditional random fields (and related methods) that treat entity recognition as a token-by-token labeling problem. In particular, since the CRF operates by assigning a single output label to each input token, we must choose to either encode citations as conventional named entities:

\[
\text{<PER>Perlman</PER> v. <BIZ>McDonald’s Corp.</BIZ>}
\]

or as structured citations:

\[
\text{<CITE>Perlman v. McDonald’s Corp.</CITE>}
\]

While having an entity type for citations would be obviously useful in the course of processing legal text, doing so in this kind of naive manner would have problematic ramifications with respect to training entity-recognition models. In particular, features that would otherwise predict that “McDonald’s” should be labeled BIZ will now also predict that it should be labeled CITE – for example, the presence of the corporate designator “Corp.”. As a result, the CRF learning algorithm will now treat this feature as less accurate in predicting the BIZ label, which in turn leads to degraded named entity recognition, since the model will now be ignoring the potential contribution of this key feature.

Our approach, and this remains work in progress, is to divide the task of detecting citations into two separate processes. During named-entity recognition, we recognize the base types of entities in citations, so “Perlman” is encoded as PER, “McDonald’s” as BIZ, and so forth. To identify the use of these instances as part of a citation, a subsequent processing stage attempts to parse these forms into citations using heuristic pattern-matching rules. For instance, the pattern “PER v. BIZ” additionally marks “Perlman v. McDonald’s Corp.” as a CITE. This two-stage approach avoids the problematic effects of trying to capture both levels of markup in the context of the CRF. As to the remainder of the citation, the West Law index for example, these components are captured as an additional entity type during entity recognition.

4. BEYOND ENTITIES: FINDING SURGERY SCENARIOS
Many analytic tasks can be completed using entities alone. One task for which that is not true arose in the context of our medical liability complaints. These complaints come from a consolidated multi-district action that gathered many individual lawsuits related to the same class of medical devices. In order to better understand the landscape of the legal action, the judge presiding over the multi-district litigation (MDL) wanted to know a number of facts about each individual complaint. Key among these were the
details of the surgery that lead to plaintiff being implanted with one of these devices, in particular: the surgery date, the surgeon in charge, the hospital or clinic where the surgery occurred, as well as the city and state where the hospital is located.

This turned out to be a complicated problem. While the bulk of the legal actions that ended up in the MDL were filed through a readily-parsed PDF form, some 5,000 complaints entered the MDL as lengthy legal briefs, with an average length of 24 pages. The District Court estimated that for someone to read through these 5,000 filings and capture the relevant information would have taken many months of effort. With this line of work, we attempted to provide substantially the same information by treating the problem as scenario extraction.

What we mean by a scenario is a complex structured event that is typically expressed through multiple facts stated in multiple sentences. In the case of these surgery scenarios, one sentence might name the surgeon in the context of describing a doctor visit. A subsequent sentence might identify the date of implantation, and another sentence might provide the name and address of the implantation clinic. All of these separate facts must be collated to capture the full surgical event scenario. Further, the narrative of the complaint might then go on to list follow-up diagnostic visits or follow-up surgeries to remove the failing device, none of which correspond to the original surgical event. Figure 4 shows an especially involved instance of this kind of narrative.

To capture these surgical events, we used an approach that became common during the Message Understanding Conferences, a series of joint evaluation exercise for information extraction systems [12]. This approach represents scenarios as templates, with each fact that makes up a facet of the scenario filling a specific slot of the template. Figure 5 shows partial templates corresponding to the first five sentences in Figure 4, with the template slots filled accordingly. To fill them, we rely on a series of cascaded detectors, an architecture that proved its mettle as part of the MUC events. The architecture has three main stages:

- The first stage of the cascade consists of generalized entity detection. This need not be restricted to the output of an entity recognition tool alone, but might also include phrases found through grammatical parsing.
- Stage two consists of the application of slot-filling patterns: for instance, the pattern ‘saw ... DOC,’ where DOC matches a doctor name, identifies the filler of the doctor slot of a partial doctor visit (DRV) template.
- When consistent to do so, these partial templates are then merged with each other by a third processing stage, yielding larger templates that aggregate the results of separate slot-filling patterns.

4.1 Processing details

To apply this approach to the medical liability complaints, we defined a simple template with the following content fields:

- type: either a doctor visit (DRV) or surgery (SRG).
- doctors: one or more surgeons or visited physicians.
- hospitals: one or more medical facilities.
- loc, state: the municipality and state for the facility(ies).
- date, vdate: dates of surgery or of doctor visits.

Beyond those information-bearing fields, the templates also have fields that capture ancillary information such as the sentence identifier(s) for the template and the like.

<table>
<thead>
<tr>
<th>Type</th>
<th>DRV In Sept. 2009, Ms. Kennedy saw Dr. Jim Boswell</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doctor</td>
<td>Jim Boswell</td>
</tr>
<tr>
<td>Hospital</td>
<td>Townsville Hospital</td>
</tr>
<tr>
<td>Location</td>
<td>Townsville</td>
</tr>
<tr>
<td>State</td>
<td>AL</td>
</tr>
<tr>
<td>Vdate</td>
<td>Sept. 2009</td>
</tr>
<tr>
<td>Date</td>
<td>Oct. 28 2009</td>
</tr>
</tbody>
</table>

Figure 4: Doctor visits and surgeries in an extended narrative, with sentence-by-sentence template types (triggers italicized)

<table>
<thead>
<tr>
<th>Type</th>
<th>SRG Surgery was at Townsville Hospital, Townsville AL</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRG At the time he implanted the phlogiston, Dr. Boswell …</td>
<td></td>
</tr>
<tr>
<td>DRV</td>
<td>Around Nov. 9 2009, Ms. Kennedy visited Dr. Boswell</td>
</tr>
<tr>
<td>DRV Dr. Boswell referred Ms. Kennedy to Dr. Sam Johnson</td>
<td></td>
</tr>
<tr>
<td>SRG</td>
<td>On Jan. 13 2010, she underwent surgery by Dr. Johnson</td>
</tr>
<tr>
<td>SRG</td>
<td>Dr. Boswell assisted with this surgery</td>
</tr>
</tbody>
</table>

Figure 5: template for the first five sentences in Fig. 4

For this particular task, the fillers of the content slots of the template could be entirely identified based on the output of the jCarafe entity tagger. Location, state, and date, for example are directly identified by the jCarafe entity model. To capture doctors and hospitals, however, we needed to post-process the entity output and identify which person names were the names of doctors, and which businesses names were those of hospitals.

There are two reasons we chose to identify doctors and hospitals as a post-process, rather than doing so directly in the entity model by, say, adding DOC and HOSP entities. First, it would cause unnecessary statistical ambiguity between persons labeled as doctors and those labeled simply as persons, and likewise for hospitals and businesses (the same issue as with citations). In addition, the names of doctors and hospitals are relatively rare in the data, each occurring only in only about 3% of person and business names respectively. Reliably capturing these entities statistically through the jCarafe model would therefore require manual annotation of much more training data than would be possible in practice.

We implemented this post-process through pattern-matching rule sequences similar to Brill’s transformations [2]. The rules apply in sequence, as with decision lists, but without early termination, so that errors made by a rule early in the sequence can be corrected by a subsequent rule. These heuristic rules were developed by hand, based on keyword-in-context analyses of person and business names, as identified by jCarafe. By sorting the potential predictive contexts by frequency of occurrence of context tokens, it was possible to review a very large number of cases with minimal effort. Identifying doctors turned out to have a very simple solution for these data, since they were almost universally referred to with a small number of honorifics (“Dr.”) and a small number of post-name credentials (“M.D.” and the like). Similarly, medical facilities could be readily identified through a small number of headwords (“clinic”, “hospital”, etc.).

As noted, the names of doctors and hospitals are relatively rare in these data, making it impractical to compile the kind of entity-marked corpus that is typically required to assess accuracy. We estimate that even at 230,000 words, our entire corpus of entity-marked complaints might only contain some 20-30 doctor names, and just over 30 hospitals. Instead of re-reading all 230,000 words
of the complaints just to mark those few entities, we relied on an approximate process driven by person and business names. In particular, we reviewed all person and business names from a 200,000-word subsample of the complaints, and assessed the accuracy of the heuristic classifier relative to those entities. While these person and business names were not manually adjudicated but were identified by the jCarafe model, the accuracy of the model for this data set is very high, at F=94 and F=95 respectively. Relative to these jCarafe entities, F score for both doctors and hospitals in this 200,000-word sample were both found to be perfect, at F=100. Spot-checking and separate analyses of other portions of the complaints corpus shows that this level of accuracy is not consistently found throughout the data set, but those recall and precision errors that we found in this way were rare.

Turning to the second-stage of template extraction, we note that much of the effort required to fill template slots is eliminated by having this expanded set of entities that also includes doctors and hospitals. In particular, since the slots are strongly typed to accept only entities that align with this expanded repertoire of entities, all that we are required to do is to decide when a sentence represents a doctor visit or a surgery event. With that determination made, the relevant entities present in the sentence are simply mapped to the appropriately-typed slots.

Identifying these relevant sentences is done by matching trigger terms that indicate the occurrence of an event. For doctor visit templates, this includes such patterns as ‘saw DOC’, ‘DOC diagnosed’ and the like. For surgery templates, we used verb patterns like ‘DOC ... implanted’ as well as event nominalizations like ‘surgical procedure’. When one or more of these triggers matches a sentence, we create a new template instance, and map the doctor, hospital, municipality, state and date entities of the sentence to their respective slots (if the sentence contains no entities of those types, it is not necessary to generate a template).

In the first sentence in Figure 4, for example, the ‘saw DOC’ pattern is triggered by “saw Dr. Jim Boswell,” where “Jim Boswell” is first identified as a person by jCarafe, and is then further categorized as a doctor by the heuristic classifier. In this instance the trigger match causes a DRV template to be generated, with the doctor slot filled by “Jim Boswell”; and the vdate slot filled by the date entity ‘September of 2009’.

Because a single sentence may not always capture all the information related to a surgery event, it is often necessary to aggregate multiple partially-filled templates to produce a final complete record of the event. To do so, we traverse the sequence of templates produced by the slot-fill rules, and where two adjacent templates can be consistently understood as denoting the same event, we merge the templates together, aggregating their slot values into a single slot of the merged template.

The key to doing this is determining consistency. At the level of template types, we allow two DRV templates to merge, likewise two SRG templates, and also allow a DRV template to merge to a subsequent SRG template, since surgery events in in these data are frequently preceded by doctor visits. If two templates have their doctor slots filled, we check whether those doctor names were identified as co-referential by the normalization stage, and if so we consider them to be consistent with each other, which licenses the templates to merge, all other things being equal. For example, the reference to “Dr. Boswell” in the second and third sentences of Figure 3 licenses those templates to merge with the first one as regards the doctor slot.

We use the same co-referential determination to decide whether two hospital names are mutually consistent and license template merging. Lastly, for dates, we use chronological overlap to tell when templates are temporally consistent. Note that by having separate vdate slots for visits and date slots for surgery events, it is not necessary for visits and surgeries to occur at the same time.

4.2 Corpus analysis and evaluation

This data set poses significant challenges as far as evaluating system accuracy. These issues all stem from the fact that the relevant phenomena occur only sparsely in the data. Preliminary analysis of the data showed that on average there were only one or fewer surgery events per filing; some filings might have more than one event, but many only noted that surgery had taken place and provided no reportable elements that we would seek to map to templates (no doctors, hospitals, surgery dates, and so forth). Because filings are so long, 24 pages on average or 5,500 words, it is impractical to build gold-standard data for surgery templates.

As a point of contrast, consider the seminal MUC3/MUC4 data set, which covered terrorist attacks and state-sponsored violence [1]. This data set also averages only around one template per document, but the documents themselves only average 270 words in length. The entire corpus of 1,000 MUC3/4 documents, with its roughly 1,000 event templates, is thus just slightly longer than our entire entity-marked complaint corpus, for which our preliminary analysis suggests there might be no more than 40 or so surgery events. For this reason, we did not deem it useful to further annotate the entity-marked data set for surgery events. Indeed, this effort would require revisiting over 200,000 words of text, and would neither yield enough event-marked data to train statistical methods, nor even to provide a convincing data set for building a heuristic rule-based extraction system by hand, as we did here.

What we ended up doing instead, was to approximate the typical creation of gold-standard data by reducing the 5,000 complaints down to those sentences for which we had detected doctors and hospitals using jCarafe coupled with the heuristic classifier. This reduction can be justified by the observation that sentences that are maximally informative of surgery events (or office visits) will contain the names of doctors or hospitals. From the perspective of optimizing recall of these informative sentences, it is therefore only necessary to capture the vocabulary of the trigger phrases in these sentences. But what about less-informative sentences that do not contain doctors or hospital? Are we not missing something by failing to model their trigger terms? Preliminary analysis showed that in practice, the less-informative sentences use only a subset of the trigger terms that are used in the more informative ones, chiefly language that one can paraphrase as “Plaintiff was implanted with the Defendant’s device.”

We used this approach to subsample a large portion of the data set, from which we then developed a heuristic classifier for trigger phrases based on the same kind of rule sequence architecture as we had used for the doctor and hospital detection. We proceeded through the data set from front to back, manually adding new rules to the trigger classifier until we had captured the potential triggers for roughly the first half of the 5,000 complaints (leaving some hard-to-process cases uncovered).

Two more points must be noted about this approach. First is the fact that it requires accurate detection of doctors and hospitals, since these are the linchpin of the approximation. As we noted above, the same issue of sparsity applies to measuring the accuracy of the heuristic classifier. However, we also noted that
subsampling suggests very high recall and precision, limited chiefly by the equally high accuracy of person name detection.

The second point regards precision. In particular, by directing us at sentences likely to contain relevant events, the subsampling method chiefly allows us to improve recall for these events. However, trigger patterns developed for these sentences may by side-effect also lead to the reporting of irrelevant events. To a large extent, these kinds of false positives prevented by suppressing template generation for trigger sentences with no reportable entities. We did identify a few situations, however, that consistently led to false positives among reported surgery events. The first involved recurring references to a Food and Drug Administration report that cited doctor opinions on these devices; these reliably resulted in surgery templates with the cited physicians in the doctor field. In another instance, problematic OCR output led through a chain of errors to surgery templates being reported for some sentences with addresses involving “Dr.” (for “Drive” not “doctor”). We addressed both classes of error through special-case rules in the event trigger rule sequence.

With all this in place, we set about estimating accuracy for surgical event extraction by running the entire set of complaints through our system and then manually inspecting system output. To estimate recall, we intercepted all the instances of doctors and hospitals that did not somehow get mapped to a template or get rejected by our special-case rules. There were fewer than 200 of these among the 3,669 doctors and 4,493 hospitals that had been identified by the coupling of jCarafe and the heuristic classifier. Even if every single one of these 200 instances ought to have generated its own template, this still suggests a very small recall error approximation of around 2.5%. To estimate precision, we manually reviewed the output of system output for several 200,000-word tranches of the full complaints data set. We looked for both improperly-generated templates as well as templates that ought to have been merged with neighbor templates, and counted both as precision errors. We found on average one of these false positives out of every 25 templates, for an equally small precision error approximation of 4%.

It is worth noting that recall and precision on more typical scenario extraction tasks from the MUC era are much lower than this, with F scores rarely reported over F=60. This is partly an artifact of the present task being defined with slots that are closely aligned to entities that can readily be detected with a CRF. It is also a consequence of the formulaic writing of these complaints that attaches relatively little importance, and less detail, to the actual facts of the surgical interventions. In other words, this is a relatively easy task, as scenario extraction goes.

Lastly, as a sanity check, we considered a non-standard metric of coverage, by which we mean the fraction of complaints for which we generated some kind of template. We found one or more template naming a doctor for 42.7% of complaints; templates naming hospitals (and possibly a doctor) for 61.6% of complaints; and templates naming neither hospital nor doctor, but only a date, for 26.2% of complaints. Overall, this amounts to some kind of surgery event being reported for 87.8% of complaints, which aligns informally with our preliminary analysis of the data.

Clearly, none of these metrics has quite the same gravitas as precision, recall, and F-measure based on gold-standard manual annotation. Especially as our recall and precision approximations are themselves based on the approximate identification of doctor and hospital entities. However, given the two alternatives: either manually annotating an inadequately small corpus, or putting the same level of effort into deriving an approximate but higher-coverage scenario extraction system, we felt the latter spoke better to the practical needs of the courts.

5. DISCUSSION

As the US Courts transition to electronic record-keeping, there will be increasing opportunities to support the Judiciary’s work through modern language-based analytics. We have suggested but a handful of applications here that may improve efficiency and help the courts better understand their caseload. Indeed, knowing more of the content of what has entered the legal system could enable the courts better to prepare for expected workloads and allocate resources appropriately.

An even more exciting opportunity is to enable greater public access to the law [10]. The courts have noted a growing trend towards self-representation among the general public. For this to succeed, members of the public need better to understand which parts of the legal code are appropriate to their legal actions, a problem sometimes referred to as “getting from barking dog to nuisance.” One way this might be enabled by techniques such as ours is through a conventional search tool in which terms like “barking dog” might be typed, but with results indexed by the legal code, as detected by an entity recognition model. This might allow self-represented individuals discover the portion of the legal code under which they might want to file a complaint – getting from barking dogs to nuisance.

A further extension of this work, would be to port these methods from case law to statutory law, especially regulations. Part of the growth of public interest in access to the law comes from business or individual entities seeking to understand the ramifications of rules and regulations, and in particular those times when non-compliance might be financially beneficial. While we would be loath to advocate cheering for attempts to circumvent regulations at the expense of the public good, enabling better understanding of regulation through computational language methods would be of value not just to potential scofflaws, but also to regulators. Indeed, these methods would equally well enable regulators to find and then patch the same regulatory loopholes in the code that might be of interest to entities seeking to avoid regulation.

As regards the technologies described here, it is clear that they remain very much at early stages of application to the work of the courts. One conclusion in particular bears noting, namely, for entity recognition, stylistic variation among kinds of filings makes for poor generalization. More work is needed to identify productive ways to accelerate the adaptation from one kind of filing to another. As far as going beyond entities to scenarios such the surgical events described here, the issue of sparseness will need to be addressed. We have shown shortcuts here for both engineering and evaluation, but more tasks of this kind will need to be attempted to know if these methods generalize.

All that being said, this domain remains an exciting one for applications of language-processing methods, and one that we hope will gather growing interest from the computational linguistic community.

---

3 See, for example, the Free Law Project: https://free.law
REFERENCES


