Automatic Categorization of Legal Texts with Word2Vec and Semi-Supervised Learning

Hao Peng
Indiana University
Bloomington, IN
penghao@umail.iu.edu

Kristin Day
Indiana University
Bloomington, IN
kricrone@iu.edu

Sanjana Pukalay
Indiana University
Bloomington, IN
spukalay@umail.iu.edu

Abstract

Automatic categorization of legal texts is a first step toward providing free and meaningful public access to legal information. Currently, access to legal information is expensive, difficult to search and generally not available to the public. Because of hand labeling, private business charges a premium for access to searchable databases and those services are only provided business-to-business. Automating the process of legal text categorization is the first step in creating a searchable legal database without the cost of employing hand labelers.

1 Introduction

Full text searches require as a first step indexing of the full text documents to be searched. This step allows for reasonable processing times for searches. If every document had to be searched in its entirety using a database of millions of documents, the search processing time would not allow for many searches and each search would be costly. Document indexing resolves this issue allowing for quicker access to relevant information.

Historically, hand indexing has been relied on for preparing documents for search databases. But, as Dabney points out in (Dabney, 1986, p. 6), ". . . human indexing is fraught with error and uncertainty, so we have sought other ways to extract information from written texts."

2 Related Work

2.1 Text Analytics

Text Mining, a subsection of Big Data Analytics, utilizes large amounts of textual data from various sources including social media, literature, and legal documents in order to discover and visualize previously unseen features. In text mining, information is retrieved from its source (ie a social media website) and then stored so that information extraction (IR) methodologies are used along with natural language processing (NLP), machine learning and statistics (Liu et al., 2015, p. 197). A number of text mining applications in the legal domain have been performed. (Moens, 2001) proposed text mining methodologies as a novel means to approaching information retrieval in the legal domain, and also proposed (Moens, 2005) a retrieval model utilizing Extensible Markup Language (XML), a human- and machine-readable means of storing, transporting and accessing data, to exploit the relatively standard structure of legal documents. Other projects have included a neural network approach to the classification, clustering and retrieval of legal documents (Chou and Hsing, 2010) as well as (Chen et al., 2013).

2.2 Types of Information Retrieval Systems

Different types of Information Retrieval systems have been proposed to date that can be classified into three broad categories. First is the traditional database system in which information is retrieved by querying a database. Second is the Browsing/Navigation System which gives a UI interface and is more user friendly in the sense that it helps the user find what he/she is looking for through a GUI-based environment. Third is the question-answer system which is a very sophisticated platform developed using deep learning and machine learning techniques. For instance, we can ask the system - "What is the maximum driving speed allowed on a freeway" (Moens, 2001) and it will be able to tell us the exact answer from all the data that it has.

In the research of (Uijttenbroek et al., 2007), the team developed "fingerprints" for full legal text indexing. The fingerprints were divided into two categories: case-based and code-based. The case-based fingerprints were indexes of judicial opinions (otherwise known as ”case law”) while the code-based fingerprints were indexes of law (or ”codes”). As one would expect, the code-based fingerprints performed better with non-interpretive concepts while the case-based fingerprints performed better with interpretive concepts. “This is in agreement with the intuition that the meaning of interpretive concepts is defined by case law.” (Uijttenbroek et al., 2007, p. 299)

Some of the relevant research on this topic comes from the medical field. Much like the attorneys, doctors often need to access medical information databases. The most commonly used is PubMed according to (Lee et al., 2006). The PubMed database is limited in the specificity of searches that can be performed. A search for a specific drug may return thousands of results and busy professionals do not have time to sort through them all. More precise retrieval of relevant information is required both in the field of medicine and in law.

2.3 Legal Document Structure

Legal documents can be of different types (see a more detailed description below in the “Data” section) but this study will focus on judicial opinions which are used to clarify laws and fill in gaps that laws do not strictly address. The major challenge faced is the legal language. It is quite different from the way we talk in our everyday lives. Legal language is very diverse and contains especially long sentences and subclauses. But, the legal documents seem to have a general structure: they tend to start with an argument, address each issue and leads to conclusion on the issues to resolve the case.

2.4 Development of Modeling

The development of any model involves all the essential general steps such as: data extraction, data cleaning, data preprocessing, data indexing, data mining, data analysis and data visualization. Most of the time, data needs to be extracted from relevant sources since legal datasets are not readily available. A lot of data cleaning and preprocessing needs to be done. The most popular approaches associated with natural language index terms are lexical analysis, removal of stop words, stemming, formation of index phrases, replacement by thesaurus class terms and assigning different weights to different terms based on their relevance. There are some well-known models which are used for retrieval purposes. Some of them are Boolean Model, Vector Space Model and Probabilistic Models. And, some of the techniques used to develop these models are Support Vector Machines, Naive Bayes, Decision Tree, K-Nearest Neighbors and Hidden Markov Model. (Moens, 2001) One researcher developed a “probabilistic latent semantic indexing model” that outperformed other latent semantic indexing work for automatic text classification. (Hofmann, 1999)

2.5 Clustering

Another important task that has been performed is clustering. Self Organizing Maps (Schweighofer et al., 2001) are a very efficient way of clustering similar documents together and assist in cluster analysis of high dimensional data. Neural Networks are extensively used for this task as well. Recently, a lot of advancement has occurred in the field of machine learning and deep learning which helps in many applications of Sentiment Analysis and related fields.

2.6 Labeling

Some of the labelling has to be done manually which is expensive and requires domain expertise. The high cost of hand labeling makes the process infeasable in most cases, but, to some extent, hand labeling seems inevitable because of the absence of a preprocessed legal datasets. With further development and study in this field, labeled legal data is expected to become available.

2.7 Classification of Feature Selection

The classification of feature vectors representing the interpretation of legal documents improves the search for similar or related documents, the interpretation of these documents as well as the navigation within the text corpus. The project KONTERM (Bochereau et al., 1991) workstation is to provide a hybrid application of methods of legal knowledge representation assisting lawyers in their task of managing presently high quantities of legal information contained in natural language documents. Neural networks have gained some attention for their ability to encapsulate legal knowledge. This might be due to the fact that knowledge-based approaches were awarded only limited success in highly narrow domains. Neural network applications in the legal domain face the complexity of legal texts which require a high number of neurons. Consequently, neural network processing of legal texts requires long training times. As in other applications, the time-consuming training-process is commonly regarded as the major obstacle of real-world large-scale neural network applications.

2.8 MARILOG

MARILOG (Merkl and Schweighofer, 1997) is a set of expert systems and legal decisional aids intended to help mayors in decision making. It consists of legal knowledge bases, data files, Statistical files, text editors, banks of legal texts and dictionaries.

We can see that a lot of work has been done in the legal domain to date, but not many have been very successful because of the structure of data and lack of domain expertise. Some work has been in the field of Automatic Text Classification as well. Neural networks with backpropagation have been employed for the task.

We propose a unique solution for Automatic Text Classification of Legal Documents using a small set of hand labeled judicial opinions with 16 unique classifiers and employing a semi-supervised learning methodology using Word2Vec.

3 Data - Collection and Labeling

For the purpose of this study, the following definitions will apply:

- Statutes - One statute is one law.
- Judicial Opinion - A judicial opinion is a judge's ruling on a legal issue in a case. Each case may have one or more judicial opinions associated with it. A judicial opinion serves to clarify the scope and meaning of a statute. Additionally, it is important to note that the terms "decisions" and "opinions" may be used interchangeably throughout this work.
- Common Law - The term common law refers to judge made law. While legislators (elected lawmakers) create statutes, there are times when certain factual situations fall within a gap between statutes. In these situations, judges are required to make decisions that create rules not covered by the statutes. These decisions comprise the "common law".

Level 1	**Level 2**	**Level 3**
Family	Substantive	Adoption
Civil	Procedure	SummaryJudgment
Civil	Procedure	StatuteOfLimitations
Family	Substantive	GrandparentVisitation
atty_ethics		
Employment	Procedure	SummaryJudgment

Table 1: Label Excerpt

3.1 Legal Corpus

Two corpuses of judicial opinions were used for this study: 1) Indiana Supreme Court decisions, and 2) decisions from all six California appellate courts and from the California Supreme Court. Only those opinions, for all courts in both states that were available online through the respective court websites were used in this study. The Indiana website used was `http://www.in.gov/judiciary/opinions/archsup.html` and the California website used was `http://www.courts.ca.gov/opinions-slip.htm`. In all, 1087 opinions were used from Indiana and 1667 opinions from California. The opinions were scraped from the respective websites in .pdf format and were converted to .txt format using the pdftotext tool and commands in terminal. Before sentences were extracted from these documents, the text was cleaned by removing the captions, punctuation and numbers from the text using regular expressions in Python. The joined corpuses resulted in a Word2Vec model with 827,301 words.

3.2 Labeled Data

Of the total 2,742 judicial opinions from Indiana and California, 250 decisions were hand labeled to three hierarchical granularities. The first level of labels was very general and a total of 16 unique labels were assigned. The second label gave more detail and the third label was the most detailed. An excerpt from the labels table is shown in Figure 1 below.

The labeler had roughly four years of experience working as a civil litigation attorney. The labeling was, by admission of the labeler, highly subjective. Often, other attorneys were consulted during the labeling process to determine the correct label. The time required to perform this task was roughly 10 hours (about 2 hours per 50 judicial opinions).

Only Indiana Supreme Court decisions were hand labeled. The corpus of Supreme Court decisions was chosen for hand labeling because Supreme Court decisions generally focus on very narrow topics of law unlike lower court cases that might addresses many legal issues in one opinion. For this initial analysis, more focused judicial opinions were desired.

4 Method and Data Splitting

We trained Bag of words (BoW) model in both semi-supervised and supervised learning.

We also trained Bag of means (BoM) with word2vec embedding model in semi-supervised and supervised learning.

In both contexts, the supervised learning model is trained with a Random Forest algorithm, while the semi-supervised one is trained with Label Propagation provided in "scikit-learn" package `http://scikit-learn.org/stable/modules/label_propagation.html`.

5 Experiments and Evaluation

There are $1,087$ legal case text files for Indiana and $1,655$ for California in our dataset. We are only able to manually label 250 out of all $2,742$ cases due to the limited time and resource. There are That's why we also add semi-supervised learning in our experiment.

By training and testing in four different ways, we expect BoM to outperform BoW in both supervised and semi-supervised learning.

The 250 cases were hand labeled into 3 levels of granularity. The first level has 16 unique labels. As we have limited labeled data, we are going to use only the Level 1 labels. An example of the labels is shown in Table 1.

For supervised learning, our dataset size is just 250 labeled data points. while for semi-supervised learning, the dataset size grows to $2,742$.

5.1 Models using Bag of Words

The vocabulary is set to $3,000$ most frequently used words. In supervised learning, we trained a Random Forest model on 200 data point and tested on 50 data point. As expected, we achieved 100% accuracy in training and 0 for testing. This model is significantly overfitted because our data has $3,000$ columns while only 200 rows.

In semi-supervised learning, we trained a Label Propagation model on $2,742$ data points with 250 labeled ones and tested on the same label data points. Out of expectation, it only classified 1 case file correctly out of 250, which is really a bad result. This suggests that Bag of Words is not a good representation method when the labeled dataset size is small.

This also indicated that when using BoW methods on a small dataset, supervised learning is better than semi-supervised.

5.2 Models using Bag of Means

We first trained a word2vec embedding model on all $2,742$ case files. In our training, the vector dimensionality is set to 300, minimum word count is set to 40, and the context window is set to 10 with sub sampling rate set as 0.001. At last, we obtained a word2vec model with $827,301$ words, and each word is embedded as a 300 dimensional vector. With the help of this newly obtained word2vec model, we represented each data point as a means of all the words' vectors in that case file.

In supervised learning, we again trained a Random Forest model with the same setting as before. It also gives us the same bad result as it trained on the previous BoW model, which is 100% accuracy for training and 0 for testing. The model is also significantly overfitted. Thus, we can't come to a fairly conclusion when comparing the BoM method with BoW one on supervised learning with small-sized dataset.

While in semi-supervised learning, we trained a Label Propagation model in the same way as before. It turns out to perform really well. The model classified 106 case files out of 250, compared to only 1 for that trained on BoW. This suggests that BoM method is far better than the BoW method when using semi-supervised learning with limited labeled data in the whole dataset.

6 Future Work

As stated in the introduction, automatic classification of legal texts if a first step toward a larger goal: creating an affordable system for meaningful public access to legal information. The most important next step in this research is to gather more data. Not only are more judicial opinions and codes needed from all of the states, but also other sources may prove useful in achieving the overall goal of bringing legal knowledge to the public. One source rich in legal information are legal blogs or "blawgs". According to (Conrad et al., 2009, p. 167), ". . . a growing number of professionals are taking an interest in legal blogs [] that increasingly provide useful information, information that is typically composed by legal scholars, lawyers or students of law." Such sources may help to bridge the gap between highly technical legal language of codes and judicial opinions, and the everyday language of the American people. Futhermore, as (Conrad and Schilder, 2007, p. 231) points out, "blawgs" are a promising source of data for sentiment analysis related to the sentiment of legal professionals about new laws or opinions. "*This is a great decision* conveys clear sentiment, but *The announcement of this decision produced a great amount of media attention* is neutral." The possibilities of text analytics applications in the legal domain are as limited as the human imagination.

6.1 Identifying User Intent

Concurrent work is in process that endeavors to identify user intent in layperson legal inquiries. This work endeavors to take a layperson legal inquiry and translate it into legal terms that will then be used to query the legal text database. In related future work, artificial intelligence systems will be developed to not only tranlsate layperson terminology into legal terminology, but to also ask follow-up questions of the user to determine intent and further refine search results.

6.2 Automatic Legal Text Classification

Future work in automatic legal text classification will include gathering more data both from social media sites that involve layperson legal queries, and from legal databases. Larger corpuses will allow for better training of Word2Vec models.

A greater corpus of hand-labeled legal data by multiple labelers would be required to further investigate the efficacy of supervised learning methods. Finding inter-labeler agreement would be very valuable both in determining the viability of the models produced and in convincing the legal community that any resulting systems perform as well if not better than human indexed documents. This step may be difficult to complete as hand labeling of legal texts is time consuming and those with sufficient training to perform the task do not work inexpensively.

Additionally, experiments with other methods of automatic classification will be conducted and the results compared to the those herein.

6.3 Creating Understandable Summaries with Returned Search Results

It is not expected that untrained users will know or properly understand legal terminology. That is the purpose of determining user intent upon query input. But, this concept also applies to returning legal content that the user can reasonably understand and use.

Thusly, future work will endeavor to gather all relevant legal texts and information to be returned to the user, but will also provide an automated summary of the topic in plain English. Links to the original documents will be included for the user's review, but a plain English summary would, in theory, be very helpful in assisting the user's initial understanding of the subject matter searched.

As with other tasks outlined herein, hand summarization is costly. Developing a system that can reasonably summarize legal topics in plain English would significantly reduce the cost of providing this information to the public while increasing public access and understanding of legal materials.

7 Conclusion

Although the semi-supervised Word2Vec BoM method far out-performed the other methods, it still only achieved a 67% accuracy. More experiments are needed to finely tune the parameters set in the Word2Vec BoM semi-supervised method. Additionally, gathering more legal data will asist in better training the Word2Vec model.

To really determine an appropriate baseline for these experiments, it would be necessary to get hand labeled data from multiple, skilled labelers. This is a very costly proposition. But, it is important to understand human agreement on classification of legal texts. As (Dabney, 1986, p. 6) stated, ". . . human indexing is fraught with error and uncertainty . . . ." To know whether a neural network can match or outperform human labeling, we first have to understand the disagreement among human labelers.

Acknowledgments

The efforts of Jordan Reifsteck and Nithipat Tanmanagul in a related project were essential to the success of this research and are acknowledged herein.

References

Laurent Bochereau, Danièle Bourcier, and Paul Bourgine. 1991. Extracting legal knowledge by means of a multilayer neural network application to municipal jurisprudence. In *Proceedings of the 3rd international conference on Artificial intelligence and law*, pages 288–296. ACM.

Y.-L. (1) Chen, Y.-H. (1) Liu, and W.-L. (2) Ho. 2013. A text mining approach to assist the general public in the retrieval of legal documents. *Journal of the American Society for Information Science and Technology*, 64(2):280–290.

S. Chou and T.-P. Hsing. 2010. *Text mining technique for Chinese written judgment of criminal case.*, volume 6122 LNCS of *Lecture Notes in Computer Science*. Dept. of Information Management, National Central University.

Jack G Conrad and Frank Schilder. 2007. Opinion mining in legal blogs. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 231–236. ACM.

Jack G Conrad, Jochen L Leidner, Frank Schilder, and Ravi Kondadadi. 2009. Query-based opinion summarization for legal blog entries. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 167–176. ACM.

Daniel P. Dabney. 1986. The curse of thamus: an analysis of full-text legal document retrieval. *Law Library Journal*, 78:5 – 40.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

Minsuk Lee, James J Cimino, Hai Ran Zhu, Carl Sable, Vijay Shanker, John W Ely, and Hong Yu. 2006. Beyond information retrieval-medical question answering. In *AMIA*.

Yi-Hung Liu, Yen-Liang Chen, and Wu-Liang Ho. 2015. Predicting associated statutes for legal problems. *Information Processing and Management*, 51:194 – 211.

Dieter Merkl and Erich Schweighofer. 1997. The exploration of legal text corpora with hierarchical neural networks: a guided tour in public international law. In *Proceedings of the 6th international conference on Artificial intelligence and law*, pages 98–105. ACM.

M.-F. Moens. 2001. Innovative techniques for legal text retrieval. *Artificial Intelligence and Law*, 9(1):29–57.

M.-F. Moens. 2005. Combining structured and unstructured information in a retrieval model for accessing legislation. In *Proceedings of the International Conference on Artificial Intelligence and Law*, number 10th International Conference on Artificial Intelligence and Law, Proceedings of the Conference - ICAIL 2005, pages 141–145, Katholieke Universiteit Leuven.

Erich Schweighofer, Andreas Rauber, and Michael Dittenbach. 2001. Automatic text representation, classification and labeling in european law. In *Proceedings of the 8th international conference on Artificial intelligence and law*, pages 78–87. ACM.

E.M. Uijttenbroek, A.R. Lodder, M.C.A. Klein, G.R. Wildeboer, W. Van Steenbergen, R.L.L. Sie, P.E.M. Huygen, and F. Van Harmelen. 2007. *Retrieval of case law to provide layman with information about liability: Preliminary results of the BEST-project.*, volume 4884 LNAI of *Lecture Notes in Computer Science*. Centre of Electronic Dispute Resolution - CEDIRE.ORG, VU University Amsterdam.

Group members

Hao Peng

M.S. in Data Science — Indiana University Bloomington Bloomington, Indiana, USA pengh1992@gmail.com

Personal Page: http://pages.iu.edu/ penghao/ GIthub Page: https://github.com/hao-app

Group Duties: Web crawling, modeling, experiments and methodology and evaluation sections of the final paper.

Absences: I have never missed the class. Only have been later for the class two times. I have talked to Dr. Mageed beforehand. Thanks for your dedicated effort on this course!

Photo: Hao Peng

Kristin Day

Kristin Day holds a law degree (J.D.) from Indiana University and practiced as a civil litigation attorney in California for nearly four years. She is now pursuing a Master of Science in Data Science at Indiana University.

Group Duties: Planned and organized all group work. Prepared, merged group contributions and formatted final paper. Coordinated work between two teams in related research.

Absences: No absences.

Photo: Kristin Day

Sanjana Pukalay

Sanjana is a Data Enthusiast originally from India with a background in Computer Science. Her curiosity and passion make her want to pursue a career in the field of Data Science. She likes reading, meeting people from different cultural backgrounds and travelling.

Group Duties: Data extraction, data preprocessing and literature review section of the final paper.

Absences: No absences.

Photo: Sanjana Pukalay