Text Mining for Quality Control of Court Records*

Eric O. Scott
The MITRE Corporation
7525 Colshire Drive
McLean, VA
escott@mitre.org

Haleh Vafaie
The MITRE Corporation
7525 Colshire Drive
McLean, VA
hvafaie@mitre.org

Zelal Gungordu
The MITRE Corporation
7525 Colshire Drive
McLean, VA
zgungordu@mitre.org

Charles E. Horowitz
The MITRE Corporation
7525 Colshire Drive
McLean, VA
chorowitz@mitre.org

Bradford C. Brown
The MITRE Corporation
7525 Colshire Drive
McLean, VA
bcbrown@mitre.org

ABSTRACT

Attorneys across the United States use government-provided electronic databases to submit docket entries and associated case files for processing and archival in public judicial records. Data entry errors in these repositories, while rare, can disrupt the court process, confuse the public record, or breach privacy and confidentiality. Docket quality assurance is thus a high priority for the courts, but manual review remains resource-intensive.

We have developed a prototype application of text mining and human language technologies to partially automate quality assurance review of electronic court documents. This solution uses document classification and named entity recognition to extract metadata directly from documents. Discrepancies between the extracted metadata and the user-provided metadata indicate a possible data entry error.

On two independent samples of publicly available court documents, we find that for a small number of classes with a sufficient number of training documents, the document class can be automatically classified with greater than 94% accuracy in one case, but only 81% in the other. Our attempts to extract case numbers and the names of parties from documents via a conditional random field model met with less success. Future work with more extensive training data is necessary to more accurately evaluate both applications.

Categories and Subject Descriptors

I.5.4 [**Information Systems**]: Pattern Recognition—*Applications, Text processing*

General Terms

Algorithms, Performance

Keywords

Document Classification, Information Extraction, Named Entity Recognition, Text Mining, Quality Assurance

1. INTRODUCTION

American courts have long had a legal and constitutional obligation to provide the public with access to judicial proceedings, but recent history has seen changes to what documents must be released and how they are made available. The way court documents are processed was radically altered in the late 1990's by the advent of electronic filing systems. In particular, all federal appellate, district and bankruptcy courts have adopted local implementations of the Case Management / Electronic Case Files system (CM/ECF) [7, 9]. Files from over 200 CM/ECF databases and over 40 million cases are centrally indexed at `www.pacer.gov`, where researchers and the public can download documents in PDF format for a fee.

A searchable database is only as good as its metadata, and court clerks are forced to allocate substantial manual resources to ensure the correctness of submitted information. A typical CM/ECF submission consists of a docket entry, a PDF document containing unstructured text, and some user-provided metadata detailing the type of document uploaded, the case number, and the names of parties, attorneys, and the judge involved with the case. Errors or omissions in any of these fields, or pairing them with the wrong document, can confound attempts to retrieve the document through a search interface such as PACER.

This report summarizes preliminary results we have obtained with a prototype application of machine learning to automatically identify both the type of an uploaded court document and any case number or party information contained in the document text. The ability to extract this information reliably could allow courts to automatically accept some fraction of submissions without manually reviewing the metadata, and/or to prioritize documents in a quality assurance queue.

2. METHODS

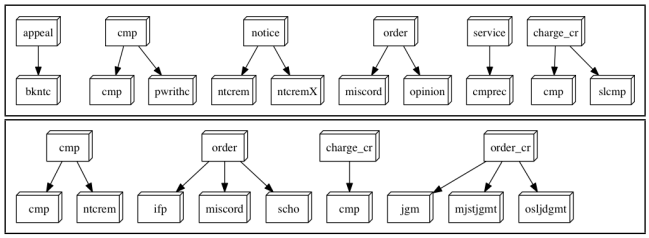

Figure 1: Top: The structure of the document classes in corpus A, showing 10 subtypes arranged into 6 supertypes. Bottom: Similarly, corpus B consisted of 9 subtypes arranged into 4 supertypes.

2.1 Data

We conducted experiments on two independent corpora of documents and associated metadata that were provided courtesy of two U.S. District Courts, referred to here as districts A and B, respectively. Each document corresponds to a unique docket event. The events were chosen from a diversity of criminal and civil cases. Furthermore, about 75% of the PDFs were only images of scanned documents, and did not contain computer-readable text.[1] These were omitted from the study.

Documents in CM/ECF are organized into a two-layered event-type hierarchy, as shown in Figure 1. Each document has both a supertype and a subtype, and each database has a dozen or more supertypes and several dozen subtypes – the exact ontology differs by local implementation. Our final dataset for district A contained 1,438 example documents taken from 6 supertypes: appeal, criminal charge, civil complaint, notice, civil order and service. Similarly, 1,176 examples were in the sample from district B, taken from 4 supertypes: criminal charge, criminal order, civil complaint, and civil order. The event types used in both corpora were chosen because these were the only types for which we had a non-negligible number of examples.

The distribution of examples across types was highly imbalanced – corpus A contained 1,072 civil complaints, for instance, but only 19 services. Attempts to reduce the imbalance via Kennard-Stone sampling had a negligible impact on classification results, so the numbers reported below are based on the original sample [5].

2.2 Models

The errors we are attempting to detect fall naturally into two distinct tasks: identifying to which event supertype and/or subtype the document belongs (classification), and identifying specific information in the document text (named entity recognition).

For classification, we filter common stopwords out of each document, and perform stemming via Porter's algorithm. Tokens which appear in less than 1% or more than 99% of the example documents are removed (about 1% of the tokens). On the order of $5 \cdot 10^4$ unique tokens remain in each dataset after these steps. The processed unigrams are then represented as TF-IDF document vectors.

[1]Text was extracted from the remaining PDFs with Apache PDFBox, available from `https://pdfbox.apache.org/`.

The vectors serve as training instances for a support vector machine, with naïve Bayes and k-nearest-neighbor classifiers serving for comparison. For the former, we used the C-SVC mode of the popular `libSVM` library, which permits multiclass learning [3].

We tuned the parameters of each algorithm at a coarse level. We tested $k = 1$ and $k = 5$ for k-NN, and ran the SVM with both radial basis function and sigmoid kernels. The SVM's C parameter was held constant at 0, and we tested the three values 0, 10 and 25 for γ. The best performing choice of parameters is reported.

For the named entity recognition (NER) task, we used version 2.0.1 of the open source MITRE Annotation Toolkit (MAT) to manually tag text that identifies case numbers, defendants, plaintiffs, attorneys and law firms in 55 documents chosen arbitrarily from corpus B.[2] 75% of the tagged documents were used as a training set to build a conditional random field (CRF) model [6, 8].

For this preliminary study, we used MAT's default feature configuration, which includes, *inter alia*, word prefixes and suffixes and several n-grams. Since the choice of features used for NER is as important as the choice of model [10], it would be fruitful for future work to experimentally adjust the feature set to the judicial domain (especially for highly stylized entities such as case numbers).

3. RESULTS

To evaluate the classification algorithms, we measured the accuracy, macro-averaged precision, macro-averaged recall and Cohen's κ from 10-fold cross-validation. The macro-averaged precision and recall tell us about classification performance when all classes are considered equally important, even if some only had a few test instances. By contrast, Cohen's κ takes differences in class size into account, correcting for classifications that occur merely by chance [4]. This metric is especially useful in multiclass cases with imbalanced data, which is our case here.

Separate classification models were trained to detect subtypes and supertypes. The results of these experiments are given in Tables 1 and 2, respectively. In all cases the SVM with a sigmoid kernel performed better than naïve Bayes or k-NN. Subtypes prove difficult for the models to distinguish, as none of the performance metrics came out higher than 0.81. The 4 supertypes in corpus B proved easiest to classify, the SVM achieving 94.5% accuracy and macro-averaged precision. For both supertypes and subtypes in corpus A, however, Cohen's κ reveals that much of the moderately high performance on of the classifiers is due to chance. This may be a result of the larger number of classes and more extreme imbalance found in corpus A. Attempts to compensate for the imbalanced data by applying Adaboost, Kennard-Stone sampling, and undersampling of over-represented classes all had an adverse impact on classifier performance (not reported).

[2]The MITRE Annotation Toolkit is available from `http://mat-annotation.sourceforge.net`, and the associated jCarafe CRF engine is available at `https://github.com/wellner/jcarafe` (Accessed 23 June, 2014). Both are released under a BSD-style license.

Corpus	Model	Acc.	κ	Prec.	Rec.
A	k-NN ($k = 1$)	0.782	0.308	0.643	0.389
	Naïve Bayes	0.696	0.368	0.421	0.463
	SVM ($\gamma = 0.0$)	0.810	0.432	0.696	0.439
B	k-NN ($k = 1$)	0.503	0.262	0.426	0.411
	Naïve Bayes	0.629	0.526	0.562	0.529
	SVM ($\gamma = 10.0$)	0.764	0.696	0.778	0.729

Table 1: Classification results on event subtypes. Corpus A had 10 such subtypes, and corpus B had 9.

Corpus	Model	Acc.	κ	Prec.	Rec.
A	k-NN ($k = 1$)	0.789	0.302	0.773	0.392
	Naïve Bayes	0.717	0.350	0.424	0.428
	SVM ($\gamma = 0.0$)	0.819	0.441	0.935	0.525
B	k-NN ($k = 1$)	0.693	0.420	0.758	0.586
	Naïve Bayes	0.871	0.785	0.833	0.796
	SVM ($\gamma = 10.0$)	0.945	0.911	0.945	0.896

Table 2: Classification results on event supertypes. Corpus A had 6 such types, and corpus B had 4.

To evaluate the named entity extraction model, we measure the precision and recall with which words that belong to an entity name are extracted from the documents in the test set. To assess how performance improves as the size of the annotated training corpus is expanded, we trained 30 models on a different subset of the data. Each run selects 41 random training documents from the corpus of 55, and feeds them to the model in a random order. The remaining 14 documents are used as a test set.[3]

The corpus as a whole contained over 1,000 examples of defendant entities, 600 plaintiffs, 160 attorneys, and 70 law firms, but only 15 case numbers. In general, performance for defendants was better than the less well-represented entities. Figures 2 and 3 show the progression of precision and recall as training examples are presented to the model. For all entity types, increasing the corpus size improves recall steadily, but there is no discernable improvement in precision. Surprisingly high precision recall is achieved for case numbers given the very small number of examples we have for that entity.

4. DISCUSSION

Online access to court records raises a number of philosophical issues and privacy concerns which were debated at length throughout CM/ECF's adoption [1, 2, 11, 12, 13]. The Judicial Conference responded with a set of rules which state that, *inter alia*, the 600,000+ attorneys[4] who submit case files to CM/ECF systems are responsible for redacting certain personal information from documents before their pub-

[3]Training and test sets selected by random resampling are not independent, since they share many of the same documents. Our intent is to control for the noise introduced by the order of document presentation, not to make general statistical claims about the problem domain (which would require truly independent trials).

[4]C.f. `http://www.uscourts.gov/annualreport_2011/Key_Studies_Projects_And_Programs.aspx` (Accessed 22 June, 2014).

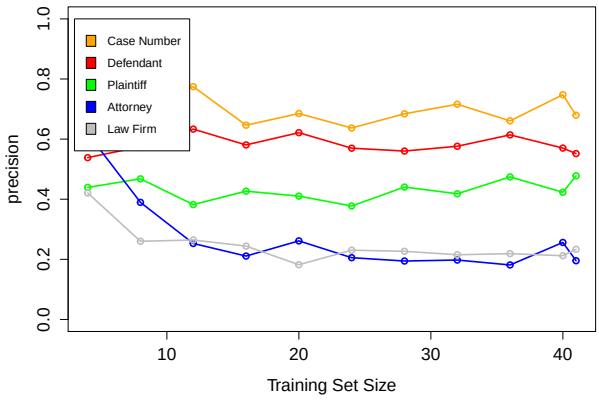

Figure 2: Precision in entity recognition as a function of the number of documents the CRF model has been trained on. Mean of 30 runs, in which documents are chosen in a random order.

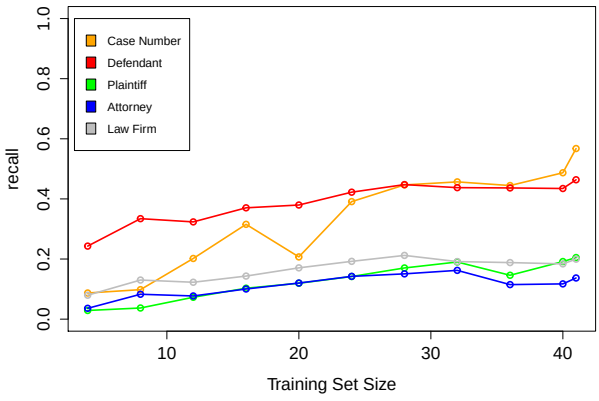

Figure 3: Recall in entity recognition as a function of the number of documents the CRF model has been trained on. Mean of 30 runs, in which documents are chosen in a random order.

lication.[5]

The more general problem of quality control for court documents is less discussed, but remains of tantamount operational importance to the courts. Commercial software packages for automated redaction, ranging from simple regex templates to sophisticated pattern recognition solutions, are available to aid attorneys' compliance with the rules. By contrast, quality assurance of the submitted metadata must currently proceed with little to no automated support.

We demonstrate a promising method of applying statistical text mining tools to the metadata extraction problem. The moderate precision and recall achieved here can likely be substantially improved with more training data and further experimentation with feature representations, etc.

5. ACKNOWLEDGMENTS

This work was sponsored by the Administrative Office of the United States Courts and carried out at the Judicial Engineering and Modernization Center, a federally funded research and development center (FFRDC) operated by The MITRE Corporation. Approved for Public Release; Distribution Unlimited. Case Number 14-2510.

6. REFERENCES

[1] G. Barber. Personal information in government records: Protecting the public interest in privacy. *Louis U. Pub. L. Rev.*, 25:63, 2006.

[2] M. Caughey. Keeping attorneys from trashing identities: Malpractice as backstop protection for clients under the United States Judicial Conference's policy on electronic court records. *Wash. L. Rev.*, 79:407, 2004.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[4] N. Japkowicz and M. Shah. *Evaluating Learning Algorithms*. Cambridge University Press, 2011.

[5] R. W. Kennard and L. A. Stone. Computer aided design of experiments. *Technometrics*, 11(1):137–148, 1969.

[6] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceeding of the 18th International Conference on Machine Learning*, pages 282–289, 2001.

[7] J. T. Matthias. E-filing expansion in state, local, and federal courts 2007. In C. Flango, C. Campbell, and N. Kauder, editors, *Future Trends in State Courts 2007*. National Center for State Courts, Williamsburg, VA, 2007.

[8] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning*, pages 188–191. Association for Computational Linguistics, 2003.

[9] D. Schanker. E-filing in state appellate courts: An appraisal. Technical report, National Conference of Apellate Court Clerks, Williamsburg, VA, February 2010.

[10] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning*, volume 4, pages 142–147. Association for Computational Linguistics, 2003.

[11] M. Whiteman. Appellate court briefs on the web: Electronic dynamos or legal quagmire? *Law library journal*, 97:467, 2005.

[12] P. A. Winn. Online court records: Balancing judicial accountability and privacy in an age of electronic information. *Wash. L. Rev.*, 79:307, 2004.

[13] R. Winters. Controversy and compromise on the way to electronic filing. In T. Peters, N. Kauder, C. Campbell, and C. Flango, editors, *Future Trends in State Courts 2005*. National Center for State Courts, Williamsburg, VA, 2007.

[5]The latest version of the federal judiciary's privacy policy is described at `http://www.uscourts.gov/RulesAndPolicies/JudiciaryPrivacyPolicy.aspx` (Accessed 22 June, 2014).