An eRulemaking Corpus:
Identifying Substantive Issues in Public Comments

Claire Cardie, Cynthia Farina, Matt Rawding, and Adil Aijaz

Faculty of Computing and Information Science, Law School, Information Science, Dept. of Computer Science
Cornell University, Ithaca, NY 14850, USA
cardie@cornell.edu, cynthia-farina@lawschool.cornell.edu, mdr36@cornell.edu, aa362@cornell.edu

Abstract

We describe the creation of a corpus that supports a real-world hierarchical text categorization task in the domain of electronic rulemaking (eRulemaking). Features of the task and of the eRulemaking domain engender both a non-traditional text categorization corpus and a correspondingly difficult machine learning task. Interannotator agreement results are presented for a group of six annotators. We also briefly describe the results of experiments that apply standard and hierarchical text categorization techniques to the eRulemaking data sets. The corpus is the first in a series of related sentence-level text categorization corpora to be developed in the eRulemaking domain.

1. Introduction

Each year regulatory agencies in the U.S. issue more than 4,000 new rules (Kerwin, 2005). By law, many of these must be created through a complex process known as *notice and comment (N&C) rulemaking*: the agency (e.g., the Department of Commerce or Department of Transportation) drafts a proposed rule and then exposes the proposal, any underlying data, and its legal and policy rationale to public comment. The agency's fundamental legal obligation, in turn, is to review all the comments received and, if it chooses to adopt the proposed rule, to issue a statement that responds to significant criticisms made in the comments and explains why it rejected alternative suggested approaches (Strauss et al., 2003).

Electronic rulemaking (eRulemaking) refers to the use of information technology to support any step in the rulemaking process. For example, comments on proposed rules can now be submitted electronically through the use of government-wide web portals like 'regulations.gov'. However, an agency may receive anywhere from dozens, to hundreds of thousands, of comments for a particular rule, and the analysis of the comments to determine the issues each raises is still performed manually in an often slow and painful "issue annotation" process: agency analysts identify each snippet of text in the comments that addresses an important issue and organize the snippets in in terms of these issues. Issue annotation of the comment precedes in-depth comment analysis and facilitates the writing of the response.

This paper presents a corpus developed to study the issue categorization task in eRulemaking. We describe the creation and annotation of the corpus, focusing on those characteristics of the task and of the eRulemaking domain that engender both a non-traditional text categorization corpus and a correspondingly difficult machine learning task. We then present and discuss the results of an interannotator agreement study for a group of six annotators.

Finally, we summarize related work, in which we use the corpus to train a text categorization system for issue categorization of public comments: given the comments submitted for a proposed rule, the automated system determines for each sentence in each comment, which of a set of predefined issues of substance it raises, if any (Cardie et al., 2008).

This corpus is the first in a series of sentence-level text categorization corpora to be developed by the Cornell eRulemaking Initiative (CeRI).

2. Related Work

In recent years, researchers have begun to investigate a range of methods from natural language processing, information retrieval, and machine learning for a number of e-rulemaking sub-tasks. Yang & Callan (2005; 2006), for example, extend duplicate detection methods from information retrieval to handle "e-postcard campaigns" — e-mail campaigns organized by special interest groups that supply constituents with electronic form letters for submission during the comment period. Kwon et al. (2006; 2007) investigate the use of natural language processing methods to identify the main claims of a comment and then categorize them according to whether they support the proposed rule, oppose the proposed rule, or are proposing a new idea.

Most relevant for the current paper is the work of Kwon et al. (2006; 2007) on topic categorization of public comments. Our work differs from theirs in that we categorize sentences in public comments according to a large set of rule-specific issues rather than a small set of general topics. We also investigate hierarchical categorization techniques in addition to standard flat text categorization methods. Finally, we create a corpus to use as a gold standard for the evaluation of automatic text categorization techniques. The creation of the corpus is described in the following section.

3. Corpus Creation

Working with analysts from the Federal Transit Authority (FTA) in the Department of Transportation, we identified two interlinked sets of comments, both involving a group of guidance "circulars" the agency proposed to issue. Such circulars are a type of document on which the FTA frequently seeks public comments. Here, the proposed advice involved grants under three federal statutes that fund

funding

JARC

planning

procedural

evaluation

AdminExpen
FundAppor
FundTransf
GEN_EligActiv
EligGrantees
TechAsstTrain
MobilMgt
DesRecip
Match
HowPlansRelate
GOMB
None

JARCPriors
JARC_EligActiv
JARCMisc

CompSelect
FairNEquit

NoFedEval
Perf_1_Effic
Perf_2_Effect
Perf_3_Satis
Cost
EvalNOverMisc

DefinNew
BeyondADA
AndOr
CompParatrans
OtherNFreeElig
NFreeMisc

PlanElements
LeadAg
OutreachEff
StakeholderParticip
AgPartnerParticip
TransProvidrParticip
MultPartic
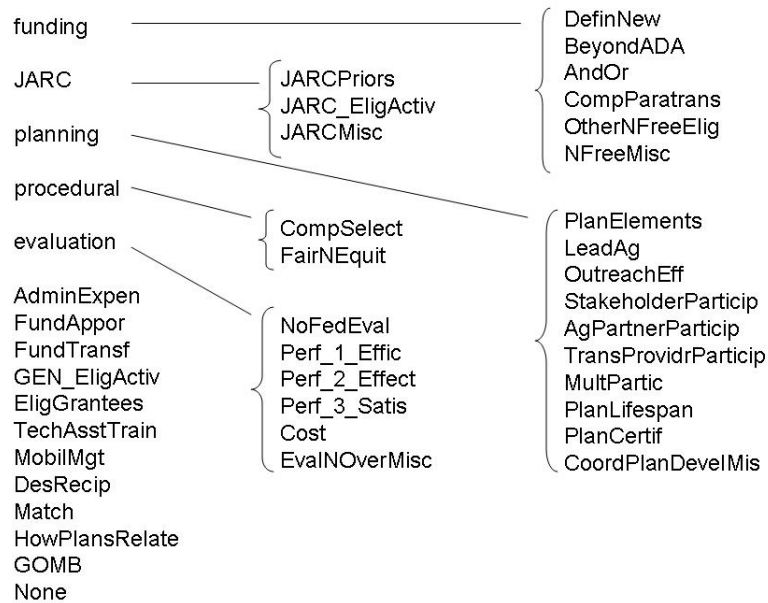PlanLifespan
PlanCertif
CoordPlanDevelMis

Figure 1: Issue Hierarchy. There are 17 top-level categories and 39 leaf categories.

local transportation services for the elderly, disabled persons, and low income persons commuting to work.[1] FTA had been seeking public input at several stages of developing this guidance.

We used comments from the final two comment periods: March 15–May 22, 2006[2] and September 6–November 6, 2006[3]. Based on the judgment of the agency official primarily responsible that comments from both periods raised the same issues, we treat them as a single set. A total of 290 comments were submitted (211 + 79). Many of the comments were not submitted electronically. When scanned by the agency, several became image-based PDFs that could not be converted to machine-readable form. Also, some commentors filed comments with identical text; we retained only a single version of such duplicate comments. As a result of these adjustments, we were left with 267 comments. These comprise the CeRI FTA Grant Circulars Corpus.

Next, we constructed a list of 38 issues likely to be raised in the comments. This list was derived by consulting both the actual issue summaries prepared by the FTA analyst when she reviewed the comments, and the Federal Register notice seeking comments, which explained the proposal in detail and highlighted various aspects. The issues are organized into a shallow categorization hierarchy in which the 38 issues are leaf nodes. Seventeen form the first level; five of these expand into two or more sub-issues at level two. The issue hierarchy, expressed in the abbreviated form used

within the annotation tool[4], is shown in Figure 1. NONE is a special category (shown as the 39th "issue"). It is automatically assigned to sentences deemed by the annotator to address none of the rule-specific issues.

The annotation team comprised six law students in their final year of study. They were deliberately selected because of their general academic performance and, particularly, their work with Farina in a course on the federal regulatory process. However, none of the annotation team, nor anyone else involved in the project, had expertise in the substantive areas or regulatory programs involved in the guidance.

After an initial three-week training period in which all students annotated the same comments and then discussed their selections as a group, they began annotation. Sporadic follow-up discussion occurred throughout the annotation period about the meaning and/or scope of specific issues, with clarifying information then being circulated to the entire group. The students annotated comments according to the 39 fine-grained issues.

The annotation tool allows for annotation at the word-, phrase-, sentence-, or paragraph- level. After an initial period of individual annotator discretion, it was determined that annotation would occur at the sentence level. As a result, all issue annotations are automatically projected to sentences. In addition, the fine-grained issue annotations can be converted to their corresponding top-level issue as needed for any of our analyses. Finally, any sentences the student annotator left unmarked are automatically assigned the label NONE. Annotators were free to assign more than one issue to a single span of text. Multiple annotations, however, were rare (4% of sentences in the corpus).

In all, there are 11,094 sentences in the corpus. On average, there are 41.55 sentences per comment. The shortest

---

[1]Docket No. FTA-2006-24037: Elderly Individuals and Individuals With Disabilities, Job Access and Reverse Commute, and New Freedom Programs: Coordinated Public Planning Guidance for FY 2007 and Proposed Circulars.

[2]FTA-2006-24037-002.

[3]FTA-2007-24037-0222.

[4]Mitre's Callisto.

comment has one sentence; the largest has 1420 sentences.

4. Features of the Public Comment Categorization Task

Formulation of the comment categorization problem as a text categorization task for both people and machines raises a number of non-standard and/or difficult issues for text categorization algorithms. We enumerate and discuss these below.

Sentence-level Categorization. Although most text categorization tasks make decisions on entire documents, issues in the e-rulemaking domain are expressed, and annotated, at the sentence level or below. This is problematic for automated methods because categorization of short texts is known to be quite a bit harder than categorization of longer texts (Hatzivassiloglou et al., 1999; Zelikovitz and Hirsh, 2000; Sahami and Heilman, 2006).

Fairly Large, Hierarchical Issue Set. Proposed rules, guidance and other documents that generate a sufficient amount of public comment to warrant the help of automatic issue categorization almost invariably raise a large number of issues. The 38-issue list we used for this corpus appears to be within the range we expect in future corpora. Hence, the e-rulemaking domain will typically present a large multi-class text categorization problem, which is generally more difficult than a binary classification problem. For one thing, because of the substantial skew in frequency with which issues are discussed (see below), insufficient numbers of training examples are likely to occur for some issues. In addition, at least some portion of the issues is likely to be hierarchically related. As discussed in the next section, the hierarchical nature of categories can both help and complicate the process of training accurate text classifiers (see, e.g., Dumais and Chen (2000)).

Multiple Issues per Sentence. Typically, the lengthier comments submitted to the agency are written by lawyers or other persons well-experienced in the legal and/or substantive regulatory domain. They tend to contain long, complex sentences. These stylistically dense sentences may also be packed with meaning, and so may be annotated with multiple issues. Handling such sentences might call for (1) phrase-, rather than sentence-, level annotation (by both the human annotators and the text categorization algorithms); (2) expansion of the issue set to include new labels that cover multiple issues; or (3) changes in the text categorization algorithm. Yet any of these would likely cause a corresponding drop in performance.

The NONE Category. The NONE category is likely to be difficult for the machine learning algorithms in part because the associated comment sections can cover a wide variety of topics. Commentors often raise a variety of points for or against the proposal or the entire process about which they feel strongly but which the agency does not consider germane. Some of these non-germane topics will appear frequently and predictably; but many will be random and unpredictable.

Multiple Gold Standards There are at least three types of gold standards one could generate from public comment

issue categorization corpora like the FTA Grant Circulars Corpus. The first is an "aggregate" gold standard comprised of comments whose annotations have been reconciled by a pair of annotators. The second type would more closely approximate what we understand, from our agency partners, to be real-world agency practice. When more than one analyst reviews a comment set to find, extract, and organize the issue references for subsequent analysis and preparation of the accompanying final statement, these analysts typically divide the issues among themselves: each reads all the comments, taking responsibility for collecting material as to his or her allotted issues. As a result, there typically is not more than one "annotator" per issue in the real-world. The gold standard under this annotation scheme would then be the union of the issue-specific annotations of each analyst.

We have adopted yet a third strategy for creating a gold standard for the purposes of this paper. In particular, we are interested in investigating the ability of the text categorization algorithms to learn to duplicate the annotations produced by an arbitrary agency analyst. As a result, we treat the annotations of each annotator as a separate gold standard, producing six separate corpora.

Skewed Distribution Across Issues. The FTA Grant Circulars data exhibits substantial skew in terms of the distribution of sentences that address each issue, further complicating the learning task. Table 1 shows the distribution of issue annotations across sentences in the aggregate gold standard described just above. No rule-specific issue (NONE) was selected for fully 51% of the 11696 sentences in the gold standard. No other first-level or second-level category approaches this level of coverage. Discounting NONE, distribution of the remaining 16 top-level issues is still problematic, with coverage ranging from 0.2% (32 sentences) for GEN_ELIGACTIV to 10.2% (1193 sentences) for PLANNING. Our agency partners indicate that this is standard for most rulemakings.

Our only attempt in the current work to deal with the skewed category distribution is to treat NONE as a special category in the hierarchical categorization algorithm (see Section 6.).

Domain Knowledge Slippage. Proposed rules, guidance and other documents on which agencies seek public comment often deal with issues that cannot be adequately understood without fairly sophisticated legal, scientific and/or technical knowledge. We believe the extraordinary demands for domain expertise posed by these kinds of text may introduce a real, but difficult to estimate, degree of confusion among non-expert annotators when an aggregate gold standard is used. Even after their initial period of training and group annotation, the upper-level law students annotating the FTA Grant Circulars Corpus struggled to establish nuances of meaning, as well as the precise scope, of many of the 38 issues. Further exacerbating these direct consequences of the lack of domain knowledge, many of the commentors were, like the agency, well-acquainted with the statutes, programs and policies involved. This shared knowledge enabled them to shortcut formal references and explanations that would have helped non-experts

Issue	Coverage (%)
funding	8.7
DefinNew	1.1
BeyondADA	1.8
AndOr	1.5
CompParatrans	0.6
OtherNFreeElig	3.2
NFreeMisc	0.5
JARC	4.6
JARCPriors	0.5
JARC_EligActiv	3.9
JARCMisc	0.2
planning	10.2
PlanElements	2.1
LeadAg	0.6
OutreachEff	1.3
StakeholderParticip	0.5
AgPartnerParticip	0.7
TransProvidrParticip	2.0
MultPartic	0.7
PlanLifespan	0.7
PlanCertif	0.3
CoordPlanDevelMisc	1.3
procedural	6.2
CompSelect	6.0
FairNEquit	0.2
evaluation	6.0
NoFedEval	0.3
Perf_1_Effic	1.1
Perf_2_Effect	0.8
Perf_3_Satis	0.8
Cost	0.2
EvalNOverMisc	2.8
AdminExpen	1.2
FundAppor	0.9
FundTransf	0.6
GEN_EligActiv	0.2
EligGrantees	0.3
TechAsstTrain	1.1
MobilMgt	1.5
DesRecip	2.8
Match	0.5
HowPlansRelate	2.8
GOMB	0.9
NONE	51.0

Table 1: Issue Distribution. Table shows the percentage of sentences (in the aggregate gold standard) that are labeled with each rule-specific issue.

	Agreement (%)	Coverage (%)
39 issues	46.4	100
17 top-level issues	64.7	100
5 hierarchical issues	69.3	35.7
5 hierarchical issues plus NONE	68.4	86.7

Table 2: Interannotator agreement scores when annotating w.r.t. different subsets of the hierarchical issue set. The table also shows the percentage of sentences in the CeRI FTA Circular Corpus that each issue set covers.

all of the substantive issues that will arise in the comments even before the comments begin to arrive. Oftentimes, the proposed rule itself lays out the set of issues that the agency would like feedback on. Unexpected issues, however, sometimes arise, and existing issues might need to be further subdivided during the annotation process. We have ignored these complications in our current study.

Variation in Comment Quality, Scope and Form. Since comments are posted by entities ranging from law firms and trade or professional associations — both of which tend to have expertise in the area of the proposed rule — to relatively non-expert members of the public, the comments themselves vary in their clarity and their use of legal and technical terminology.

Knowledge Transfer Across Rulemakings. For text categorization techniques to be a feasible solution for rule-specific issue categorization, the amount of manually annotated training data (i.e., comments annotated by the rulewriters and analysts themselves) should be kept to a minimum. For this reason, text categorization methods that allow for inductive transfer across related rulemakings will need to be employed and developed (Silver et al., 2005) so that new rulemakings can benefit from previous rulemakings. We have also left this issue for future work.

5. Interannotator Agreement Results

This section present the results of an interannotator agreement study for which we randomly selected 146 of the 267 available comments. Each of the selected comments had an average of 2.66 annotators. Because there can be multiple issues per sentence and the annotators covered different numbers and subsets of the documents, we currently measure interannotator agreement using a basic agreement (AGR) measure (rather than Fleiss' kappa)[6]: for all pairs of annotators across all comments that were annotated by both annotators, we calculate the percentage of sentences for which the annotators assign overlapping issue labels. In most cases, this amounts to checking for an exact issue match (since 96% of the sentences are assigned a single issue). Table 2 shows the AGR score across all pairs of annotators for the full set of 39 issues, the top-level of the issue hierarchy (17 issues), the five hierarchical issues, and the five hierarchical issues plus NONE. Along with the AGR

make categorization decisions.[5] Thus it is likely to be very difficult to obtain training sets with high levels of agreement across large issue sets for these kinds of texts using student or other non-expert annotators. We currently do not try to identify or correct for domain knowledge slippage.

Dynamically Changing Issue Set. According to our agency collaborators, their analysts can determine virtually

---

[5]Such "repeat players" are a feature of virtually every rulemaking and typically write the longest, most issue-laden and — according to agency rulewriters — "useful" comments.

[6]In current work, we have moved to the more reliable Cohen's and Fleiss' kappa for measuring interannotator agreement (Krippendorff, 2004).
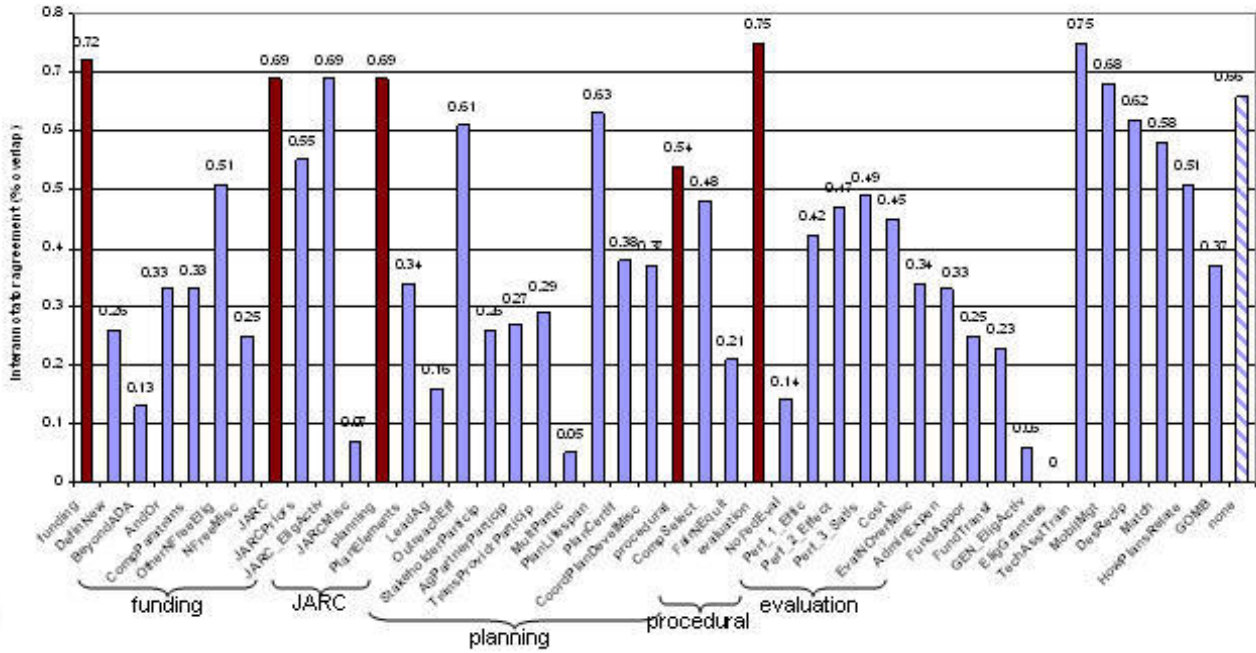
Figure 2: Issue-by-Issue Interannotator Agreement Results.

scores, we show the coverage of each issue set across all sentences of the corpus.

When calculated across the full set of 39 issues (38 issues plus NONE), interannotator agreement scores are quite low (see row 1 of the table), indicating either that more training is required or that there is inherent difficulty in interpreting the meaning and applicability of each issue. The latter possibility was raised in the Section 4.

Annotation according to just the 17 top-level issues (row 2) ameliorates the problem to some degree — agreement increases to 64.7% across all sentences in the corpus. Even higher levels of agreement (69.3%) can be obtained if annotation is limited to just the five hierarchical issues at the top-level (row 3) although this issue subset covers only 35.7% of the sentences in the corpus. Annotating these five issues as well as NONE's, however, allows for agreement scores approaching 70% and sentence coverage of 86.7%.[7]

Figure 2 shows the interannotator agreement scores on an issue-by-issue basis. It is clear from the chart that there is wide variation in annotators' abilities to identify various issues. Not surprisingly, however, the agreement scores for the five hierarchical categories are among the best that we encounter for this task (these are the dark bars in the Figure): for four of five hierarchical issues, agreement scores are at 69% or above, and in all but one case, the scores for the hierarchical issue are substantially higher than those of each of its associated sub-issues. Many of the top-level non-hierarchical issues also score reasonably well (see the righthandside of the chart), indicating that automatic text categorization methods might be able to achieve reasonable levels of accuracy for an important subset of the issues.

---

[7]Coverage is measured on the "aggregate gold standard" described in Section 4.

6. Text Categorization Experiments

In spite of the rather low interannotator agreement scores and the difficulties raised in Section 4., we have made progress in applying text categorization techniques to the CeRI FTA Circular Corpus. We investigate *flat* and *hierarchical* text categorization methods, where *flat* refers to direct categorization of sentences according to the 39 issues; *hierarchical* categorization exploits the issue hierarchy in an attempt to make the assignment of issues to sentences easier for the learning algorithms. For hierarchical classification, we train multiple classifiers, one for each internal node in the issue hierarchy of Figure 1. In addition, we found that treating the NONE class specially improves performance.

none-classifier: distinguishes sentences that address NONE of the issues from those that address *some* issue.

level-1 classifier: distinguishes among the remaining 16 top-level issues (i.e., excluding NONE)

level-2 classifiers: one classifier is created for each of the five hierarchical classes from level-1 (FUNDING, JARC, PLANNING, PROCEDURAL, and EVALUATION) to distinguish among its leaf classes.

Test sentences are processed by first applying the none-classifier. If the sentence is deemed non-NONE, then the level-1 and possibly a level-2 classifier is applied depending on the issue specificity required. This approach is similar in spirit to methods employed in previous work (Dumais and Chen, 2000; Koller and Sahami, 1997) although we do not try to reduce the feature set size.

	Macro averages across annotators
flat categorization 39 issues	0.48
hierarchical categorization 39 issues	0.45
flat categorization 17 level-1 issues	0.61
hierarchical categorization 17 level-1 issues	0.59

Table 3: Summary of Flat and Hierarchical Categorization Results. Results are 5-fold cross-validation accuracies.

Following the real-world eRulemaking setting that we are trying to emulate, we create six gold standards, one for each annotator. During training and testing, we treat sentences with multiple issues as separate instances, one for each assigned issue. As a result, we will get at most one of the alternative instances correct in the test data. We investigated SVMs, naive Bayes, and CRFs under a variety of parameter settings and using 5-fold cross-validation. We report here only the results for SVMs. Additional experiment details can be found in Cardie et al. (2008).

6.1. Results

Results are summarized in Table 3. These represent macro-averages of 5-of-cross-validation results across three separate gold standards selected randomly from the set of six. As in most previous research on hierarchical classification, we also find that flat categorization consistently outperforms hierarchical categorization (see Dumais and Chen (2000) for a discussion). For both flat and hierarchical categorization, the learning algorithms significantly outperform random predictions, which achieve accuracies of 5.9% for 17 issues, and 2.6% for 39 issues. Our results outperform a classifier that always selects the most frequent issue that appears in the training set.[8] Moreover, performance approaches our current interannotator agreement results, a promising indication that improvements in individual and inter-rater reliability in the training data will produce similar gains for automated text categorization techniques. For additional details, see Cardie et al. (2008).

7. Conclusion

We have presented a a new e-rulemaking corpus — the CeRI FTA Circular Corpus. The corpus is manually annotated at the sentence level according to rule-specific issues in an attempt to duplicate the current practices of agency analysts. We discuss a number of task-specific complications associated with the creation of the corpus, and report on an initial use of the corpus to train flat and hierarchical text categorization classifiers to predict the issue discussed in each sentence of public comments submitted in response to a U.S. Federal Transit Authority circular on

planned transportation initiatives for individuals with disabilities.

Given our promising results, we will continue to investigate text categorization and related methods that have the potential to reduce the time devoted to repetitive text analysis tasks in the eRulemaking domain. In particular, we are creating similar sentence-level text categorization corpora for additional rules and plan to compare the rule-specific issue annotations of our law student annotators with those of agency rulewriters and analysts. As part of our ongoing investigations, we have also applied active learning methods (e.g., Cohn et al. (1994), Freund et al. (1997)) to the CeRI FTA Circular Corpus in an attempt to further minimize the amount of manual effort required to analyze public comments in the e-rulemaking domain (Purpura et al., 2008).

8. Acknowledgments

9. References

Claire Cardie, Cynthia Farina, Adil Aijaz, Matt Rawding, and Stephen Purpura. 2008. A Study in Rule-Specific Issue Categorization for e-Rulemaking. In *Proceedings of the Ninth Annual International Conference on Digital Government Research*. to appear.

D. Cohn, L. Atlas, and R. Ladner. 1994. Improving generalization with active learning. *Machine Learning*, 15(2):201–221.

Susan Dumais and Hao Chen. 2000. Hierarchical classification of web content. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 256–263, New York, NY, USA. ACM.

Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. 1997. Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133–168.

V. Hatzivassiloglou, J. Klavans, and E. Eskin. 1999. Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*, pages 203–212, University of Maryland, College Park, MD. Association for Computational Linguistics.

C. Kerwin. 2005. The state of rulemaking in the federal government. Technical report, Transcript Panel 1.

Daphne Koller and Mehran Sahami. 1997. Hierarchically classifying documents using very few words. In Douglas H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 170–178, Nashville, US. Morgan Kaufmann Publishers, San Francisco, US.

K. Krippendorff. 2004. *Content analysis: An introduction to its methodology (2nd Ed.)*. Sage Publications.

Namhee Kwon and Eduard Hovy. 2007. Information acquisition using multiple classifications. In *Proceedings of the Fourth International Conference on Knowledge Capture (K-CAP 2007)*.

---

[8]The performance of the most frequent issue baseline varies for each of the three annotation sets — from 24% accuracy for annotator3 to 35% for annotator1 and annotator2.

Namhee Kwon, Eduard Hovy, and Stuart Shulman. 2006. Multidimensional text analysis for erulemaking. In *Proceedings of the 7th Annual International Conference on Digital Government Research*.

Stephen Purpura, Claire Cardie, and Jesse Simons. 2008. Active Learning for e-Rulemaking: Public Comment Categorization. In *Proceedings of the Ninth Annual International Conference on Digital Government Research*. to appear.

Mehran Sahami and Timothy D. Heilman. 2006. A web-based kernel function for measuring the similarity of short text snippets. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 377–386, New York, NY, USA. ACM.

Danny Silver, Goekhan Bakir, Kristin Bennett, Rich Caruana, Massimiliano Pontil, Stuart Russell, and Prasad Tadepalli. 2005. *Inductive Transfer : 10 Years Later*. NIPS 2005 Workshop.

P. Strauss, T. Rakoff, and C. Farina. 2003. *Administrative Law*. 10th edition.

H. Yang and J. Callan. 2005. Near-duplicate detection for erulemaking. In *Proceedings of the Fifth National Conference on Digital Government Research*.

H. Yang and J. Callan. 2006. Near-duplicate detection by instance-level constrained clustering. In *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Sarah Zelikovitz and Haym Hirsh. 2000. Improving short text classification using unlabeled background knowledge. In Pat Langley, editor, *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 1183–1190, Stanford, US. Morgan Kaufmann Publishers, San Francisco, US.