The U.S. Policy Agenda Legislation Corpus Volume 1 - a Language Resource from 1947 - 1998

Stephen Purpura, John Wilkerson, Dustin Hillard

Information Science, Dept. of Political Science, Dept. of Electrical Engineering
Cornell University, University of Washington, University of Washington
sp559@cs.cornell.edu, {jwilker, hillard}@u.washington.edu

Abstract

We introduce the corpus of United States Congressional bills from 1947 to 1998 for use by language research communities. The U.S. Policy Agenda Legislation Corpus Volume 1 (USPALCV1) includes more than 375,000 legislative bills annotated with a hierarchical policy area category. The human annotations in USPALCV1 have been reliably applied over time to enable social science analysis of legislative trends. The corpus is a member of an emerging family of corpora that are annotated by policy area to enable comparative parallel trend recognition across countries and domains (legislation, political speeches, newswire articles, budgetary expenditures, web sites, etc.). This paper describes the origins of the corpus, its creation, ways to access it, design criteria, and an analysis with common supervised machine learning methods. The use of machine learning methods establishes a baseline proposed modeling for the topic classification of legal documents.

1. Introduction

In this paper, we introduce the corpus of United States Congressional bills for use as a language resource. For each of approximately 375,000 bills offered as legislation from 1947 to 1998, the corpus contains the title and/or a short description of the bill, its sponsor, and its progress through the legislative process, along with other substantive details. In addition, the corpus has been manually annotated according to a two-level hierarchical topic categorization scheme (known as the Policy Agenda Annotation Scheme) that covers 20 major topics and 226 fine-grained topics.

After placing the work in context with related work, this paper describes the corpus and its creation, and reports inter-annotator agreement results. High inter-annotator agreement levels have been achieved: 0.9 and 0.8 Kappa values for the major topic and sub-topic hierarchy levels, respectively. To facilitate a discussion about the unique aspects of the corpus, we apply a collection of standard automated text categorization techniques to the corpus to predict both the major topic and subtopic associated with each bill. These initial benchmark experiments show that automated techniques are able to achieve performance similar to human annotators. We next discuss the phenomenon of "topic drift" that can occur for corpora, like the United States Congressional bills corpus, that are created and extended over a long period of time. Finally, we investigate active learning as a semi-automated strategy for combating topic drift in temporally grounded on-line corpora.

2. Related Work

For decades, language researchers and information scientists have constructed test corpora (Robertson and Walker, 1997) in (MacMullen, 2003). These collections usually consist of documents (titles, abstracts, or full-text articles), a set of standardized queries made by experts and relevance judgments (MacMullen, 2003). Examples of test corpora include the TREC data sets, Reuters RCV1 (Rose et al., 2002) and, more recently, Claire Cardie, Cynthia Farina, Matt Rawding, Adil Aijaz, and Stephen Purpura (2008).[1]

Using these prior works as a guide, this work describes the creation of a test corpora which includes the titles of all bills introduced in the United States Congress during a 50 year period. Each bill title has been labeled with a mutually exclusive relevance judgment so that queries can easily be constructed and tested. The queries are derived from the specified topic annotation scheme. An example query is: "'Produce a list of all of the environmental legislation introduced from 1970 through 1998."' In addition, bills have been marked to identify examples of topic drift because managing topic drift is a critical problem which human and machine learning systems must address for corpora with temporal consistency concerns. Together, these attributes make this resource a unique reference data set.

3. The Motivation for Corpus Creation

One of the systemic outputs of the United States Congress is proposed legislation. Congressional bills are recorded by the Library of Congress and researchers examine them to study legislative trends over time as well as to explore finer questions, such as the substance of environmental bills introduced in 1968 or the characteristics of the sponsors of environmental legislation.

Each bill is identified by a unique bill number, which is assigned sequentially as a bill is introduced on the floor of the Congress. In recent years, the rich history of a bill can be examined via the Internet in the THOMAS system (http://thomas.loc.gov/). THOMAS includes a topic indexing system called LIV (Legislative Indexing Vocabulary).[2] While the LIV enables topical search in THOMAS, it is often insufficient for social science research because its con-

[1] While this work does not attempt to align itself with the expanding research on ontologies, the authors recognize that future research could adapt this test corpora into an ontology.

[2] THOMAS was developed by Bruce Croft and Robert Cook with assistance from Dean Wilder. A description of it can be found at http://csdl.tamu.edu/DL95/papers/croft/croft.html

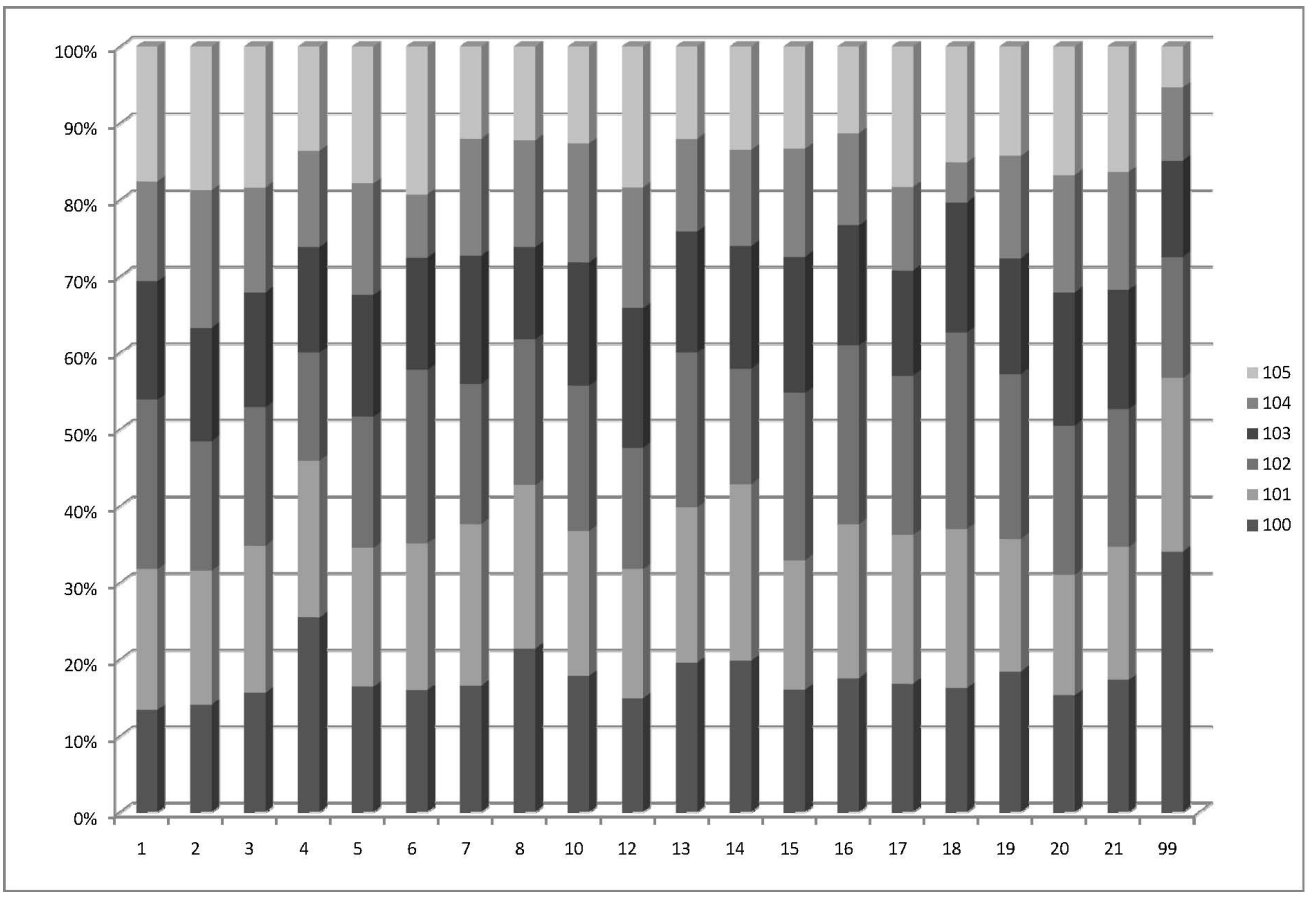

Figure 1: The Percentage of Bills per Congressional Session versus Major Category. Each change in shading is a 2 year duration Congressional session from the 100th through the 105th Congresses. For example, the trend in the decline of private bills (topic 99) over the 12 year period from 1987 - 1998 can be determined from the stacked vertical bar on the far right of the graph.

temporary focus exhibits two problems of "topic drift," or the assignment of similar events to different topics as users' conceptions of what those events are about changes. For example, consider how difficult it would be for an organization to compare budget data if the definitions of expenditure classes changed annually. Unless the changes to definitions were readily apparent, it would be impossible to compare the amount of money spent on, say, child welfare between one period and the next. The shifting definitions that classify the expenditures in different categories might lead us to believe that wild shifts occurred in Congressional appropriations from year-to-year.

To the social science researcher, the benefits of maintaining inter-temporal reliability with a topic coding scheme are significant because they help avoid confusion and save time searching for related material.[3] People who believe in the use of ontologies for the standardization of semantic web services might see the parallel between defining the semantics, or intended meaning, of a category across time. Adler and Wilkerson (2008) use the Congressional Bills Project database to study the impact of congressional reforms. To accomplish this, they needed to trace the impact of changes in a specific set of congressional committee reforms. The reforms altered bill referrals within a specific set of issue jurisdictions. Had Adler and Wilkerson attempted to use the LIV system to search for environmental legislation, they would have had to individually inspect about 100,000 bills identified as related to 'environmental legislation'. Instead, the fact that all of the bills during the years of interest had already been annotated according to the Policy Agenda Annotation Scheme's topic categories allowed them to reduce the number of bills that needed to be individually inspected from about 100,000 to 'just' 8,000.

THOMAS' LIV indexing system is not the only search system which exhibits this problem. Lexis-Nexis' legislative topic indexing system has the same problems. The result is that (when using these systems) the researcher must expend significant effort constructing many queries to find documents, and these methods are not considered reliable for distinguishing complex events. A keyword search that is too narrow in scope (e.g. "renewable energy") will omit relevant events ("solar"), while one that is too broad (e.g. "energy") will generate unwanted "false positives" ("refineries"). Although it is theoretically possible to create sufficiently discriminating keyword search commands, to date, human-centered annotation practices are preferred in many

[3]See Adler, E. Scott and John Wilkerson, Congressional Bills Project: 1947-1998, NSF 00880066 and 00880061. www.congressionalbills.org

Category	Description
1	Macroeconomics
2	Civil Rights, Minority Issues, Civil Liberties
3	Health
4	Agriculture
5	Labor, Employment, and Immigration
6	Education
7	Environment
8	Energy
10	Transportation
12	Law, Crime, and Family Issues
13	Social Welfare
14	Community Development and Housing Issues
15	Banking, Finance, Domestic Commerce
16	Defense
17	Space, Science, Technology, Communications
18	Foreign Trade
19	International Affairs and Foreign Aid
20	Government Operations
21	Public Lands and Water Management
99	Private Legislation

Table 1: The Major Topics of the Congressional Bills Project

Category	Description
300	General
301	Comprehensive health care reform
302	Insurance reform, availability, and cost
321	Regulation of drug industry, medical devices, and clinical labs
322	Facilities construction, regulation, and payments
323	Provider and insurer payment and regulation
324	Medical liability, fraud and abuse
325	Health Manpower & Training
331	Prevention, communicable diseases and health promotion
332	Infants and children
333	Mental health and mental retardation
334	Long-term care, home health, terminally ill, and rehabilitation services
335	Prescription drug coverage and costs
336	Other or multiple benefits and procedures
341	Tobacco Abuse, Treatment, and Education
342	Alcohol Abuse and Treatment
343	Controlled and Illegal Drug Abuse, Treatment, and Education
344	Drug and Alcohol or Substance Abuse Treatment
398	Research and development
399	Other

Table 2: The Subtopics of the Health Major Topic (3)

situations because humans can better appreciate the context in which words are used.

4. Corpus Creation

The problems associated with reliably searching for and classifying government documents for social research led to the creation of the Policy Agendas project[4] and the Congressional Bills project. The Congressional Bills Project received funding from the National Science Foundation to work with the Library of Congress to make available in electronic form information about federal public and private bills introduced since 1947.

In addition to information available form THOMAS, each bill has been annotated, by hand, with a topic code from the Policy Agendas Annotation Scheme. This scheme assigns a mutually exclusive, hierarchical classification. Table 1 lists the 20 major topics of this system. Each major topic has additional partitions, for a total of 226 subtopics. For example, topic 3 (health) includes 20 subtopics which are listed in Table 2. Another example is topic 7 (environment) which includes 12 subtopics such as 'species and forest protection,' 'recycling,' and 'drinking water safety.'[5] It is important to emphasize that this scheme partitions the legislative agenda by issue area rather than by program. Thus, the subject categories remain valid even as programs come and go. Related projects have or are applying the same topic system to executive, judicial, media and public opinion data since WWII, and to U.S. state legislatures, nations in the European Union, and Canada.

When the team annotates each bill, the key focus is topic assignment that assures inter-temporal reliability. Human annotators examine each bill's title (1973-98) or short description (1947-72) and place it into one of the 226 subtopics. Although the human annotation team will refer to the full text of the bill when appropriate, the use of bill titles and short descriptions as a proxy for the entire bill content is practically motivated. It is much less text, and, by the parliamentary rules of the House, the bill title must indicate the primary topic of the legislation. The parliamentary requirements assure that the bill title is suitable for quickly assigning a bill to a committee for consideration and review. In past research, we have verified that the use of the bill title as a proxy for full bill content is reasonable for the purposes of assigning a primary topic with inter-temporal reliability (Hillard et al., 2007).

The annotation teams are supervised by four project directors and many annotation team members have worked on the project over the years. Each is trained using a six week training protocol that begins by annotating 100 bills per week. These 'training bills' have been annotated in the past. After four weeks of this training, the prospective team member is given a test which they must pass with high inter-rater agreement. Many hours of annotation by trained graduate and undergraduate students have been invested in the project, with observed inter-rater agreement of Cohen's Kappa (Cohen, 1968) approaching 0.9 at the major topic level and 0.8 at the subtopic level.

During a decade of human annotation, temporal inconsistencies in the annotation process have been found(Baumgartner et al., 1998). These examples have allowed us to construct test scenarios for observing topic drift within the data set.

5. Data Location on the Web

Since this work is intended to introduce a test corpora, the data extracts, reference queries, and additional supporting

[4] http://www.policyagendas.org/

[5] Additional details about these topic categories and the coding process can be reviewed online at http://www.policyagendas.org/codebooks/topicindex.html/

	SVM	Maxent	Boostexter	Naive Bayes	Ensemble
Major topic N=20	88.7% (.881)	86.5% (.859)	85.6% (.849)	81.4% (.805)	89.0% (.884)
Subtopic N=226	81.0% (.800)	78.3% (.771)	73.6% (.722)	71.9% (.705)	81.0% (.800)

Table 3: Humans versus Machine Agreement for Five Model Types

Congress Train	Congress Test	(1) N of Bills in Test Set	(2) % of Bills Classifiers Agree	(3) % agreement when Classifiers Agree	(4) % agreement when Classifiers Disagree	(5) % agreement Entire Ensemble	(6) % agreement Best Individual Classifier
99th	100th	8508	61.5	89.7	59.3	78.0	78.3
100th	101th	9248	62.1	93.0	61.5	81.1	80.8
101th	102th	9602	62.4	90.3	61.1	79.3	79.3
102th	103th	7879	64.8	90.1	60.2	79.6	79.5
103th	104th	6543	62.4	89.0	57.5	77.1	76.6
104th	105th	7529	60.0	87.4	58.9	76.0	75.6
	Mean	8218	62.2	89.9	59.7	78.5	78.4

Table 4: Machine Learning Prediction Performance when Classifiers Agree and Disagree

documentation are available for download and research use. The data extracts used for machine learning experiments are available at http://www.congressionalbills.org/corpus and http://www.stephenpurpura.com. The extracts are formatted in XML and in a legacy file format which enables import into database programs such as Microsoft Access.
Additionally, the Congressional Bills web site[6] keeps online up-to-date versions of the data sets. The human annotations for the underlying data continuously improve and new Congressional Sessions are added. These improvements will be rolled into a test corpora through new volumes and controlled revisions that will also be linked from http://www.congressionalbills.org/corpus/.

6. Initial Machine Learning Experiments

Prior to Purpura and Hillard (2006), the Congressional Bills team had little confidence that machines could easily learn to replicate human annotations for the Congressional Bills Project. While Purpura and Hillard (2006) demonstrated that machine learning might allow relatively inexpensive replication of the performance of human annotators, it failed to provide a method for the human annotation team to follow for actually applying the machine learning technology while managing error. In this section, we update the experiments of Purpura and Hillard (2006) and elaborate on the challenges for machines to learn to replicate the performance of human annotators in labeling subsequent Congressional legislation.
The goal for a machine learning system is that, given the same input available to humans, a machine learning system should classify a bill into 1 of 226 categories of the Policy Agenda Annotation Scheme. We exploit the natural hierarchy of the categories by first building a classification system to determine the major category, and then building a child system for each of the major categories that decides among the subcategories within the major class that is decided by the first level of classification. This is the simplification approach advocated by Koller and Sahami (1997). Unlike other research, such as Dumais and Chen (2000) and Claire Cardie, Cynthia Farina, Matt Rawding, Adil Aijaz, and Stephen Purpura (2008), which shows that flat classification usually exceeds the performance of hierarchical classification, we note that hierarchical classification was chosen over flat classification after empirical testing demonstrated its advantage when using the same features.

6.1. Text Pre-processing

Input to text categorization systems is usually preprocessed to create word/term vectors for each training and test instance (Salton and McGill, 1983). In addition, the word-based feature vectors are associated with a corresponding weight vector that ascribes a different weight to each word. Before creating word vectors, we remove non-word tokens, map text to lower case, and then apply the Porter Stemming Algorithm described in Porter (1980).
Weighting strategies such as *tf-idf* (i.e. term frequency multiplied by inverse document frequency) have been shown to be generally effective, but specialized weighting schemes often provide improvements (Papineni, 2001). After empirical testing of various weighting schemes on the training data, this work adopts a term weighting strategy related to mutual information, which is the ratio of sentence-based word frequency and the overall frequency of the word across the corpus. Equation 1 for the feature value w_i is shown:

$$w_i = log\left(\frac{p(w|t)}{p(w)}\right) \quad (1)$$

In equation 1, the top term, $p(w|t)$, is the probability of a word in a particular bill title (the number of occurrences in each bill title, divided by the number of total words in the title). The denominator term $p(w)$ is the average probability of a word across all titles (the number of occurrences of this word in all bill titles, divided by the total number of words in all bill titles).
Finally, only words with $w_i > 0$ are included in the bill title-based term vectors.[7]

[6] http://www.congressionalbills.org

[7] The run-svm-text.pl script from Purpura and Hillard (2006)

6.2. Classifiers and Parameters

Existing research indicates that combining the decisions of multiple statistical systems (a.k.a. ensemble learning) usually improves final results (Brill and Wu, 1998; Dietterich, 2000; Curran, 2002). For the ensembles, we employ three modeling approaches that are freely available to the research community: a Support Vector Machine (SVM), a Maximum Entropy classifier, and a boosting classifier. For SVM classification, we use SVMlight(Joachims, 1998); we use the Bow toolkit for Maximum Entropy classification(McCallum, 1996); and the Boostexter tool for the AdaBoost.MH algorithm (Schapire and Singer, 2000). In addition to the classifiers used in the ensemble, we also compare performance of our systems against the performance of the Naive Bayes classifier in the Bow toolkit.

For the experiments here, we did not learn the optimal parameter settings for each classifier based on a validation set. Rather, we ran each algorithm under a number of parameter settings and selected the settings that provided the best performance on a portion of the corpus when the classifier was used in isolation, i.e. not in an ensemble.

To support multi-class classification with SVMlight, we used the run-svm-text.pl script that implements pairwise voting instead of a one vs. the rest voting schemes.

6.3. Discussion and Results

The results of experiments are presented in Table 3, and are based on using 187,000 randomly sampled records to predict 187,000 randomly sampled unlabeled cases.[8] Agreement is computed based on a comparison of predictions of machine to previously assigned predictions of humans. Cohen's Kappa measure is presented in parentheses.

This experiment benefits from a few key aspects of the corpus that are worth noting. As reported in Stephen Purpura, Claire Cardie, and Jesse Simons (2008), 120,927 records of the 375,517 records in the data set are near duplicates. The relatively large number of near duplicates is caused by systemic factors in the United States Congress. First, multiple bills with substantially the same bill title, yet different bill text, may be introduced in the Congress for a variety of reasons. Second, program re-authorizations regularly occur and the titles of these bills intentionally enable legislators to associate the reauthorizations with previous legislation. Third, in the early years of the period covered by the corpus, the Congress artificially limited the number of bill co-sponsors. In their wisdom, legislators realized that they could publicly signal their association as a co-sponsor of the bill by simply reintroducing (largely) the same bill with different co-sponsors.

In addition to the large number of near duplicates, the corpus is sequential in nature. The Policy Agenda projects (which include the Congressional Bills project), always acts in a historical research mode because they annotate instances (bills) after they are introduced. However, the amount of data available at any moment in time is limited because researchers cannot predict into the future. This experiment benefits from using instances from each Congress in the training set. During previous experiments for Hillard et al. (2007), results suggest that accuracy always substantially improves (at least 5%) when predicting the labels of the ith Congress if even a relatively small number of randomly selected instances from the ith Congress are included in the training set. This implies that *some* human annotation for the bills of each Congressional session will yield payoffs in higher accuracy in predicting the class labels of the rest of the bills in any Congressional session.

But from this experiment, our conclusion is that machine learning assistance is promising. With annotated bills from every Congressional period, the agreement between humans and machine is very good.

6.4. Bill Sequencing and Topic Drift

Since our previous experiment does not deal with sequencing or topic drift, in this section we begin to outline more of the known challenges researchers will face when they approach the task of using the Policy Agenda scheme to annotate the bills from previously unseen Congressional sessions. As mentioned in the previous section, since approximately 10,000 bills are introduced in every 2-year duration Congress, new bills will always be available for annotation. These new bills must be labeled sequentially.[9] In addition to dealing with sequencing, topic drift will certainly occur. With the Congressional legislation, topic drift primarily takes two forms. First, the topics covered in the bills during any Congressional session change. Intuitively, this is because national problems rise and fall in priority. Second, the language associated with topics changes over time. This condition is the most dangerous for managing automated labeling reliability and temporal consistency because it can be difficult for people to identify (Soroka et al., 2006; Baumgartner et al., 2002).

While the Policy Agendas annotation scheme is designed to capture the primary topics of legislation in one sense, specific programs come and go. The result can be a problem for machine learning system interested in predicting the correct class label of previously unseen program legislation. An easy example to consider is the introduction of legislation related to the Internet.

In the period 1947 to 1998, forty one bill titles mention the Internet. The first two mentions are almost certainly data entry errors, as they occur during the $85th$ Congress (1965 - 1967). 'Internet' was typed when 'Internal' was intended, as these bills mention changes to the Internal Revenue Service code. As Table 5 shows, the remaining 39 bill titles occur during the $104th$ and $105th$ Congresses (1995 - 1998) and are scattered across major categories.

To a certain degree, the rise and fall of specific legislative topics is as predictable as topic change in Reuters newswire articles. The other words in the bill titles help a human or

performs the pre-processing steps described above and is available for download from www.stephenpurpura.com.

[8] These results are also reported in Hillard et al. (2008) which discusses the general problem of conducting temporally consistent mixed-method social science research with quantitative and qualitative requirements and information retrieval or extraction methods.

[9] Bills must be labeled sequentially in the sense that if we annotate all of the bills prior to today, within the next month or so there will probably be new bills to label. The actual content of these new bills will be unknown, even if somewhat predictable.

Congress	Major Category	Frequency
89	Education	1
89	Space, Science, Technology, Communications	1
104	Civil Rights, Minority Issues, Civil Liberties	2
104	Health	1
104	Transportation	1
104	Government Operations	1
105	Civil Rights, Minority Issues, Civil Liberties	4
105	Education	2
105	Law, Crime, and Family Issues	5
105	Social Welfare	1
105	Banking, Finance, Domestic Commerce	4
105	Space, Science, Technology, Communications	12
105	Government Operations	6

Table 5: Frequency of the term 'Internet' in Bill Titles by Category

a machine learning system place the bill in context with a class. An example bill title from the bills which mention 'Internet' is Senate bill number 2648 from the $105th$ Congress: 'A bill to protect children with respect to the Internet, to increase the criminal and civil penalties associated with certain crimes relating to children, and for other purposes'. Unsurprisingly, this bill is a member of the class 'Law, Crime, and Family Issues'.

Despite successful managed cases such as Senate bill 2648, a bill's language cannot always lead to correct classification without additional information. For this reason, in the electronic corpus we have marked bills through a combination of human annotation and machine learning when the bill title language indicates an example of topic drift. The machine learning systems mentioned in this paper are used to identify bills which are either incorrectly marked or marked with low confidence. This subset was then evaluated by human annotators to produce a non-exhaustive list of 'marker bills' which can be used to empirically assess the performance of both human and machine learning systems at correctly identifying and labeling bills which are examples of topic drift.

To experimentally address the constraints of sequencing and topic drift, we build a system which overcompensates by asking humans to annotate any bill which might be a case of topic drift. We identify possible topic drift cases as those bills where any of our ensemble classifiers disagree. Table 4 shows the results of using the nth Congress to predict the categories of the bills of the $nth + 1$ Congress. When all 3 of the classifiers in the machine learning ensemble agree on a prediction, the system predicts the topic of a bill with 90% accuracy or roughly the same as humans. When classifiers disagree the overall accuracy drops (in part due to topic drift which is captured differently by the different classifiers), and we then ask humans to annotate the bills. The resulting simple active learning method is explained in detail in Hillard et al. (2008).

In this sense, this active learning experiment achieves a key goal of the Congressional Bills project team. It is conservative, in that it begins to realize when it is making mistakes which would critically impact the usefulness of the underlying data in social research. But it still saves time and effort. However, it is also clear that this initial experiment is just a baseline. Application of research in active learning improvements, natural language processing, and topic drift management should yield further reductions in the amount of work still needed to be performed by humans.

7. Conclusion

The corpus of United States Congressional bills (USPALCV1) is a unique asset for language researchers and information scientists. A human annotated corpus of more than 375,000 documents, it now also includes 'test scenarios' for managing topic drift over time. In publishing these baseline performance estimates, we hope to encourage language researchers to download and investigate the corpus for the purpose of significantly improving upon the methods outlined in this paper.

In addition to the USPALCV1 corpus, social researchers around the world are generating parallel corpora using the same Policy Agendas coding scheme. Data sets with parallel annotations can be made available for newswire articles, budgetary expenditures, and speeches. Researchers in other countries are also using the annotation scheme to annotate similarly diverse data sets. As these data sets are transformed into reference corpora, language researchers can devise a multitude of experiments to test theories, classification model performance, machine translation, and the usefulness of language models.

8. Acknowledgements

Special thanks to Claire Cardie for her helpful comments and Thorsten Joachims for his counsel.

9. References

E. Scott Adler and John Wilkerson. 2008. Intended consequences: Juridictional reform and issue control in the u.s. house of representatives. *Legislative Studies Quarterly*, 33(1):85 – 112.

Baumgartner, Jones, and Macleod. 1998. Lessons from the trenches: Quality, reliability, and usability in a new data source. *The Political Methodologist*.

F. Baumgartner, B. Jones, and J. Wilkerson, 2002. *Policy Dynamics*, chapter 2. University of Chicago Press.

Eric Brill and Jun Wu. 1998. Classifier combination for improved lexical disambiguation. In *Proc. ACL*, pages 191–195.

Claire Cardie, Cynthia Farina, Matt Rawding, Adil Aijaz, and Stephen Purpura. 2008. A Study in Rule-Specific Issue Categorization for e-Rulemaking. In *Proceedings of the 9th Annual International Conference on Digital Government Research*.

J. Cohen. 1968. Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, 70(4):213–220.

J. Curran. 2002. Ensemble methods for automatic thesaurus extraction. *Proc. Empirical Methods in Natural Language Processing*, pages 222–229.

T. Dietterich. 2000. Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857:1–15.

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.

D. Hillard, S. Purpura, and J. Wilkerson. 2007. An active learning framework for classifying political text. In *Midwest Political Science Association 65th Annual National Conference*.

Dustin Hillard, Stephen Purpura, and John Wilkerson. 2008. Computer assisted topic classification for mixed methods social science research. *Journal of Information Technology and Politics*, 4(4).

T. Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proc. European Conference on Machine Learning*.

D. Koller and M. Sahami. 1997. Hierarchically classifying documents using very few words. *Proc. Int. Conf. on Machine Learning*, pages 170–178.

W. John MacMullen. 2003. Requirements definition and design criteria for test corpora in information science. Technical report, University of North Carolina at Chapel Hill, March.

A. McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow.

K. Papineni. 2001. Why inverse document frequency? In *Proceedings of the North American Association for Computational Linguistics, NAACL*, pages 25–32.

M. F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130 – 137.

Stephen Purpura and Dustin Hillard. 2006. Automated Classification of Congressional Legislation. In *Proceedings of the 7th Annual International Conference on Digital Government Research*.

S.E. Robertson and S. Walker. 1997. Laboratory experiments with okapi: participation in the trec programme. *Journal of Documentation*, 53:20 – 34.

Tony Rose, Mark Stevenson, and Miles Whitehead. 2002. The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*.

G. Salton and M.J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

R. E. Schapire and Y. Singer. 2000. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168.

S. Soroka, C. Wlezien, and I. McLean. 2006. Public Expenditure in the UK: How Measures Matter. *Journal of the Royal Statistical Society*, pages 255–271.

Stephen Purpura, Claire Cardie, and Jesse Simons. 2008. Active Learning for e-Rulemaking: Public Comment Categorization. In *Proceedings of the 9th Annual International Conference on Digital Government Research*.