The U.S. Policy Agenda Legislation Corpus Volume 1
A Language Resource from 1947 – 1998

Data files available at: http://www.congressionalbills.org/corpus

Stephen Purpura – Information Science Program – Cornell University
John Wilkerson – Political Science – University of Washington
Dustin Hillard – Electrical Engineering – University of Washington

Critical Descriptive Points:

- 40 years of U.S. Congressional bill titles
- Annotated using Policy Agenda Method
- 20 major categories/226 subcategories
- 375,517 bills including 120,927 duplicates

Critical Questions:

- Can machines reduce the work involved in annotating bills?
- Can ensembles improve upon baseline?
- Is language shift a significant problem?

Results:

Machines can reduce the work to annotate the bills, but care must be exercised to detect the machine mistakes due to language shift and confusion. Ensembles can significantly help with this problem.

The Top Level Categories of the Policy Agendas Project

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 Laws
15	Banking, Finance, Domestic Commerce
16	Defense
17	Space, Science, Technology, and Communications
18	Foreign Trade
19	International Affairs and Foreign Aid
20	Government Operations
21	Public Lands and Water Management
99	Private Legislation

Baseline (bag of words) Text Categorization with Linear Classifiers (Table 3)

	SVM	Maxent	Boostexter	Naïve Bayes	Ensemble
Major Topics (n=20)	88.7% (0.881)	86.5% (0.859)	85.6% (0.849)	81.4% (0.805)	89.0% (0.884)
Subtopics (n=226)	81.0% (0.800)	78.3% (0.771)	73.6% (0.722)	71.9% (0.705)	81.0% (0.800)

Machine Learning Prediction Performance when Classifiers Agree and Disagree (Table 4)

Congress Session used in Train	Congress Session used in 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	101st	9248	62.1	93.0	61.5	81.1	80.8
101st	102nd	9602	62.4	90.3	61.1	79.3	79.3
102nd	103rd	7879	64.8	90.1	60.2	79.6	79.5
103rd	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