Get out the vote: Determining support or opposition from Congressional floor-debate transcripts

Matt Thomas, Bo Pang, and Lillian Lee
Department of Computer Science, Cornell University
Ithaca, NY 14853-7501
mattthomas84@gmail.com, pabo@cs.cornell.edu, llee@cs.cornell.edu

Original version: Proceedings of EMNLP 2006, pp. 327–335. This revision of December 7, 2006 includes minor updates to four data points that were originally slightly off due to a single incorrect threshold value. All changes are indicated by both the color red and by ~~strikethroughs~~ or [bracketed comments], except for the reformatting of the bibliography.

Abstract

We investigate whether one can determine from the transcripts of U.S. Congressional floor debates whether the speeches represent support of or opposition to proposed legislation. To address this problem, we exploit the fact that these speeches occur as part of a discussion; this allows us to use sources of information regarding relationships between discourse segments, such as whether a given utterance indicates agreement with the opinion expressed by another. We find that the incorporation of such information yields substantial improvements over classifying speeches in isolation.

1 Introduction

> *One ought to recognize that the present political chaos is connected with the decay of language, and that one can probably bring about some improvement by starting at the verbal end.* — Orwell, "Politics and the English language"

We have entered an era where very large amounts of politically oriented text are now available online. This includes both official documents, such as the full text of laws and the proceedings of legislative bodies, and unofficial documents, such as postings on weblogs (blogs) devoted to politics. In some sense, the availability of such data is simply a manifestation of a general trend of "everybody putting their records on the Internet".[1] The online accessibility of politically oriented texts in particular, however, is a phenomenon that some have gone so far as to say will have a potentially society-changing effect.

In the United States, for example, governmental bodies are providing and soliciting political documents via the Internet, with lofty goals in mind: *electronic rulemaking* (eRulemaking) initiatives involving the "electronic collection, distribution, synthesis, and analysis of public commentary in the regulatory rulemaking process", may "[alter] the citizen-government relationship" (Shulman and Schlosberg, 2002). Additionally, much media attention has been focused recently on the potential impact that Internet sites may have on politics[2], or at least on political journalism[3]. Regardless of whether one views such claims as clear-sighted prophecy or mere hype, it is obviously important to help people understand and analyze politically oriented text, given the importance of enabling informed participation in the political process.

Evaluative and persuasive documents, such as a politician's speech regarding a bill or a blogger's commentary on a legislative proposal, form a particularly interesting type of politically oriented text. People are much more likely to consult such evaluative statements than the actual text of a bill or law under discussion, given the dense nature of legislative language and the fact that (U.S.) bills often reach several hundred pages in length (Smith, Roberts, and Vander Wielen,

[1] It is worth pointing out that the United States' Library of Congress was an extremely early adopter of Web technology: the THOMAS database (http://thomas.loc.gov) of congressional bills and related data was launched in January 1995, when Mosaic was not quite two years old and Altavista did not yet exist.

[2] E.g., "Internet injects sweeping change into U.S. politics", Adam Nagourney, *The New York Times*, April 2, 2006.

[3] E.g., "The End of News?", Michael Massing, *The New York Review of Books*, December 1, 2005.

2005). Moreover, political opinions are explicitly solicited in the eRulemaking scenario.

In the analysis of evaluative language, it is fundamentally necessary to determine whether the author/speaker supports or disapproves of the topic of discussion. In this paper, we investigate the following specific instantiation of this problem: we seek to determine from the transcripts of U.S. Congressional floor debates whether each "speech" (continuous single-speaker segment of text) represents support for or opposition to a proposed piece of legislation. Note that from an experimental point of view, this is a very convenient problem to work with because we can automatically determine ground truth (and thus avoid the need for manual annotation) simply by consulting publicly available voting records.

Task properties Determining whether or not a speaker supports a proposal falls within the realm of *sentiment analysis*, an extremely active research area devoted to the computational treatment of subjective or opinion-oriented language (early work includes Wiebe and Rapaport (1988), Hearst (1992), Sack (1994), and Wiebe (1994); see Esuli (2006) for an active bibliography). In particular, since we treat each individual speech within a debate as a single "document", we are considering a version of *document-level sentiment-polarity classification*, namely, automatically distinguishing between positive and negative documents (Das and Chen, 2001; Pang, Lee, and Vaithyanathan, 2002; Turney, 2002; Dave, Lawrence, and Pennock, 2003).

Most sentiment-polarity classifiers proposed in the recent literature categorize each document independently. A few others incorporate various measures of inter-document similarity between the texts to be labeled (Agarwal and Bhattacharyya, 2005; Pang and Lee, 2005; Goldberg and Zhu, 2006). Many interesting opinion-oriented documents, however, can be linked through certain relationships that occur in the context of evaluative *discussions*. For example, we may find textual[4] evidence of a high likelihood of *agreement* between two speakers, such as explicit assertions ("I second that!") or quotation of messages in emails or postings (see Mullen and Malouf (2006) but cf. Agrawal et al. (2003)). Agreement evidence can be a powerful aid in our classification task: for example, we can easily categorize a complicated (or overly terse) document if we find within it indications of agreement with a clearly positive text.

Obviously, incorporating agreement information provides additional benefit only when the input documents are relatively difficult to classify individually. Intuition suggests that this is true of the data with which we experiment, for several reasons. First, U.S. congressional debates contain very rich language and cover an extremely wide variety of topics, ranging from flag burning to international policy to the federal budget. Debates are also subject to digressions, some fairly natural and others less so (e.g., "Why are we discussing this bill when the plight of my constituents regarding this other issue is being ignored?")

Second, an important characteristic of persuasive language is that speakers may spend more time presenting evidence in support of their positions (or attacking the evidence presented by others) than directly stating their attitudes. An extreme example will illustrate the problems involved. Consider a speech that describes the U.S. flag as deeply inspirational, and thus contains only positive language. If the bill under discussion is a proposed flag-burning ban, then the speech is *supportive*; but if the bill under discussion is aimed at rescinding an existing flag-burning ban, the speech may represent *opposition* to the legislation. Given the current state of the art in sentiment analysis, it is doubtful that one could determine the (probably topic-specific) relationship between presented evidence and speaker opinion.

Qualitative summary of results The above difficulties underscore the importance of enhancing standard classification techniques with new information sources that promise to improve accuracy, such as inter-document relationships between the documents to be labeled. In this paper, we demonstrate that the incorporation of agreement modeling can provide substantial improvements over the application of support vector machines (SVMs) in isolation, which represents the state of the art in the individual classification of documents. The en-

[4]Because we are most interested in techniques applicable across domains, we restrict consideration to NLP aspects of the problem, ignoring external problem-specific information. For example, although most votes in our corpus were almost completely along party lines (and despite the fact that same-party information is easily incorporated via the methods we propose), we did not use party-affiliation data. Indeed, in other settings (e.g., a movie-discussion listserv) one may not be able to determine the participants' political leanings, and such information may not lead to significantly improved results even if it were available.

	total	train	test	development
speech segments	3857	2740	860	257
debates	53	38	10	5
average number of speech segments per debate	72.8	72.1	86.0	51.4
average number of speakers per debate	32.1	30.9	41.1	22.6

Table 1: Corpus statistics.

hanced accuracies are obtained via a fairly primitive automatically-acquired "agreement detector" and a conceptually simple method for integrating isolated-document and agreement-based information. We thus view our results as demonstrating the potentially large benefits of exploiting sentiment-related discourse-segment relationships in sentiment-analysis tasks.

2 Corpus

This section outlines the main steps of the process by which we created our corpus (download site: www.cs.cornell.edu/home/llee/data/convote.html).

GovTrack (http://govtrack.us) is an independent website run by Joshua Tauberer that collects publicly available data on the legislative and fundraising activities of U.S. congresspeople. Due to its extensive cross-referencing and collating of information, it was nominated for a 2006 "Webby" award. A crucial characteristic of GovTrack from our point of view is that the information is provided in a very convenient format; for instance, the floor-debate transcripts are broken into separate HTML files according to the subject of the debate, so we can trivially derive long sequences of speeches guaranteed to cover the same topic.

We extracted from GovTrack all available transcripts of U.S. floor debates in the House of Representatives for the year 2005 (3268 pages of transcripts in total), together with voting records for all roll-call votes during that year. We concentrated on debates regarding "controversial" bills (ones in which the losing side generated at least 20% of the speeches) because these debates should presumably exhibit more interesting discourse structure.

Each debate consists of a series of *speech segments*, where each segment is a sequence of uninterrupted utterances by a single speaker. Since speech segments represent natural discourse units, we treat them as the basic unit to be classified. Each speech segment was labeled by the vote ("yea" or "nay") cast for the proposed bill by the person who uttered the speech segment.

We automatically discarded those speech segments belonging to a class of formulaic, generally one-sentence utterances focused on the yielding of time on the house floor (for example, "Madam Speaker, I am pleased to yield 5 minutes to the gentleman from Massachusetts"), as such speech segments are clearly off-topic. We also removed speech segments containing the term "amendment", since we found during initial inspection that these speeches generally reflect a speaker's opinion on an amendment, and this opinion may differ from the speaker's opinion on the underlying bill under discussion.

We randomly split the data into training, test, and development (parameter-tuning) sets representing roughly 70%, 20%, and 10% of our data, respectively (see Table 1). The speech segments remained grouped by debate, with 38 debates assigned to the training set, 10 to the test set, and 5 to the development set; we require that the speech segments from an individual debate all appear in the same set because our goal is to examine classification of speech segments in the context of the surrounding discussion.

3 Method

The support/oppose classification problem can be approached through the use of standard classifiers such as support vector machines (SVMs), which consider each text unit in isolation. As discussed in Section 1, however, the conversational nature of our data implies the existence of various relationships that can be exploited to improve cumulative classification accuracy for speech segments belonging to the same debate. Our classification framework, directly inspired by Blum and Chawla (2001), integrates both perspectives, optimizing its labeling of speech segments based on both individual speech-segment classification scores and preferences for groups of speech segments to receive the same label. In this section, we discuss the specific classification framework that we adopt

and the set of mechanisms that we propose for modeling specific types of relationships.

3.1 Classification framework

Let $s_1, s_2, \ldots, s_n$ be the sequence of speech segments within a given debate, and let $\mathcal{Y}$ and $\mathcal{N}$ stand for the "yea" and "nay" class, respectively. Assume we have a non-negative function $\mathit{ind}(s, C)$ indicating the degree of preference that an individual-document classifier, such as an SVM, has for placing speech-segment s in class C. Also, assume that some pairs of speech segments have *weighted links* between them, where the non-negative *strength* (weight) $str(\ell)$ for a link ℓ indicates the degree to which it is preferable that the linked speech segments receive the same label. Then, any class assignment $c = c(s_1), c(s_2), \ldots, c(s_n)$ can be assigned a *cost*

$$\sum_{s} \mathit{ind}(s, \overline{c}(s)) + \sum_{s,s' \colon c(s) \neq c(s')} \; \sum_{\ell \text{ between } s,s'} str(\ell),$$

where $\overline{c}(s)$ is the "opposite" class from $c(s)$. A *minimum-cost* assignment thus represents an optimum way to classify the speech segments so that each one tends not to be put into the class that the individual-document classifier disprefers, but at the same time, highly associated speech segments tend not to be put in different classes.

As has been previously observed and exploited in the NLP literature (Pang and Lee, 2004; Agarwal and Bhattacharyya, 2005; Barzilay and Lapata, 2005), the above optimization function, unlike many others that have been proposed for graph or set partitioning, can be solved *exactly* in an provably efficient manner via methods for finding minimum cuts in graphs. In our view, the contribution of our work is the examination of new types of relationships, not the method by which such relationships are incorporated into the classification decision.

3.2 Classifying speech segments in isolation

In our experiments, we employed the well-known classifier SVM^{light} to obtain individual-document classification scores, treating $\mathcal{Y}$ as the positive class and using plain unigrams as features.[5] Following standard practice in sentiment analysis (Pang, Lee, and Vaithyanathan, 2002), the input to SVM^{light} consisted of normalized presence-of-feature (rather than frequency-of-feature) vectors. The *ind* value for each speech segment s was based on the signed distance $d(s)$ from the vector representing s to the trained SVM decision plane:

$$\mathit{ind}(s, \mathcal{Y}) \stackrel{\text{def}}{=} \begin{cases} 1 & d(s) > 2\sigma_s; \\ \left(1 + \frac{d(s)}{2\sigma_s}\right)/2 & |d(s)| \leq 2\sigma_s; \\ 0 & d(s) < -2\sigma_s \end{cases}$$

where σ_s is the standard deviation of $d(s)$ over all speech segments s in the debate in question, and $\mathit{ind}(s, \mathcal{N}) \stackrel{\text{def}}{=} 1 - \mathit{ind}(s, \mathcal{Y})$.

We now turn to the more interesting problem of representing the preferences that speech segments may have for being assigned to the same class.

3.3 Relationships between speech segments

A wide range of relationships between text segments can be modeled as positive-strength links. Here we discuss two types of constraints that are considered in this work.

Same-speaker constraints: In Congressional debates and in general social-discourse contexts, a single speaker may make a number of comments regarding a topic. It is reasonable to expect that in many settings, the participants in a discussion may be convinced to change their opinions midway through a debate. Hence, in the general case we wish to be able to express "soft" preferences for all of an author's statements to receive the same label, where the strengths of such constraints could, for instance, vary according to the time elapsed between the statements. Weighted links are an appropriate means to express such variation.

However, if we assume that most speakers do not change their positions in the course of a discussion, we can conclude that all comments made by the same speaker must receive the same label. This assumption holds by fiat for the ground-truth labels in our dataset because these labels were derived from the single vote cast by the speaker on the bill being discussed.[6] We can implement this assumption via links whose weights are essentially infinite. Although one can also implement

[5] SVM^{light} is available at svmlight.joachims.org. Default parameters were used, although experimentation with different parameter settings is an important direction for future work (Daelemans and Hoste, 2002; Munson, Cardie, and Caruana, 2005).

[6] We are attempting to determine whether a speech segment represents support or not. This differs from the problem of determining what the speaker's actual opinion is, a problem that, as an anonymous reviewer put it, is complicated by "grandstanding, backroom deals, or, more innocently, plain change of mind ('I voted for it before I voted against it')".

this assumption via concatenation of same-speaker speech segments (see Section 4.3), we view the fact that our graph-based framework incorporates both hard and soft constraints in a principled fashion as an advantage of our approach.

Different-speaker agreements In House discourse, it is common for one speaker to make reference to another in the context of an agreement or disagreement over the topic of discussion. The systematic identification of instances of agreement can, as we have discussed, be a powerful tool for the development of intelligently selected weights for links between speech segments.

The problem of agreement identification can be decomposed into two sub-problems: identifying references and their targets, and deciding whether each reference represents an instance of agreement. In our case, the first task is straightforward because we focused solely on by-name references.[7] Hence, we will now concentrate on the second, more interesting task.

We approach the problem of classifying references by representing each reference with a word-presence vector derived from a window of text surrounding the reference.[8] In the training set, we classify each reference connecting two speakers with a positive or negative label depending on whether the two voted the same way on the bill under discussion[9]. These labels are then used to train an SVM classifier, the output of which is subsequently used to create weights on *agreement links* in the test set as follows.

Let $d(r)$ denote the distance from the vector representing reference r to the agreement-detector SVM's decision plane, and let σ_r be the standard deviation of $d(r)$ over all references in the debate in question. We then define the strength agr of the *agreement link* corresponding to the reference as:

$$agr(r) \stackrel{\text{def}}{=} \begin{cases} 0 & d(r) < \theta_{\text{agr}}; \\ \alpha \cdot d(r)/4\sigma_r & \theta_{\text{agr}} \leq d(r) \leq 4\sigma_r; \\ \alpha & d(r) > 4\sigma_r. \end{cases}$$

The free parameter α specifies the relative importance of the agr scores. The threshold θ_{agr} controls the precision of the agreement links, in that values of θ_{agr} greater than zero mean that greater confidence is required before an agreement link can be added.[10]

Agreement classifier ("reference⇒agreement?")	Devel. set	Test set
majority baseline	81.51	80.26
Train: no amdmts; $\theta_{\text{agr}} = 0$	84.25	81.07
Train: with amdmts; $\theta_{\text{agr}} = 0$	**86.99**	**80.10**

Table 2: Agreement-classifier accuracy, in percent. "Amdmts"="speech segments containing the word 'amendment'". Recall that boldface indicates results for development-set-optimal settings.

4 Evaluation

This section presents experiments testing the utility of using speech-segment relationships, evaluating against a number of baselines. All reported results use values for the free parameter α derived via tuning on the development set. In the tables, **boldface** indicates the development- and test-set results for the *development-set-optimal* parameter settings, as one would make algorithmic choices based on development-set performance.

4.1 Preliminaries: Reference classification

Recall that to gather inter-speaker agreement information, the strategy employed in this paper is to classify by-name references to other speakers as to whether they indicate agreement or not.

To train our agreement classifier, we experimented with undoing the deletion of amendment-related speech segments in the training set. Note that such speech segments were *never* included in the development or test set, since, as discussed in Section 2, their labels are probably noisy; however, including them in the *training* set allows the

[7] One subtlety is that for the purposes of mining agreement cues (but *not* for evaluating overall support/oppose classification accuracy), we temporarily re-inserted into our dataset previously filtered speech segments containing the term "yield", since the yielding of time on the House floor typically indicates agreement even though the yield statements contain little relevant text on their own.

[8] We found good development-set performance using the 30 tokens before, 20 tokens after, and the name itself.

[9] Since we are concerned with references that potentially represent relationships between speech segments, we ignore references for which the target of the reference did not speak in the debate in which the reference was made.

[10] Our implementation puts a link between just one arbitrary pair of speech segments among all those uttered by a given pair of apparently agreeing speakers. The "infinite-weight" same-speaker links propagate the agreement information to all other such pairs.

Agreement classifier	Precision (in percent): Devel. set	Test set
$\theta_{\text{agr}} = 0$	86.23	82.55
$\theta_{\text{agr}} = \mu$	**89.41**	~~88.47~~ **88.50**

Table 3: Agreement-classifier precision.

classifier to examine more instances even though some of them are labeled incorrectly. As Table 2 shows, using more, if noisy, data yields better agreement-classification results on the development set, and so we use that policy in all subsequent experiments.[11]

An important observation is that precision may be more important than accuracy in deciding which agreement links to add: false positives with respect to agreement can cause speech segments to be incorrectly assigned the same label, whereas false negatives mean only that agreement-based information about other speech segments is not employed. As described above, we can raise agreement precision by increasing the threshold θ_{agr}, which specifies the required confidence for the addition of an agreement link. Indeed, Table 3 shows that we can improve agreement precision by setting θ_{agr} to the (positive) mean agreement score μ assigned by the SVM agreement-classifier over all references in the given debate[12]. However, this comes at the cost of greatly reducing agreement accuracy (development: 64.38%; test: ~~66.18~~ 66.34%) due to lowered recall levels. Whether or not better speech-segment classification is ultimately achieved is discussed in the next sections.

4.2 Segment-based speech-segment classification

Baselines The first two data rows of Table 4 depict baseline performance results. The #("support") − #("oppos") baseline is meant to explore whether the speech-segment classification task can be reduced to simple lexical checks. Specifically, this method uses the signed difference between the number of words containing the stem "support" and the number of words containing the stem "oppos" (returning the majority class if the difference is 0). No better than 62.67% test-set accuracy is obtained by either baseline.

[11] Unfortunately, this policy leads to inferior *test-set* agreement classification. Section 4.5 contains further discussion.

[12] We elected not to explicitly tune the value of θ_{agr} in order to minimize the number of free parameters to deal with.

Support/oppose classifer ("speech segment⇒yea?")	Devel. set	Test set
majority baseline	54.09	58.37
#("support") − #("oppos")	59.14	62.67
SVM [speech segment]	70.04	66.05
SVM + same-speaker links	79.77	67.21
SVM + same-speaker links . . .		
+ agreement links, $\theta_{\text{agr}} = 0$	**89.11**	**70.81**
+ agreement links, $\theta_{\text{agr}} = \mu$	87.94	~~71.16~~ 70.81

Table 4: Segment-based speech-segment classification accuracy, in percent.

Support/oppose classifer ("speech segment⇒yea?")	Devel. set	Test set
SVM [speaker]	71.60	70.00
SVM + agreement links . . .		
with $\theta_{\text{agr}} = 0$	**88.72**	**71.28**
with $\theta_{\text{agr}} = \mu$	84.44	~~76.05~~ 76.16

Table 5: Speaker-based speech-segment classification accuracy, in percent. Here, the initial SVM is run on the concatenation of all of a given speaker's speech segments, but the results are computed over speech segments (not speakers), so that they can be compared to those in Table 4.

Using relationship information Applying an SVM to classify each speech segment in isolation leads to clear improvements over the two baseline methods, as demonstrated in Table 4. When we impose the constraint that all speech segments uttered by the same speaker receive the same label via "same-speaker links", both test-set and development-set accuracy increase even more, in the latter case quite substantially so.

The last two lines of Table 4 show that the best results are obtained by incorporating agreement information as well. The highest test-set result, ~~71.16~~ 70.81%, is obtained by using ~~a high precision threshold to determine which agreement links to add~~ agreement links, with the $\theta_{\text{agr}} = 0$ policy performing better than $\theta_{\text{agr}} = \mu$ on the development set. ~~While the development set results would induce us to utilize the standard threshold value of 0, which is sub-optimal on the test set, the $\theta_{\text{agr}} = 0$ agreement link policy still achieves noticeable improvement over not using agreement links (test set: 70.81% vs. 67.21%).~~

4.3 Speaker-based speech-segment classification

We use speech segments as the unit of classification because they represent natural discourse units. As a consequence, we are able to exploit relationships at the speech-segment level. However, it is interesting to consider whether we really need to consider relationships specifically between speech segments themselves, or whether it suffices to simply consider relationships between the *speakers* of the speech segments. In particular, as an alternative to using same-speaker links, we tried a *speaker-based* approach wherein the way we determine the initial individual-document classification score for each speech segment uttered by a person p in a given debate is to run an SVM on the concatenation of *all* of p's speech segments within that debate. (We also ensure that agreement-link information is propagated from speech-segment to speaker pairs.)

How does the use of same-speaker links compare to the concatenation of each speaker's speech segments? Tables 4 and 5 show that, not surprisingly, the SVM individual-document classifier works better on the concatenated speech segments than on the speech segments in isolation. However, the effect on overall classification accuracy is less clear: the development set favors same-speaker links over concatenation, while the test set does not.

But we stress that the most important observation we can make from Table 5 is that once again, the addition of agreement information leads to substantial improvements in accuracy.

4.4 "Hard" agreement constraints

Recall that in in our experiments, we created finite-weight agreement links, so that speech segments appearing in pairs flagged by our (imperfect) agreement detector can potentially receive different labels. We also experimented with *forcing* such speech segments to receive the same label, either through infinite-weight agreement links or through a speech-segment concatenation strategy similar to that described in the previous subsection. Both strategies resulted in clear degradation in performance on both the development and test sets, a finding that validates our encoding of agreement information as "soft" preferences.

4.5 On the development/test set split

We have seen several cases in which the method that performs best on the development set does not yield the best test-set performance. However, we felt that it would be illegitimate to change the train/development/test sets in a post hoc fashion, that is, after seeing the experimental results.

Moreover, and crucially, it is very clear that using agreement information, encoded as preferences within our graph-based approach rather than as hard constraints, yields substantial improvements on both the development and test set; this, we believe, is our most important finding.

5 Related work

Politically-oriented text Sentiment analysis has specifically been proposed as a key enabling technology in eRulemaking, allowing the automatic analysis of the opinions that people submit (Shulman et al., 2005; Cardie et al., 2006; Kwon, Shulman, and Hovy, 2006). There has also been work focused upon determining the political leaning (e.g., "liberal" vs. "conservative") of a document or author, where most previously-proposed methods make no direct use of relationships between the documents to be classified (the "unlabeled" texts) (Laver, Benoit, and Garry, 2003; Efron, 2004; Mullen and Malouf, 2006). An exception is Grefenstette et al. (2004), who experimented with determining the political orientation of websites essentially by classifying the concatenation of all the documents found on that site.

Others have applied the NLP technologies of near-duplicate detection and topic-based text categorization to politically oriented text (Yang and Callan, 2005; Purpura and Hillard, 2006).

Detecting agreement We used a simple method to learn to identify cross-speaker references indicating agreement. More sophisticated approaches have been proposed (Hillard, Ostendorf, and Shriberg, 2003), including an extension that, in an interesting reversal of our problem, makes use of sentiment-polarity indicators within speech segments (Galley et al., 2004). Also relevant is work on the general problems of dialog-act tagging (Stolcke et al., 2000), citation analysis (Lehnert, Cardie, and Riloff, 1990), and computational rhetorical analysis (Marcu, 2000; Teufel and Moens, 2002).

We currently do not have an efficient means

to encode *disagreement* information as hard constraints; we plan to investigate incorporating such information in future work.

Relationships between the unlabeled items Carvalho and Cohen (2005) consider sequential relations between different types of emails (e.g., between requests and satisfactions thereof) to classify messages, and thus also explicitly exploit the structure of conversations.

Previous sentiment-analysis work in different domains has considered inter-document similarity (Agarwal and Bhattacharyya, 2005; Pang and Lee, 2005; Goldberg and Zhu, 2006) or explicit inter-document references in the form of hyperlinks (Agrawal et al., 2003).

Notable early papers on graph-based semi-supervised learning include Blum and Chawla (2001), Bansal, Blum, and Chawla (2002), Kondor and Lafferty (2002), and Joachims (2003). Zhu (2005) maintains a survey of this area.

Recently, several alternative, often quite sophisticated approaches to *collective classification* have been proposed (Neville and Jensen, 2000; Lafferty, McCallum, and Pereira, 2001; Getoor et al., 2002; Taskar, Abbeel, and Koller, 2002; Taskar, Guestrin, and Koller, 2003; Taskar, Chatalbashev, and Koller, 2004; McCallum and Wellner, 2004). It would be interesting to investigate the application of such methods to our problem. However, we also believe that our approach has important advantages, including conceptual simplicity and the fact that it is based on an underlying optimization problem that is provably and in practice easy to solve.

6 Conclusion and future work

In this study, we focused on very general types of cross-document classification preferences, utilizing constraints based only on speaker identity and on direct textual references between statements. We showed that the integration of even very limited information regarding inter-document relationships can significantly increase the accuracy of support/opposition classification.

The simple constraints modeled in our study, however, represent just a small portion of the rich network of relationships that connect statements and speakers across the political universe and in the wider realm of opinionated social discourse. One intriguing possibility is to take advantage of (readily identifiable) information regarding interpersonal relationships, making use of speaker/author affiliations, positions within a social hierarchy, and so on. Or, we could even attempt to model relationships between topics or concepts, in a kind of extension of collaborative filtering. For example, perhaps we could infer that two speakers sharing a common opinion on evolutionary biologist Richard Dawkins (a.k.a. "Darwin's rottweiler") will be likely to agree in a debate centered on Intelligent Design. While such functionality is well beyond the scope of our current study, we are optimistic that we can develop methods to exploit additional types of relationships in future work.

Acknowledgments We thank Claire Cardie, Jon Kleinberg, Michael Macy, Andrew Myers, and the six anonymous EMNLP referees for valuable discussions and comments. We also thank Reviewer 1 for generously providing additional *post hoc* feedback, and the EMNLP chairs Eric Gaussier and Dan Jurafsky for facilitating the process (as well as for allowing authors an extra proceedings page. . .). This paper is based upon work supported in part by the National Science Foundation under grant no. IIS-0329064 [Addition:], an Alfred P. Sloan Research Fellowship, and Google Anita Borg Memorial Scholarship funds. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied, of any sponsoring institutions, the U.S. government, or any other entity.

References

Agarwal, Alekh and Pushpak Bhattacharyya. 2005. Sentiment analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified. In *Proceedings of the International Conference on Natural Language Processing (ICON)*.

Agrawal, Rakesh, Sridhar Rajagopalan, Ramakrishnan Srikant, and Yirong Xu. 2003. Mining newsgroups using networks arising from social behavior. In *Proceedings of WWW*, pages 529–535.

Bansal, Nikhil, Avrim Blum, and Shuchi Chawla. 2002. Correlation clustering. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS)*, pages 238–247. Journal version in *Machine Learning Journal*, special issue on theoretical advances in data clustering, 56(1-3):89–113 (2004).

Barzilay, Regina and Mirella Lapata. 2005. Collective

content selection for concept-to-text generation. In *Proceedings of HLT/EMNLP*, pages 331–338.

Blum, Avrim and Shuchi Chawla. 2001. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of ICML*, pages 19–26.

Cardie, Claire, Cynthia Farina, Thomas Bruce, and Erica Wagner. 2006. Using natural language processing to improve eRulemaking. In *Proceedings of Digital Government Research (dg.o)*.

Carvalho, Vitor and William W. Cohen. 2005. On the collective classification of email "speech acts". In *Proceedings of SIGIR*, pages 345–352.

Daelemans, Walter and Véronique Hoste. 2002. Evaluation of machine learning methods for natural language processing tasks. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 755–760.

Das, Sanjiv and Mike Chen. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*.

Dave, Kushal, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW*, pages 519–528.

Efron, Miles. 2004. Cultural orientation: Classifying subjective documents by cociation [sic] analysis. In *Proceedings of the AAAI Fall Symposium on Style and Meaning in Language, Art, Music, and Design*, pages 41–48.

Esuli, Andrea. 2006. Sentiment classification bibliography. liinwww.ira.uka.de/bibliography/Misc/Sentiment.html.

Galley, Michel, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd ACL*, pages 669–676.

Getoor, Lise, Nir Friedman, Daphne Koller, and Benjamin Taskar. 2002. Learning probabilistic models of relational structure. *Journal of Machine Learning Research*, 3:679–707. Special issue on the Eighteenth ICML.

Goldberg, Andrew B. and Jerry Zhu. 2006. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *TextGraphs: HLT/NAACL Workshop on Graph-based Algorithms for Natural Language Processing*.

Grefenstette, Gregory, Yan Qu, James G. Shanahan, and David A. Evans. 2004. Coupling niche browsers and affect analysis for an opinion mining application. In *Proceedings of RIAO*.

Hearst, Marti. 1992. Direction-based text interpretation as an information access refinement. In Paul Jacobs, editor, *Text-Based Intelligent Systems*. Lawrence Erlbaum Associates, pages 257–274.

Hillard, Dustin, Mari Ostendorf, and Elizabeth Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of HLT-NAACL*.

Joachims, Thorsten. 2003. Transductive learning via spectral graph partitioning. In *Proceedings of ICML*, pages 290–297.

Kondor, Risi Imre and John D. Lafferty. 2002. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of ICML*, pages 315–322.

Kwon, Namhee, Stuart Shulman, and Eduard Hovy. 2006. Multidimensional text analysis for eRulemaking. In *Proceedings of Digital Government Research (dg.o)*.

Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*.

Lehnert, Wendy, Claire Cardie, and Ellen Riloff. 1990. Analyzing research papers using citation sentences. In *Program of the Twelfth Annual Conference of the Cognitive Science Society*, pages 511–18.

Marcu, Daniel. 2000. *The theory and practice of discourse parsing and summarization*. MIT Press.

McCallum, Andrew and Ben Wellner. 2004. Conditional models of identity uncertainty with application to noun coreference. In *Proceedings of NIPS*.

Mullen, Tony and Robert Malouf. 2006. A preliminary investigation into sentiment analysis of informal political discourse. In *Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs*, pages 159–162.

Munson, Art, Claire Cardie, and Rich Caruana. 2005. Optimizing to arbitrary NLP metrics using ensemble selection. In *Proceedings of HLT-EMNLP*, pages 539–546.

Neville, Jennifer and David Jensen. 2000. Iterative classification in relational data. In *Proceedings of the AAAI Workshop on Learning Statistical Models from Relational Data*, pages 13–20.

Pang, Bo and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, pages 271–278.

Pang, Bo and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, pages 115–124.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86.

Purpura, Stephen and Dustin Hillard. 2006. Automated classification of congressional legislation. In *Proceedings of Digital Government Research (dg.o)*.

Sack, Warren. 1994. On the computation of point of view. In *Proceedings of AAAI*, page 1488. Student abstract.

Shulman, Stuart, Jamie Callan, Eduard Hovy, and Stephen Zavestoski. 2005. Language processing technologies for electronic rulemaking: A project highlight. In *Proceedings of Digital Government Research (dg.o)*, pages 87–88.

Shulman, Stuart and David Schlosberg. 2002. Electronic rulemaking: New frontiers in public participation. Prepared for the Annual Meeting of the American Political Science Association.

Smith, Steven S., Jason M. Roberts, and Ryan J. Vander Wielen. 2005. *The American Congress*. Cambridge University Press, fourth edition.

Stolcke, Andreas, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Taskar, Ben, Pieter Abbeel, and Daphne Koller. 2002. Discriminative probabilistic models for relational data. In *Proceedings of UAI*, Edmonton, Canada.

Taskar, Ben, Vassil Chatalbashev, and Daphne Koller. 2004. Learning associative Markov networks. In *Proceedings of ICML*.

Taskar, Ben, Carlos Guestrin, and Daphne Koller. 2003. Max-margin Markov networks. In *Proceedings of NIPS*.

Teufel, Simone and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Turney, Peter. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the ACL*, pages 417–424.

Wiebe, Janyce M. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

Wiebe, Janyce M. and William J. Rapaport. 1988. A computational theory of perspective and reference in narrative. In *Proceedings of the ACL*, pages 131–138.

Yang, Hui and Jamie Callan. 2005. Near-duplicate detection for eRulemaking. In *Proceedings of Digital Government Research (dg.o)*.

Zhu, Jerry. 2005. Semi-supervised learning literature survey. Computer Sciences Technical Report TR 1530, University of Wisconsin-Madison. Available at http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf; has been updated since the initial 2005 version.