Recognizing Stances in Online Debates

Swapna Somasundaran
Dept. of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
swapna@cs.pitt.edu

Janyce Wiebe
Dept. of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
wiebe@cs.pitt.edu

Abstract

This paper presents an unsupervised opinion analysis method for *debate-side classification*, i.e., recognizing which stance a person is taking in an online debate. In order to handle the complexities of this genre, we mine the web to learn associations that are indicative of opinion stances in debates. We combine this knowledge with discourse information, and formulate the debate side classification task as an Integer Linear Programming problem. Our results show that our method is substantially better than challenging baseline methods.

1 Introduction

This paper presents a method for *debate-side classification*, i.e., recognizing which stance a person is taking in an online debate posting. In online debate forums, people debate issues, express their preferences, and argue why their viewpoint is right. In addition to expressing positive sentiments about one's preference, a key strategy is also to express negative sentiments about the other side. For example, in the debate *"which mobile phone is better: iPhone or Blackberry,"* a participant on the iPhone side may explicitly assert and rationalize why the iPhone is better, and, alternatively, also argue why the Blackberry is worse. Thus, to recognize stances, we need to consider not only which opinions are positive and negative, but also what the opinions are about (their *targets*).

Participants directly express their opinions, such as *"The iPhone is cool,"* but, more often, they mention associated aspects. Some aspects are particular to one topic (e.g., Active-X is part of IE but not Firefox), and so distinguish between them. But even an aspect the topics share may distinguish between them, because people who are positive toward one topic may value that aspect more. For example, both the iPhone and Blackberry have keyboards, but we observed in our corpus that positive opinions about the keyboard are associated with the pro Blackberry stance. Thus, we need to find distinguishing aspects, which the topics may or may not share.

Complicating the picture further, participants may concede positive aspects of the opposing issue or topic, without coming out in favor of it, and they may concede negative aspects of the issue or topic they support. For example, in the following sentence, the speaker says positive things about the iPhone, even though he does not prefer it: *"Yes, the iPhone may be cool to take it out and play with and show off, but past that, it offers nothing."* Thus, we need to consider discourse relations to sort out which sentiments in fact reveal the writer's stance, and which are merely concessions.

Many opinion mining approaches find negative and positive words in a document, and aggregate their counts to determine the final document polarity, ignoring the targets of the opinions. Some work in product review mining finds aspects of a central topic, and summarizes opinions with respect to these aspects. However, they do not find distinguishing factors associated with a preference for a stance. Finally, while other opinion analysis systems have considered discourse information, they have not distinguished between concessionary and non-concessionary opinions when determining the overall stance of a document.

This work proposes an unsupervised opinion analysis method to address the challenges described above. First, for each debate side, we mine the web for opinion-target pairs that are associated with a preference for that side. This information is employed, in conjunction with discourse information, in an Integer Linear Programming (ILP) framework. This framework combines the individual pieces of information to arrive at debate-side

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 226–234,
Suntec, Singapore, 2-7 August 2009.

classifications of posts in online debates.

The remainder of this paper is organized as follows. We introduce our debate genre in Section 2 and describe our method in Section 3. We present the experiments in Section 4 and analyze the results in Section 5. Related work is in Section 6, and the conclusions are in Section 7.

2 The Debate Genre

In this section, we describe our debate data, and elaborate on characteristic ways of expressing opinions in this genre. For our current work, we use the online debates from the website http://www.convinceme.net.[1]

In this work, we deal only with dual-sided, dual-topic debates about named entities, for example iPhone vs. Blackberry, where $topic_1$ = iPhone, $topic_2$ =Blackberry, $side_1$ = pro-iPhone, and $side_2$=pro-Blackberry.

Our test data consists of posts of 4 debates: Windows vs. Mac, Firefox vs. Internet Explorer, Firefox vs. Opera, and Sony Ps3 vs. Nintendo Wii. The iPhone vs. Blackberry debate and two other debates, were used as development data.

Given below are examples of debate posts. Post 1 is taken from the iPhone vs. Blackberry debate, Post 2 is from the Firefox vs. Internet Explorer debate, and Post 3 is from the Windows vs. Mac debate:

(1) While the iPhone may appeal to younger generations and the BB to older, there is no way it is geared towards a less rich population. In fact it's exactly the opposite. It's a gimmick. The initial purchase may be half the price, but when all is said and done you pay at least $200 more for the 3g.

(2) In-line spell check...helps me with big words like onomatopoeia

(3) Apples are nice computers with an exceptional interface. Vista will close the gap on the interface some but Apple still has the prettiest, most pleasing interface and most likely will for the next several years.

2.1 Observations

As described in Section 1, the debate genre poses significant challenges to opinion analysis. This subsection elaborates upon some of the complexities.

Multiple polarities to argue for a side. Debate participants, in advocating their choice, switch back and forth between their opinions towards the sides. This makes it difficult for approaches that use only positive and negative word counts to decide which side the post is on. Posts 1 and 3 illustrate this phenomenon.

Sentiments towards both sides (topics) within a single post. The above phenomenon gives rise to an additional problem: often, conflicting sides (and topics) are addressed within the same post, sometimes within the same sentence. The second sentence of Post 3 illustrates this, as it has opinions about both Windows and Mac.

Differentiating aspects and personal preferences. People seldom repeatedly mention the topic/side; they show their evaluations indirectly, by evaluating aspects of each topic/side. *Differentiating* aspects determine the debate-post's side.

Some aspects are unique to one side/topic or the other, e.g., "3g" in Example 1 and "inline spell check" in Example 2. However, the debates are about topics that belong to the same domain and which therefore share many aspects. Hence, a purely ontological approach of finding "has-a" and "is-a" relations, or an approach looking only for product specifications, would not be sufficient for finding differentiating features.

When the two topics do share an aspect (e.g., a keyboard in the iPhone vs. Blackberry debate), the writer may perceive it to be more positive for one than the other. And, if the writer values that aspect, it will influence his or her overall stance. For example, many people prefer the Blackberry keyboard over the iPhone keyboard; people to whom phone keyboards are important are more likely to prefer the Blackberry.

Concessions. While debating, participants often refer to and acknowledge the viewpoints of the opposing side. However, they do not endorse this rival opinion. Uniform treatment of all opinions in a post would obviously cause errors in such cases. The first sentence of Example 1 is an instance of this phenomenon. The participant concedes that the iPhone appeals to young consumers, but this positive opinion is opposite to his overall stance.

[1] http://www.forandagainst.com and http://www.createdebate.com are other similar debating websites.

DIRECT OBJECT Rule: dobj(opinion, target) In words: The target is the direct object of the opinion Example: I love$_{opinion1}$ Firefox$_{target1}$ and defended$_{opinion2}$ it$_{target2}$
NOMINAL SUBJECT Rule: nsubj(opinion, target) In words: The target is the subject of the opinion Example: IE$_{target}$ breaks$_{opinion}$ with everything.
ADJECTIVAL MODIFIER Rule: amod(target, opinion) In words: The opinion is an adjectival modifier of the target Example: The annoying$_{opinion}$ popup$_{target}$
PREPOSITIONAL OBJECT Rule: if prep(target1,IN) $\Rightarrow$ pobj(IN, target2) In words: The prepositional object of a known target is also a target of the same opinion Example: The annoying$_{opinion}$ popup$_{target1}$ in IE$_{target2}$ ("popup" and "IE" are targets of "annoying")
RECURSIVE MODIFIERS Rule: if conj(adj2, opinion$_{adj1}$) $\Rightarrow$ amod(target, adj2) In words: If the opinion is an adjective (adj1) and it is conjoined with another adjective (adj2), then the opinion is tied to what adj2 modifies Example: It is a powerful$_{opinion(adj1)}$ and easy$_{opinion(adj2)}$ application$_{target}$ ("powerful" is attached to the target "application" via the adjective "easy")

Table 1: Examples of syntactic rules for finding targets of opinions

3 Method

We propose an unsupervised approach to classifying the stance of a post in a dual-topic debate. For this, we first use a web corpus to learn preferences that are likely to be associated with a side. These learned preferences are then employed in conjunction with discourse constraints to identify the side for a given post.

3.1 Finding Opinions and Pairing them with Targets

We need to find opinions and pair them with targets, both to mine the web for general preferences and to classify the stance of a debate post. We use straightforward methods, as these tasks are not the focus of this paper.

To find opinions, we look up words in a subjectivity lexicon: all instances of those words are treated as opinions. An opinion is assigned the prior polarity that is listed for that word in the lexicon, except that, if the prior polarity is positive or negative, and the instance is modified by a negation word (e.g., "not"), then the polarity of that instance is reversed. We use the subjectivity lexicon of (Wilson et al., 2005),[2] which contains approximately 8000 words which may be used to express opinions. Each entry consists of a subjective word, its prior polarity (positive ($^+$), negative ($^-$), neutral (*)), morphological information, and part of speech information.

To pair opinions with targets, we built a rule-based system based on dependency parse information. The dependency parses are obtained using the Stanford parser.[3] We developed the syntactic rules on separate data that is not used elsewhere in this paper. Table 1 illustrates some of these rules. Note that the rules are constructed (and explained in Table 1) with respect to the grammatical relation notations of the Stanford parser. As illustrated in the table, it is possible for an opinion to have more than one target. In such cases, the single opinion results in multiple opinion-target pairs, one for each target.

Once these opinion-target pairs are created, we mask the identity of the opinion word, replacing the word with its polarity. Thus, the opinion-target pair is converted to a polarity-target pair. For instance, "pleasing-interface" is converted to $interface^+$. This abstraction is essential for handling the sparseness of the data.

3.2 Learning aspects and preferences from the web

We observed in our development data that people highlight the aspects of topics that are the bases for their stances, both positive opinions toward aspects of the preferred topic, and negative opinions toward aspects of the dispreferred one. Thus, we decided to mine the web for aspects associated with a side in the debate, and then use that information to recognize the stances expressed in individual posts.

Previous work mined web data for aspects associated with topics (Hu and Liu, 2004; Popescu et al., 2005). In our work, we search for aspects associated with a topic, but particularized to polarity. Not all aspects associated with a topic are

[2] Available at http://www.cs.pitt.edu/mpqa.

[3] http://nlp.stanford.edu/software/lex-parser.shtml.

	$side_1$ (pro-iPhone)		$side_2$ (pro-blackberry)	
$term^p$	$P(iPhone^+\|term^p)$	$P(blackberry^-\|term^p)$	$P(iPhone^-\|term^p)$	$P(blackberry^+\|term^p)$
$storm^+$	0.227	0.068	0.022	0.613
$storm^-$	0.062	0.843	0.06	0.03
$phone^+$	0.333	0.176	0.137	0.313
$e\text{-}mail^+$	0	0.333	0.166	0.5
$ipod^+$	0.5	0	0.33	0
$battery^-$	0	0	0.666	0.333
$network^-$	0.333	0	0.666	0
$keyboard^+$	0.09	0.12	0	0.718
$keyboard^-$	0.25	0.25	0.125	0.375

Table 2: Probabilities learned from the web corpus (iPhone vs. blackberry debate)

discriminative with respect to stance; we hypothesized that, by including polarity, we would be more likely to find useful associations. An aspect may be associated with both of the debate topics, but not, by itself, be discriminative between stances toward the topics. However, *opinions* toward that aspect might discriminate between them. Thus, the basic unit in our web mining process is a polarity-target pair. Polarity-target pairs which explicitly mention one of the topics are used to anchor the mining process. Opinions about relevant aspects are gathered from the surrounding context.

For each debate, we downloaded weblogs and forums that talk about the main topics (corresponding to the sides) of that debate. For example, for the iPhone vs. Blackberry debate, we search the web for pages containing "iPhone" and "Blackberry." We used the Yahoo search API and imposed the search restriction that the pages should contain both topics in the http URL. This ensured that we downloaded relevant pages. An average of 3000 documents were downloaded per debate.

We apply the method described in Section 3.1 to the downloaded web pages. That is, we find all instances of words in the lexicon, extract their targets, and mask the words with their polarities, yielding polarity-target pairs. For example, suppose the sentence *"The interface is pleasing"* is in the corpus. The system extracts the pair "pleasing-interface," which is masked to "positive-interface," which we notate as $interface^+$. If the target in a polarity-target pair happens to be one of the topics, we select the polarity-target pairs in its vicinity for further processing (the rest are discarded). The intuition behind this is that, if someone expresses an opinion about a topic, he or she is likely to follow it up with reasons for that opinion. The sentiments in the surrounding context thus reveal factors that influence the preference or dislike towards the topic. We define the vicinity as the same sentence plus the following 5 sentences.

Each unique target word $target_i$ in the web corpus, i.e., each word used as the target of an opinion one or more times, is processed to generate the following conditional probabilities.

$$P(topic_j^q|target_i^p) = \frac{\#(topic_j^q, target_i^p)}{\#target_i^p} \quad (1)$$

where $p = \{^+, ^-, ^*\}$ and $q = \{^+, ^-, ^*\}$ denote the polarities of the target and the topic, respectively; $j = \{1, 2\}$; and $i = \{1...M\}$, where M is the number of unique targets in the corpus. For example, $P(Mac^+|interface^+)$ is the probability that "interface" is the target of a positive opinion that is in the vicinity of a positive opinion toward "Mac."

Table 2 lists some of the probabilities learned by this approach. (Note that the neutral cases are not shown.)

3.2.1 Interpreting the learned probabilities

Table 2 contains examples of the learned probabilities. These probabilities align with what we qualitatively found in our development data. For example, the opinions towards "Storm" essentially follow the opinions towards "Blackberry;" that is, positive opinions toward "Storm" are usually found in the vicinity of positive opinions toward "Blackberry," and negative opinions toward "Storm" are usually found in the vicinity of negative opinions toward "Blackberry" (for example, in the row for $storm^+$, $P(blackberry^+|storm^+)$ is much higher than the other probabilities). Thus, an opinion expressed about "Storm" is usually the opinion one has toward "Blackberry." This is expected, as Storm is a type of Blackberry. A similar example is $ipod^+$, which follows the opinion toward the iPhone. This is interesting because an

iPod is not a phone; the association is due to preference for the brand. In contrast, the probability distribution for "phone" does not show a preference for any one side, even though both iPhone and Blackberry are phones. This indicates that opinions towards phones in general will not be able to distinguish between the debate sides.

Another interesting case is illustrated by the probabilities for "e-mail." People who like e-mail capability are more likely to praise the Blackberry, or even criticize the iPhone — they would thus belong to the pro-Blackberry camp.

While we noted earlier that positive evaluations of keyboards are associated with positive evaluations of the Blackberry (by far the highest probability in that row), negative evaluations of keyboards, are, however, *not* a strong discriminating factor.

For the other entries in the table, we see that criticisms of batteries and the phone network are more associated with negative sentiments towards the iPhones.

The possibility of these various cases motivates our approach, in which opinions and their polarities are considered when searching for associations between debate topics and their aspects.

3.3 Debate-side classification

Once we have the probabilities collected from the web, we can build our classifier to classify the debate posts.

Here again, we use the process described in Section 3.1 to extract polarity-target pairs for each opinion expressed in the post. Let N be the number of instances of polarity-target pairs in the post. For each instance I_j ($j = \{1...N\}$), we look up the learned probabilities of Section 3.2 to create two scores, w_j and u_j:

$$w_j = P(topic_1^+|target_i^p) + P(topic_2^-|target_i^p) \quad (2)$$

$$u_j = P(topic_1^-|target_i^p) + P(topic_2^+|target_i^p) \quad (3)$$

where $target_i^p$ is the polarity-target type of which I_j is an instance.

Score w_j corresponds to $side_1$ and u_j corresponds to $side_2$. A point to note is that, if a target word is repeated, and it occurs in different polarity-target instances, it is counted as a separate instance each time — that is, here we account for tokens, not types. Via Equations 2 and 3, we interpret the observed polarity-target instance I_j in terms of debate sides.

We formulate the problem of finding the overall side of the post as an Integer Linear Programming (ILP) problem. The side that maximizes the overall side-score for the post, given all the N instances I_j, is chosen by maximizing the objective function

$$\sum_{j=1}^{N}(w_j x_j + u_j y_j) \quad (4)$$

subject to the following constraints

$$x_j \in \{0,1\},\ \forall j \quad (5)$$

$$y_j \in \{0,1\},\ \forall j \quad (6)$$

$$x_j + y_j = 1,\ \forall j \quad (7)$$

$$x_j - x_{j-1} = 0,\ j \in \{2..N\} \quad (8)$$

$$y_j - y_{j-1} = 0,\ j \in \{2..N\} \quad (9)$$

Equations 5 and 6 implement binary constraints. Equation 7 enforces the constraint that each I_j can belong to exactly one side. Finally, Equations 8 and 9 ensure that a single side is chosen for the entire post.

3.4 Accounting for concession

As described in Section 2, debate participants often acknowledge the opinions held by the opposing side. We recognize such discourse constructs using the Penn Discourse Treebank (Prasad et al., 2007) list of discourse connectives. In particular, we use the list of connectives from the Concession and Contra-expectation category. Examples of connectives in these categories are "while," "nonetheless," "however," and "even if." We use approximations to finding the arguments to the discourse connectives (*ARG1* and *ARG2* in Penn Discourse Treebank terms). If the connective is mid-sentence, the part of the sentence prior to the connective is considered conceded, and the part that follows the connective is considered non-conceded. An example is the second sentence of Example 3. If, on the other hand, the connective is sentence-initial, the sentence is split at the first comma that occurs mid sentence. The first part is considered conceded, and the second part is considered non-conceded. An example is the first sentence of Example 1.

The opinions occurring in the conceded part are interpreted in reverse. That is, the weights corresponding to the sides w_j and u_j are interchanged in equation 4. Thus, conceded opinions are effectively made to count towards the opposing side.

4 Experiments

On http://www.convinceme.net, the html page for each debate contains side information for each post ($side_1$ is blue in color and $side_2$ is green). This gives us automatically labeled data for our evaluations. For each of the 4 debates in our test set, we use posts with at least 5 sentences for evaluation.

4.1 Baselines

We implemented two baselines: the OpTopic system that uses topic information only, and the OpPMI system that uses topic as well as related word (noun) information. All systems use the same lexicon, as well as exactly the same processes for opinion finding and opinion-target pairing.

The OpTopic system This system considers only explicit mentions of the topic for the opinion analysis. Thus, for this system, the step of opinion-target pairing only finds all $topic_1^+$, $topic_1^-$, $topic_2^+$, $topic_2^-$ instances in the post (where, for example, an instance of $topic_1^+$ is a positive opinion whose target is explicitly $topic_1$). The polarity-topic pairs are counted for each debate side according to the following equations.

$$score(side_1) = \#topic_1^+ + \#topic_2^- \quad (10)$$

$$score(side_2) = \#topic_1^- + \#topic_2^+ \quad (11)$$

The post is assigned the side with the higher score.

The OpPMI system This system finds opinion-target pairs for not only the topics, but also for the words in the debate that are significantly related to either of the topics.

We find semantic relatedness of each noun in the post with the two main topics of the debate by calculating the Pointwise Mutual Information (PMI) between the term and each topic over the entire web corpus. We use the API provided by the Measures of Semantic Relatedness (MSR)[4] engine for this purpose. The MSR engine issues Google queries to retrieve documents and finds the PMI between any two given words. Table 3 lists PMIs between the topics and the words from Table 2.

Each noun k is assigned to the topic with the higher PMI score. That is, if
$PMI(topic_1,\text{k}) > PMI(topic_2,\text{k}) \Rightarrow \text{k}= topic_1$
and if
$PMI(topic_2,\text{k}) > PMI(topic_1,\text{k}) \Rightarrow \text{k}= topic_2$
Next, the polarity-target pairs are found for the post, as before, and Equations 10 and 11 are used to assign a side to the post as in the OpTopic system, except that here, related nouns are also counted as instances of their associated topics.

[4] http://cwl-projects.cogsci.rpi.edu/msr/

word	iPhone	blackberry
storm	0.923	0.941
phone	0.908	0.885
e-mail	0.522	0.623
ipod	0.909	0.976
battery	0.974	0.927
network	0.658	0.961
keyboard	0.961	0.983

Table 3: PMI of words with the topics

4.2 Results

Performance is measured using the following metrics: Accuracy ($\frac{\#Correct}{\#Total\ posts}$), Precision ($\frac{\#Correct}{\#guessed}$), Recall ($\frac{\#Correct}{\#relevant}$) and F-measure ($\frac{2*Precision*Recall}{(Precision+Recall)}$).

In our task, it is desirable to make a prediction for all the posts; hence $\#relevant = \#Total\ posts$. This results in Recall and Accuracy being the same. However, all of the systems do not classify a post if the post does not contain the information it needs. Thus, $\#guessed \leq \#Total\ posts$, and Precision is not the same as Accuracy.

Table 4 reports the performance of four systems on the test data: the two baselines, our method using the preferences learned from the web corpus (OpPr) and the method additionally using discourse information to reverse conceded opinions.

The OpTopic has low recall. This is expected, because it relies only on opinions explicitly toward the topics.

The OpPMI has better recall than OpTopic; however, the precision drops for some debates. We believe this is due to the addition of noise. This result suggests that not all terms that are relevant to a topic are useful for determining the debate side.

Finally, both of the OpPr systems are better than both baselines in Accuracy as well as F-measure for all four debates.

The accuracy of the full OpPr system improves, on average, by 35 percentage points over the OpTopic system, and by 20 percentage points over the

OpPMI system. The F-measure improves, on average, by 25 percentage points over the OpTopic system, and by 17 percentage points over the OpPMI system. Note that in 3 out of 4 of the debates, the full system is able to make a guess for all of the posts (hence, the metrics all have the same values).

In three of the four debates, the system using concession handling described in Section 3.4 outperforms the system without it, providing evidence that our treatment of concessions is effective. On average, there is a 3 percentage point improvement in Accuracy, 5 percentage point improvement in Precision and 5 percentage point improvement in F-measure due to the added concession information.

	OpTopic	OpPMI	OpPr	OpPr + Disc
Firefox Vs Internet explorer (62 posts)				
Acc	33.87	53.23	64.52	66.13
Prec	67.74	60.0	64.52	66.13
Rec	33.87	53.23	64.52	66.13
F1	45.16	56.41	64.52	66.13
Windows vs. Mac (15 posts)				
Acc	13.33	46.67	66.67	66.67
Prec	40.0	53.85	66.67	66.67
Rec	13.33	46.67	66.67	66.67
F1	20.0	50.00	66.67	66.67
SonyPs3 vs. Wii (36 posts)				
Acc	33.33	33.33	56.25	61.11
Prec	80.0	46.15	56.25	68.75
Rec	33.33	33.33	50.0	61.11
F1	47.06	38.71	52.94	64.71
Opera vs. Firefox (4 posts)				
Acc	25.0	50.0	75.0	100.0
Prec	33.33	100	75.0	100.0
Rec	25.0	50	75.0	100.0
F1	28.57	66.67	75.0	100.0

Table 4: Performance of the systems on the test data

5 Discussion

In this section, we discuss the results from the previous section and describe the sources of errors.

As reported in the previous section, the OpPr system outperforms both the OpTopic and the OpPMI systems. In order to analyze why OpPr outperforms OpPMI, we need to compare Tables 2 and 3. Table 2 reports the conditional probabilities learned from the web corpus for polarity-target pairs used in OpPr, and Table 3 reports the PMI of these same targets with the debate topics used in OpPMI. First, we observe that the PMI numbers are intuitive, in that all the words, except for "e-mail," show a high PMI relatedness to both topics. All of them are indeed semantically related to the domain. Additionally, we see that some conclusions of the OpPMI system are similar to those of the OpPr system, for example, that "Storm" is more closely related to the Blackberry than the iPhone.

However, notice two cases: the PMI values for "phone" and "e-mail" are intuitive, but they may cause errors in debate analysis. Because the iPhone and the Blackberry are both phones, the word "phone" does not have any distinguishing power in debates. On the other hand, the PMI measure of "e-mail" suggests that it is not closely related to the debate topics, though it is, in fact, a desirable feature for smart phone users, even more so with Blackberry users. The PMI measure does not reflect this.

The "network" aspect shows a comparatively greater relatedness to the blackberry than to the iPhone. Thus, OpPMI uses it as a proxy for the Blackberry. This may be erroneous, however, because negative opinions towards "network" are more indicative of negative opinions towards iPhones, a fact revealed by Table 2.

In general, even if the OpPMI system knows what topic the given word is more related to, it still does not know what the opinion towards that word *means* in the debate scenario. The OpPr system, on the other hand, is able to map it to a debate side.

5.1 Errors

False lexicon hits. The lexicon is word based, but, as shown by (Wiebe and Mihalcea, 2006; Su and Markert, 2008), many subjective words have both objective and subjective senses. Thus, one major source of errors is a false hit of a word in the lexicon.

Opinion-target pairing. The syntactic rule-based opinion-target pairing system is a large source of errors in the OpPr as well as the baseline systems. Product review mining work has explored finding opinions with respect to, or in conjunction with, aspects (Hu and Liu, 2004; Popescu et al., 2005); however, in our work, we need to find

information in the other direction – that is, given the opinion, what is the opinion about. Stoyanov and Cardie (2008) work on opinion co-reference; however, we need to identify the specific target.

Pragmatic opinions. Some of the errors are due to the fact that the opinions expressed in the post are pragmatic. This becomes a problem especially when the debate post is small, and we have few other lexical clues in the post. The following post is an example:

(4) The blackberry is something like $150 and the iPhone is $500. I don't think it's worth it. You could buy a iPod separate and have a boatload of extra money left over.

In this example, the participant mentions the difference in the prices in the first sentence. This sentence implies a negative opinion towards the iPhone. However, recognizing this would require a system to have extensive world knowledge. In the second sentence, the lexicon does hit the word "worth," and, using syntactic rules, we can determine it is negated. However, the opinion-target pairing system only tells us that the opinion is tied to the "it." A co-reference system would be needed to tie the "it" to "iPhone" in the first sentence.

6 Related Work

Several researchers have worked on similar tasks. Kim and Hovy (2007) predict the results of an election by analyzing forums discussing the elections. Theirs is a supervised bag-of-words system using unigrams, bigrams, and trigrams as features. In contrast, our approach is unsupervised, and exploits different types of information. Bansal et al. (2008) predict the vote from congressional floor debates using agreement/disagreement features. We do not model inter-personal exchanges; instead, we model factors that influence stance taking. Lin at al (2006) identify opposing perspectives. Though apparently related at the task level, perspectives as they define them are not the same as opinions. Their approach does not involve any opinion analysis. Fujii and Ishikawa (2006) also work with arguments. However, their focus is on argument visualization rather than on recognizing stances.

Other researchers have also mined data to learn associations among products and features. In their work on mining opinions in comparative sentences, Ganapathibhotla and Liu (2008) look for user preferences for one product's features over another's. We do not exploit comparative constructs, but rather probabilistic associations. Thus, our approach and theirs are complementary. A number of works in product review mining (Hu and Liu, 2004; Popescu et al., 2005; Kobayashi et al., 2005; Bloom et al., 2007) automatically find features of the reviewed products. However, our approach is novel in that it learns and exploits associations among opinion/polarity, topics, and aspects.

Several researchers have recognized the important role discourse plays in opinion analysis (Polanyi and Zaenen, 2005; Snyder and Barzilay, 2007; Somasundaran et al., 2008; Asher et al., 2008; Sadamitsu et al., 2008). However, previous work did not account for concessions in determining whether an opinion supports one side or the other.

More sophisticated approaches to identifying opinions and recognizing their contextual polarity have been published (e.g., (Wilson et al., 2005; Ikeda et al., 2008; Sadamitsu et al., 2008)). Those components are not the focus of our work.

7 Conclusions

This paper addresses challenges faced by opinion analysis in the debate genre. In our method, factors that influence the choice of a debate side are learned by mining a web corpus for opinions. This knowledge is exploited in an unsupervised method for classifying the side taken by a post, which also accounts for concessionary opinions.

Our results corroborate our hypothesis that finding relations between aspects associated with a topic, but particularized to polarity, is more effective than finding relations between topics and aspects alone. The system that implements this information, mined from the web, outperforms the web PMI-based baseline. Our hypothesis that addressing concessionary opinions is useful is also corroborated by improved performance.

Acknowledgments

This research was supported in part by the Department of Homeland Security under grant N000140710152. We would also like to thank Vladislav D. Veksler for help with the MSR engine, and the anonymous reviewers for their helpful comments.

References

Nicholas Asher, Farah Benamara, and Yvette Yannick Mathieu. 2008. Distilling opinion in discourse: A preliminary study. In *Coling 2008: Companion volume: Posters and Demonstrations*, pages 5–8, Manchester, UK, August.

Mohit Bansal, Claire Cardie, and Lillian Lee. 2008. The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. In *Proceedings of COLING: Companion volume: Posters*.

Kenneth Bloom, Navendu Garg, and Shlomo Argamon. 2007. Extracting appraisal expressions. In *HLT-NAACL 2007*, pages 308–315, Rochester, NY.

Atsushi Fujii and Tetsuya Ishikawa. 2006. A system for summarizing and visualizing arguments in subjective documents: Toward supporting decision making. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 15–22, Sydney, Australia, July. Association for Computational Linguistics.

Murthy Ganapathibhotla and Bing Liu. 2008. Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 241–248, Manchester, UK, August.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *AAAI-2004*.

Daisuke Ikeda, Hiroya Takamura, Lev-Arie Ratinov, and Manabu Okumura. 2008. Learning to shift the polarity of words for sentiment classification. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)*.

Soo-Min Kim and Eduard Hovy. 2007. Crystal: Analyzing predictive opinions on the web. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1056–1064.

Nozomi Kobayashi, Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2005. Opinion extraction using a learning-based anaphora resolution technique. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05), poster*, pages 175–180.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on? Identifying perspectives at the document and sentence levels. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-2006)*, pages 109–116, New York, New York.

Livia Polanyi and Annie Zaenen. 2005. Contextual valence shifters. In *Computing Attitude and Affect in Text*. Springer.

Ana-Maria Popescu, Bao Nguyen, and Oren Etzioni. 2005. OPINE: Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 32–33, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

R. Prasad, E. Miltsakaki, N. Dinesh, A. Lee, A. Joshi, L. Robaldo, and B. Webber, 2007. *PDTB 2.0 Annotation Manual*.

Kugatsu Sadamitsu, Satoshi Sekine, and Mikio Yamamoto. 2008. Sentiment analysis based on probabilistic models using inter-sentence information. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May.

Benjamin Snyder and Regina Barzilay. 2007. Multiple aspect ranking using the good grief algorithm. In *Proceedings of NAACL-2007*.

Swapna Somasundaran, Janyce Wiebe, and Josef Ruppenhofer. 2008. Discourse level opinion interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 801–808, Manchester, UK, August.

Veselin Stoyanov and Claire Cardie. 2008. Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 817–824, Manchester, UK, August. Coling 2008 Organizing Committee.

Fangzhong Su and Katja Markert. 2008. From word to sense: a case study of subjectivity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*, Manchester, UK, August.

Janyce Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of COLING-ACL 2006*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT-EMNLP*, pages 347–354, Vancouver, Canada.