Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference

Automated Classification of Stance in Student Essays: An Approach Using Stance Target Information and the Wikipedia Link-Based Measure

Adam Faulkner
The Graduate Center,
The City University of New York
afaulkner@gc.cuny.edu

Abstract

We present a new approach to the automated classification of document-level argument stance, a relatively under-researched sub-task of Sentiment Analysis. In place of the noisy online debate data currently used in stance classification research, a corpus of student essays annotated for essay-level stance is constructed for use in a series of classification experiments. A novel set of features designed to capture the stance, stance targets, and topical relationships between the essay prompt and the student's essay is described. Models trained on this feature set showed significant increases in accuracy relative to two high baselines.

1 Introduction

This work presents a new approach to the automated classification of document-level argument stance. To date, opinion classification has dominated Sentiment Analysis (SA) research, while sub-tasks of SA such as stance classification (SC) remain under-researched. Traditional approaches to opinion mining cannot be easily ported to SC tasks since there are marked differences between opinion-bearing and stancetaking language. These differences are partly due to the very different objects being evaluated by these two registers. While opinion-bearing language is found mainly in reviews and deals with positive or negative evaluations of entities such as movies, books, and gadgets, writers use stancetaking language in argumentative essays and debates to evaluate the truth or likelihood of propositions such as "God exists" or "Money is the root of all evil." SC features, therefore, must capture somewhat different properties of language than those captured in opinion mining research. In particular, features designed to capture the targets of stancetaking must somehow incorporate information regarding full propositions, rather than entities.

The corpus used in this work consists of argumentative essays written by students in response to prompts such as (1).

(1) The prison system is outdated. No civilized society should punish its criminals: it should rehabilitate them.

When arguing *for* the propositions in this prompt, a student might extract segments of the prompt text and modify those segments with stancetaking words. In (2), the student uses the negative opinion word *old-fashioned* to describe the prompt phrase *prison system* and endorses the target proposition *the prison system is old-fashioned* using the positive stance words *would say*. As example (2) shows, the stance polarity and opinion polarity of a sentence needn't overlap: one can use a negative opinion word such as *old-fashioned* as part of a *for* (=positive) stanced argument. Alternatively, a student might articulate a *for* argument using words that are topically related to the prompt text words *prison* and *criminals*, such as the use of *prisoners* in (3). In (3), the target proposition *prisoners BE rehabilitated* is endorsed with a *for* stance word *should*.

(2) Initially I would say that the prison system is old-fashioned.

(3) Prisoners should be rehabilitated.

One goal of this work is to capture such patterns as classification features.

Currently, there exist no professionally annotated corpora for document-level SC and only a handful of approaches have been developed. In this work, we try to fill these lacunae in the SC research by describing the construction of a corpus annotated for stance at the document level using crowdsourced annotators. We also describe the creation of a novel set of features motivated by linguistic research on stancetaking language. These features are designed to capture the linguistic phenomena evident in examples (2) and (3). We evaluated our classification models using two baselines: a bag-of-words model, and the model presented in Somasundaran and Wiebe (2010), which was also designed to capture stance target information. A series of machine learning experiments conducted using these models showed significant increases in accuracy relative to both baselines. Applications for the models presented in this paper include stance classification of argumentative text varieties that are similar in structure and language to argumentative essays: eRulemaking data, political blogs, and position statements.

2 Related work

The most significant work in document-level stance classification is Somasundaran and Wiebe (2010), Anand et al.

(2011), Walker et al. (2012), and Hasan and Ng (2013a; 2013b). These researchers utilized text collected from online debate forums dealing with issues such as abortion and the existence of God. Users are invited to create debate propositions that can be evaluated via the *for* or *against* distinction such as "Abortion should be legal" and "God exists." These are then debated by users in posts that are self-tagged as *for* or *against*. Somasundaran and Wiebe's approach to capturing stance target information involved tagging all content words in each debate post with the stance polarity of the sentence in which it is located. Sentence-level stance was determined by summing likelihood scores associated with sentence words matched to a stance lexicon compiled from stance annotated spans in the Multi-Perspective Question Answering corpus (MPQA) (Wiebe, Wilson, and Cardie 2005). Working with a corpus of 2232 debate posts, their highest scoring system had an accuracy of .63 using a combination of stance and opinion tagged content words.

The system of Anand et al. was trained on a debates corpus of 4772 posts. An accuracy of .64 was achieved using ngrams, syntactic dependencies, and various other features including sentence length and pronominal form occurrence. Anand et al. take a coarse-grained approach to capturing stance targets. Each post is reduced to *(opinion, target)* dependency tuples, such as *(overwhelming, evidence)*, which were then generalized to tuple members' parts of speech (POS) to create tuples such as *(JJ, evidence)* and *(overwhelming,NN)*. Hasan and Ng (2013a; 2013b) experimented with the feature of Anand et al., "extra-linguistic constraints"' such as the stance of an immediately preceding post (an approach to capturing inter-post relations is also presented in Walker et al. 2012), the writer's stance towards other debate topics, and features based on frame semantics. The highest accuracy reported was .74 for a set of debates on abortion.

An issue left unaddressed by these researchers is whether online debate data are truly representative of argumentative text. The language of these debates is often highly emotion-laden (a feature of opinion-bearing, rather than stancetaking language), sometimes consists of only one or two sentences, and displays few of the text organization features typical of argumentative text (premise-conclusion sequences with associated discourse markers such as *therefore*, *if...then*, and *because*). In this study, we address this point using quintessentially argumentative data, student argumentative essays.

3 Corpus description and annotation steps

The test bed for the experiments reported here is the International Corpus of Learner English (ICLE) of Granger (2003). ICLE is a collection of largely argumentative essays written by non-native speakers of English, with each essay responding to one of 14 separate essay prompts. We chose this corpus for the current study because of its size, diversity of topics, and availability. Additionally, these essays maintain the argumentative scenario described in linguistic research dealing with stancetaking language and typified by online debate forums: a proposition is given, and speakers/writers are asked to argue *for* or *against* that proposition. An alternative student essays corpus, described in Römer and O'Donnell (2011), is written by native speakers of English but does not maintain this scenario. Essays responding to 7 of 14 prompts were chosen. These 7 prompts contained unambiguous use of propositions similar to those found in the corpus of Anand et al. and Somasundaran and Wiebe. Additionally, there were at least 100 responses to each of these prompts which allowed us to maintain a relatively uniform distribution of topics. We manually pruned any essays that were glaringly incoherent.

Prompt topic	#essays	%*for*	%*ag.*	%*neith.*
Draft	126	.74	.19	.07
Television	156	.83	.13	.04
Feminism	153	.27	.64	.09
Money/evil	103	.49	.42	.09
Prison	126	.74	.19	.07
Science	376	.12	.86	.02
University	279	.51	.45	.04
Total	**1319**	**.45**	**.50**	**.05**

Table 1: Distribution of essay prompts provided to annotators along with percentages of gold standard *for, against*, and *neither* tags for each prompt.

All of our annotations were collected using non-expert annotators recruited from the online crowdsourcing service, Crowdflower (CF). [1] Crowdsourcing services such as CF have proven to be reliable sources of non-expert annotations of short texts in SA (Mellebeek et al. 2010). All 1319 essays were posted to CF and five unique annotators were recruited to annotate each essay. For each essay, the essay prompt associated with that essay was included on the same screen. Annotators were asked to read both the essay prompt and the essay in their entirety and to tag each essay as displaying a *for* or *against* stance toward the given prompt. If the essay did not display a stance toward the prompt, annotators were asked to tag it as *neither*. Gold-standard annotation was performed by the author. Table 1 gives the distribution of essays by prompt topic along with percentages of gold-standard *for*, *against*, and *neither* tags.

Interannotator agreement was calculated between the gold-standard and the CF-tagged corpus. Annotation tags for the CF-tagged corpus were determined using majority voting and random tie-breaking. Agreement adjusted for chance was calculated using Cohen's κ which was .68 for the entire corpus. This score compares favorably with the .72 κ score of Mellebeek et al. for a similar crowdsourced document-level sentiment annotation task.

4 Classification features

Part-of-speech generalized dependency subtrees

Our first set of features captures word-level information regarding stance polarity together with information regarding the proposition targeted by that stance word. In SA, word-level polarity information (e.g, *great* has a positive po-

[1] http://www.crowdflower.com/

larity, while *terrible* has negative polarity) is usually captured using manually or automatically compiled lexicons. Since these resources are meant for use in SA systems dealing with review language, which involves entities such as movies or gadgets, the language of these lexicons tends to be adjectival (a *great* movie, a *terrible* phone). However, as argued in Hunston and Thompson (2000) and Martin and White (2005), writers and speakers take stances on propositions using evidential markers, such as modals (*ought, should*) and modal adverbs (*possibly, certainly*), which modify whole clauses, rather than adjectives such as *great* or *terrible*, which modify noun phrases. This has led researchers to make a distinction between opinion-bearing language and stancetaking language based on the semantic class of the target of the opinion or stance: opinions take entities as targets while stances take propositions as targets. This distinction is illustrated by the opinion-bearing sentence in (4) and the stancetaking sentences (5) and (6). The targeted entity and propositions of (4) and (5-6), respectively, are bracketed. All opinion-bearing language in (4) and stancetaking language in (5-6) is boldfaced.

(4) "Snake Eyes" is the most **aggravating** kind of [movie]: the kind that shows so much potential and then becomes **unbelievably disappointing**. (opinion=positive)

(5) This **indicates** that [our prisons are higher institutions for criminals].(stance=*for*)

(6) So we **can infer** that [the statement is very true]. (stance=*for*)

To capture evidential word occurrence in the ICLE essays we made use of two resources. The first is a lexicon of stance words created by replicating the methodology described in Somasundaran and Wiebe (2010). Using all text spans from the MPQA corpus annotated for stance, the initial ngram (up to three) of each stemmed text span was extracted and assigned the stance polarity of the span from which it was extracted. Since many of these ngrams will appear in both *for* and *against* spans, each receives a score indicating the likelihood that it is a *for* or *against* expression. This is calculated as the candidate's frequency of occurrence in a *for* (*against*) arguing span divided by its frequency of occurrence in the entire MPQA corpus. i.e., $P(\mathit{for}|\text{ngram}) = \frac{\text{\#ngram is in a } \mathit{for} \text{ span}}{\text{\# ngram is in the MPQA corpus}}$ or $P(\mathit{against}|\text{ngram}) = \frac{\text{\#ngram is in an } \mathit{against} \text{ span}}{\text{\# ngram is in the MPQA corpus}}$

The higher of these scores determines an ngram's final label as *for* (*against*).This resulted in a lexicon of 2166 *for* and 513 *against* ngrams. We manually examined the resulting ngram lexicon and found that its noisiest sections involved bigrams and trigrams. We used only the unigrams in the lexicon since these appeared to be more reliably scored as *for* (*against*). We also pruned any unigrams that were obviously not stancetaking (e.g., *thanks, anybody, pre-election, suspicions*).

To supplement this list, we also used a selection of the metadiscourse markers listed in the appendix of Hyland (2005). Markers from the following categories were used: boosters (*clearly, decidedly*), hedges (*claim, estimate*), and engagement markers (*demonstrate, evaluate*). All of these markers were adjudged positively stanced by the criteria given in Martin and White (2005) and thus were added to the list of *for* unigrams. With Hyland's metadiscourse markers added to the initial lexicon, the final lexicon consisted of 373 *for* and 80 *against* unigrams.

Our next task was to capture the targets of the stancetaking language in each ICLE essay. As mentioned, syntactically the targets of stancetaking are clausal rather than nominal. We thus cannot make use of extant approaches in opinion classification, such as Kim and Hovy (2006) and Popescu and Etzioni (2007), which use nominal material such as noun phrases as target information. Our alternative approach is motivated by Martin and White's notion that stancetaking involves a speaker/writer arguing *for* (*against*) an "attitudinal assessment" (Martin & White, 2005: 95). In general, a proposition will contain a certain amount of attitudinal (more commonly known as opinion-bearing) language and the act of stancetaking can be reduced to taking a *for* or *against* stance toward this attitudinal assessment. In (6), for example, the writer takes a *for* stance toward the attitudinal assessment *the statement is very true* using *can* and *infer*. To capture this, we perform the following steps. Using the MPQA subjectivity lexicon of opinion-bearing words (Wiebe, Wilson, and Cardie 2005), we first find any opinion-bearing word(s) in the proposition *the statement is very true*. These words will serve as *proxies* for the target proposition. In the case of (6), we find a match in the positive section of the lexicon for *true*. Sentence (6), then, can be reduced the *(stance, proposition target)* tuples *(can, true)* and *(infer, true)*.

An additional advantage of this approach is that it can capture formulaic lexico-syntactic patterns, which can then be leveraged as classification features. The pattern *can-V-true*, for example, where V is the main verb, is often found in *for*-stanced sentences such as (6 - 8).

(7) Some day our dreams **can** come **true.** → *can-come-true*

(8) I **can** only say that this statement is completely **true.** → *can-say-true*

To capture such patterns along with long distance dependencies—as occurs between *infer* and *true* in (6)— we used a dependency-parse representation. Additionally, to increase feature redundancy, we partially POS-generalized all subtrees identified in each dependency-parse. These *POS-generalized dependency subtrees* were identified in the following manner:

- **Step 1**. Using the Stanford parser (De Marneffe et al., (2008)) each sentence in a given essay is given two structural representations: a dependency parse and a phrase structure parse.
- **Step 2**. Any stancetaking and opinion-bearing language in the dependency parse is located using the stance lexicon and the MPQA subjectivity lexicon.

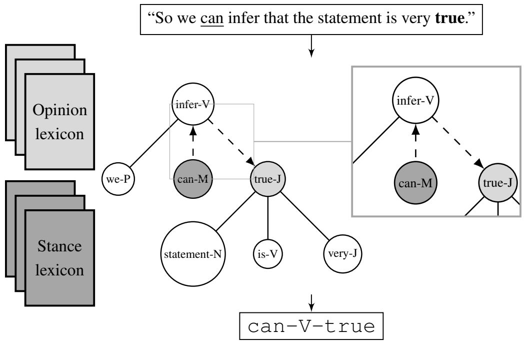


Figure 1: Scheme showing the creation of a POS-generalized stance proposition subtree using dependency parsing, lexicon lookup, tree traversal, and POS generalization.

- **Step 3**. If the immediately neighboring node of a stance word node contains a negator, the polarity of the stance word is reversed by appending *not* to the word.
- **Step 4**. Starting at the stance word node, an undirected version of the dependency tree is traversed in breadth-first fashion until a node containing an opinion-bearing word is found. The phrase structure parse is examined to see if the opinion-bearing word is located in the immediate or embedded clause of the stance word. If these syntactic restrictions have been met, the opinion-bearing word is considered a good proxy for the proposition targeted by the stance word. The subtree containing the stance and opinion-bearing nodes is returned.
- **Step 5**. The mid-nodes between the stance and opinion-bearing nodes of all sub-trees returned in steps 1-4 are then POS-generalized.

These steps are summarized in Figure 1. Examples of features generated using this procedure are *because-important, can-V-right, should-punish, not-should-fear*, and *not-express-N-N-lost*.

Prompt topic words

Our second set of features captures the relationship between the language of the essay prompt and the language of the ICLE essay responding to that prompt. That there exists a relationship between a given prompt and the language of an essay responding to that prompt is obvious from the lexical overlap (indicated by boldfaced sections) in prompt/response pairs such as (9).

(9) *Prompt:* **Most university degrees are theoretical and do not prepare students for the real world**. They are therefore of very little value.

Response: Nowadays there have been many debates on whether **most university degrees are theoretical and don't prepare students for real world** or not.

To capture essay language that is topically related to the prompt text, we use a term similarity metric that is able to capture both simple lexical overlap, as occurs in (9), along with words that are related to prompt words by cultural association. Consider the essay prompt (10), and the two sentences (11) and (12), taken from essays responding to this prompt. Restricting our attention to the boldfaced content words in the prompt text we find several words (also boldfaced) in the response sentences that are related either by lexical or cultural association to these content words and hence likely deal with the same topics as the prompt text.

(10) *Prompt*: In the words of the old song, "**Money** is the root of all **evil**."

(11) **Rich** people can go any where they want to look for the cure of their diseases, whereas the **poor** don't even be diagnosed as they can't go to a doctor.

(12) **Raskolnikov** killed the old woman because he decided that according to his theory such deed can be done.

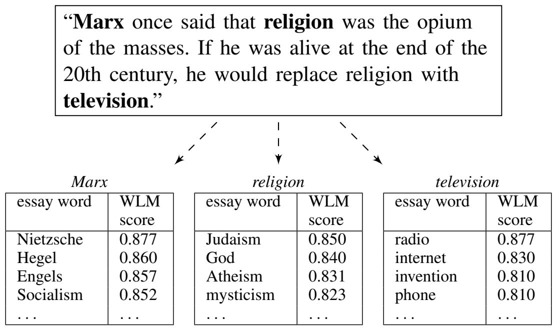
In (11), *rich* and *poor* deal with the subject of money since both these words are definitionally related to *money*: to have a lot of money is to be rich; to have very little money is to be poor. In (12), *Raskolnikov* has a cultural association with *evil* and *money* by virtue of the subject matter of the novel *Crime and Punishment*, which deals with the murder of a pawnbroker for money and relativistic notions of evil.

To capture these associations, we first experimented with standard similarity metrics such as LSA and various WordNet-based measures, but were unsatisfied with the results. Cultural kinds are grouped by association rather than by any principled measures of semantic similarity and so a semantic similarity metric based on a corpus dealing with a vast number of topics was required. Our first choice was Cilibrasi and Vitanyi's (2007) Normalized Google Distance (NGD), a web-based metric that uses Google hit counts to associate words. Unfortunately, large scale ($\geq$ 100 word pairs) use of NGD is prohibitively expensive for researchers due to Google's search API fees. Instead, we chose a Wikipedia-based similarity metric inspired by NGD, Witten and Milne's (2008) Wikipedia Link-based Measure (WLM). The typical Wikipedia[2] page contains a large network of cross-references in the form of internal (connected to another Wikipedia page) and external (connected to a page outside of Wikipedia) hyperlinks. The WLM uses this inter-link structure to score term similarity. Witten and Milne define the WLM as

$$wlm(a,b) = \frac{log(max(|A|,|B|) - log(|A \cap B|)}{log(|W|) - log(min(|A|,|B|))}$$

where a and b are Wikipedia article titles (e.g., the articles for *evil* and *Raskolnikov*), A and B are the sets of articles that backlink to a and b, and W is the count of all articles currently contained in Wikipedia (as of this writing, $\sim$ 4.3 million). As given, if $wlm(a,b) = 0$ then a and b are as semantically similar as possible and if $wlm(a,b) \geq 1$ then they are semantically dissimilar. For ease of interpretation,

[2] http://en.wikipedia.org/wiki/Main_Page



"**Marx** once said that **religion** was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with **television**."

Marx

essay word	WLM score
Nietzsche	0.877
Hegel	0.860
Engels	0.857
Socialism	0.852
...	...

religion

essay word	WLM score
Judaism	0.850
God	0.840
Atheism	0.831
mysticism	0.823
...	...

television

essay word	WLM score
radio	0.877
internet	0.830
invention	0.810
phone	0.810
...	...

Figure 2: Examples of WLM scored topic words occurring in essays responding to the *Television* prompt. Essay words are scored relative to the boldfaced words in the prompt text.

we subtract all WLM scores from 1, so that that a score of 1 means that *a* and b are as similar as possible.

For each essay in the corpus, a set of words with WLM scores ≥ 0 were created using the following procedure:

- **Step 1**. Using the stance lexicon, all stance words in a given essay were identified.
- **Step 2**. The phrase-structure representations created for our first feature set were used to identify propositions in the immediate or embedded clause of the identified stance word.
- **Step 3**. For each content word in the prompt to which the essay is responding, a WLM score was calculated relative to all content words contained in the proposition identified in Step 2.
- **Step 4**. Many topic words received WLM scores ≥ 0 for more than one prompt word. In such cases, the highest WLM score is assigned to that topic word.

Figure 2 shows three sets of high scoring topic words generated by this procedure. Each set corresponds to a boldfaced content word in the prompt.

5 Experiment setup and results

We chose two basic learning algorithms for our experiments, multinomial Naive Bayes (MNB) and SVMs (RBF kernel). Of the original set of 1319 annotated essays, the 65 essays tagged neutral were discarded,19 essays were not used since the Stanford parser found them unparseable, and an additional 100 were used as a development set. This left us with 1135 essays, divided into 564 (=49.7%) *for* and 571 (=50.3%) *against*. Rather than using the rather low majority class percentage as a baseline, we made the classification task more challenging by using two alternative models as baselines. The first baseline is a bag-of-words model, which is often difficult to beat in SA experiments. The second baseline is a model based on the feature sets described in Somasundaran and Wiebe (2010). Somasundaran and Wiebe report the results of experiments involving three different feature set types. We used their highest-scoring reported feature set type, which is a combination of two feature sets: an unordered vector of stance polarity-tagged stemmed content words (described above in section 2) and an unordered vector of opinion polarity-tagged stemmed content words.While the feature sets of Anand et al. (2011) and Hasan and Ng (2013a; 2013b) are somewhat more sophisticated than that of Somasundaran and Wiebe, we used the latter-most feature set as a baseline since it incorporates information regarding stance targets and does so by making use of a larger version of the stance lexicon also used here. By comparing our two systems, we can get a sense of whether our approach to incorporating stance target information represents an improvement over that of Somasundaran and Wiebe.

We experimented with three different feature sets. The first feature set uses the stance-proposition feature representation framework described in section 4 to represent each essay as an unordered set of POS-generalized stance-proposition subtrees. The second set represents each essay as an unordered collection of stemmed, WLM-scored topic words extracted using the procedure also described in section 4. The third set combines the first two sets. We split all essays into training and test sets using 10-fold cross-validation. All experiments were performed using the Weka machine learning toolkit of Hall et al. (2009).

We experimented with 10 different versions of the topic words and combined topic words/stance-proposition subtree feature sets, with each version containing topic words with WLM scores $\geq$ a given threshold. In Table 2, we present the highest-scoring version of topic words and combined topic words and stance-proposition subtree features relative to WLM score threshold. Table 3 provides p-value scores for each classifier's highest scoring model relative to each baseline. We used McNemar's χ^2 test to measure statistical significance (Table 3).

	Acc.	Prec.	Rec.	F
Multinomial NB				
SP trees	78.5	79.2	78.6	78.7
Topic words (WLM score $\geq$ 0)	79.0	79.9	79.0	79.1
SP trees + Topic words (WLM score $\geq$ .8)	**80.0**	**81.4**	**80.1**	**80.2**
Baseline 1: Bag of words	67.9	67.9	67.9	67.9
Baseline 2: S&W features	72.5	72.7	72.6	72.5
SVM (RBF kernel)				
SP trees	67.3	75.4	67.3	62.3
Topic words (WLM score $\geq$ 0)	81.7	81.8	81.8	81.8
SP trees + Topic words (WLM score $\geq$ 0)	**82.0**	**82.0**	**82.0**	**82.0**
Baseline 1: Bag of words	77.8	78.0	77.9	77.9
Baseline 2: S&W features	73.8	73.8	73.8	73.8

Table 2: Essay-level stance classification experiment results for the highest scoring feature set/classifier combinations.

6 Discussion

All of our models beat both baselines. The highest-accuracy model overall was an SVM classifier trained on a combination of stance-proposition subtrees and topic words (WLM-score $\geq$ .0). MNB also acheved its highest accuracy using a combination of SP trees and topic words. We observed significant improvement ($p <$ .002) in both classifier models when topic words were added to SP trees. The key role played by features capturing the relationship between prompt and response in student essays is evidence

SVM (RBF kernel)

Null hypothesis	p-*value*
Baseline 1 vs. SP trees + Topic words (WLM score $\geq$ 0)	1.423×10^{-11}
Baseline 2 vs. Topic words (WLM score $\geq$ 0)	8.568×10^{-05}

Multinomial NB

Null hypothesis	p-*value*
Baseline 1 vs.SP trees + Topic words (WLM score $\geq$ .8)	0.003133
Baseline 2 vs. SP trees + Topic words (WLM score $\geq$ .8)	2.328×10^{-08}

Table 3: Significance of model accuracy relative to both baselines. Significance was measured using McNemar's χ^2 test.

of another important difference between stancetaking and opinion-bearing language. Unlike opinion polarity, which can be recognized *a priori* as positive or negative, (*great* is always positive; *terrible* is always negative), stance polarity must always be evaluated relative to the proposition it is evaluating: the stance word *should* can be used to argue *for* the proposition *Prisoners must be rehabilitated* (e.g., *Prisoners should be reintegrated into society*) but it can also be used to argue *against* that same proposition (e.g., *Prisoners should be punished*). Topic features can help to capture these subtleties by incorporating information regarding words that are topically related to the prompt such as *reintegrated* and *punished*.

7 Conclusion and future work

We presented a new approach to document-level SC using two feature sets designed to capture the linguistic characteristics of stancetaking language. To test the effectiveness of features based on linguistic research involving argumentative language, we constructed a corpus of student essays annotated for stance at the essay level. This corpus served as a more representative example of argumentative text than the noisy online debate text currently used in SC research. We conducted classification experiments using our linguistically motivated features and beat two high baseline models, a bag-of-words model and a model based on the features of Somasundaran and Wiebe (2010).

Future work will involve the construction of a larger SC corpus using expert, rather than crowdsourced, annotators. Additionally, future experimental models will incorporate text-level structural information such as the position of stancetaking sentences in the essay. This will necessitate the creation of a sentence-level SC model in addition to a document-level model.

References

Anand, P.; Walker, M.; Abbott, R.; Tree, J. E. F.; Bowmani, R.; and Minor, M. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, 1–9. Stroudsburg, PA, USA: ACL.

Cilibrasi, R. L., and Vitanyi, P. M. 2007. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering,* 19(3):370–383.

De Marneffe, M.-C., and Manning, C. D. 2008. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, 1–8. ACL.

Granger, S. 2003. The international corpus of learner english: A new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly* 37(3):538–546.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11(1):10–18.

Hasan, K. S., and Ng, V. 2013a. Extra-linguistic constraints on stance recognition in ideological debates. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 816–821. Sofia, Bulgaria: ACL.

Hasan, K. S., and Ng, V. 2013b. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 1348–1356. Nagoya, Japan: Asian Federation of Natural Language Processing.

Hunston, S., and Thompson, G. 2000. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford University Press.

Hyland, K. 2005. *Metadiscourse: Exploring Interaction in Writing*. Continuum International Publishing Group.

Kim, S.-M., and Hovy, E. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, 1–8. ACL.

Martin, J. R., and White, P. R. 2005. *The Language of Evaluation*. Palgrave Macmillan Basingstoke and New York.

Popescu, A.-M., and Etzioni, O. 2007. Extracting product features and opinions from reviews. In *Natural language processing and text mining*. Springer. 9–28.

Römer, U., and O'Donnell, M. B. 2011. From student hard drive to web corpus (part 1): the design, compilation and genre classification of the michigan corpus of upper-level student papers (micusp). *Corpora* 6(2).

Somasundaran, S., and Wiebe, J. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 116–124. ACL.

Walker, M. A.; Anand, P.; Abbott, R.; and Grant, R. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 592–596. ACL.

Wiebe, J.; Wilson, T.; and Cardie, C. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39(2-3):165–210.

Witten, I., and Milne, D. 2008. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, 25–30.