Argument Relation Classification Using a Joint Inference Model

Yufang Hou and **Charles Jochim**
IBM Research – Ireland
{yhou|charlesj}@ie.ibm.com

Abstract

In this paper, we address the problem of argument relation classification where argument units are from different texts. We design a joint inference method for the task by modeling *argument relation classification* and *stance classification* jointly. We show that our joint model improves the results over several strong baselines.

1 Introduction

What is a good counterargument or support argument for a given argument? Despite recent advances in computational argumentation, such as argument unit (e.g., claims, premises) mining (Habernal and Gurevych, 2015), argumentative relation (e.g., support, attack) prediction between argument units from the same text (Stab and Gurevych, 2014; Nguyen and Litman, 2016), as well as assessing argument strength of essays (Persing and Ng, 2015) or predicting convincingness of Web arguments (Habernal and Gurevych, 2016), this question is still an unsolved problem.

In this work we focus on the problem of argument relation classification where argument units are from different texts, i.e., given a set of arguments related to the same topic, we aim to predict relations (e.g., *agree* or *disagree*) between any two arguments. We are aware of argumentative relations between premises and the conclusion within a structured argument. Instead, here we are interested in modeling relations among atomic argument units in dialogic argumentation. This task is important for argumentation in debates (Zhang et al., 2016), stance classification (Sridhar et al., 2015), or persuasion analysis (Tan et al., 2016), among others.

There are various different views on the meaning of "support" and "attack" in argumentation theory (Cayrol and Lagasquie-Schiex, 2005, 2013). In this paper, we use "agree" and "disagree" to represent relations between two arguments which bear a stance regarding the same topic. Specifically, if a_1 agrees with a_2 regarding the topic t then a_1 and a_2 are conflict-free. And if a_1 disagrees with a_2 then they are not conflict-free.

There is a close relationship between *argument relation classification* and *stance classification*. First, argument relation classification can benefit from knowing the stance information of arguments. Specifically, if two arguments hold different stances with regard to the same topic, then they likely disagree with each other. Likewise, two arguments that hold the same stance regarding the same topic tend to agree with each other. Secondly, stance classification can benefit from modeling relations between arguments. For instance, we would expect two arguments that disagree with each other to hold different stances.

There has been a large amount of work focusing on stance classification in on-line debate forums by integrating disagreement information between posts connected with reply links (Somasundaran and Wiebe, 2009; Murakami and Raymond, 2010; Sridhar et al., 2015). However, disagreement information is mainly used as an auxiliary variable and is not explicitly evaluated. Our goal in this paper is to examine argument relation classification in dialogic argumentation. Our task is more challenging because unlike most previous work on disagreement classification, which can explore meta information (e.g., reply links between posts are strong indicators of disagreement), we are only provided with text information (see examples in Table 1).

In this paper, we model *argument relation classification* and *stance classification* jointly. We evaluate our model on a dataset extracted from De-

Proceedings of the 4th Workshop on Argument Mining, pages 60–66

Debate Topic: *Are genetically modified foods (GM foods) beneficial?*		
Sub Topic: *Consumer safety*		
Arg (1)	Pro	Foods with poisonous allergens can be modified to reduce risks.
Arg (2)	Pro	GM crops can be fortified with vitamins and vaccines.
Arg (3)	Con	There are many instances of GM foods proving dangerous.
Sub Topic: *socio-economic impacts*		
Arg (4)	Pro	GM crops are made disease-resistant, which increases yields.
Arg (5)	Con	GM agriculture threatens the viability of traditional farming communities.
Arg (6)	Pro	GM crops generate greater wealth for farming communities.

Table 1: Examples of Debatepedia structure: arguments are organized into different sub-topics, each argument holds a stance regarding the topic.

batepedia[1]. We show that the joint model performs better than several strong baselines for *argument relation classification*. To our knowledge, this is the first work applying joint inference on argument relation classification on dialogic argumentation.

2 Related Work

Argument unit mining. Recent achievements in argument unit mining on different genres has provided us with high quality input for argument relation mining. Teufel (1999) proposed an Argumentative Zoning model for scientific text. Levy et al. (2014) and Rinott et al. (2015) extracted claims and evidences from Wikipedia respectively. Habernal and Gurevych (2015) focused on mining argument components from user-generated Web content. Lippi and Torroni (2016) extracted claims from political debates by utilizing speech features.

Argumentative relation classification. Most existing work on argumentative relation focuses on classifying relations between argument units of monologic argumentation, from a single text. One line of research (Stab and Gurevych, 2014; Persing and Ng, 2016; Nguyen and Litman, 2016) extracted argument units and predicted relations (i.e., support, attack, none) between argument units in persuasive student essays. Peldszus and Stede (2015) identified the argument structure of short texts in a bilingual corpus. In contrast, in our work the argument units are from different texts. Therefore, we do not have discourse connectives (e.g., "on the contrary" or "however") which usually are strong indicators for argument relations.

Cabrio and Villata (2012) used a textual entailment system to predict argument relations between argument pairs which are extracted from Debatepedia. An argument pair could be an argument coupled with the subtopic, or an argument coupled with another argument of the opposite stance.

Recently, Menini and Tonelli (2016) predicted agreement/disagreement relations between argument pairs of dialogic argumentation in the political domain. The authors also create a large agreement/disagreement dataset by extracting arguments from the same sub-topic of Debatepedia. However, they only consider argument pairs that share a topic keyword. We do not have such constraints (see Arg (1) and Arg (2) in Table 1). In addition, they use SVM while we do joint inference.

Stance classification. There has been an increasing interest on modeling stance in debates (e.g., congressional debates or online political forums) (Thomas et al., 2006; Somasundaran and Wiebe, 2009; Murakami and Raymond, 2010; Walker et al., 2012; Gottipati et al., 2013; Hasan and Ng, 2014). As discussed in Section 1, there is a close relationship between stance classification and argument relation classification. For instance, Sridhar et al. (2015) showed that stance classification in online debate forums can benefit from modeling disagreement of the reply links (e.g., you could assume an argument is attacking the preceding argument). In our work, we focus on modeling argument relations.

Joint inference and Markov logic networks. Markov logic networks (MLNs) (Domingos and Lowd, 2009) are a statistical relational learning framework that combine *first order logic* and *Markov networks*. They have been successfully applied to various NLP tasks such as semantic role labeling (Meza-Ruiz and Riedel, 2009), information extraction (Poon and Domingos, 2010), coreference resolution (Poon and Domingos, 2008) and bridging resolution (Hou et al., 2013). In this paper, we apply MLNs to model *argument relation classification* and *stance classification* jointly.

[1] http://www.debatepedia.org/

Hidden predicates	
$p1$	$relation(a_1, a_2, r)$
$p2$	$stance(a_1, t, s)$
Formulas	
$f1$	$\forall a_1, a_2 \in A:\ relation(a_1, a_2, r) \rightarrow relation(a_2, a_1, r)$
$f2$	$\forall a_1, a_2, a_3 \in A:\ relation(a_1, a_2, \text{“}agree''\text{)} \wedge relation(a_2, a_3, \text{“}agree''\text{)} \rightarrow relation(a_1, a_3, \text{“}agree''\text{)}$
$f3$	$\forall a_1, a_2, a_3 \in A:\ relation(a_1, a_2, \text{“}agree''\text{)} \wedge relation(a_2, a_3, \text{“}disagree''\text{)} \rightarrow relation(a_1, a_3, \text{“}disagree''\text{)}$
$f4$	$\forall a_1, a_2, a_3 \in A:\ relation(a_1, a_2, \text{“}disagree''\text{)} \wedge relation(a_2, a_3, \text{“}disagree''\text{)} \rightarrow relation(a_1, a_3, \text{“}agree''\text{)}$
$f5$	$\forall a_1, a_2 \in A:\ stance(a_1, t, s_1) \wedge stance(a_2, t, s_2) \wedge s_1 \neq s_2 \rightarrow relation(a_1, a_2, \text{“}disagree''\text{)}$
$f6$	$\forall a_1, a_2 \in A:\ stance(a_1, t, s_1) \wedge stance(a_2, t, s_2) \wedge s_1 = s_2 \rightarrow relation(a_1, a_2, \text{“}agree''\text{)}$
$f7$	$\forall a_1, a_2 \in A:\ stance(a_1, t, s_1) \wedge relation(a_1, a_2, \text{“}disagree''\text{)} \wedge s_1 \neq s_2 \rightarrow stance(a_2, t, s_2)$
$f8$	$\forall a_1, a_2 \in A:\ stance(a_1, t, s_1) \wedge relation(a_1, a_2, \text{“}agree''\text{)} \rightarrow stance(a_2, t, s_1)$
$f9$	$\forall a_1, a_2 \in A:\ localRelPrediction(a_1, a_2, r) \rightarrow relation(a_1, a_2, r)$
$f10$	$\forall a_1 \in A:\ localStancePrediction(a_1, t, s) \rightarrow stance(a_1, t, s)$

Table 2: Hidden predicates and formulas used for argument relation classification. a_1, a_2 represent arguments in the topic t. $r \in \{agree, disagree\}$, $s \in \{pro, con\}$.

3 Method

As stated in the introduction, our goal is *argument relation classification* as opposed to stance classification. Therefore, given a topic t and a set of arguments A which belongs to t, instead of finding the position (i.e., *pro* or *con*) a_i ($a_i \in A$) takes with respect to t, we want to predict the relation (i.e., *agree* or *disagree*) between a_i and a_j.

The approach we propose tries to make the best use of the topics and arguments by classifying the stances of arguments and the relations between arguments jointly, using Markov logic networks (MLNs).

More specifically, given a topic t and its argument set A we would like to find the stance s_i for each argument a_i and the relation r_{ij} between argument a_i and a_j ($a_i, a_j \in A$) jointly. Let r_{ij} be a relation assignment for an argument pair $a_i, a_j \in A$, R_A be a relation classification result for all arguments in A, R_A^n be the set of all relation classification results for A. Let s_a be a stance prediction for an argument $a \in A$, S_A be a stance prediction result for arguments in A, S_A^n be the set of all possible stance prediction results for A. Our joint inference for *argument relation classification* and *stance classification* can be represented as a log-linear model:

$$P(R_A, S_A | A; w) = \frac{\exp(w \cdot \Phi(A, R_A, S_A))}{\sum_{R_A' \in R_A^n, S_A' \in S_A^n} \exp(w \cdot \Phi(A, R_A', S_A'))}$$

where w is the model's weight vector, $\Phi(A, R_A, S_A)$ is a "global" feature vector which takes the entire relation and stance assignments for all arguments in A into account. We define $\Phi(A, R_A, S_A)$ as:

$$\begin{aligned}\Phi(A, R_A, S_A) = &\sum_{l \in F_r} \sum_{a_i, a_j \in A} \Phi_l(a_i, a_j, r_{ij}) \\ &+ \sum_{k \in F_s} \sum_{a \in A} \Phi_k(a, s_a) \\ &+ \sum_{g \in F_g} \sum_{a_i, a_j \in A} \Phi_g(r_{ij}, s_{a_i}, s_{a_j})\end{aligned}$$

where $\Phi_l(a_i, a_j, r_{ij})$ and $\Phi_k(a, s_a)$ are local feature functions for *argument relation classification* and *stance classification*, respectively. The former looks at two arguments a_i and a_j, the latter at the argument a and the stance s_a. The global feature function $\Phi_g(r_{ij}, s_{a_i}, s_{a_j})$ looks at the relation and stance assignments for a_i and a_j at the same time (see $f5 - f8$ in Table 2).

This log-linear model can be represented using Markov logic networks (MLNs). Table 2 shows formulas for modeling the problem in MLNs. $p1$ and $p2$ are hidden predicates that we predict, i.e., predicting the relation (i.e., *agree* or *disagree*) between a_1 and a_2, and deciding the stance (i.e., *pro* or *con*) of a_1. $f1$ models the symmetry of argument relation. $f2$ models the transitivity of the agree relation. $f3$ and $f4$ model agree/disagree relations among three arguments. $f5 - f8$ model mutual relation between the two hidden predicates, i.e., arguments holding the same/different stance are likely to agree/disagree with each other. $f9$ and $f10$ integrate predictions from the local classifier for argument relation classification and stance classification respectively.

4 Experiments

4.1 Dataset

Debatepedia is an encyclopedia of arguments collected from different sources on debate topics.

Each debate topic is organized hierarchically. It contains background of the topic and usually a number of subtopics, with *pro* and *con* arguments for or against each subtopic (see Table 1 for an example). An argument typically includes a claim and a few supporting evidences.

	Training	Dev	Testing
topics	607	25	25
subtopics	2512	173	176
arguments	15700	968	1037
— pro	7920	472	534
— con	7780	496	503
arg pairs from same subtopics			
agree arg pairs	28271	1713	1828
disagree arg pairs	30759	1893	2078

Table 3: Training, development and testing data.

We create a corpus by extracting all subtopics and their arguments from Debatepedia. We pair all arguments from the same subtopic and label every argument pair as "agree" (for arguments holding the same stance) or "disagree" (for arguments holding the opposite stance). In total we collect data from 657 topics. We reserve 25 topics as the development set and 25 topics as the test set, using the remaining 607 topics for the training set. Table 3 gives an overview of the whole corpus.[2]

4.2 Experimental Setup

Local argument relation classification (*localRel*). We employ logistic regression to train a local argument relation classification model using agree and disagree pairs from the training set. Our local classifier replicates, to the extent possible, the state-of-the-art local stance classifier from Walker et al. (2012) used by Sridhar et al. (2015) as well as the disagreement classifier from Menini and Tonelli (2016). We include features of unigrams, all word pairs of the concatenation of two arguments, the overall sentiment of each argument from Stanford CoreNLP (Socher et al., 2013; Manning et al., 2014), the content overlap of two arguments, as well as the number of negations in each argument using a list of negation cues (e.g., not, no, neither) from Councill et al. (2010). We also include three types of dependency features (Anand et al., 2011) which consist of triples from the dependency parse of the argument. Specifically, a basic dependency feature (rel_i, t_j, t_k) encodes the syntactic relation rel_i between words t_j and t_k. One variant is to replace the head word of the relation rel_i with its part-of-speech tag. The other variant is replacing tokens in a triple with their polarities (i.e., $+$ or $-$) using MPQA dictionary of opinion words (Wilson et al., 2005).

localStanceToRel. We again employ logistic regression to train a local stance classification model (*localStance*) using the same features as in *localRel*. We construct the training instances by pairing a topic t and all its pro/con arguments in the training set[3]. During testing, we predict two arguments agree/disagree to each other if they have the same/differences stances regarding the topic.

LSTM+attention. We adapt the attention-based LSTM model used for textual entailment in Rocktäschel et al. (2016). We use GloVe vectors (Pennington et al., 2014) with 100 dimensions trained on Wikipedia and Gigaword as word embeddings. To avoid over-fitting, we apply dropout before and after the LSTM layer with the probability of 0.1. We train the model with 60 epochs using cross-entropy loss. We use Adam for optimization with the learning rate of 0.01.

EDIT. We reimplement the approach for argument relation classification from Cabrio and Villata (2012). Specifically, we train the textual entailment system EDIT[4] on our training set using the same configuration used in Cabrio and Villata (2012). We then apply the trained model on the testing dataset.

Joint model. For our approach described in Section 3, we use the output of the two local classifiers (*localRel* and *localStance*) as the input for formulas $f9$ and $f10$ in Table 2.[5] The weights of the formulas are learned on the dev dataset. We use *thebeast*[6] to learn weights for the formulas and to perform inference. *thebeast* employs cutting plane inference (Riedel, 2008) to improve the accuracy and efficiency of MAP inference for Markov logic.

4.3 Results and Discussion

Table 4 shows the results of different approaches on *argument relation classification*. *EDIT* performs the worst among four local classifiers with an accuracy of 0.50. We think this is mainly due to the difference between the corpora, i.e., we don't

[2]The dataset and splits will be available on publication.

[3]Although *localRel* and *localStance* use the same features, we notice that logistic regression can pick up informative features for each task based on different training set (i.e., arg1-arg2 v.s. topic-arg).

[4]http://edits.fbk.eu/

[5]Another option is to predict *localRel* and *localStance* using MLNs. We leave this for future research.

[6]http://code.google.com/p/thebeast

	localRel			*localStanceToRel*			*LSTM+attention*			*EDIT*			*joint*		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
agree	52.6	61.1	56.6	71.6	58.5	64.4	55.5	56.1	56.0	76.1	47.9	58.8	63.6	63.1	63.3
disagree	70.5	62.8	66.5	55.4	68.9	61.4	62.4	61.4	61.9	27.1	56.3	36.6	67.3	67.7	67.5
Acc.		62.1			63.0			59.1			**50.0**			**65.5**	
Macro F.		61.8			63.6			58.9			**51.8**			**65.4**	

Table 4: Experimental results of argument relation classification on the testing dataset. Bold indicates statistically significant differences over the baselines using randomization test ($p < 0.01$).

pair an argument with its topic in our argument relation classification dataset.

Additionally, the results of *LSTM+attention* are worse than *localRel* and *localStanceToRel*. We suspect this is because the amount of our training data is only 1/10 of the SNLI corpus used in Rocktäschel et al. (2016). Also our dataset has a richer lexical variability.

In general, the local model *localRel* is better at predicting *disagree* than *agree*. The approach *localStanceToRel* flips this by predicting more argument pairs as *agree*. Overall, there is a small improvement in accuracy from *localRel* to *localStanceToRel*. Our joint model combines the strengths of the two local classifiers and performs significantly better than both of them in terms of accuracy and macro-average F-score (randomization test, $p < 0.01$).

5 Conclusions

We propose a joint inference model for argument relation classification on dialogic argumentation. The model utilizes the mutual support relations between *argument relation classification* and *stance classification*. We show that our joint model significantly outperforms other local models.

References

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 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,* Portland, Oregon, 24 June 2011, pages 1–9.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics,* Jeju Island, Korea, 8–14 July 2012, pages 208–212.

Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In Lluís Godo, editor, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 8th European Conference*, pages 378–389. Springer Berlin Heidelberg.

Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2013. Bipolarity in argumentation graphs: Towards a better understanding. *International Journal of Approximate Reasoning*, 54:876–899.

Isaac Councill, Ryan McDonald, and Leonid Velikovich. 2010. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing,* Uppsala, Sweden, 10 July 2010, pages 51–59.

Pedro Domingos and Daniel Lowd. 2009. *Markov Logic: An Interface Layer for Artificial Intelligence*. Morgan Claypool Publishers.

Swapna Gottipati, Minghui Qiu, Yanchuan Sim, Jing Jiang, and Noah A. Smith. 2013. Learning topics and positions from Debatepedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* Seattle, Wash., 18–21 October 2013, pages 1858–1868.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* Lisbon, Portugal, 17–21 September 2015, pages 2127–2137.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* Berlin, Germany, 7–12 August 2016, pages 1589–1599.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* Doha, Qatar, 25–29 October 2014, pages 751–762.

Yufang Hou, Katja Markert, and Michael Strube. 2013. Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, 9–14 June 2013, pages 907–917.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of the 25th International Conference on Computational Linguistics,* Dublin, Ireland, 23–29 August 2014, pages 1489–1500.

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. In *Proceedings of the 30th Conference on the Advancement of Artificial Intelligence,* Phoenix, Arizona, USA, 12–17 February 2016, pages 2979–2985.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the ACL 2014 System Demonstrations,* Baltimore, USA, 22–27 June 2014, pages 55–50.

Stefano Menini and Sara Tonelli. 2016. Agreement and disagreement: Comparison of points of view in the political domain. In *Proceedings of the 26th International Conference on Computational Linguistics,* Osaka, Japan, 11–16 December 2016, pages 2461–2470.

Ivan Meza-Ruiz and Sebastian Riedel. 2009. Jointly identifying predicates, arguments and senses using Markov logic. In *Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics,* Boulder, Col., 31 May – 5 June 2009, pages 155–163.

Akiko Murakami and Rudy Raymond. 2010. Support or oppose? classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics,* Beijing, China, 23–27 August 2010, pages 869–875.

Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* Berlin, Germany, 7–12 August 2016, pages 1127–1137.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* Lisbon, Portugal, 17–21 September 2015, pages 938–948.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* Doha, Qatar, 25–29 October 2014, pages 1532–1543.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the Joint Conference of the 53th Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing,* Beijing, China, 26–31 July 2015, pages 543–552.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* San Diego, California, 12–17 June 2016, pages 1384–1394.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, Honolulu, Hawaii, 25–27 October 2008, pages 650–659.

Hoifung Poon and Pedro Domingos. 2010. Unsupervised ontology induction from text. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July 2010, pages 296–305.

Sebastian Riedel. 2008. Improving the accuracy and efficiency of MAP inference for Markov logic. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence,* Helsinki, Finland, 9–12 July 2008, pages 468–475.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* Lisbon, Portugal, 17–21 September 2015, pages 440–450.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *Proceedings of the 4th International Conference on Learning Representations,* San Juan, Puerto Rico, 2-4 May 2016.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* Seattle, Wash., 18–21 October 2013, pages 1631–1642.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics*

and the 4th International Joint Conference on Natural Language Processing, Singapore, 2–7 August 2009.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the Joint Conference of the 53th Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing,* Beijing, China, 26–31 July 2015, pages 116–125.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* Doha, Qatar, 25–29 October 2014, pages 46–56.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th World Wide Web Conference,* Montréal, Québec, Canada, 11 – 15 April, 2016, pages 613–624.

Simone Teufel. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,* Sydney, Australia, 22–23 July 2006, pages 327–335.

Marilyn A. Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 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,* Montréal, Québec, Canada, 3–8 June 2012, pages 592–596.

Theresa Wilson, Janyce M. Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing,* Vancouver, B.C., Canada, 6–8 October 2005, pages 347–354.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* San Diego, California, 12–17 June 2016, pages 136–141.