Exploring the Role of Argument Structure in Online Debate Persuasion

Jialu Li
Cornell University
jl3855@cornell.edu

Esin Durmus
Cornell University
ed459@cornell.edu

Claire Cardie
Cornell University
cardie@cs.cornell.edu

Abstract

Online debate forums provide users a platform to express their opinions on controversial topics while being exposed to opinions from diverse set of viewpoints. Existing work in Natural Language Processing (NLP) has shown that linguistic features extracted from the debate text and features encoding the characteristics of the audience are both critical in persuasion studies. In this paper, we aim to further investigate the role of discourse structure of the arguments from online debates in their persuasiveness. In particular, we use the factor graph model to obtain features for the argument structure of debates from an online debating platform and incorporate these features to an LSTM-based model to predict the debater that makes the most convincing arguments. We find that incorporating argument structure features play an essential role in achieving the better predictive performance in assessing the persuasiveness of the arguments in online debates.

1 Introduction

The increase in availability of online argumentation platforms has provided opportunity for researchers to develop computational methods at a larger scale studying the important factors of persuasiveness such as the language use (Hidey et al., 2017; Tan et al., 2016; Zhang et al., 2016), characteristics of audience (i.e. prior beliefs, demographics) (Durmus and Cardie, 2019a, 2018) and social interactions (Durmus and Cardie, 2019b).

Prior work has showed incorporating argument structure features is important in assessing the quality of monological persuasive essays (Klebanov et al., 2016; Wachsmuth et al., 2016). Hidey et al. (2017) and Egawa et al. (2019) further collected annotations for semantic types of argument components and studied the relationship between the semantic types and persuasiveness of the arguments from online argumentative platform ChangeMyView (CMV) (Tan et al., 2016). CMV consists of discussion trees where the users interact with the original poster to change their opinion on a given topic. Although the discussion trees are of a high quality since they are monitored by moderators (Tan et al., 2016), they are not as structured since any user in the subreddit can participate in the discussions once the original post is posted. Furthermore, the persuasiveness of the posts in CMV is evaluated only by the original poster (i.e. whether they change their stance or not). In this paper, we aim to investigate the effect of argument structure in persuasion on online debates. We focus on debates from DDO corpus (Durmus and Cardie, 2019a) where debaters from two diverging sides of an issue express their opinions on a controversial topic in turns since these debates are more structured and the persuasiveness of the arguments in debates are evaluated by a larger set of audience. Moreover, this setup allows us to account for the audience characteristics when studying the effect of the argument structure on persuasion.

We first generate argument structure on DDO dataset (Durmus and Cardie, 2019a) using the model proposed by Niculae et al. (2017). We then incorporate the features extracted from argument structure to an LSTM-based model that encodes the sequence of turns from two sides (i.e. PRO vs. CON). We compare our results with the baselines proposed in (Durmus and Cardie, 2018) which extracts linguistics features from the debate text as well as features that encode characteristics of the audience. We find that incorporating argument structure features achieves significantly better results than the baselines. Our analysis further shows that argument structure features encode important strategies of persuasion, for example, we find that more convincing arguments are more likely to in-

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 8905–8912,
November 16–20, 2020.

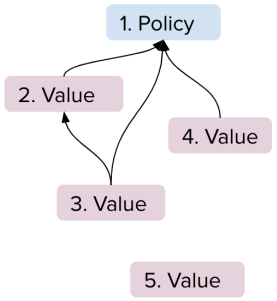
Sent No.	Debate text	Predicted Argument Structure
1	Parents should not send their children to preschool for several reasons.	
2	First and foremost, the year is better spent with a full-time parent.	
3	In addition, most children will learn very little at preschool.	
4	Contrary to claims made by preschool advocates, children are not better equipped because of preschool.	
5	They may develop social skills and hand painting skills sooner, however children that miss preschool will quickly catch up before they finish the first grade.	

Figure 1: An example of argument structure extracted from the debate text in one round from one side.

clude personal experiences of the debater and appeal to audience emotion.

2 Related Work

Analysis of discourse structure There has been a lot effort to understand the role of discourse structure in argumentation. Jiang et al. (2019) applied RST to essays written by students in K-12 schools and demonstrated its potential to provide automated feedback for essay quality. Argument structures, which can be considered as a special kind of discourse structure, have been widely analyzed in the task of automatic essay scoring and feedback (Klebanov et al., 2016; Ghosh et al., 2016; Wachsmuth et al., 2016). Furthermore, Duthie and Budzynska (2018) has studied the relationship between ethos, a specific kind of argument unit, and the dynamics of governments from the UK parliamentary debates. The role of argument structure in persuasion on online debates is much less explored, which is the main focus of this paper.

Analysis of Persuasion Prior studies on persuasion has mainly focused on understanding the role of linguistic factors (Petty et al., 1983; Chaiken, 1987; Dillard and Pfau, 2002; Gold et al., 2015). Besides, the interaction between debaters has shown to be an important cue in persuasion studies (Zhang et al., 2016; Tan et al., 2016; Wang et al., 2017). Luu et al. (2019) further found that the debater's skill estimated from debate text history is also predictive of convincing the audience. User factors are explored in previous papers (Durmus and Cardie, 2019a, 2018; Longpre et al., 2019), demonstrating the importance of characteristics and beliefs of the audience. Furthermore, Potash and Rumshisky (2017) proposed a recurrent neural network architecture with attention and annotated audience favorability to predict the winner of the debate. Villata et al. (2018) and Benlamine et al. (2017) studied the correlation of the engagement index in brain hemispheres with the persuasion strategies. Argument structures have been used to understand argumentative strategies in dialogues and news editorials (Al Khatib et al., 2017; Wang et al., 2019). A few studies have explored the impact of argument structures in predicting persuasion on CMV dataset based on statistical analysis of proposition types (Hidey et al., 2017; Egawa et al., 2019; Morio et al., 2019). In this paper, we particularly study persuasion in online debates. We propose novel argument structure features based on n-grams of the supporting relations in argument structure graph of the debate text and experiment with these using both linear and neural models.

3 Dataset

We experiment with DDO dataset (Durmus and Cardie, 2019a) which includes 77,655 debates covering 23 different topic categories. Each debate consists of multiple rounds with each round containing one utterance from the PRO side and one utterance from the CON side. Besides the text information for debates, the dataset also contains user information and votes provided by the audience on six different criteria of evaluating both the debaters. We use the criterion *"Made more convincing arguments"* as an overall signal to study the role of argument structure in predicting more convincing arguments.

4 Prediction Task

Task. We aim to predict which side (i.e. PRO vs. CON) makes more convincing arguments during a debate, and thus is more persuasive.

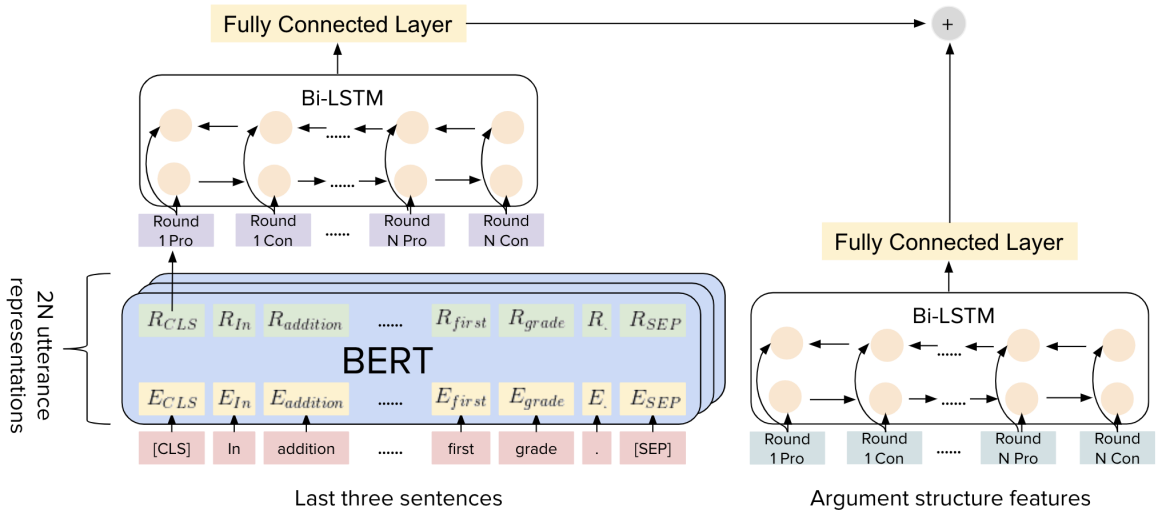


Figure 2: Model for predicting which side makes more convincing arguments

Data preprocessing. We count which side of the debate gets more votes for the criterion *"Made more convincing arguments"*. We eliminate debates if they are tied or the difference in votes is only 1.[1] The final dataset contains 2,606 debates.

4.1 Argument structure features

We apply the pre-trained model (Niculae et al., 2017) on DDO dataset to get the stucture of the arguments. We select this method since we can predict argumentative relations and classify proposition type at the same time, while the method proposed by Chakrabarty et al. (2020) mainly focuses on predicting argumentative relations. Besides, this model can model argumentative relations that do not necessarily form a tree structure which is more suitable to argumentation in the wild comparing to the models proposed in Stab and Gurevych (2017) and Peldszus and Stede (2015). We generate argument structures for the selected 2,606 debates.[2] The argument structure model outputs the proposition type for each sentence (i.e. REFERENCE, TESTIMONY, FACT, VALUE, POLICY) as well as the supporting relationship between the propositions. An example of argument structure generated on the text from one side in one round of the debate *'Preschool Is A Waste Of Time'* is shown in Figure 1. We use *Amazon Mechanical Turk* (AMT) to further evaluate the quality of the argument structure on debates by asking Turkers to classify each argument from randomly picked 30 debates into five categories: POLICY, VALUE, FACT, TESTIMONY, REFERENCE. In total, we get annotations for 1,098 sentences, and each sentence is annotated by two annotators. We find that around 64% of the output generated from the pre-trained model is consistent with either of the annotations from the Turkers. We then extract three sets of argument structure features to capture the proposition types and link between propositions:

Proposition n-gram frequency Similar to Wachsmuth et al. (2016), we obtain the frequencies of proposition unigrams, bigrams, and trigrams from the sequence of propositions. For example, (POLICY,VALUE) and (VALUE,VALUE) bigram features in Figure 1 has values 0.25,0.75 respectively.

Link n-gram frequency We extract the n-gram information from the supporting relations in argument structure graph. For example, we represent two propositions connected with a link as a bigram (i.e $a \rightarrow b$ in the graph is represented with bigram (a,b)).

Graphical representation Rahwan (2008) has found that there are five common argument structures in online environment: basic argument, convergent argument, serial argument, divergent argument, and linked argument. A typical basic argument is $a \rightarrow b$[3], while serial argument is $a \rightarrow b \,\&\, b \rightarrow c$. The simplest convergent argument

[1] Since the average number of total votes in one debate is 8, we consider difference of two or more votes as significant.

[2] Since the model takes relatively long inference time and performs worse for long debates, we eliminate all the debates with more than 40 sentences in one round from one side. We also eliminate debates where one of the debaters forfeit during the debate.

[3] a, b, c denotes propositions and $a \rightarrow b$ denotes the directed link between a and b.

Pro	Con
[...] One of the quotes I remembered clearly was, "God will give us whatever we want, as long as we don't screw up." [...] I haven't committed genocide or anything bad like that. But I've made my mistakes, and everyone has. [...] I'm not dead.	[...] If cheating on test made someone happy, that doesn't make up for their unfair advantage. [...] Multiple times it is mentioned in the bible that homosexuality is wrong, it's a sin. "You shall not lie with a male as one lies with a female; it is an abomination." [...]

Table 1: Example debate "GAY MARRIAGE" that is classified correctly after adding argument structure features.

Model	Accuracy
Majority baseline	62.62%
Linguistic+User LR	67.41%
Arg-Struct LR	69.52%
Linguistic+Arg-Struct LR	70.48%
Linguistic+User+Arg-Struct LR	70.44%
Our Model	**77.28%**
Our Model w/o All Arg-Struct	75.29%
Our Model w/o Proposition N-gram	76.21%
Our Model w/o Link N-gram	76.86%
Our Model w/o Graphical	76.95%

Table 2: Comparison with feature based Logistic Regression (LR). Arg-Struct denotes the argument structure features.

is in the form of $a \rightarrow b \,\&\, c \rightarrow b$, and a divergent argument is in the form of $a \rightarrow b \,\&\, a \rightarrow c$. Similarly, a linked argument is in the form of $a, c \rightarrow b$. We extract features to represent which of these types of arguments are used in the text of the debaters. We further classify the convergent arguments into two categories – where two propositions support one proposition (regular convergent argument) and more than two propositions support one proposition (multi convergent argument). Similarly, we classify divergent argument into regular divergent argument and multi divergent argument.

4.2 Model Architecture

We employ two separate bidirectional LSTM (Hochreiter and Schmidhuber, 1997) models to encode the argument structure features and features encoding the debate text derived from pre-trained BERT model (Devlin et al., 2019) as shown in Figure 2. LSTM modeling the debate text takes BERT representation (Devlin et al., 2019) while LSTM encoding argument structure features takes three set of argument structure features of an utterance in a round at each time step. Two fully connected layers with softmax are used to predict the output probabilities over both of these LSTM models separately. The model learns weights during training to combine these probabilities.

5 Experiments and Analysis

We compare our model with the baseline proposed by Durmus and Cardie (2019a) employing linguistic features and features encoding audience characteristics. The prediction accuracy is evaluated using 5-fold cross-validation, and the model parameters for each split are picked with 3-fold cross-validation on the training set. As shown in Table 2, incorporating argument structure features to Logistic Regression achieves significantly better performance than the baseline with linguistic and user-based features. LSTM with argument structure features achieves the best predicive performance since LSTM can better represent context and the interplay between debaters. We perform t-test over 10 runs between the model with and without argument structure features, the p-value is 0.0038, indicating a statistically significant result. Furthermore, we do ablation over different sets of argument structure features. The results show that using the sequential flow of arguments is more effective than using argumentative relations in our setting.

We further analyze what type of argument structure is more correlated with making more convincing arguments. Comparing the unigram, bigram and trigram frequencies of the propositions by more convincing vs. less convincing debaters, we find that unigram TESTIMONY ($p < 0.0001$)[4], bigram (VALUE,TESTIMONY) ($p < 0.001$), and trigram (VALUE,TESTIMONY,VALUE) ($p < 0.0001$) appear more frequently in the more convincing side. This result suggests that justifying the objective claims with personal experiences is an effective strategy as also shown in previous work (Villata et al., 2018).

[4]The p-values are calculated using the Wilcoxon signed-rank test.

Table 1 shows an example that is predicted classified by the model correctly after adding argument structure features. We observe that the side referring to their personal experiences (PRO) is voted as the side making more convincing arguments. Besides, we find that unigram POLICY ($p < 0.0001$), bigram (POLICY,VALUE) ($p < 0.005$) appear more frequently in the less convincing side suggesting that using propositions with type POLICY – which is used to specify a specific course of action to be taken – may not be a very effective strategy in online debating. Analyzing the link n-gram frequency features, we have further found that propositions with type VALUE from more convincing side are supported by a FACT ($p < 0.05$) more often. This suggests that the more convincing debaters may be using logos to support their views as also shown in previous work (Hidey et al., 2017). Finally, we observe that more convincing side tends to have more divergent arguments ($p = 0.052$). Divergent arguments involves three or more consecutive sentences most of the time. In the case of three consecutive sentences, the middle sentence supports both the other two sentences by giving explanations or evidence, and serves as a transition between two similar ideas.

We also look into some examples that are classified wrong by the model. A typical error is caused by wrong proposition type classification. For example, in the debate "Driving on public roads is a right not a privilege", sentences from PRO side "In addition, in purchasing our vehicles, we have the right to drive said vehicle." and "I appreciate the insight given by my opponent but he/she has failed to address the issue at hand." are classified as "Testimony" wrongly, which makes the model prefer PRO as the more convincing side. We believe that incorporating more accurate argument structure generation models can further improve the performance on persuasion prediction.

6 Conclusion

In this work, we explore the role of argument structure in online debate persuasion and find that incorporating argument structure features along with the linguistic features achieves the best predictive performance models. Moreover, we observe that argument structure features provide important cues about effective persuasion strategies in online debates.

Acknowledgements

We thank the anonymous reviewers for their useful feedback. This work was supported in part by NSF grants IIS-1815455. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

References

Khalid Al Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357.

Mohamed S Benlamine, Serena Villata, Ramla Ghali, Claude Frasson, Fabien Gandon, and Elena Cabrio. 2017. Persuasive argumentation and emotions: An empirical evaluation with users. In *International Conference on Human-Computer Interaction*, pages 659–671. Springer.

Shelly Chaiken. 1987. The heuristic model of persuasion. In *Social influence: the ontario symposium*, volume 5, pages 3–39. Hillsdale, NJ: Lawrence Erlbaum.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy Mckeown, and Alyssa Hwang. 2020. Ampersand: Argument mining for persuasive online discussions. *arXiv preprint arXiv:2004.14677*.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

James Price Dillard and Michael Pfau. 2002. *The persuasion handbook: Developments in theory and practice*. Sage Publications.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045, New Orleans, Louisiana. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2019a. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607, Florence, Italy. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2019b. Modeling the factors of user success in online debate. In *The World Wide Web Conference*, WWW '19, page 2701–2707, New York, NY, USA. Association for Computing Machinery.

Rory Duthie and Katarzyna Budzynska. 2018. A deep modular rnn approach for ethos mining. In *IJCAI*, pages 4041–4047.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2019. Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 422–428.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 987–996.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554.

Valentin Gold, Mennatallah El-Assady, Annette Hautli-Janisz, Tina Bögel, Christian Rohrdantz, Miriam Butt, Katharina Holzinger, and Daniel Keim. 2015. Visual linguistic analysis of political discussions: Measuring deliberative quality. *Digital Scholarship in the Humanities*, 32(1):141–158.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathleen McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Shiyan Jiang, Kexin Yang, Chandrakumari Suvarna, Pooja Casula, Mingtong Zhang, and Carolyn Rose. 2019. Applying rhetorical structure theory to student essays for providing automated writing feedback. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 163–168.

Beata Beigman Klebanov, Christian Stab, Jill Burstein, Yi Song, Binod Gyawali, and Iryna Gurevych. 2016. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 70–75.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Liane Longpre, Esin Durmus, and Claire Cardie. 2019. Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th Workshop on Argument Mining*, pages 167–176.

Kelvin Luu, Chenhao Tan, and Noah A Smith. 2019. Measuring online debaters' persuasive skill from text over time. *Transactions of the Association for Computational Linguistics*, 7:537–550.

Gaku Morio, Ryo Egawa, and Katsuhide Fujita. 2019. Revealing and predicting online persuasion strategy with elementary units. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6275–6280.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument Mining with Structured SVMs and RNNs. In *Proceedings of ACL*.

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*, pages 938–948.

Richard E Petty, John T Cacioppo, and David Schumann. 1983. Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of consumer research*, 10(2):135–146.

Peter Potash and Anna Rumshisky. 2017. Towards debate automation: a recurrent model for predicting debate winners. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2475.

Iyad Rahwan. 2008. Mass argumentation and the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1):29–37.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, volume 6.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Chenhao Tan and Lillian Lee. 2016. Talk it up or play it down?(un) expected correlations between (de-) emphasis and recurrence of discussion points in consequential us economic policy meetings. *arXiv preprint arXiv:1612.06391*.

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 international conference on world wide web*, pages 613–624. International World Wide Web Conferences Steering Committee.

Serena Villata, Sahbi Benlamine, Elena Cabrio, Claude Frasson, and Fabien Gandon. 2018. Assessing persuasion in argumentation through emotions and mental states. In *The Thirty-First International Flairs Conference*.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

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*, pages 136–141, San Diego, California. Association for Computational Linguistics.

A Appendix

A.1 Argument Structure Features Used

Proposition n-gram frequency When we eliminate proposition bigram and trigram that occur less than 3% in all training debates, five types of unigrams, eight types of bigrams and ten types of trigrams remain.

Unigram types: policy, value, fact, testimony, reference.

Bigram types: (value, value), (testimony, value), (value, testimony), (value, policy), (policy, value), (fact, value), (value, fact), (testimony, testimony).

Trigram types: (value, value, value), (testimony, value, value), (value, value, policy), (value, value, testimony), (value, testimony, value), (fact, value, value), (policy, value, value), (value, fact, value), (value, policy, value), (value, value, fact).

Link n-gram frequency When we eliminate all link bigrams that occur less than 3% of all training data, four types of link bigrams remain (i.e. (value, value), (value, policy), (fact, value), (testimony, value)).

Graphical representation There are 5 types of features for graphical representation: basic argument, regular convergent argument, regular divergent argument, multi convergent argument, multi divergent argument.

In total, the argument structure features are 32-dimensional.

A.2 Linguistic Features and User Features

The linguistic features and user features we use for the Logistic Regression based baseline is the same as the features used by Durmus and Cardie (2019a). They include hedge words (Tan and Lee, 2016), evidence words (e.g. "according to"), positive words, negative words, swear words, personal pronouns, tf-idf, argument lexicon features (Somasundaran et al., 2007), politeness marks (Danescu-Niculescu-Mizil et al., 2013), sentiment, connotation (Feng and Hirst, 2011), subjectivity (Wilson et al., 2005), modal verbs, type-token ratio (diverse word usage), and punctuation.

The user features include opinion similarity for big issues, religious and political ideology match and persuadability score (how likely a person will be persuaded) (Longpre et al., 2019).

A.3 BERT Representation Generation

We input the utterance in one round for one debater. The segment embedding for each word in the utter-

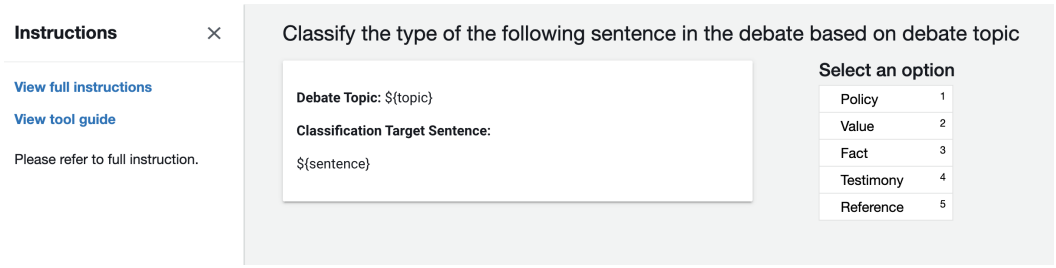


Figure 3: AMT annotation example

ance is the same, though one utterance will contain multiple sentences. Due to the maximum sequence length of 128 tokens of BERT [5], which is much shorter than the average length of utterance in one round for one debater, we truncate the debate text input and only preserve the last three sentences[6] in each round for each debater. The truncate method of choosing the first three sentences of the utterance has also been tested, but the performance of the model was around 3% lower.

A.4 Implementation Details

We use grid search to pick the hyperparameter. For the model that encodes linguistic information, we use a one-layer bidirectional LSTM with 768 dimension BERT representation input and 32 dimension hidden states. (We search in [16, 32, 64] for hidden dimension.) For the model that encodes argument structure information, we use a one-layer bidirectional LSTM with 32 dimension argument features input and 4 dimension hidden states. (We search in [16, 8, 4] for hidden dimension). We have a 0.5 dropout rate for both fully connected layers. Total number of parameters is around 100k. We use Adagrad (Duchi et al., 2011) with initial learning rate 0.005 and weight decay 0.01 to optimize the cross-entropy loss. (We also experiment with Adam with default setting, Adagrad without weight decay, learning rate between [0.001, 0.005, 0.01]). 2200 debates are used for training, 200 for validation and 206 for test set. We use early stopping to avoid overfitting, the model is trained for around 15 epochs on average. It takes less than 15 minutes to run the model on a CPU (2.7 GHz Intel Core i7). To test the stability of our results, we train and evaluate our model 10 times and take the average accuracy.

Type	# Proposition	Consistency
Policy	97	56.70%
Value	834	65.47%
Fact	85	84.71%
Testimony	79	37.97%
Reference	3	33.33%
All	1,098	64.12%

Table 3: Annotation results from Amazon Mechanical Turker.

A.5 Details on AMT result

Figure 3 shows the screenshot for a typical HIT for the Turkers. For each HIT, the turkers are given the debate topic and the sentence to be classified. They need to choose between 5 categories: Policy, Value, Fact, Testimony, Reference. The definition of these proposition types and the corresponding example are included in the full instruction.

In total, we get annotations for 1,098 sentences from seventeen annotators. The detailed results are listed in Table 3. Consistency means the generated annotations is consistent with either of the annotations from the Turkers. We also compute Inter-Annotator Agreement (IAA) using Kripendorff's alpha (Krippendorff, 1970). The Kripendorff's alpha is 0.2, indicating that annotating argument structure is still a hard task for Turkers.

[5] We use BERT-base with uncased input as the pretrained model.

[6] We also experimented using more sentences (e.g. last five sentences) in cases where the sequence length has not been maxed out has also been tested, but it doesn't show significant improvement.