Determining Relative Argument Specificity and Stance for Complex Argumentative Structures

Esin Durmus
Cornell University
ed459@cornell.edu

Faisal Ladhak
Amazon
faisall@amazon.com

Claire Cardie
Cornell University
cardie@cs.cornell.edu

Abstract

Systems for automatic argument generation and debate require the ability to (1) determine the stance of any claims employed in the argument and (2) assess the specificity of each claim relative to the argument context. Existing work on understanding claim specificity and stance, however, has been limited to the study of argumentative structures that are relatively shallow, most often consisting of a single claim that directly supports or opposes the argument thesis. In this paper, we tackle these tasks in the context of complex arguments on a diverse set of topics. In particular, our dataset consists of manually curated argument trees for 741 controversial topics covering 95,312 unique claims; lines of argument are generally of depth 2 to 6. We find that as the distance between a pair of claims increases along the argument path, determining the relative specificity of a pair of claims becomes easier and determining their relative stance becomes harder.

1 Introduction

The tasks of automatic argument generation and debate require the ability to present a diverse and comprehensive set of supporting and opposing arguments given a controversial topic. Two critical components of such systems are an ability to determine the **stance** and the **specificity** of any claims employed in the proposed argument. Consider, for example, the argument thesis (i.e., the topic) of Figure 1: (THESIS) *Would we like to live in the world of Harry Potter?* Construction of an argument in support or in opposition to this thesis necessarily requires knowing the stance of the claims that comprise it: the claim *Magic opens a lot of interesting possibilities* should be identified as a claim in support of the THESIS, and *The capacity of harm is greater when magic is involved* (HARM), as a claim in opposition. Indeed, previous work has studied this task (e.g., Bar-Haim et al. (2017); Faulkner (2014)).

It is not sufficient, however, to determine claim stance only with respect to the argument thesis. Debate and argument generation systems, in general, should also be able to determine whether two claims that address the same line of reasoning represent the same, or the opposing stance: using *Defense is also made easier through magic* to refute the HARM claim in Figure 1, for example, requires recognizing that it represents the opposite stance.

The issue of claim specificity in argumentation has been much less addressed. Existing work, however, suggests that a high degree of specificity is correlated with argument quality and persuasiveness (Carlile et al., 2018; Swanson et al., 2015). In terms of argument quality though, it is entirely possible for the presented claims to be coherent and meaningful, yet be too specific within the given discourse, and therefore be logically irrelevant (Dessalles, 2016). As a concrete example, suppose we wanted to assert a claim in support of the argument THESIS of Figure 1. While *The Unforgivable Curses are illegal...and their use is grounds for immediate life imprisonment* supports the THESIS, it is too specific a claim to introduce at this point in the argument. Namely, it doesn't flow naturally without first introducing the concept of *Unforgivable Curses*.

To date, existing work on understanding claim specificity and stance has mostly employed annotated monologic persuasive documents or discussion forums and, as a result has been limited to the study of argumentative structures that are relatively shallow, most often only consisting of claims that directly support or oppose the argument thesis (Bar-Haim et al., 2017; Faulkner, 2014).

To support the generation of diverse and potentially complex arguments on a topic of choice, we present here a dataset of manually curated
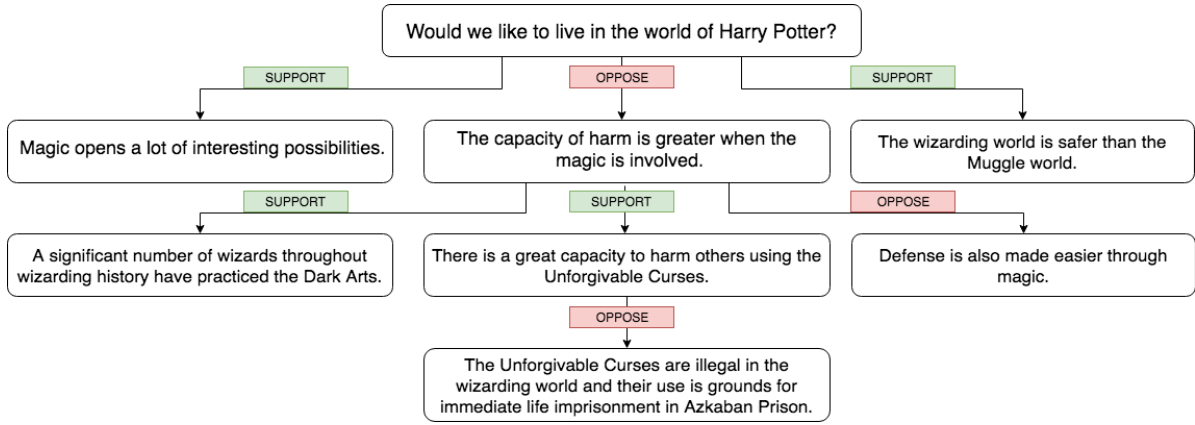
Figure 1: Partial tree for the controversial topic "Would we like to live in the world of Harry Potter?". Each claim's position towards its parent argument is indicated in the box on the edge between the claim and its parent. The full argument tree for this topic can be found at https://www.kialo.com/is-the-world-of-harry-potter-really-the-place-to-be-2415/2415.0=2415.1.

argument trees for 741 controversial topics covering 95,312 unique claims. In contrast to existing datasets, ours consists of argument trees where each root node represents the argument thesis (main claim) and every other node represents a claim that either supports or opposes its parent. Taking advantage of this relatively complex argumentative structure, we formulate two prediction tasks to study relative specificity and stance. The main contributions of our study are the following:

- We provide a publicly available dataset of argument trees consisting of a diverse set supporting and opposing claims for 741 controversial topics[1].

- We propose two novel settings to study claim specificity and stance in the context of a diverse set of supporting and opposing points.

- We control for specific aspects of the argument tree (e.g., depth, stance) in our experiments to understand their effect on claim specificity and stance detection.

2 Dataset

We extracted argument trees for 741 controversial topics from www.kialo.com[2]. Kialo is a collaborative platform where users provide supporting and opposing claims for each claim related to a controversial issue. Besides providing the claims themselves, users also help to improve the quality of

existing claims by suggesting edits, and rating the quality of claims. This process of collaborative editing helps to create a high quality, diverse set of supporting and opposing points for each controversial topic[3].

The dataset includes diverse set of controversial topics. Each controversial topic is represented by a **thesis** and tagged to be related to pre-defined generic categories such as *Politics*, *Ethics*, *Society* and *Technology*[4]. Figure 2(a) shows the number of controversial topics with the given pre-defined categories. The controversial topics' theses include: "A free Press is necessary to democracy.", "All drugs should be legalised.", "A society with no gender would be better.", "Hate speech should be banned", etc.

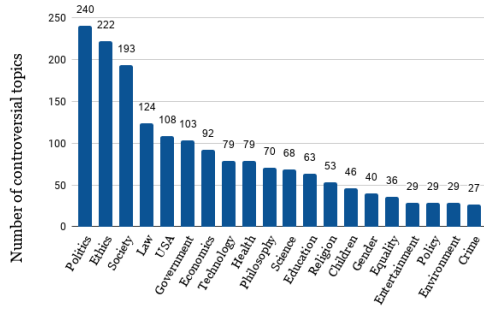### 2.1 Structure of the arguments

The arguments for each controversial topic are represented as trees. The root node of each such tree represents the **thesis** of the controversial topic. Every other node in the tree represents a **claim** that either **supports** or **opposes** its parent claim. Figure 1 shows a partial argument tree for the thesis "Would we like to live in the world of Harry Potter?". We see that besides the supporting and opposing claims for the thesis, there are supporting and opposing claims for the claims at different depths. With this structure, we can identify indirect support/oppose relationships even between nodes without parent-child relationships if they

---

[1] The dataset will be made publicly available at http://www.cs.cornell.edu/ esindurmus/.
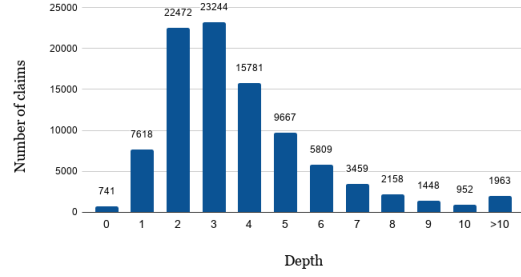
[2] This covers all controversial topics on the website at the time we collected the data.

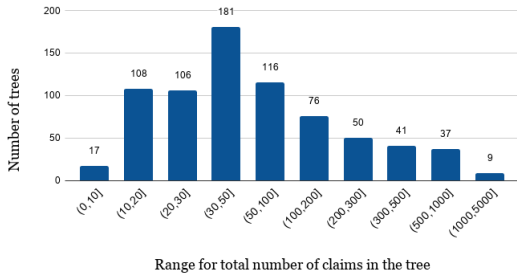[3] The data is crawled from this website in accordance with the terms and conditions.

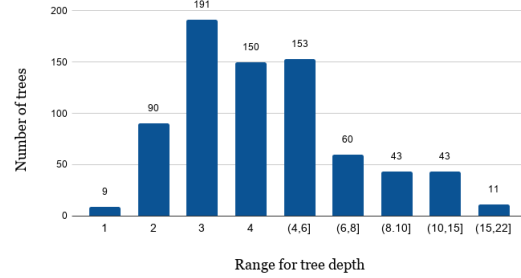[4] Note that a controversial topic can be relevant to multiple pre-defined categories.

(a) Number of controversial topics with the given predefined categories. Note that a controversial topic could be related to multiple pre-defined categories.

(b) Number of claims at given depths. The majority of the claims lie at the depth 3 or higher.

(c) Number of trees with given range of total number of claims. For the majority of trees, the argument tree has more than 30 claims in the tree. Average number of claims per argument tree is 127.

(d) Number of trees with given range of depth. For the majority of trees, the depth of the argument tree is 4 or higher, and average depth per argument tree is 5.

are on the same **argument path**. For example, the claim "Defense is also made easier through magic" indirectly supports the thesis, since it is in opposition with its parent "The capacity of harm is greater when the magic is involved", which is an opposing claim to the thesis. Another observation is that as we go deeper along an argument path, the claims get more specific, since each claim aims to either support or oppose its parent. For example, while the claim "The capacity of harm is greater when the magic is involved" refers to the general harms that can be caused by magic, one of its child claims "There is a great capacity to harm others using the Unforgivable Curses" is more specific as it refers to harm via a particular set of *curses* in *magic*.

2.2 Data Statistics

The dataset consists of argument trees for 741 controversial topics comprised of $95,312$ unique claims. The distribution of argument trees with the given range of total claims, and depth is shown in Figures 2(c) and 2(d) respectively. We see that for the majority of trees, the depth is 4 or higher, and the number of claims is greater than 30.

Figure 2(b) shows the total number of claims

at a given depth. We see that only $7,618$ out of $95,312$ claims are directly supporting or opposing the theses of the controversial topics. The majority of the claims lie at the depth 3 or higher. This shows that the dataset has a rich set of supporting and opposing claims for not only for the theses, but for claims at different depths of the tree.

In total, there are 44,572 claims that are supporting and 50,740 claims that are opposing their parent claims. 90% of claims consist of 1 (61%) to 2 (29%) sentences and average number of tokens per claim is 30.

3 Claim Specificity

Determining the relative specificity of arguments is an important step towards being able to generate logically relevant arguments in a given discourse (Dessalles, 2016). For a system that disregards the relative specificity of claims, it is entirely possible to generate coherent and meaningful, yet logically irrelevant claims, when the generated claims are either too generic or specific for the given argument discourse.

In this work, we determine the relative specificity between a pair of claims that are along the
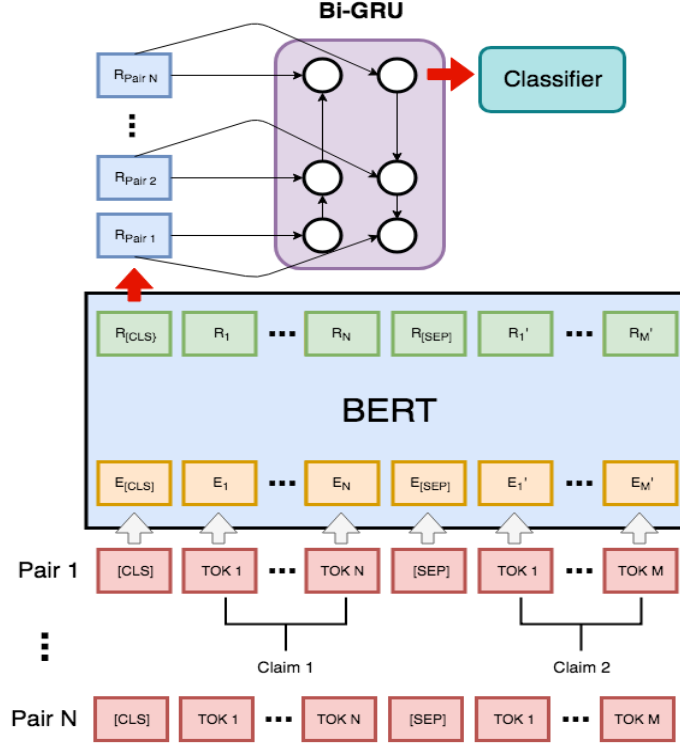
Figure 2: Hierarchical model for stance classification. A pre-trained BERT model is used to encode pairs of claims, which are then fed into a bi-directional GRU, to encode the path. In the figure, E_i represents the input embedding for token TOK_i, R_i represents the contextual representation for token TOK_i from the final layer in the BERT model, and $R_{pair\ i}$ is the representation of Pair i.

same argument path from the thesis to a given leaf claim. We note that specificity always increases along a given path, as each child claim is addressing some aspect of its parent claim, by either supporting or opposing, and therefore by definition has to be more specific. While an increase in depth is correlated with an increase in specificity for claims within a given argument path, this correlation does not necessarily hold for claims across different argument paths within a tree[5]. One important note is that we use the path information only as a way to reliably generate specificity labels, without requiring human annotations. The task of relative specificity detection itself does not require any path information to be present, nor do we make any assumptions in our models about the availability of path information.

For this task, given a pair of claims, we want the model to determine whether the second claim is more specific than the first claim. We note that unlike in stance prediction, we never provide the

path information between a pair of claims, as this would be equivalent to giving the gold label as input to the model, since given the path, the relative specificity is deterministic.

3.1 Results and Analysis

Baseline. We experiment with feature-based Logistic Regression (LR) model that incorporates all the features that are shown to be effective in determining sentence specificity (Louis and Nenkova, 2012). For example, this feature list includes polarity of the claims (Wilson et al., 2005), number of personal pronouns in the claims, and length of the claims since (Louis and Nenkova, 2012) shows that generic sentences have stronger polarity, less number of personal pronouns and are shorter in length. While Ko et al. (2019) has also looked at the task of specificity prediction, we cannot directly apply their models to our data, since their annotation scheme requires each sentence to be labelled as general or specific, whereas we argue that specificity is relative.

Fine-tuned BERT. We compare our baselines with a fine-tuned BERT model (Devlin et al., 2018). BERT is a pre-trained deep bidirectional

---

[5]We also cannot guarantee that these claims are completely irrelevant and specificity comparison is not applicable. We would need human annotation for these cases to be able to make any claims for the relative specificity.

	Train	Development	Test
Specificity	196,474	77,599	79,394
Stance	159,726	60,891	65,732

Table 1: Number of examples (claim pairs) in each split for claim specificity and claim stance tasks.

transformer model that can encode sentences into dense vector representations. It is trained on large un-annotated corpora such as Wikipedia and the BooksCorpus (Zhu et al., 2015) using two different learning objectives, namely masked language model and next sentence prediction. These learning objectives together allow the model to learn representations that can be easily fine-tuned to achieve state-of-the-art performance for a wide range of natural language processing tasks.

For relative specificity detection, we feed the pair of claims as a single sequence with the special [SEP] token between the claims, and a [CLS] token at the beginning of the sequence, as shown in Figure 2, into a pre-trained BERT model[6]. In addition, we indicate each token in the first claim (as well as the [CLS] and [SEP] tokens) as belonging to sentence A, and each token in the second claim as belonging to sentence B, which is used by the BERT model to add the appropriate learned sentence embedding to each token. Note that this approach of packing a pair of claims into a single sequence is consistent with the input representation from (Devlin et al., 2018), for tasks where the input is a pair of sequences. We then take the output of the [CLS] token from the final layer of the BERT model, and feed it into a classification layer. We fine-tune[7] this architecture for relative specificity detection.

We split our data into train, development and test sets, by topic, which ensures that all nodes from the same tree are confined to a single split. We split the data in this way in order to encourage our models to learn more domain independent features, that are applicable across the diverse set of controversial topics. Number of examples in each split for each task is shown in Table 1.

Table 2 compares the performance of the dif-

ferent models for relative specificity, across three different settings. In the first setting, we evaluate the models across all claim pairs that occur in the same argument path in a given tree. We then control for the distance between the pair, in the second setting, by evaluating only across pairs of nodes that are distance 1 from each other, i.e. have a parent-child relationship. Finally, we control for the stance, in the third setting, and evaluate across pairs of claims that have the same stance relative to their parent.

Analysis. Consistent with previous work (Li and Nenkova, 2015), we find that length is highly predictive of specificity and more specific claims are longer than more generic claims. Across all settings, the fine-tuned BERT model achieves the best performance. As expected, the performance degrades, for all models, as we control for distance and stance, since the claims get more similar in language, for both cases.

Table 4 shows the top weighted words by BOW model for each class. We find that connectives (such as also, but, because, when) are associated more with arguments with higher specificity as they are mostly used to add more specific information to the claims as also found by Lugini and Litman (2017), whereas concept words (such as society, world, gender) have higher association with more generic arguments since these words represents the concepts of the controversial topics that people argue about.

We further evaluate our models for the claim pairs with distance values 2 to 5 as shown in Table 3. We find that BERT model is consistently the best performing model for all distance pairs. As we increase the distance, the models achieve higher prediction performance despite having less training examples for higher distance values.

4 Claim Stance Detection

It is not sufficient for debate and argument generation systems to determine the claim stance only with respect to the argument thesis; it is also necessary to determine the stance between any pair of claims that address the same line of reasoning. An argument generation system, for example, may need to generate arguments that oppose some of the opponent's previous claims while supporting some of its own previous claims during the debate which would require to determine the stance between any candidate claims and the claims in the

---

[6]Specifically, we use the BERT-Base (Uncased) model, which contains 12 layers of bidirectional transformers, with a hidden size of 768 units and 12 attention heads (for a total of 110M parameters).

[7]For all fine-tuning experiments with BERT, we used a learning rate of $2e^{-5}$. We ran the fine-tuning jobs for a maximum of 5 epochs, and used the validation performance for early stopping.

Model	All pairs	Distance one	Same stance
Majority	50.14	50.25	49.97
Length	74.94	64.67	69.62
Bag of Words (BOW) LR	77.10	66.01	70.43
Feature-based LR	78.18	67.06	72.03
BOW + Feature based LR	79.12	67.54	73.14
Fine-tuned BERT	**84.91**	**74.51**	**80.23**

Table 2: Accuracy numbers for argument specificity, across the different settings.

Model	d=1	d=2	d=3	d=4	d=5
Length	64.67	76.40	80.22	80.40	79.69
BOW + Feature based LR	67.54	79.98	84.46	85.14	85.66
Fine-tuned BERT	**74.51**	**85.57**	**89.30**	**90.57**	**91.62**

Table 3: Accuracy numbers for argument specificity at distance 2-5.

More generic	More specific
should	also
society	but
gender	only
world	because
humans	at
rights	when
would	even
government	that

Table 4: Words associated with more generic and specific arguments.

previous argument discourse.

In this work, given a claim A at depth d and claim B at depth $> d$ along the same argument path, we determine whether B (in)directly SUPPORTS or OPPOSES A (stance). If A and B do not have parent-child relationship, we determine whether B indirectly SUPPORTS or OPPOSES A by considering support/oppose relationship of each parent-child claims between A and B. Following the example shown in Figure 1, the claim "The capacity of harm is greater when the magic is involved" is directly supported by the claim "There is a great capacity to harm others using the Unforgivable Curses", with a **direct** parent-child relationship. However, the argument "The Unforgivable Curses are illegal in the wizarding world and their use is grounds for immediate life imprisonment in Azkaban Prison" is **indirectly opposing** the same claim, by rebutting it's parent, which presents a supporting point for the claim.

4.1 Results and Analysis

We experiment with a feature-based Logistic Regression model and a fine-tuned BERT model (Devlin et al., 2018) using the same strategy to split the data into train, development and test sets as in Section 3.1.

Baseline. Our feature-based model employs features shown to be effective in stance detection tasks (Mohammad et al., 2016) such as bag of words, word match, sentiment match, document embedding similarity, and MPQA subjectivity features (Wilson et al., 2005)[8]. We cannot evaluate the model from Sun et al. (2018) as a baseline, as that requires additional annotations for argument phrases for the given topics. Similarly, we cannot evaluate the model from Bar-Haim et al. (2017) as a baseline, since it would require additional annotations for target phrases in each claim, polarity towards the target phrases, and consistent/contrastive labels between the target phrases of two claims.

Fine-tuned BERT. We feed a pair of claims into a pre-trained BERT model, in the same manner as detailed above for relative specificity detection, and take the output of the [CLS] token from final layer and feed it into a classifier. We fine-tune this model for relative stance detection.

Fine-tuned BERT with path (simple). In this model, we incorporate path information in a very naïve manner. For a given pair of claims A and

---

[8]For Featured-based LR with path, we concatenate the all claims along an argument path, and extract features from this concatenated sequence.

Model	Distance one	All pairs
Majority	44.63	49.48
Feature-based LR	63.02	55.10
Feature-based LR with path	61.27	54.70
Fine-tuned BERT	74.84	64.08
Fine-tuned BERT with path (simple)	76.77	66.22
Fine-tuned BERT with path (hierarchical)	**77.46**	**68.55**

Table 5: Accuracy numbers for argument stance detection, across the different settings.

B, where A is a predecessor of B, we concatenate the path of claims starting from B up to A with each claim separated by the special [SEP] token. We indicate each token from claim B as belonging to sentence A, and the tokens from all other claims in the path, including claim A, are indicated as belonging to sentence B. We note that this way of processing the input is similar to how (Devlin et al., 2018) processed their input for the QA task. Similar to the previous model, we feed the output of the [CLS] token from the final layer into a classifier. We then fine-tune this model for relative stance classification.

Fine-tuned BERT with path (hierarchical). We hypothesize that the task of determining relative stance becomes easier, if we can follow along the argument path and determine the relative stance between parent-child claims. We incorporate this inductive bias into the model by constructing a hierarchical architecture for relative stance classification, as shown in Figure 2. First, we feed each parent-child pair along an argument path as a single sequence into the BERT encoder, separated by the [SEP] token, and take the representation of the [CLS] token from final layer of the BERT model, as the pair representations. We then feed the sequence of pair representations into a bidirectional Gated Recurrent Unit (GRU) (Cho et al., 2014), to get the path representation. In our experiments, we used a single bidirectional GRU layer with 128 units. The output of the last token from the forward GRU, and the output of the first token from the backward GRU are concatenated together to get the final path representation. We then feed this into a classifier to predict relative stance. We fine-tune this architecture for relative stance classification.

Table 5 compares the performance of the different models for argument stance detection, across two different settings. In the first setting, we evaluate the models only across pairs of claims that are distance 1 from each other, i.e. in a parent-child relationship. In the second setting, we evaluate the model across all pairs that occur in the same argument path in a given tree **with** and **without** incorporating the claims along the argument path between these pair of claims.

Analysis. We find that the fine-tuned BERT models perform much better than the feature based models and baselines, across both the settings. Also, as we hypothesized, having the argument path information is useful for determining relative stance between claims that do not have a parent-child relationship, as the BERT models with path information consistently perform better in the second setting, with the hierarchical BERT model being the best. In our dataset, an argument path from the tree is the best approximation that we have for an argumentative discourse, and as such our results suggest that considering discourse level context is useful in determining relative stance between two claims. However, as shown by our results, our models can still be employed when there is limited or no discourse information.

The performance degrades significantly[9] in the second setting, where we include claim pairs with all the distances, implying that it is easier to determine the stance relative to the parent, than claims that are further on the same path.

We do a more fine grained analysis of the performance of the fine-tuned BERT models, at different distances, which we present in Table 6. As expected, performance degrades for all models as the distance between the pair of claims increases. We find that at distance d=4 Fine-tuned BERT model that incorporates path information in a simple manner performs similarly to the model without path information. The hierarchical model, however, performs significantly better, which further justifies our choice to treat the argument path

---

[9]We measure the significance performing t-test.

4636

	d=1	d=2	d=3	d=4
Number of examples	21,451	19,940	14,947	9,394
Fine-tuned BERT	74.84	60.69	58.34	55.88
Fine-tuned BERT with path (simple)	76.77	65.10	59.12	55.80
Fine-tuned BERT with path (hierarchical)	**77.46**	**67.74**	**62.51**	**59.51**

Table 6: Accuracy for relative stance at distance 1-4.

context as a hierarchical rather than a flat representation.

5 Related Work

Argumentation Generation. Previous work in argument generation has focused on generating summaries of opinionated text (Wang and Ling, 2016), rebuttals for a given argument (Jitnah et al., 2000), paraphrases from predicate/argument structure (Kozlowski et al., 2003), generation via sentence retrieval (Sato et al., 2015) and developing argumentative dialogue agents (Le et al., 2018; Rakshit et al., 2017). The work on developing argumentative dialogue agents, in particular, has employed mostly social media data such as IAC (Walker et al., 2012c) to design retrieval-based or generative models to make argumentative responses to the users. These models, however, employ very limited context in generating the claims, and there is no notion of generating a claim with a particular stance or the appropriate level of specificity within the context. Furthermore, these models are trained on social media conversations, which can be noisy, and as noted by Rakshit et al. (2017), many sentences either do not express an argument or cannot be understood out of context. In contrast, our dataset explicitly provides the sequence of claims in an argument path that leads to any particular claim, which can enable an argument generation system to generate relevant claims, with a particular stance and at the right level of specificity. Recent work by Hua and Wang (2018) studies the task of generating claims of a different stance for a given statement, however their context is limited to the given statement and they do not take specificity into account.

Stance Detection. Previous work on claim stance detection has studied the important linguistic features to determine the stance of a claim relative to a thesis/main claim (Somasundaran and Wiebe, 2009, 2010; Walker et al., 2012a,b; Hasan and Ng, 2013; Sridhar et al., 2014; Thomas et al., 2006; Yessenalina et al., 2010; Burfoot et al.,

2011; Kwon et al., 2007; Faulkner, 2014; Bar-Haim et al., 2017). Some of these studies have shown that simple linear classifiers with uni-gram and n-gram features are effective for this task (Somasundaran and Wiebe, 2010; Hasan and Ng, 2013; Mohammad et al., 2016). However, in our setting, since we try to predict the stance between all pairs of claims on an argument path, rather than simply claims that are directed towards the thesis or the parent claim, we find that the models with a hierarchical representation of the argument path, i.e. the context, significantly outperform these baselines.

Argument Structure and Quality. There has been tremendous amount of work in computational argumentation mining focusing on determining argumentative components (Mochales and Moens, 2011; Stab and Gurevych, 2014; Nguyen and Litman, 2015) and argument structure in text (Palau and Moens, 2009; Biran and Rambow, 2011; Feng and Hirst, 2011; Lippi and Torroni, 2015; Park and Cardie, 2014; Peldszus and Stede, 2015; Niculae et al., 2017; Rosenthal and McKeown, 2015), and understanding the argument quality dimensions (Wachsmuth et al., 2017; Carlile et al., 2018) and the characteristics of persuasive arguments (Kelman, 1961; Burgoon et al., 1975; Chaiken, 1987; Tykocinskl et al., 1994; Chambliss and Garner, 1996; Durmus and Cardie, 2018; Dillard and Pfau, 2002; Cialdini, 2007; Durik et al., 2008; Tan et al., 2014; Marquart and Naderer, 2016; Durmus and Cardie, 2019). Existing work on claim specificity and stance detection has mostly employed datasets extracted from monologic documents that include more shallow support/oppose structures (Bar-Haim et al., 2017; Faulkner, 2014). Although there has been some work on constructing argument structure datasets using news sources (Reed et al., 2008), micro-texts (Peldszus, 2014) and user comments (Park and Cardie, 2018), these structures tend to be shallower and include fewer opposing claims since they employ existing monologic texts that are rel-

atively short. In contrast, the dataset we provide is constructed with the goal of providing supporting and opposing claims for each of the claim presented in an argument tree. Therefore, these argument tree structures are deeper and have more balanced number of supporting and opposing claims.

6 Conclusion

We present a new dataset of manually curated argument trees, which can open interesting avenues of research in argumentation. We use this dataset to study methods for determining claim stance and relative claim specificity for complex argumentative structures. We find that it is easier to predict stance for claims that have a parent-child relationship, where as relative specificity is more difficult to predict in the same case. For future work, it may be interesting to understand which other models would be effective in claim specificity and stance detection tasks. Besides, developing techniques to incorporate the claim stance and specificity detection models in argument generation to generate more coherent and consistent arguments is another interesting research direction to be explored.

Acknowledgments

References

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261. Association for Computational Linguistics.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 162–168. IEEE.

Clinton Burfoot, Steven Bird, and Timothy Baldwin. 2011. Collective classification of congressional floor-debate transcripts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,*

pages 1506–1515. Association for Computational Linguistics.

Michael Burgoon, Stephen B Jones, and Diane Stewart. 1975. Toward a message-centered theory of persuasion: Three empirical investigations of language intensity1. *Human Communication Research*, 1(3):240–256.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631. Association for Computational Linguistics.

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

Marilyn J. Chambliss and Ruth Garner. 1996. Do adults change their minds after reading persuasive text? *Written Communication*, 13(3):291–313.

Kyunghyun Cho, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Robert B. Cialdini. 2007. Influence: The psychology of persuasion.

Jean-Louis Dessalles. 2016. A cognitive approach to relevant argument generation. In *Principles and Practice of Multi-Agent Systems*, pages 3–15, Cham. Springer International Publishing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Amanda M Durik, M Anne Britt, Rebecca Reynolds, and Jennifer Storey. 2008. The effects of hedges in persuasive arguments: A nuanced analysis of language. *Journal of Language and Social Psychology*, 27(3):217–234.

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. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2019. Modeling the factors of user success in online debate. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2701–2707.

Adam Faulkner. 2014. Automated classification of stance in student essays: An approach using stance target information and the wikipedia link-based measure. In *FLAIRS Conference*.

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-Volume 1*, pages 987–996. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356. Asian Federation of Natural Language Processing.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230. Association for Computational Linguistics.

Nathalie Jitnah, Ingrid Zukerman, Richard McConachy, and Sarah George. 2000. Towards the generation of rebuttals in a bayesian argumentation system. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*.

Herbert C Kelman. 1961. Processes of opinion change. *Public opinion quarterly*, 25(1):57–78.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Domain agnostic real-valued specificity prediction. In *AAAI*.

Raymond Kozlowski, Kathleen F. McCoy, and K. Vijay-Shanker. 2003. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In *Proceedings of the Second International Workshop on Paraphrasing*.

Namhee Kwon, Liang Zhou, Eduard Hovy, and Stuart W. Shulman. 2007. Identifying and classifying subjective claims. In *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains*, dg.o '07, pages 76–81. Digital Government Society of North America.

Dieu-Thu Le, Cam Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130. Association for Computational Linguistics.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *AAAI*.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining.

Annie Louis and Ani Nenkova. 2012. General versus specific sentences: automatic identification and application to analysis of news summaries. *Technical Reports (CIS)*.

Luca Lugini and Diane Litman. 2017. Predicting specificity in classroom discussion. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–61. Association for Computational Linguistics.

Franziska Marquart and Brigitte Naderer. 2016. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*, pages 231–242. Springer Fachmedien Wiesbaden, Wiesbaden.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41. Association for Computational Linguistics.

Huy Nguyen and Diane Litman. 2015. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28, Denver, CO. Association for Computational Linguistics.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured svms and rnns. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995. Association for Computational Linguistics.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. ACM.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2018. A Corpus of eRulemaking User Comments for Measuring Evaluability of Arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Andreas Peldszus. 2014. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 88–97, Baltimore, Maryland. Association for Computational Linguistics.

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, Lisbon, Portugal. Association for Computational Linguistics.

Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn A. Walker. 2017. Debbie, the debate bot of the future. *CoRR*, abs/1709.03167.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *LREC 2008*.

Sara Rosenthal and Kathy McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *SIGDIAL Conference*, pages 168–177.

Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. 2015. End-to-end argument generation system in debating. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 109–114. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 226–234. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 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*, pages 116–124. Association for Computational Linguistics.

Dhanya Sridhar, Lise Getoor, and Marilyn Walker. 2014. Collective stance classification of posts in online debate forums. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 109–117. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409. Association for Computational Linguistics.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226. Association for Computational Linguistics.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. *arXiv preprint arXiv:1405.1438*.

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*, pages 327–335. Association for Computational Linguistics.

Orit Tykocinskl, E Tory Higgins, and Shelly Chaiken. 1994. Message framing, self-discrepancies, and yielding to persuasive messages: The motivational significance of psychological situations. *Personality and Social Psychology Bulletin*, 20(1):107–115.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187. Association for Computational Linguistics.

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. 2012a. 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*, pages 592–596. Association for Computational Linguistics.

Marilyn A. Walker, Pranav Anand, Rob Abbott, Jean E. Fox Tree, Craig Martell, and Joseph King. 2012b. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4):719 – 729. 1) Computational Approaches to Subjectivity and Sentiment Analysis 2) Service Science in Information Systems Research : Special Issue on PACIS 2010.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012c. A corpus for research on deliberation and debate. In *LREC*.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In

Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 47–57. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ainur Yessenalina, Yisong Yue, and Claire Cardie. 2010. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1046–1056. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27.