AMPERSAND: Argument Mining for PERSuAsive oNline Discussions

Tuhin Chakrabarty[1], Christopher Hidey[1], Smaranda Muresan[1,2]
, Kathleen Mckeown[1] and Alyssa Hwang[1]
[1]Department of Computer Science, Columbia University
[2]Data Science Institute, Columbia University

{tuhin.chakrabarty, smara, a.hwang}@columbia.edu
{chidey, kathy}@cs.columbia.edu

Abstract

Argumentation is a type of discourse where speakers try to persuade their audience about the reasonableness of a claim by presenting supportive arguments. Most work in argument mining has focused on modeling arguments in monologues. We propose a computational model for argument mining in online persuasive discussion forums that brings together the micro-level (argument as product) and macro-level (argument as process) models of argumentation. Fundamentally, this approach relies on identifying relations between components of arguments in a discussion thread. Our approach for relation prediction uses contextual information in terms of fine-tuning a pre-trained language model and leveraging discourse relations based on Rhetorical Structure Theory. We additionally propose a candidate selection method to automatically predict what parts of one's argument will be targeted by other participants in the discussion. Our models obtain significant improvements compared to recent state-of-the-art approaches using pointer networks and a pre-trained language model.

1 Introduction

Argument mining is a field of corpus-based discourse analysis that involves the automatic identification of argumentative structures in text. Most current studies have focused on monologues or *micro-level* models of argument that aim to identify the structure of a single argument by identifying the argument components (classes such as "claim" and "premise") and relations between them ("support" or "attack") (Somasundaran et al., 2007; Stab and Gurevych, 2014; Swanson et al., 2015; Feng and Hirst, 2011; Habernal and Gurevych, 2017; Peldszus and Stede, 2015). *Macro-level* models (or dialogical models) and rhetorical models which focus on the *process* of argument in a dialogue (Bentahar et al., 2010) have received less attention.

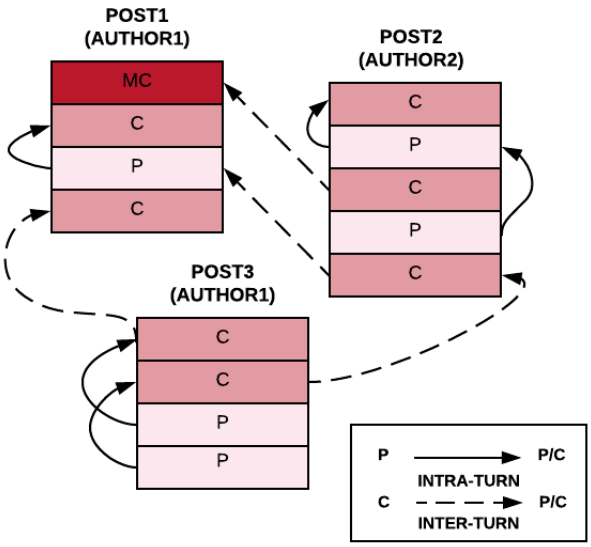

Figure 1: Annotated Discussion Thread (C: Claim, P: Premise, MC: Main Claim)

We propose a novel approach to automatically identify the argument structure in persuasive dialogues that brings together the micro-level and the macro-level models of argumentation. Our approach identifies argument components in a full discussion thread and two kinds of argument relations: *inter-turn relations* (argumentative relations to support or attack the other person's argument) and *intra-turn* relations (to support one's claim). Figure 1 shows a thread structure consisting of multiple posts with argumentative components (main claims, claims or premises) and both intra- and inter-turn relations. We focus here on the relation identification task (i.e, predicting the existence of a relation between two argument components).

To learn to predict relations, we **annotate argumentative relations** in threads from the Change

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 2933–2943,
Hong Kong, China, November 3–7, 2019.

My View (CMV) subreddit. In addition to the CMV dataset we also **introduce two novel distantly-labeled data-sets** that incorporate the macro- and micro-level context (further described in Section 3). We also **propose a new transfer learning approach** to fine-tune a pre-trained language model (Devlin et al., 2019) on the distant-labeled data (Section 4) and demonstrate improvement on both argument component classification and relation prediction (Section 5). We further show that **using discourse relations** based on Rhetorical Structure Theory (Mann and Thompson, 1988) improves the results for relation identification both for inter-turn and intra-turn relations. Overall, our approach for argument mining **obtains statistically significant improvement** over a state-of-the-art model based on pointer networks (Morio and Fujita, 2018) and a strong baseline using a pre-trained language model (Devlin et al., 2019). We make our data,[1] code, and trained models publicly available.[2]

2 Related Work

Argument mining on monologues Most prior work in argument mining (AM) has focused on monologues or essays. Stab and Gurevych (2017) and Persing and Ng (2016) used pipeline approaches for AM, combining parts of the pipeline using integer linear programming (ILP) to classify argumentative components. Stab and Gurevych (2017) represented relations of argument components in essays as tree structures. Both Stab and Gurevych (2017) and Persing and Ng (2016) propose joint inference models using ILP to detect argumentative relations. They however rely on handcrafted (structural, lexical, indicator, and discourse) features. We on the other hand use a transfer learning approach for argument mining in discussion forums.

Recent work has examined neural models for argument mining. Potash et al. (2017) use pointer networks (Vinyals et al., 2015) to predict both argumentative components and relations in a joint model. For our work, we focus primarily on relation prediction but conduct experiments with predicting argumentative components in a pipeline. A full end-to-end neural sequence tagging model was developed by Eger et al. (2017) that predicts argumentative discourse units (ADUs), components, and relations at the token level. In contrast, we assume we have ADUs and predict components and relations at that level.

Argument mining on discussion forums (online discussion) Most computational work related to argumentation as a process has focused on the detection of agreement and disagreement in online interactions (Abbott et al., 2011; Sridhar et al., 2015; Rosenthal and McKeown, 2015; Walker et al., 2012). However, these approaches do not identify the argument components that the (dis)agreement has scope over (i.e., what has been targeted by a disagreement or agreement move). Also in these approaches, researchers predict the type of relation (e.g. agreement) given that a relationship exists. On the contrary, we predict both argument components as well as the existence of a relation between them. Boltužić and Šnajder (2014) and Murakami and Raymond (2010) address relations between complete arguments but without the micro-structure of arguments as in Stab and Gurevych (2017). Ghosh et al. (2014) introduce a scheme to annotate inter-turn relations between two posts as "target-callout", and intra-turn relations as "stance-rationale". However, their empirical study is reduced to predicting the type of inter-turn relations as agree/disagree/other. Our computational model on the other hand handles both macro- and micro- level structures of arguments (argument components and relations).

Budzynska et al. (2014) focused on building foundations for extracting argument structure from dialogical exchanges (radio-debates) in which that structure may be implicit in the dynamics of the dialogue itself. By combining recent advances in theoretical understanding of inferential anchors in dialogue with grammatical techniques for automatic recognition of pragmatic features, they produced results for illocutionary structure parsing which are comparable with existing techniques acting at a similar level such as rhetorical structure parsing. Furthermore, Visser et al. (2019) presented US2016, the largest publicly available set of corpora of annotated dialogical argumentation. Their annotation covers argumentative relations, dialogue acts and pragmatic features.

Although Niculae et al. (2017) tried relation prediction between arguments in user comments on web discussion forums using structured SVM

[1] https://github.com/chridey/change-my-view-modes

[2] https://github.com/tuhinjubcse/AMPERSAND-EMNLP2019

and RNN the work closest to our task is that of Morio and Fujita (2018). They propose a pointer network model that takes into consideration the constraints of thread structure. Their model discriminates between types of arguments (e.g., claims or premises) and both intra-turn and inter-turn relations, simultaneously. Their dataset, which is not publically available, is three times larger than ours, so instead we focus on transfer learning approaches that take advantage of discourse and dialogue context and use their model as our baseline.

3 Data

3.1 Labeled Persuasive Forum Data

To learn to predict relations, we use a corpus of 78 threads from the CMV subreddit annotated with argumentative components (Hidey et al., 2017). The authors annotate *claims* ("propositions that express the speakers stance on a certain matter"), and *premises* ("propositions that express a justification provided by the speaker in support of a claim"). In this data, the *main claim*, or the central position of the original poster, is always the title of the original post.

We expand this corpus by annotating the argumentative relations among these propositions (both inter-turn and intra-turn) and extend the corpus by annotating additional argument components (using the guidelines of the authors) for a total of 112 threads. It is to be noted that the claims, premises, and relations have been annotated jointly. Our annotation results on the argumentative components for the additional threads were similar to Hidey et al. (2017) (IAA using Kripendorff alpha is 0.63 for claims and 0.65 for premises). The annotation guidelines for relations are available in the aforementioned Github repository.

Intra-turn Argumentative Relations As in previous work (Morio and Fujita, 2018), we restrict intra-turn relations to be between a premise and another claim or premise, where the premise either supports or attacks the claim or other premise. Evidence in the form of a premise is either support or attack. Consider the following example:

> [Living and studying overseas is an irreplaceable experience.]$_{0:\text{CLAIM}}$ [One will struggle with loneliness]$_{1:\text{PREMISE:ATTACK:0}}$ [but those difficulties will turn into valuable experiences.]$_{2:\text{PREMISE:ATTACK:1}}$ [Moreover, one will learn to live without depending on anyone.]$_{3:\text{PREMISE:SUPPORT:0}}$

The example illustrates that the argumentative component at index 0 is a claim, and is followed by attacking one's own claim, and in turn attacking that premise (an example of the rhetorical move known as prolepsis or prebuttal). They conclude the argument by providing supporting evidence for their initial claim.

Inter-turn Argumentative Relations Inter-turn relations connect the arguments of two participants in the discussion (agreement or disagreement). The argumentative components involved in inter-turn relations are claims, as the nature of dialogic argumentation is a difference in stance. An *agreement* relation expresses an agreement or positive evaluation of another user's claim whereas a *disagreement/attack* relation expresses disagreement. We further divide disagreement/attack relations into *rebuttal* and *undercutter* types to distinguish when a claim directly challenges the truth of the other claim or challenges the reasoning connecting the other claim to the premise that supports it, respectively. We also allowed annotators to label claims as *partial* agreement or disagreement/attack if the response concedes part of the claim, depending on the order of the concession.

In the following example, User A makes a claim and supports it with a premise. User B agrees with the claim but User C disputes the idea that there even is global stability. Meanwhile, User D disagrees with the reasoning that the past is always a good predictor of current events.

> A: [I think the biggest threat to global stability comes from the political fringes.]$_{0:\text{CLAIM}}$ [It has been like that in the past.]$_{1:\text{PREMISE}}$
>
> B: [Good arguments.]$_{2:\text{AGREEMENT:0}}$
>
> C: [The only constant is change.]$_{3:\text{REBUTTAL:0}}$
>
> D: [What happened in the past has nothing to do with the present.]$_{4:\text{UNDERCUTTER:1}}$

We obtain moderate agreement for relation annotations, similar to other argumentative tasks

(Morio and Fujita, 2018). The Inter-Annotator Agreement (IAA) with Kripendorff's α is 0.61 for relation presence and 0.63 for relation types.

In total, the dataset contains 380 turns of dialogue for 2756 sentences. There were 2741 argumentative propositions out of which 1205 are claims and 1536 are premises, with an additional 799 non-argumentative propositions. 66% of relations were in support, 26% attacking, and 8% partial. As we found that most intra-turn relations were in support and inter-turn relations were attacking, due to the dialogic nature of the data, for our experiments we only predicted whether a relation was present and not the relation type.

Overall, there are 7.06 sentences per post for our dataset, compared to 4.19 for Morio and Fujita (2018). This results in a large number of possible relations, as all pairs of argumentative components are candidates. The resulting dataset is very unbalanced (only 4.6% of 27254 possible pairs have a relation in the intra-turn case with only 3.2% of 26695 for inter-turn), adding an extra challenge to modeling this task.

3.2 Distant-Labeled Data

As the size of this dataset is small for modern deep learning approaches, we leverage distant-labeled data from Reddit and use transfer learning techniques to fine-tune a model on the appropriate context — micro-level for intra-turn relations and macro-level (dialogue) for inter-turn relations.

Micro-level Context Data In order to leverage transfer learning methods, we need a large dataset with distant-labeled relation pairs. In previous work, Chakrabarty et al. (2019) collected a distant-labeled corpus of opinionated claims using sentences containing the internet acronyms IMO (in my opinion) or IMHO (in my humble opinion) from Reddit. We leverage their data to model relation pairs by considering the following sentence (i.e., a premise) as well (when it was present), resulting in a total of 4.6 million comments. We denote this dataset as **IMHO+context**. The following example shows an opinionated claim (in bold) backed by a supporting premise (in italics).

> **IMHO, Calorie-counting is a crock what you have to look at is how wholesome are the foods you are eating.** *Refined sugar is worse than just empty calories - I believe your body uses a lot of nutrients up just processing and digesting it.*

We assume in this data that a relation exists between the sentence containing IMHO and the following one. While a premise does not always directly follow a claim, it does so frequently enough that this noisy data should be helpful.

Macro-level Context Data While the IMHO data is useful for modeling context from consecutive sentences from the same author, inter-turn relations are of a dialogic nature and would benefit from models that consider that macro-level context. We take advantage of a feature of Reddit: when responding to a post, users can easily *quote* another user's response, which shows up in the metadata. Particularly in CMV, this feature is used to highlight exactly what part of someone's argument a particular user is targeting. In the example in Table 1, the response contains an exact quote of a claim in the original post. We collect 95,406 threads from the full CMV subreddit between 2013-02-16 and 2018-09-05 and find pairs of posts where the quoted text in the response is an exact match for the original post (removing threads that overlap with the labeled data). This phenomenon occurs a minority of the time, but we obtain 19,413 threads. When the quote feature is used, posters often respond to multiple points in the original text, so for the 19,413 threads we obtain 97,636 pairs. As most language model fine-tuning is performed at the sentence level, we take the quoted text and the following sentence as our distant-labeled inter-post pairs. We refer to this dataset as **QR**, for quote-response pairs.

4 Methods

Identifying argumentative components is a necessary precursor to predicting an argumentative relation. In our data, we require the "source" of an intra-turn relation to be a premise that supports or attacks a "target" (a premise or another claim). For inter-turn relations, the source is always a claim that agrees or disagrees with a target. Thus, we model this process as a pipeline: perform three-way classification on claims, premises, and non-arguments and then predict if an outgoing relation exists from the source premise/claim to a target premise/claim. In predicting these relations, we consider all possible source-target pairs of premises and argumentative components within

CMV: A rise in female representation in elected government isn't a good or bad thing. According to this new story, a record number of women are seeking office in this year's US midterm elections. While some observers hail this phenomenon as a step in the right direction, I don't think it's good thing one way or the other: a politician's sex has zero bearing on their ability to govern or craft effective legislation. As such...
"I don't think it's good thing one way or the other: a politician's sex has zero bearing on their ability to govern or craft effective legislation" Nobody is saying that women are better politicians than men, and thus, more female representation is inherently better for our political system. Rather, the argument is that...

Table 1: Two posts from the CMV sub-reddit where a claim is targeted by the response

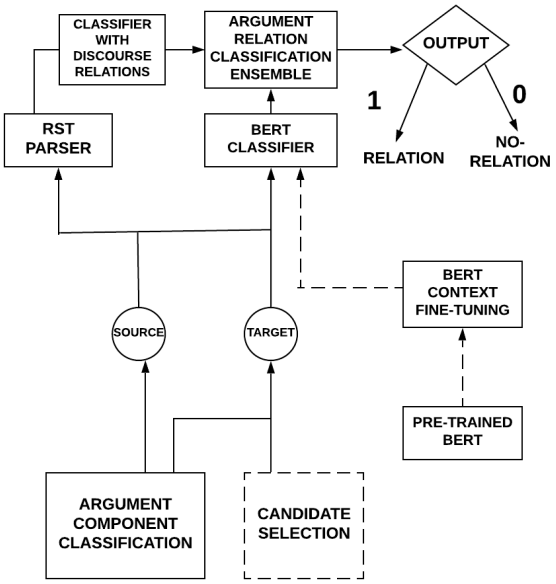

Figure 2: Schematic of our pipeline involving various components. The Candidate Selection Component is only used for Inter-turn relation identification

a single post (for intra-turn) and claims from one post and argumentative components from another post (for inter-turn).

As our data set is relatively small, we leverage recent advances in transfer learning for natural language processing. BERT (Devlin et al., 2019) has shown excellent performance on many semantic tasks, both for single sentences and pairs of sentences, making this model an ideal fit for both argument component classification and relation prediction. The BERT model is initially trained with a multi-task objective (masked language modeling and next-sentence prediction) over a 3.3 billion word English corpus. In the standard use, given a pre-trained BERT model, the model can be used for transfer learning by fine-tuning on a domain-specific corpus using a supervised learning objective.

We take an additional step of fine-tuning on the relevant distant-labeled data from Section 3.2 before again fine-tuning on our data from Section 3.1. We hypothesize that this novel use of BERT will help because the data is structured such that the sentence and next sentence pair (for intra-turn) or quote and response pair (for inter-turn) will encourage the model to learn features that improve performance on detecting argumentative components and relations. For argumentative components, we use this fine-tuned BERT model to classify a single proposition (as BERT can be used for both single inputs and input pairs). For relations (both intra-turn and inter-turn), we predict the relation between a pair of propositions. In the relation case, we postulate that additional components besides BERT can help: adding features derived from RST structure and pruning the space of possible pairs by selecting candidate target components. The pipeline is illustrated in Figure 2.

Argument Component Classification We fine-tune a BERT model on the **IMHO+context** dataset described in Section 3.2 using both the masked language modeling and next sentence prediction objectives. We then again fine-tune BERT to predict the argument components for both sources and targets, as indicated in Figure 2. Given the predicted argumentative components, we consider all possible valid pairs either within a post (for intra-turn) or across (inter-turn) and make a binary prediction of whether a relation is present.

Context Fine-Tuning For intra-turn relation prediction we use the same fine-tuned BERT model on **IMHO+context** that we used for argument component classification, as premises often immediately follow claims so this task is a noisy analogue to the task of interest. We then fine-tune on the relation prediction task on all possible pairs, using the labeled relations in the CMV data. For inter-turn relation prediction, we fine-tune first on the **QR** dataset, where the dialogue context more closely represents our labeled inter-post relations. Then, we fine-tune on inter-turn relation predic-

tion using all possible pairs as training. This process is indicated in Figure 2, where we use the appropriate context fine-tuning to obtain a relation classifier.

Discourse Relations Rhetorical Structure Theory was originally developed to offer an explanation of the coherence of texts. Musi et al. (2018) and, more recently Hewett et al. (2019), showed that discourse relations from RST often correlate with argumentative relations. We thus derive features from RST trees and train a classifier using these features to predict an argumentative relation. To extract features from a pair of argumentative components, we first concatenate the two components so that they form a single text input. We then use a state-of-the-art RST discourse parser (Ji and Eisenstein, 2014)[3] to create parse trees and take the predicted discourse relation at the root of the parse tree as a categorical feature in a classifier. There are 28 unique discourse relations predicted in the data, including *Circumstance*, *Purpose*, and *Antithesis*. We use a one-hot encoding of these relations as features and train an XGBoost Classifier (Chen and Guestrin, 2016) to predict whether an argument relation exists. This classifier with discourse relations, as indicated in Figure 2, is then ensembled with our predictions from the BERT classifier by predicting a relation if either one of the classifiers predicts a relation.

Candidate Target Selection For inter-turn relations, we take additional steps to reduce the number of invalid relation pairs. Predicting an argumentative relation is made more difficult by the fact that we need to consider all possible relation pairs. However, some argumentative components may contain linguistic properties that allow us to predict when they are targets even without the full relation pair. Thus, if we can predict the targets with high recall, we are likely to increase precision as we can reduce the number of false positives. Our candidate selection component, which identifies potential targets (as shown in Figure 2), consists of two sub-components: an extractive summarizer and a source-target constraint.

First, we use the **QR** data to train a model to identify candidate targets using techniques from extractive summarization, with the idea that targets may be salient sentences or propositions. We treat the quoted sentences as gold labels, resulting in 19,413 pairs of document (post) and summary (quoted sentences). For the example provided in Table 1, this would result in one sentence included in the summary. Thus, for a candidate argument pair A $\rightarrow$ B, where B is the quoted sentence in Table 1, if B is not extracted by the summarization model we predict that there is no relation between A and B. An example of target selection via extractive summarization is shown in Figure 3.

We use a state-of-the-art extractive summarization approach (Liu, 2019) for extracting the targets. The authors obtain sentence representations from BERT (Devlin et al., 2019), and build several summarization specific layers stacked on top of the BERT outputs, to capture document-level features for extracting summaries. We refer the reader to (Liu, 2019) for further details. We select the best summarization model on a held out subset using recall at the top K sentences.

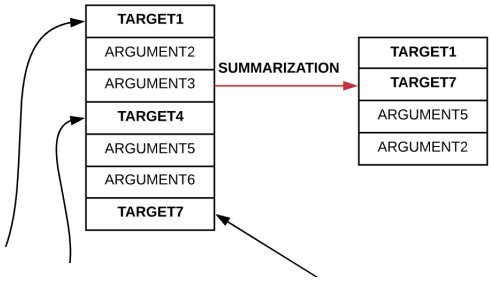

Figure 3: Schematic of our Target Extraction Approach

Second, in addition to summarization, we take advantage of a dataset-specific constraint: a target cannot also be a source unless it is related to the main claim. In other words, if B is a predicted target in A $\rightarrow$ B, we predict that there is no relation for B $\rightarrow$ A except when A is a main claim. In the CMV data, the main claim is always the title of the original Reddit post, so it is trivial to identify.

Method	C	P	NA
Stab and Gurevych (2017) + EWE	56.0	65.9	69.6
Morio and Fujita (2018)	54.2	68.5	73.2
Chakrabarty et al. (2019)	57.8	70.8	70.5
BERT Devlin et al. (2019)	62.0	72.2	71.3
IMHO Context Fine-Tuned BERT	**67.1**	**72.5**	**75.7**

Table 2: F-scores for 3-way Classification: Claim (C), Premise (P), Non-Argument (NA)

5 Experiments and Results

We use the CMV data from Section 3.1 for training and testing, setting aside 10% of the data for test.

[3]We use Wang et al. (2018) for segmentation of text into elementary discourse units.

Method	Precision		Recall		F-Score	
	Gold	Pred	Gold	Pred	Gold	Pred
All relations	5.0	-	100.0	-	9.0	-
Menini et al. (2018)	7.0	5.9	82.0	80.0	13.0	11.0
Menini et al. (2018) + RST Features	7.4	6.1	83.0	81.0	13.7	11.4
RST Features	6.3	5.7	79.5	77.0	11.8	10.6
Morio and Fujita (2018)	10.0	-	48.8	-	16.6	-
BERT Devlin et al. (2019)	12.0	11.0	67.0	60.0	20.3	18.5
IMHO Context Fine-Tuned BERT	14.3	13.2	69.0	65.0	23.7	21.8
+ RST Ensemble	**16.7**	**15.5**	**73.0**	**70.2**	**27.2**	**25.4**

Table 3: Results for Intra-turn Relation Prediction with Gold and Predicted Premises

Hyper-parameters are discussed in Appendix A.

5.1 Argumentative Component Classification

For baseline experiments on argumentative component classification we rely on the handcrafted features used by Stab and Gurevych (2017): lexical (unigrams), structural (token statistics and position), indicator (*I, me, my*), syntactic (POS, modal verbs), discourse relation (PDTB), and word embedding features. Hidey et al. (2017) show that emotional appeal or pathos is strongly correlated with persuasion and appears in premises. This motivated us to augment the work of Stab and Gurevych (2017) with emotion embeddings (Agrawal et al., 2018) which capture emotion-enriched word representations and show improved performance over generic embeddings (denoted in the table as EWE).

We also compare our results to several neural models - a joint model using pointer networks (Morio and Fujita, 2018), a model that leverages context fine-tuning (Chakrabarty et al., 2019), and a BERT baseline (Devlin et al., 2019) using only the pre-trained model without our additional fine-tuning step.

Table 2 shows that our best model gains statistically significant improvement over all the other models ($p < 0.001$ with a Chi-squared test). To compare directly to the work of Chakrabarty et al. (2019), we also test our model on the binary claim detection task and obtain a Claim F-Score of 70.0 with fine-tuned BERT, which is a 5-point improvement in F-score over pre-trained BERT and a 12-point improvement over Chakrabarty et al. (2019). These results show that fine-tuning on the appropriate context is key.

5.2 Relation Prediction

For baseline experiments on relation prediction, we consider prior work in macro-level argument mining. Menini et al. (2018) predict argumentative relations between entire political speeches from different speakers, which is similar to our dialogues. We re-implement their model using their features (lexical overlap, negation, argument entailment, and argument sentiment, among others). As with component classification, we also compare to neural models for relation prediction - the joint pointer network architecture (Morio and Fujita, 2018) and the pre-trained BERT (Devlin et al., 2019) baseline.

As the majority of component pairs contain no relation, we could obtain high accuracy by predicting that all pairs have no relation. Instead, we want to measure our performance on relations, so we also include an "all-relation" baseline, where we always predict that there is a relation between two components, to indicate the difficulty of modeling such an imbalanced data set. In the test data, for intra-turn relations there are 2264 relation pairs, of which only 174 have a relation, and for inter-turn relations there are 120 relation pairs, compared to 2381 pairs with no relation.

As described in Section 3.1, for intra-turn relations, the source is constrained to be a premise whereas for intra-turn, it is constrained to be a claim. We thus provide experiments using both gold claims/premises and predicted ones.

Intra-turn Relations We report the results of our binary classification task in Table 3 in terms of Precision, Recall and F-score for the "true" class, i.e., when a relation is present. We report results given both gold premises and predicted premises (using our best model from 5.1). Our best re-

Method	Precision		Recall		F-Score	
	Gold	Pred	Gold	Pred	Gold	Pred
All relations	5.0	-	100.0	-	9.0	-
Menini et al. (2018)	5.9	4.8	82.0	80.0	11.0	9.0
Menini et al. (2018) + RST Features	6.4	4.9	83.0	80.0	11.8	9.3
RST Features	5.1	3.8	80.0	77.0	9.6	7.2
Morio and Fujita (2018)	7.6	-	40.0	-	12.7	-
BERT Devlin et al. (2019)	8.8	7.9	76.0	70.0	15.8	14.1
QR Context Fine-Tuned BERT	11.0	10.0	75.3	72.5	19.1	17.6
+ RST Features	11.0	12.2	79.0	75.5	21.2	19.1
+ Extractive Summarizer	16.0	14.5	**79.4**	**75.6**	26.8	24.3
+ Source $\neq$ Target Constraint	**18.9**	**17.5**	79.0	74.0	**30.5**	**28.3**

Table 4: Results for Inter-Turn Relation Prediction with Gold and Predicted Claims

sults are obtained from ensembling the RST classifier with BERT fine-tuned on **IMHO+context**, for statistically significant ($p < 0.001$) improvement over all other models. We obtain comparable performance to previous work on relation prediction in other argumentative datasets (Niculae et al., 2017; Morio and Fujita, 2018).

Inter-turn Relations As with intra-turn relations, we report F-score on the "true" class in Table 4 for both gold and predicted claims. Our best results are obtained by fine-tuning the BERT model on the appropriate context (in this case the **QR** data) and ensembling the predictions with the RST classifier. We again obtain statistically significant ($p < 0.001$) improvement over all baselines.

Our methods for candidate target selection obtain further improvement. Using our extractive summarizer, we found that we obtained the best target recall of 62.7 at $K = 5$ (the number of targets to select). This component improves performance by approximately 5 points in F-score by reducing the search space of relation pairs. By further constraining targets to only be a source when targeting a main claim, we obtain another 4 point gain.

Window Clipping We also conduct experiments showing the performance for intra-turn relation prediction when constraining the relations to be within a certain distance in terms of the number of sentences apart. Often, in persuasive essays or within a single post the authors use premises to back/justify claims they immediately made. As shown in Figure 4, this behavior is also reflected in our dataset where the distance between the two arguments in the majority of the relations is +1.

We thus limit the model's prediction of a relation to be within a certain window and predict "no relation" for any pairs outside of that window. Table 5 shows that this window clipping on top of our best model improves F-score by only limiting the context where we make predictions. As our models are largely trained on discourse context and the next sentence usually has a discourse relation, we obtain improved performance as we narrow the window size. While we see a drop in recall the precision improves compared to our previous results in Table 3. It is also important to note that window clipping is only beneficial once we have a high recall, low precision scenario because when we predict everything at a distance of +1 as a relation we obtain low F-scores.

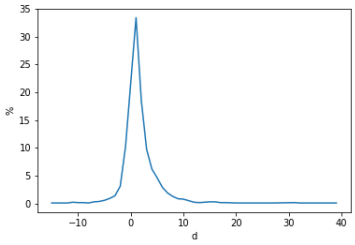

Figure 4: Distances d between Intra-Turn Relations

6 Qualitative Analysis

Role of Context We retrieve examples from the **IMHO+context** and **QR** data using TF-IDF similarity to pairs of argument components from our data that were predicted incorrectly by pre-trained BERT but correctly by the respective fine-tuned model. The first two rows in Table 6 show a relation between a claim and premise in the IMHO+Context and the CMV data respectively

Method	Window	Precision		Recall		F-Score	
		Gold	Pred	Gold	Pred	Gold	Pred
All relations	0 TO +1	5.0	4.0	31.0	25.0	8.7	6.9
Best Model	0 TO +5	19.5	17.1	**70.0**	**67.0**	30.5	27.2
	0 TO +4	21.4	19.5	67.0	65.0	32.2	30.0
	0 TO +3	25.2	23.3	61.1	58.0	35.6	33.2
	0 TO +2	32.5	29.8	50.0	48.0	39.3	36.8
	0 TO +1	**41.5**	**39.1**	47.0	42.0	**44.1**	**40.3**

Table 5: Intra-Turn Relation Prediction with Varying Window Settings

while the last four rows show a relation between a claim and premise in the QR data and the CMV data. The model learns discriminative discourse relations from the **IMHO+context** data and correctly identifies this pair. The last four rows show rebuttal from the QR and the CMV data respectively, where the model learns discriminative dialogic phrases (highlighted in bold).

Context	Pair
IMHO	IMHO, you should not **quantify** it as good or bad. **Tons of people** have monogamous relationships without issue.
CMV	[how would you even **quantify** that] [there are **many people** who want close relationships without romance]
QR	[It might be that egalitarians,anti-feminists, MRAs & redpillers, groups that I associate with opposing feminism - might be in fact very **distinct** & **different** groups, but I don't know that] [**I do see** all four of these as **distinct** groups].
CMV	[I may have a different stance on seeing no **difference** between companion animals and farm animals.] [**I do see distinction** between a pet and livestock]
QR	[Of course you **intend to kill** the person if you draw your **weapon**, if you can reasonably assume that they have a **weapon**] [**I don't think** some of them would **start killing**].
CMV	[So i thought, why would a police officer even use **firearms** if he/she doesn't **intend to kill**?] [**I dont think**, police are allowed to **start killing** someone with their gun if they don't intend to .]

Table 6: CMV and Context Examples

Role of Discourse We also provide examples that are predicted incorrectly by BERT but correctly by our classifier trained with RST features. For the first example in Table 7 the RST parser predicts an *Evaluation* relation, which is an indicator of an argumentative relation according to our model. For the second example the RST parser predicts *Antithesis*, which is correlated with attack relations (Musi et al., 2018), and is predicted correctly by our model.

Discourse	Argument1	Argument2
Evaluation	The only way your life lacks meaning is if you give it none to begin with	Life is ultimately meaningless and pointless.
Antithesis	Joseph was just a regular Jew without the same kind of holiness as the other two	Aren't Mary and Joseph, two holy people especially perfect virgin Mary, both Jews? Wasn't Jesus a Jew?

Table 7: Predicted Discourse Relations in CMV

7 Conclusion

We show how fine-tuning on data-sets similar to the task of interest is often beneficial. As our data set is small we demonstrated how to use transfer learning by leveraging discourse and dialogue context. We show how the structure of the fine-tuning corpus is essential for improved performance on pre-trained language models. We also showed that predictions that take advantage of RST discourse cues are complementary to BERT predictions. Finally, we demonstrated methods to reduce the search space and improve precision.

In the future, we plan to experiment further with language model fine-tuning on other sources of data. We also plan to investigate additional RST features. As the RST parser is not perfect, we want to incorporate other features from these trees that allow us to better recover from errors.

Acknowledgements

The authors thank Fei-Tzin Lee and Alexander Fabbri for their helpful comments on the initial draft of this paper and the anonymous reviewers for helpful comments.

References

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E Fox Tree, Robeson Bowmani, and Joseph King.

2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on LSM*, pages 2–11.

Ameeta Agrawal, Aijun An, Papagelis Chen, and Manos. 2018. Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artif. Intell. Rev.*, 33(3):211–259.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards argument mining from dialogue.

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019. IMHO fine-tuning improves claim detection. 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 558–563, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Jacob Devlin, Ming-Wei Chang, Lee Kenton, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 17th Annual Meeting of the North American Association for Computational Linguistics*.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.*, pages 11–22.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *ACL*, pages 987–996.

Debanjan Ghosh, Smaranda Muresan, Wacholder Nina, Mark Aakhus, and Matthew. Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *In Proceedings of the First Workshop on Argument Mining, hosted by the 52nd Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL*, pages 39–48.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Freya Hewett, Roshan Prakash Rane, Nina Harlacher, and Manfred Stede. 2019. The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103, Florence, Italy. Association for Computational Linguistics.

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 *In Proceedings of the 4th Workshop on Argument Mining. EMNLP*, pages 11–21.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,*, pages 13–24.

Yang Liu. 2019. Fine-tune bert for extractive summarization. In *https://arxiv.org/abs/1903.10318*.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Stefano Menini, Elena Cabrio, Sara Tonelli, and Villata Serena. 2018. Never retreat, never retract: Argumentation analysis for political speeches. In *Association for the Advancement of Artificial Intelligence*.

Gaku Morio and Katsuhide Fujita. 2018. End-to-end argument mining for discussion threads based on parallel constrained pointer architecture. In *In Proceedings of the 5th Workshop on Argument Mining. EMNLP*, pages 11–21.

Akiko Murakami and Rudy Raymond. 2010. Support or oppose?: Classifying positions in online debates from reply activities and opinion expressions. In *In Proceedings of the 23rd International Conference on Computational Linguistics, ArgMining@ACL*, pages 869–875.

Elena Musi, Tariq Alhindi, Stede Manfred, Leonard Kriese, Smaranda Muresan, and Andrea Rocci. 2018. A multi-layer annotated corpus of argumentative text: From argument schemes to discourse relations. In *Proceedings of Language Resources and Evaluation Conference (LREC 2018)*.

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, Vancouver, Canada. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *EMNLP*, pages 938–948.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *In Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*, pages 1384–1394.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Heres my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1364–1373.

Sara Rosenthal and Kathleen McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *16th Annual Meeting of the SIGDIAL*.

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.

Dhanya Sridhar, James R Foulds, Bert Huang, Lise Getoor, and Marilyn A Walker. 2015. Joint models of disagreement and stance in online debate. In *ACL*, pages 116–125.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *EMNLP*.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. In *Computational Linguistics, pages in press, preprint available at arXiv:1604.07370*.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the SIGDIAL*, pages 217–226.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28*, pages 2692–2700.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2019. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, pages 812–817.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967.