“Sharks are not the threat humans are”: Argument Component Segmentation in School Student Essays

Tariq Alhindi *
Department of Computer Science
Columbia University
tariq@cs.columbia.edu

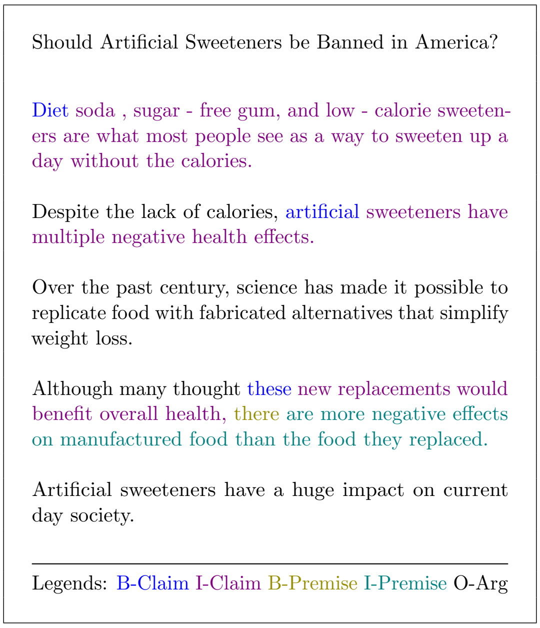
Debanjan Ghosh
Educational Testing Service
dghosh@ets.org

Abstract

Argument mining is often addressed by a pipeline method where segmentation of text into argumentative units is conducted first and proceeded by an argument component identification task. In this research, we apply a token-level classification to identify claim and premise tokens from a new corpus of argumentative essays written by middle school students. To this end, we compare a variety of state-of-the-art models such as discrete features and deep learning architectures (e.g., BiLSTM networks and BERT-based architectures) to identify the argument components. We demonstrate that a BERT-based multi-task learning architecture (i.e., token and sentence level classification) adaptively pretrained on a relevant unlabeled dataset obtains the best results.

1 Introduction

Computational argument mining focuses on subtasks such as identifying the Argumentative Discourse Units (ADUs) (Peldszus and Stede, 2013), their nature (i.e., claim or premise), and the relation (i.e., support/attack) between them (Ghosh et al., 2014; Wacholder et al., 2014; Stab and Gurevych, 2014, 2017; Stede and Schneider, 2018; Nguyen and Litman, 2018; Lawrence and Reed, 2020). Argumentation is essential in academic writing as it enhances the logical reasoning, as well as, critical thinking capacities of students (Ghosh et al., 2020). Thus, in recent times, argument mining has been used to assess students’ writing skills in essay scoring and provide feedback on the writing (Song et al., 2014; Somasundaran et al., 2016; Wachsmuth et al., 2016; Zhang and Litman, 2020).



Should Artificial Sweeteners be Banned in America?

Diet soda , sugar - free gum, and low - calorie sweeteners are what most people see as a way to sweeten up a day without the calories.

Despite the lack of calories, artificial sweeteners have multiple negative health effects.

Over the past century, science has made it possible to replicate food with fabricated alternatives that simplify weight loss.

Although many thought these new replacements would benefit overall health, there are more negative effects on manufactured food than the food they replaced.

Artificial sweeteners have a huge impact on current day society.

Legends: B-Claim I-Claim B-Premise I-Premise O-Arg

Figure 1: Excerpt from an annotated essay with Claim Premise segments in BIO notation

While argument mining literature has addressed students writing in the educational context, so far, it has primarily addressed college level writing (Blanchard et al., 2013; Persing and Ng, 2015; Beigman Klebanov et al., 2017; Eger et al., 2017) except for a very few ones (Attali and Burstein, 2006; Lugini et al., 2018; Correnti et al., 2020). Instead, in this paper, we concentrate on identifying arguments from essays written by *middle school students*. To this end, we perused a new corpus of 145 argumentative essays written by middle school students to identify the argument components. These essays are obtained from an Educational app - *Writing Mentor* - that operates as a Google-docs Add-on.[1]

Normally, research that investigates college

*Work performed during internship at ETS

[1] https://mentormywriting.org

Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pages 210–222
April 20, 2021 ©2021 Association for Computational Linguistics

students writing in the context of argument mining apply a pipeline of subtasks to first detect arguments at the token-level text units and subsequently classify the text units to argument components (Stab and Gurevych, 2017). However, middle school student essays are vastly different from college students' writing (detailed in Section 3). We argue they are more difficult to analyze through the pipeline approach due to run-on sentences, unsupported claims, and the presence of several claims in a sentence. Thus, instead of segmenting the text into argumentative/non-argumentative units first, we conduct a token-level classification task to identify the type of the argument component (e.g., B/I tokens from claims and premises) directly by joining the first and the second subtask in a single task. Figure 1 presents an excerpt from an annotated essay with their corresponding gold annotations of claims (e.g., "Diet soda . . . the calories") and premises (e.g., "there are . . . replaced"). The legends represent the tokens by the standard BIO notations.

We propose a detailed experimental setup to identify the argument components using both feature-based machine learning techniques and deep learning models. For the former, we used several structural, lexical, and syntactic features in a sequence classification framework using the Conditional Random Field (CRF) classifier (Lafferty et al., 2001). For the latter, we employ a BiLSTM network and, finally, a transformer architecture - BERT (Devlin et al., 2019) with its pretrained and task-specific fine-tuned models. We achieve the best result from a particular BERT architecture (7.5% accuracy improvement over the discrete features) that employs a joint multitask learning objective with an uncertainty-based weighting of two task-specific losses: (a) the *main task* of token-level sequence classification, and (b) the *auxiliary task* of sentence classification (i.e., whether a sentence contains argument or not). We make the dataset (student essays) from our research publicly available.[2]

2 Related Work

The majority of the prior work on argument mining addressed the problems of argument segmentation, component, and relation identification modeled in a pipeline of subtasks (Peldszus and Stede, 2015; Stab and Gurevych, 2017; Potash et al., 2017; Niculae et al., 2017) except a few research (Schulz et al., 2019). However, most of the research assumes the availability of segmented argumentative units and do the subsequent tasks such as the classification of argumentative component types (Biran and Rambow, 2011; Stab and Gurevych, 2014; Park and Cardie, 2014), argument relations (Ghosh et al., 2016; Nguyen and Litman, 2016), and argument schemes (Hou and Jochim, 2017; Feng and Hirst, 2011).

Previous work on argument segmentation includes approaches that model the task as a sentence classification to argumentative or non-argumentative sentences (Moens et al., 2007; Palau and Moens, 2009; Mochales and Moens, 2011; Rooney et al., 2012; Lippi and Torroni, 2015; Ajjour et al., 2017; Chakrabarty et al., 2019), or by defining heuristics to identify argumentative segment boundaries (Madnani et al., 2012; Persing and Ng, 2015; Al-Khatib et al., 2016). Although we conduct segmentation, we focus on the token-level classification to directly identify the argument component's type. This setup is related to Schulz et al. (2018) where authors analyzed students' diagnostic reasoning skills via token level identification. Our joint model using BERT is similar to (Eger et al., 2017). However, we set the main task as the token-level classification where the auxiliary task of argumentative sentence identification assists the main task to attain a better performance.

As stated earlier, most of the research on argumentative writing in an educational context focuses on identifying argument structures (i.e., argument components and their relations) (Persing and Ng, 2016; Nguyen and Litman, 2016) as well as to predict essays scores from features derived from the essays (e.g., number of claims and premises, number of supported claims, number of dangling claims) (Ghosh et al., 2016). Related investigations have also examined the challenge of scoring a certain dimension of essay quality, such as relevance to the prompt (Persing and Ng, 2014), opinions and their targets (Farra et al., 2015), argument strength (Persing and Ng, 2015) among others.

[2] https://github.com/EducationalTestingService/argument-component-essays

Majority of the above research are conducted in the context of college-level writing. For instance, Nguyen and Litman (2018) investigated argument structures in TOEFL11 corpus (Blanchard et al., 2013) which was also the main focus of (Ghosh et al., 2016). Beigman Klebanov et al. (2017) and Persing and Ng (2015) analyzed writing of university students and Stab and Gurevych (2017) used data from "essayforum.com", where college entrance examination is the largest forum. Although, writing quality in essays by young writers has been addressed (Attali and Burstein, 2006; Attali and Powers, 2008; Deane, 2014), identification of arguments was not part of these studies. Computational analysis of arguments from school students is in infancy except for a few research (Lugini et al., 2018; Afrin et al., 2020; Ghosh et al., 2020). We believe our dataset (Section 3) will be useful for researchers working at the intersection of argument mining and education.

3 Data

We obtained a large number of English essays (over 10K) through the *Writing Mentor* Educational App. This App is a Google Docs add-on designed to provide instructional writing support, especially for academic writing. The add-on provides students to write argumentative or narrative essays and receive feedback on their writings. We selected a subset of 145 argumentative essays for the annotation purpose. Essays were either self-labeled as "argumentative" or annotators identified their argumentative nature from the titles (e.g., "Should Artificial Sweeteners be Banned in America ?").[3] Essays covered various social issues related to climate change, veteran care, effects of wars, whether sharks are dangerous or not, etc. We denote this corpus as $ARG2020$ in the remaining sections of the paper. We employed three expert annotators (with academic and professional background in Linguistics and Education) to identify the argument components. The annotators were instructed to read sentences from the essays and identify the *claims* (defined as, "a potentially arguable statement that indicates a person is arguing for or arguing against something. Claims are not clarification or elaboration statements.") that the argument is in reference to. Next, once the claims are identified, the annotators annotated the *premises* (defined as, "reasons given by either for supporting or attacking the claims making those claims more than mere assertions").[4] Earlier research has addressed college level writing, and even such resources are scarce except for a few corpora (Stab and Gurevych, 2017) (denoted as $SG2017$ in this paper). On the contrary, $ARG2020$ is based on middle school students writing, which differs from college level writing $SG2017$ in several aspects briefly discussed in the next paragraph.

First, we notice that essays in $SG2017$ maintain distinct paragraphs such as the introduction (initiates the major claim in the essay), the conclusion (summarizes the arguments), and a few paragraphs in between that express many claims and their premises. However, essays written by middle school students do not always comply with such writing conventions to keep a concrete introduction and conclusion paragraph, rather, they write many short paragraphs (7-8 paragraphs on average) per essay while each paragraph contains multiple claims. Second, in general, claims in college essays in $SG2017$ are justified by one or multiple premises, whereas $ARG2020$ has many unsupported claims. For instance, the excerpt from the annotated essay in Figure 1 contains two unsupported claims (e.g., "Diet soda, sugar ... without the calories" and "artificial sweeteners ... health effects"). Third, middle school students often put opinions (e.g., "Sugar substitutes produce sweet food without feeling guilty consequences") or matter-of-fact statements (e.g., "Even canned food and dairy products can be artificially sweetened") that are not argumentative claims but structurally they are identical to claims. Fourth, multiple claims frequently appear in a single sentence that are separated by discourse markers or commas. Fifth, many essays contain run-on sentences (e.g., "this is hard on the family, they have a hard time adjusting") that make the task of parsing even tricky. We argue these reasons

[3] Other metadata reveal that middle school students write these essays. However, we did not use any such information while annotating the essays.

[4] Definitions are from (Stab and Gurevych, 2017) and Argument: Claims, Reasons, Evidence - Department of Communication, University of Pittsburgh (https://bit.ly/396Ap3H)

Corpora	Split	Essays	B-Claim	I-Claim	B-Premise	I-Premise	O-Arg	Total
*ARG*2020	*training*	100	1,780	21,966	317	3,552	51,478	79,093
	dev	10	171	1,823	32	371	4,008	6,405
	test	35	662	8,207	92	1,018	14,987	24,966

Table 1: Token counts of each category in the *training*, *dev*, and *test* sets of *ARG2020*

make identifying argument claims and premises from *ARG*2020 more challenging.

The annotators were presented with specific guidelines and examples for annotation. We conducted a pilot task first where all the three annotators annotated ten essays and exchanged their notes for calibration. Following that, we continued pair-wise annotation tasks (30 essays for each pair of annotators), and finally, individual annotators annotated the remaining essays. Since the annotation task involves identifying each argumentative component's words, we have to account for fuzzy boundaries (e.g., in-claim vs. not-in-claim tokens) to measure the IAA. We considered the Krippendorff's α (Krippendorff, 2004) metric to compute the IAA. We measure the α between each pair of annotators and report the average. For *claim* we have a modest agreement of 0.71 that is comparable to (Stab and Gurevych, 2014) and for *premise*, we have a high agreement of 0.90.

Out of the 145 essays from *ARG*2020 we randomly assign 100 essays for *training*, 10 essays for *dev*, and the remaining 35 essays for *test*. Table 1 represents the data statistics in the standard BIO format. We find the number of claims is almost six times the number of premises showing that the middle school students often fail to justify their proposed claims. We keep identifying opinions and argumentative relations (support/attack) as future work.

4 Experimental Setup

Majority of the argumentation research first segment the text in argumentative and non-argumentative segments and then identify the structures such as components and relations (Stab and Gurevych, 2017). Petasis (2019) mentioned that the granularity of computational approaches addressing the second task of argument component identification is diverse because some approaches consider detecting components at the clause level (e.g., approaches focused on the *SG*2017 corpus (Stab and Gurevych, 2014, 2017; Ajjour et al., 2017; Eger et al., 2017)) and others at the sentence levels (Chakrabarty et al., 2019; Daxenberger et al., 2017). We avoided both approaches for the following two reasons. First, middle school student essays often contain run-on sentences, and it is unclear how to handle clause level annotations because parsing might be inaccurate. Second, around 62% of the premises in the *training* set appears to be in the same sentence as their claims. This makes sentence classification to either claim or premise impractical (Figure 1 contains one such example). Thus, instead of relying on the pipeline approach, we tackle the problem by identifying argument components from the token-level classification akin to Schulz et al. (2019). Our unit of sequence tagging is a sentence, unlike a passage (Eger et al., 2017). We apply a five-way token-classification (or sequence tagging) task while using the standard BIO notation for the claim and premise tokens (See Table 1). Any token that is not "B-Claim", "I-Claim", "B-Premise", or "I-Premise" is denoted as "O-Arg". As expected, the number of "O-Arg" tokens is much larger than the other categories (see Table 1).

We explore three separate machine learning approaches well-established for studying token-based classification. First, we experiment with the sequence classifier Conditional Random Field (CRF) that exploits state-of-the-art discrete features. Second, we implement a BiLSTM network (with and without CRF) based on the BERT embeddings. Finally, we experiment with the fine-tuned BERT models with/without multitask learning setting.

4.1 Feature-based Models

Akin to (Stab and Gurevych, 2017) we experiment with three groups of discrete features: *structural*, *syntactic* and *lexico-syntactic* with some modifications. In addition, we experiment with embedding features extracted from the contextualized pre-trained language model of BERT.

Discrete Features For each token in a given essay, we extract structural features that include token position (e.g., the relative and absolute position of the token in the sentence, paragraph and, essay from the beginning of the essay) and punctuation features (e.g., whether the token is, preceded, or succeeded by punctuation). Such position features have shown to be useful in identifying claims and premises against sentences that do not contain any argument (Stab and Gurevych, 2017). We also extract syntactic features for each token that include part-of-speech tag of the token and normalized length to the lowest common ancestor (LCA) of the token and its preceding (and succeeding) token in the parse tree. In contrast with (Stab and Gurevych, 2017), we use dependency parsing as the base for the syntactic features rather than constituency parsing. Finally, we extract lexico-syntactic features (denoted as *lexSyn* in Table 2) that include the dependency relation governing the token in the dependency parse tree and the token itself, plus its governing dependency relation as another feature. This is also different than (Stab and Gurevych, 2017) where the authors used lexicalized-parse tree (Collins, 2003) to generate their lexico-syntactic features. These features are effective in identifying the argumentative discourse units. We also observed that using dependency parse trees as a basis for the lexico-syntactic features yields better results than constituency parse trees in our pilot experiments.

Embedding Features from BERT BERT (Devlin et al., 2019), a bidirectional transformer model, has achieved state-of-the-art performance in many NLP tasks. BERT is initially trained on the tasks of masked language modeling (MLM) and next sentence prediction (NSP) over very large corpora of English Wikipedia and BooksCorpus. During its training, a special token "[CLS]" is added to the beginning of each training instance, and the "[SEP]" tokens are added to indicate the end of utterance(s) and separate, in case of two utterances.

Pretrained BERT ("bert-base-uncased") can be used directly by extracting the token representations' embeddings. We use the average embeddings of the top four layers as suggested in Devlin et al. (2019). For tokens with more than one word-piece when running BERT's tokenizer, their final embeddings feature is the average vector of all of their word-pieces. This feature yields a 768D-long vector that we use individually as well as in combination with the other discrete features in our experiments. We utilize the sklearn-crfsuite tool for our CRF experiments.[5]

4.2 BiLSTM-CRF Models

To compare our models with standard sequence tagging models for argument segmentation (Petasis, 2019; Ajjour et al., 2019; Hua et al., 2019), we experiment with the BiLSTM-CRF sequence tagging model introduced by Ma and Hovy (2016) using the flair library (Akbik et al., 2019). We use the standard BERT ("bert-base-uncased") embeddings (768D) in the embedding layer and projected to a single-layer BiLSTM of 256D. BiLSTMs provide the context to the token's left and right, which proved to be useful for sequence tagging tasks. We train this model with and without a CRF decoder to see its effect on this task. The CRF layer considers both the output of the BiLSTM layer and the other neighboring tokens' labels, which improves the accuracy of the modeling desired transitions between labels (Ma and Hovy, 2016).

4.3 Transformers Fine-tuned Models

Pre-trained BERT can also be used for transfer learning by fine-tuning on a downstream task, i.e., claim and premise token identification task where training instances are from the labeled dataset *ARG*2020. We denote this model as $BERT_{bl}$. Besides fine-tuning with the labeled data, we also experiment with a *multitask learning* setting as well as conducted *adaptive pretraining* (Gururangan et al., 2020), that is continued pretraining on unlabeled corpora that can be task and domain relevant. We discuss the settings below.

Transformers Multitask Learning Multitask learning aims to leverage useful information in multiple related tasks to improve the performance of each task (Caruana, 1997). We treat the *sequence labeling* task of five-way token-level argument classification as the *main* task while we adopt the binary task

[5] https://sklearn-crfsuite.readthedocs.io

of *sentence-level argument* identification (i.e., whether the candidate sentence contains an argument (Ghosh et al., 2020) as the *auxiliary* task. Here, if any sentence in the candidate essay contains claim or premise token(s), the sentence is labeled as the positive category (i.e., argumentative), otherwise non-argumentative. We hypothesize that this auxiliary task of identifying argumentative sentences in a multitask setting could be useful for the main task of token-level classification.

We deploy two classification heads - one for each task - and the relevant gold labels are passed to them. For the auxiliary task, the learned representation for the "[CLS]" token is passed to the classification head. The two losses from these individual heads are added and propagated back through the model. This allows BERT to model the nuances of both tasks and their interdependence simultaneously. However, instead of simply adding the losses from the two tasks, we employ *dynamic weighting* of task-specific losses during the training process, based on the homoscedastic uncertainty of tasks, as proposed in Kendall et al. (2018):

$$L = \sum_t \frac{1}{2\sigma_t^2} L_t + \log \sigma_t^2 \tag{1}$$

where L_t and σ_t depict the task-specific loss and its variance (updated through backpropagation), respectively, over the training instances. We denote this model as $BERT_{mt}$.

Adaptive Pre-training Learning We adaptively pretrained BERT over two unlabeled corpora. First, we train on a *task relevant* Reddit corpus of 5.5 million opinionated claims that was released by Chakrabarty et al. (2019). These claims are self-labeled by the acronym: IMO/IMHO (in my (humble) opinion), which is commonly used in Reddit. We denote this model as $BERT_{IMHO}$. Next, we train on a *task and domain relevant* corpus of around 10K essays that we obtained originally (See section 3) from the *Writing Mentor* App, excluding the annotated set of *ARG*2020 essays. We denote this model as $BERT_{essay}$. Figure 2 displays the use of the *adaptive pretraining* step (in orange block) and the two classification heads (in green blocks) employed for the multitask variation.

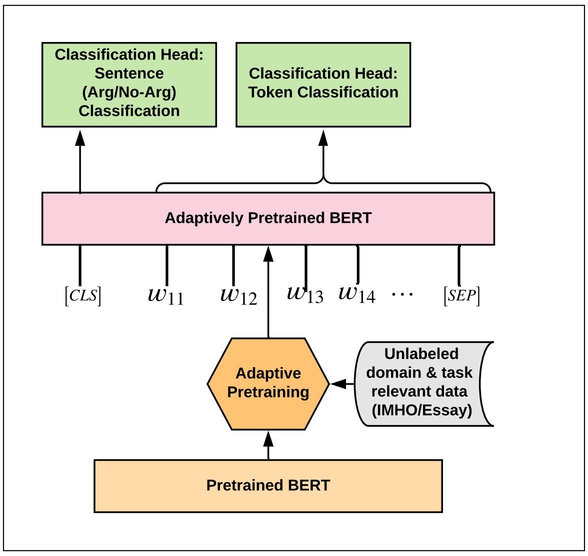

Classification Head: Sentence (Arg/No-Arg) Classification
Classification Head: Token Classification
Adaptively Pretrained BERT
[CLS]
w_{11}
w_{12}
w_{13}
w_{14}
···
[SEP]
Adaptive Pretraining
Unlabeled domain & task relevant data (IMHO/Essay)
Pretrained BERT

Figure 2: BERT fine-tuning with adaptive pretraining on unlabeled data from a relevant domain followed by fine-tuning on the labeled dataset with the multitask variation.

For brevity, the parameter tuning description for all the models and experiments - discrete feature-based and deep-learning ones (e.g., CRF, BiLSTM, BERT) is in the supplemental material.

5 Results and Discussion

We present our experiments' results using the CRF, BiLSTM, and BERT models under different settings. We report the individual F1, Accuracy, and Macro-F1 (abbrev. as "Acc." and "F1") scores for all the categories in Table 2 and Table 3.

We apply the discrete features (structural, syntactic, lexico-syntactic ("lexSyn")) together and individually to the CRF model. We observe the structural and syntactic features do not perform well individually, especially in the case of premise tokens (See Table 5 in Appendix A.3) and therefore, we only report the results of all discrete features (Discrete* in Table 2) and individually only the performance of the lexSyn features. Stab and Gurevych (2017) noticed that structural features are effective to identify argument components, especially from the introduction and conclusion sections of the college level essays because they contain few argumentatively relevant content. On the contrary, as stated earlier, school student es-

	CRF						
Features	B-Claim	I-Claim	B-Premise	I-Premise	O-Arg	Acc.	F1
lexSyn	.395	.530	.114	.176	.768	.673	.397
Discrete*	.269	.504	0	.013	.695	.595	.296
Embeddings	.401	.560	.048	.139	.769	.676	.384
Embeddings+lexSyn	.482	.610	.134	.180	.764	.682	.434
Embeddings+Discrete*	.434	.593	.055	.152	.762	.676	.399
	BiLSTM						
Setup	B-Claim	I-Claim	B-Premise	I-Premise	O-Arg	Acc.	F1
BiLSTM	.556	.680	.239	.438	.797	.735	.542
BiLSTM-CRF	.558	.676	.199	.378	.789	.727	.520
	BERT						
Setup	B-Claim	I-Claim	B-Premise	I-Premise	O-Arg	Acc.	F1
BERT$_{bl}$	.563	.674	.274	.425	.795	.728	.546
BERT$_{bl_{IMHO}}$	.571	.681	.304	.410	.795	.730	.540
BERT$_{bl_{essay}}$	.564	.676	.261	.406	.792	**.747**	.561
BERT$_{mt}$	.567	.685	.242	.439	.805	.741	.548
BERT$_{mt_{IMHO}}$	.562	.684	.221	.413	.794	.731	.534
BERT$_{mt_{essay}}$	.580	.702	.254	.427	.810	**.752**	**.574**

Table 2: F1 scores for Claim and Premise Token Detection on the test set. Underlined: highest Accuracy/F1 in group. **Bold**: highest Accuracy/F1 overall. *Discrete: includes structural, syntactic, and lexSyn features.

says do not always comply with such writing conventions. Table 2 displays that the lexSyn feature independently performs better by almost 8% accuracy than the combination of the other discourse features. This correlates to the findings from prior work on $SG2017$ (Stab and Gurevych, 2017) where the lexSyn features reached the highest F1 on a similar corpus. Next, we augment the embedding features from the BERT pre-trained model with the discrete features and notice a marginal improvement in the accuracy score (less than 1%) over the performance of lexSyn features. This improvement is achieved from the higher accuracy in detecting the claim terms (e.g., Embedding+Discrete* achieves around 17% and 10%, an improvement over Discrete* features in the case of B-Claim and I-Claim, respectively). However, the accuracy of detecting the premise tokens is still significantly low. We assume that this could be due to the low frequency of premises in the *training* set, which seems to be more challenging for the CRF model to learn useful patterns from the pre-trained embeddings. On the contrary, the O-Arg token is the most frequent in the essays and that is reflected in the overall high accuracy scores for the O-Arg tokens (i.e., over 76% on average).

The overall performance(s) improve when we apply the BiLSTM networks on the *test* data. Accuracy improves by 5.3% in the case of BiLSTM against the Embeddings+lexSyn features. However, results do not improve when we augment the CRF classifier on top of the LSTM networks (BiLSTM-CRF). Instead, the performance drops by 0.8% accuracy (See Table 2). On related research, Petasis (2019) have conducted extensive experiments with the BiLSTM-CRF architecture with various types of embeddings and demonstrated that only the specific combination of embeddings (e.g., GloVe+Flair+BERT) achieves higher performance than BiLSTM-only architecture, but we leave such experiments for future work.

In the case of BERT based experiments, we observe BERT$_{bl}$, obtains an accuracy of 73% that is comparable to the BiLSTM performance. In terms of the individual categories, we observe BERT$_{bl}$ achieves around 7.5% improvement over the BiLSTM-CRF classifier for the B-Premise tokens. We also observe that the two adaptive-pretrained models (e.g., BERT$_{IMHO}$ and BERT$_{essay}$) perform better than the BERT$_{bl}$ where BERT$_{essay}$ achieves the best accuracy of 74.7%, a 2% improvement over BERT$_{bl}$. Although BERT$_{IMHO}$ was trained on a much larger corpus than BERT$_{essay}$, we assume since BERT$_{essay}$ was trained on a *domain relevant* corpus it achieves the highest F1. Likewise, in the case of multitask models, we observe BERT$_{mt}$ performs better than BERT$_{bl}$ by 1.3%. This shows that using argumentative sentence identification as an auxiliary task is beneficial for token-level classification. With regards to the adaptive-pretrained models, akin to the BERT$_{bl}$ based experiments, we observe BERT$_{mt_{essay}}$ perform best by achieving the highest accuracy over 75%.

Argument Segmentation We choose the five-way token-level classification of argument component over the standard pipeline approach because the standard level of granularity (sentence or clause-based) is not applicable to our *training* data. In order to test the benefit of the five-way token-level classification, we also compare it against the traditional approach of segmentation of argumentative units into argumentative and non-argumentative tokens. We again follow the standard BIO notation for a three-way token classification setup (B-Arg, I-Arg, and O-Arg) for argument segmentation. In this setup, the B-Claim and B-Premise classes are merged into B-Arg, and I-Claim and I-Premise are merged into I-Arg, while the O-Arg class remains unchanged. The results of all of our models on this task are shown in Table 3. We notice similar patterns (except for BERT$_{mt_{IMHO}}$ that performs better than BERT$_{mt}$ this time) in this three-way classification task as we saw in the five-way classification. The best model remains to be the BERT$_{mt_{essay}}$ with 77.3% accuracy, which is an improvement of 2-3% over the BiLSTM and other BERT-based architecture.

In summary, we have two main observations from Table 2 and Table 3. First, the best model in Table 3 reports only about 3% improvement over the result from Table 2 which shows that the five-way token-level classification is comparable against the standard task of argument segmentation. Second, the accuracy of the argument segmentation task is much lower than the accuracy of college-level essay corpus *SG*2017 (Stab and Gurevych, 2017) reported accuracy of 89.5%). This supports the challenges of analyzing middle school student essays.

5.1 Qualitative Analysis

Since we have explored three separate machine learning approaches with a variety of experiments, we analyze the results obtained from the BERT$_{mt_{essay}}$ model that has performed the best (Table 2). According to the confusion matrix, there are three major sources of errors: (a) around 2500 “O-Arg” tokens are wrongly classified as “I-Claim” (b) 2162 “I-Claim” tokens are wrongly classified as “O-Arg”, and (c) 273 “I-Premise” tokens are erroneously classified as “I-Claim”. Here, (a) and (b) are not surprising given these are the two categories with the largest number of tokens. For (c) we looked at a couple of examples, such as “because of [Walmart ’s goal of saving money]$_{premise}$, [customers see value in Walmart that is absent from other retailers]$_{claim}$”. Here, the premise tokens are wrongly classified as O-Arg tokens. This is probably because the premise appears before the claim, which is uncommon in our *training* set. We notice some of the other sources of errors, and we discuss them as follows:

	CRF				
Features	B-Arg	I-Arg	O-Arg	Acc.	F1
lexSyn	.385	.518	.768	.683	.557
Discrete	.288	.493	.710	.625	.497
Embeddings	.379	.596	.767	.699	.581
Emb+lexSyn	.468	.622	.768	.708	.619
Emb+Discrete	.381	.599	.767	.699	.582
	BiLSTM				
Setup	B-Arg	I-Arg	O-Arg	Acc.	F1
BiLSTM	.546	.730	.792	.759	.689
BiLSTM-CRF	.553	.707	.793	.752	.684
	BERT				
Setup	B-Arg	I-Arg	O-Arg	Acc.	F1
BERT$_{bl}$	.567	.698	.795	.750	.687
BERT$_{bl_{IMHO}}$	.558	.717	.778	.744	.684
BERT$_{bl_{essay}}$	.567	.707	.795	.754	.690
BERT$_{mt}$	.555	.702	.803	.758	.688
BERT$_{mt_{IMHO}}$	.568	.719	.804	.764	.700
BERT$_{mt_{essay}}$	.563	.735	.811	**.773**	**.710**

Table 3: F1 scores for Argument Token Detection on the test set. Underlined: highest Accuracy/F1 in group. **Bold**: highest Accuracy/F1 overall.

non-arguments classified as arguments: This error occurs often, but it is more challenging for *opinions* or *hypothetical examples* that resemble arguments but are not necessarily arguments. For instance, the opinion “that actually makes me feel good afterward . . . ” and the hypothetical example “Next , you will not be eating due to your lack of money” are similar to an argument, and the classifier erroneously classifies them as claim. In the future, we plan to include the labeled *opinions* during training to investigate how the model(s) handle opinions vs. arguments during the classification.

missing multiple-claims from a sentence: In many examples, we observe multiple claims appear in a single sentence, such as: "[Some coral can recover from this]$_{claim}$ though [for most it is the final straw .]$_{claim}$". During prediction, the model predicts the first claim correctly but then starts the second claim with an "I-Claim" label, which is an impossible transition from "Arg-O" (i.e., does not enforce well-formed spans). Besides, the model starts the second claim wrongly at the word "most" rather than "for". This indicates the model's inability to distinguish discourse markers such as "though" as potential separators between argument components. This could be explained by the fact that markers such as "though" or "because" are frequently part of an argument claim. Such as, in "[those games do not seem as violent even though they are at the same level]$_{claim}$", "though" is labeled as "I-Claim'.

investigating run-on sentences: Some sentences contain multiple claims, which are written as one sentence via a comma-splice run-on such as "[Humans in today 's world do not care about the consequences]$_{claim}$, [only the money they may gain .]$_{claim}$" which has two claims in the gold annotations but it was predicted as one long claim by our best model. Another example is "[The oceans are also another dire need in today's environment]$_{claim}$, each day becoming more filled with trash and plastics.", in which the claim is predicted correctly in addition to another predicted claim starting at the word *"each"*. The model tends to over predicts claims when a comma comes in the middle of the sentence followed by a noun. However, in the former example, the adverb "only" that has a "B-Claim" label follows the comma rather than the more frequent nouns. Such instances add more complexity to understand and model argument structures in middle school student writing.

effect of the multitask learning: We examined the impact of multitask learning and notice two characteristics. First, as expected, the multitask model can identify claims and premises that are missed by the single task model(s), such as: "[many more negative effects that come with social media . . . "]$_{claim}$" that was correctly identified by the multitask model. Second, the clever handling of the back-propagation helps the multitask model to reduce false positives to be more precise. Many non-argumentative sentences, such as: "internet's social networks help teens find communities . . . " and opinions, such as: "take $1.3 billion off $11.3 billion the NCAA makes and give it to players" are wrongly classified as claims by the single task models but are correctly classified as non-argumentative by the multitask model.

6 Conclusion

We conduct a token-level classification task to identify the type of the argument component tokens (e.g., claims and premises) by combining the argument segmentation and component identification in one single task. We perused a new corpus collected from essays written by middle school students. W Our findings show that a multitask BERT performs the best with an absolute gain of 7.5% accuracy over the discrete features. We also conducted an in-depth comparison against the standard segmentation step (i.e., classifying the argumentative vs. non-argumentative units) and proposed a thorough qualitative analysis.

Middle school student essays often contain run-on sentences or unsupported claims that make the task of identifying argument components much harder. We achieve the best performance using a multitask framework with an adaptive pretrained model, and we plan to continue to augment other tasks (e.g., opinion and stance identification) under a similar multitask framework (Eger et al., 2017). We plan to generate personalized feedback for the students (e.g., which are the supported claims in the essay?) that is useful in automated writing assistance.

Acknowledgments

The authors would like to thank Tuhin Chakrabarty, Elsbeth Turcan, Smaranda Muresan and Jill Burstein for their helpful comments and suggestions.

References

Tazin Afrin, Elaine Lin Wang, Diane Litman, Lindsay Clare Matsumura, and Richard Correnti. 2020. Annotation and classification of evidence and reasoning revisions in argumentative writing. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 75–84, Seattle, WA, USA → Online. Association for Computational Linguistics.

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. 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 2922–2932, Hong Kong, China. Association for Computational Linguistics.

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan. The COLING 2016 Organizing Committee.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning and Assessment*, 4(3).

Yigal Attali and Don Powers. 2008. A developmental writing scale. *ETS Research Report Series*, 2008(1):i–59.

Beata Beigman Klebanov, Binod Gyawali, and Yi Song. 2017. Detecting Good Arguments in a Non-Topic-Specific Way: An Oxymoron? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 244–249, Vancouver, Canada. Association for Computational Linguistics.

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

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

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.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational linguistics*, 29(4):589–637.

Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, Diane Litman, Zahra Rahimi, and Zahid Kisa. 2020. Automated scoring of students' use of text evidence in writing. *Reading Research Quarterly*, 55(3):493–520.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066.

Paul Deane. 2014. Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks. *ETS Research Report Series*, (1):1–23.

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.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22.

Noura Farra, Swapna Somasundaran, and Jill Burstein. 2015. Scoring persuasive essays using opinions and their targets. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 64–74.

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, Portland, Oregon, USA. Association for Computational Linguistics.

Debanjan Ghosh, Beata Beigman Klebanov, and Yi Song. 2020. An exploratory study of argumentative writing by young students: A transformer-based approach. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 145–150, Seattle, WA, USA. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the first workshop on argumentation mining*, pages 39–48.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics.

Yufang Hou and Charles Jochim. 2017. Argument relation classification using a joint inference model. In *Proceedings of the 4th Workshop on Argument Mining*, pages 60–66, Copenhagen, Denmark. Association for Computational Linguistics.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality and quantity*, 38:787–800.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Luca Lugini, Diane Litman, Amanda Godley, and Christopher Olshefski. 2018. Annotating student talk in text-based classroom discussions. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 110–116.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Nitin Madnani, Michael Heilman, Joel Tetreault, and Martin Chodorow. 2012. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–28, Montréal, Canada. Association for Computational Linguistics.

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

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230.

Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137.

Huy V Nguyen and Diane J Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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.

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*, ICAIL '09, page 98–107, New York, NY, USA. Association for Computing Machinery.

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.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

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.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552.

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

Georgios Petasis. 2019. Segmentation of argumentative texts with contextualised word representations. In *Proceedings of the 6th Workshop on Argument Mining*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

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

Niall Rooney, Hui Wang, and Fiona Browne. 2012. Applying kernel methods to argumentation mining. In *FLAIRS Conference*, volume 172.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41, New Orleans, Louisiana. Association for Computational Linguistics.

Claudia Schulz, Christian M Meyer, and Iryna Gurevych. 2019. Challenges in the automatic analysis of students' diagnostic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6974–6981.

Swapna Somasundaran, Brian Riordan, Binod Gyawali, and Su-Youn Yoon. 2016. Evaluating argumentative and narrative essays using graphs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1568–1578.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

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

Manfred Stede and Jodi Schneider. 2018. Argumentation mining. *Synthesis Lectures on Human Language Technologies*, 11(2):1–191.

Nina Wacholder, Smaranda Muresan, Debanjan Ghosh, and Mark Aakhus. 2014. Annotating multiparty discourse: Challenges for agreement metrics. In *Proceedings of LAW VIII-The 8th Linguistic Annotation Workshop*, pages 120–128.

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, Osaka, Japan. The COLING 2016 Organizing Committee.

Haoran Zhang and Diane Litman. 2020. Automated topical component extraction using neural network attention scores from source-based essay scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8569–8584, Online. Association for Computational Linguistics.

A Appendix

A.1 Parameter Tuning

CRF experiment: For the CRF model, we search over the two regularization parameters c1 and c2 by sampling from exponential distributions with 0.5 scale for c1 and 0.05 scale for c2 using a 3 cross-validation over 50 iterations, which takes about 20 minutes of run-time. The final values are 0.8 for c1 and 0.05 for c2 for the best CRF model that uses LexSyn and BERT embeddings features.

BiLSTM experiment: For BiLSTM networks based experiments we searched the hyper parameters over the *dev* set. Particularly we experimented with different mini-batch size (e.g., 16, 32), dropout value (e.g., 0.1, 0.3, 0.5, 0.7), number of epochs (e.g., 40, 50, 100 with early stopping), hidden state of sized-vectors (256). Embeddings were generated using BERT ("bert-base-uncased") (768 dimensions). After tuning we use the following hyper-parameters for the *test* set: mini-batch size of 32, number of epochs = 100 (stop between 30-40 epochs), and dropout value of 0.1. The model has one BiLSTM layer with size 256 of the hidden layer.

BERT based models: We use the *dev* partition for hyperparameter tuning (batch size of 8, 16, 32, 48), run for 3,5,6 epochs, learning rate of 3e-5) and optimized networks with the Adam optimizer. The training partitions were fine-tuned for 5 epochs with batch size = 16. Each training epoch took between 08:46 $\sim$ 9 minutes over a K-80 GPU with 48GB vRAM.

A.2 Results of Discourse Feature Groups

We show below the results of using each of the three feature groups individually: structural, syntactic and lexical-syntactic. As mentioned in the results section of the paper, we can see below that the structural and syntactic features do not do well when used individually. Therefore, they were excluded from further experimentation with BERT embeddings. Only the LexSyn features were tested individually with the embeddings.

Features	Structural	Syntactic	LexSyn	All Discrete
B-claim	.294	.218	.395	.269
I-claim	.461	.390	.530	.504
B-premise	0	0	.114	0
I-premise	.018	.009	.176	.013
O-Arg	.655	.745	.768	.695
Accuracy	.560	.625	.673	.595
Macro F1	.285	.272	.397	.296

Table 4: Accuracy and F1 scores for Claim and Premise Token Detection on the test set for each group of the discrete features in the CRF model.