Argumentation Mining on Essays at Multi Scales

Hao Wang[1*], **Zhen Huang**[2*], **Yong Dou**[3], **Yu Hong**[4]
[1,2,3] School of Computer Science, National University of Defense Technology, Changsha, Hunan, China
[4] School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu, China
[1,2,3]{hao.wang, huangzhen, yongdou}@nudt.edu.cn, [4]{tianxianer}@gmail.com

Abstract

Argumentation mining on essays is a new challenging task in natural language processing, which aims to identify the types and locations of argumentation components. Recent research mainly models the task as a sequence tagging problem and deal with all the argumentation components at word level. However, this task is not scale-independent. Some types of argumentation components which serve as core opinions on essays or paragraphs, are at essay level or paragraph level. Sequence tagging method conducts reasoning by local context words, and fails to effectively mine these components. To this end, we propose a multi-scale argumentation mining model, where we respectively mine different types of argumentation components at corresponding levels. Besides, an effective coarse-to-fine argumentation fusion mechanism is proposed to further improve the performance. We conduct a serial of experiments on the Persuasive Essay dataset (PE 2.0). Experimental results indicate that our model outperforms existing models on mining all types of argumentation components.

1 Introduction

Argumentation mining (AM) is a challenging task in natural language processing (Lippi and Torroni, 2016). Recent research mainly involves independent sentences (Habernal and Gurevych, 2016; Bar-Haim et al., 2017; Daxenberger et al., 2017; Niven and Kao, 2019; Reimers et al., 2019) and also essays (Levy et al., 2014; Rinott et al., 2015; Rinott et al., 2015; Lippi and Torroni, 2016; Habernal and Gurevych, 2017; Eger et al., 2017; Chernodub et al., 2019; Petasis, 2019). In this paper, we focus on argumentation mining on essays. Argumentation mining on essays aims to identify the types and locations of argumentation components from essay text (Lippi and Torroni, 2016). Typically, there are three argumentation types, namely **major claims** (MC), **claim** (C) and **premises** (P).

Previous research (Levy et al., 2014; Rinott et al., 2015) takes *sentences* as the smallest argumentative unit, and handles this task in a rough way. They firstly split the essay into several sentences, and adopt a sentence classification model to select and reserve sentences which may be promising to contain argumentation components. Then they further identify the exact boundaries of argumentation components in those sentences. These pipeline approaches fail to conduct effective argumentation mining, since they ignore the argumentation structure of the essay and only handle the task at sentence level.

Recent research (Eger et al., 2017; Chernodub et al., 2019) focuses on end-to-end neural models. They boil the task down to a sequence tagging problem, and handle it at *word level* instead of *sentence level*. Typically, neural network is employed as encoder for text representation, and Conditional Random Field (CRF) is employed as decoder to make final prediction. This *word-level* sequence tagging method can simultaneously identify the types and locations of all argumentation components.

However, as shown in Figure 1, it can be observed that different types of argumentation components are at different levels:

• Major claims serve the whole essay as the core opinions. They can be straightly proposed at the beginning of the essay, or summarized in the end. They are at essay level.

Proceedings of the 28th International Conference on Computational Linguistics, pages 5480–5493
Barcelona, Spain (Online), December 8-13, 2020

International tourism is now more common than ever before

The last decade has seen an increasing number of tourists traveling to visit natural wonder sights, ancient heritages and different cultures around the world. While some people might think that this international tourism has negative effects on the destination countries, I would contend that it has contributed to the economic development as well as preserved the culture and environment of the tourist destinations**[MC]**.

Firstly, international tourism promotes many aspects of the destination country's economy in order to serve various demands of tourists**[P]**. Take Cambodia for example, a large number of visitors coming to visit the Angkowat ancient temple need services like restaurants, hotels, souvenir shops and other stores**[P]**. These demands trigger related business in the surrounding settings which in turn create many jobs for local people improve infrastructure and living standard**[P]**. Therefore tourism has clearly improved lives in the tourist country**[C]**.

Secondly $\cdots$

To conclude, as far as I am concerned, international tourism has both triggered economic development and maintained cultural and environment values of the tourist countries**[MC]**. In addition, the authorities should adequately support these sustainable developments.

Figure 1: Argumentation components in a persuasive essay

• Claims serve specific paragraphs as the core statements. They can appear anywhere in a paragraph, either proposed at the beginning, or summarized in the end, and also given in the middle. They are at paragraph level.

• Premises serve as all kinds of evidences to give reasons for major claims and claims. They can be logical statements, survey results, typical examples, public thoughts, expert suggestions, etc. They are at word level.

Moreover, sequence tagging method utilizes classical CRF model to capture sophisticated dependency in a word-by-word way. Such method is thus appropriate to integrate local word-level information, but unsuitable for inference on long-distance text at essay level or paragraph level. To this end, we argue that different types of argumentation components should be handled at different levels.

In this paper, we propose a multi-scale argumentation mining model. In order to mine major claims, we design essay-level argumentation extraction submodule based on multi-span extraction strategy. Besides, to mine claims, we design paragraph-level argumentation extraction submodule based on randomized extraction strategy. As for mining premises, we follow the word-level sequence tagging method. Finally, a coarse-to-fine argumentation fusion mechanism is proposed to further improve the performance.

We carry out a serial of experiments on the Persuasive Essays dataset (PE 2.0) (Stab and Gurevych, 2017). The experimental results indicate that our model can significantly improve the performance as compared to state-of-the-art models, where our model respectively achieves 8.92% absolute improvement on overall performance, 14.89% absolute improvement on mining major claims, and 11.05% absolute improvement on mining claims. Moreover, we compare the performance of (i) multi-span extraction and randomized extraction (ii) argumentation extraction and argumentation tagging, which allow us to validate the effectiveness of our strategies of processing different types of argumentation components at their corresponding levels.

The organization of this paper is as follows. Firstly we give a detailed explanation to our multi-scale argumentation mining model in Section 2. Then in Section 3, we introduce our experiments. The detailed experimental results are displayed and analyzed in Section 4. In Section 5, we give a brief overview of related work about argumentation mining on essays. Finally we draw our conclusion in Section 6.

2 Multi-scale Argumentation Mining Model

An overview of our multi-scale argumentation mining model is shown in Figure 2. For major claims, we design an essay-level argumentation extraction submodule based on multi-span extraction strategy

in Section 2.1, where the whole essay is taken as the input of BERT encoder, and a pointer network is utilized to score each word and thus score all the candidate spans. By these scores and a set of reasonable rules, we rank and filter the candidate spans to select result spans. For claims, we design a paragraph-level argumentation extraction submodule based on randomized extraction strategy in Section 2.2, where each paragraph is respectively taken as the input of BERT encoder to mine result spans, and result spans of each paragraph are gathered as the result spans of the corresponding essay. For premises, we design a word-level argumentation tagging submodule in Section 2.3, where the whole essay is taken as the input of BERT encoder, and CRF is utilized as decoder to obtain the tag sequence with the highest sequence score. Finally, a coarse-to-fine argumentation fusion mechanism in Section 2.4 is utilized to obtain the final results, since there may exist some overlaps on result spans of different argumentation types.

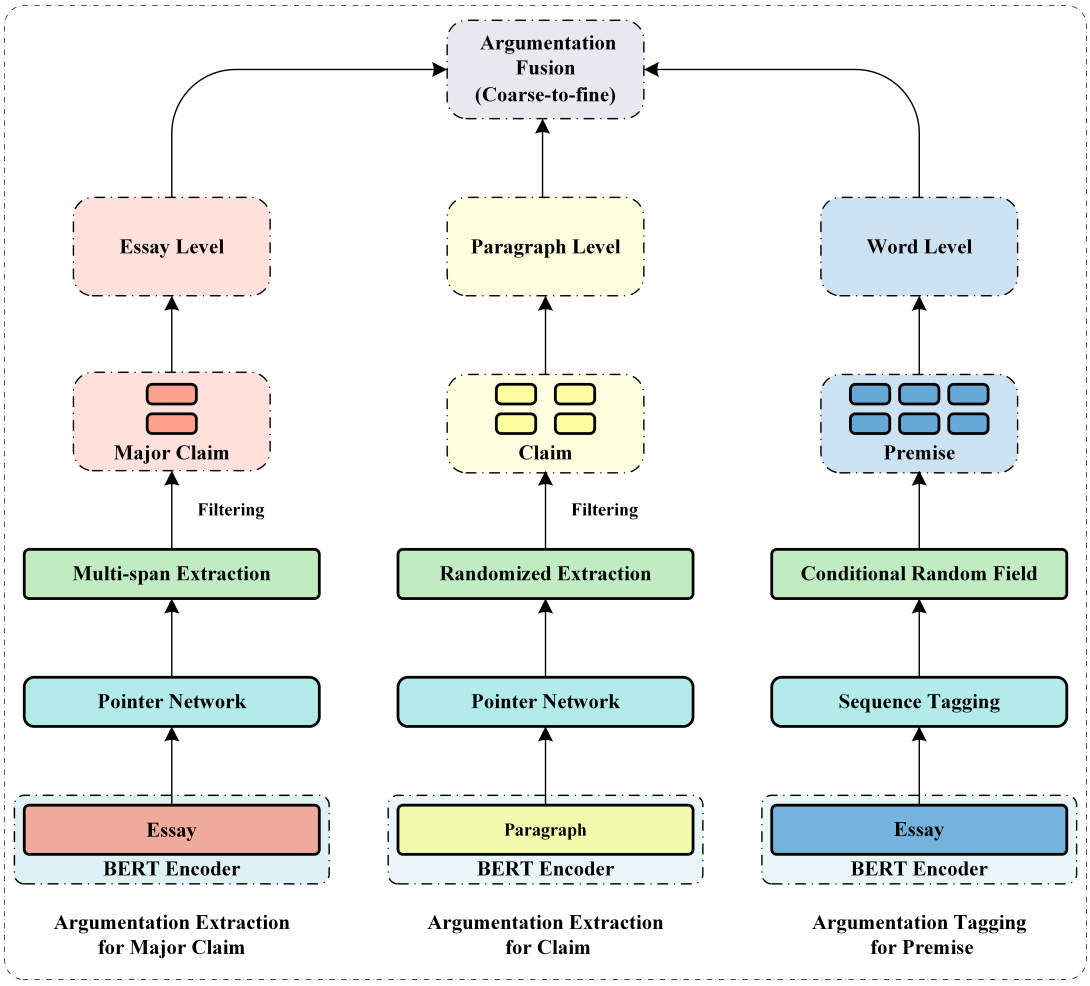

Figure 2: Multi-scale Argumentation Mining Model

2.1 Essay-level Argumentation Extraction for Major Claim

Major claims are at essay level. For each essay, let $E = \{w_1, \cdots, w_{l_{es}}\}$ denotes the essay. To mine major claims, the input sequence is:

$$\mathbb{E} = [CLS]\ E\ [SEP] \tag{1}$$

where l_{es} is the length of the essay. The sequence is encoded with BERT encoder (Devlin et al., 2019):

$$H = \mathrm{BERT}(\mathbb{E}) \tag{2}$$

Through the multi-head self-attention mechanism, BERT can perceive and more heavily weight the attentive words in the essay. This allows the model to capture essay context by multi-layer transformers.

Then inspired by pointer networks (Vinyals et al., 2015), for each word w_i in the essay, its embedding H_i is utilized to score the word through a linear layer:

$$[score_i^s,\ score_i^e] = W_{score} H_i \tag{3}$$

where $score_i^s$ is the start score for the word to be the start of a major claim span, while $score_i^e$ is the end score. Then the cross entropy loss of start position and end position are respectively calculated, and the sum of start loss and end loss is employed as the final loss:

$$\mathcal{L}_I = -\sum_{i=1}^{l_{es}} y_i^s \log(\mathrm{softmax}(score_i^s)) - \sum_{i=1}^{l_{es}} y_i^e \log(\mathrm{softmax}(score_i^e)) \tag{4}$$

where y_i^s is the start label for w_i to be the start of a major claim span, 1 for golden start word while 0 for non-start word, and y_i^e is the end label.

Moreover, as shown in Figure 1, there are some occasions where an essay contains more than one major claim spans. Actually, each essay has at least one major claim span, and two major claim spans in usual, where one is straightly proposed at the beginning, while another is summarized in the end. Hence, we adopt a **multi-span extraction strategy** during training, where all major claims in an essay are admitted. It indicates that start label y^s and end label y^e may be multi-one-hot labels:

$$\sum_{i=1} y_i^s >= 1 \qquad \sum_{i=1} y_i^e >= 1 \tag{5}$$

When prediction, all candidate spans are ranked according to their corresponding probability. The probability for a span starting from w_i and ending at w_j is defined as Equation 6:

$$p_{i,j} = \frac{\exp^{score_i^s} \exp^{score_j^e}}{\sum_{m=1}^{l_{es}} \sum_{n=m}^{l_{es}} \exp^{score_m^s} \exp^{score_n^e}} \tag{6}$$

Then we propose a set of reasonable rules, which are based on common sense, to filter apparently wrong and overlapped candidate spans. The rules are explained in detail in Appendix 1. Finally we reserve top K as result spans for each essay.

2.2 Paragraph-Level Argumentation Extraction for Claim

Claims are at paragraph level. For each essay, firstly we respectively mine claims from each paragraph, and then gather the results for subsequent argumentation fusion on essays. Specifically, for each paragraph, let $P = \{w_1, \cdots, w_{l_{pa}}\}$ denotes the paragraph. The input sequence is:

$$\mathbb{P} = [CLS]\ P\ [SEP] \tag{7}$$

where l_{pa} is the length of the paragraph. The sequence is also encoded with BERT encoder for contextualized embedding:

$$H = \mathrm{BERT}(\mathbb{P}) \tag{8}$$

Then similar to the submodule for major claim in Section 2.1, the start and end score of a word comes from its embedding:

$$[score_i^s,\ score_i^e] = W'_{score} H_i \tag{9}$$

and the sum of the cross entropy loss of start position and end position is adopted as final loss:

$$\mathcal{L}_{II} = -\sum_{i=1}^{l_{pa}} y_i^s \log(\mathrm{softmax}(score_i^s)) - \sum_{i=1}^{l_{pa}} y_i^e \log(\mathrm{softmax}(score_i^e)) \tag{10}$$

Besides, as shown in Figure 1, a paragraph may contain one claim span, or none. Moreover, there are very few occasions where a paragraph contains more than one claim spans. Taking this into account, we

adopt a **randomized extraction strategy**. It means that, if a paragraph contains more than one claim spans, then in each training epoch, only one span is admitted and other spans are ignored. The admitted one is randomly chosen in each epoch. Thus start label y^s and end label y^e may be one-hot labels for paragraphs which have at least one claim span, and full-zero labels for paragraphs which does not contain any claim span:

$$\sum_{i=1} y_i^s <= 1 \qquad \sum_{i=1} y_i^e <= 1 \tag{11}$$

Similarly, during prediction, all candidate spans are ranked according to span probability:

$$p_{i,j} = \frac{\exp^{score_i^s} \exp^{score_j^s}}{\sum_{m=1}^{l_{pa}} \sum_{n=m}^{l_{pa}} \exp^{score_m^s} \exp^{score_n^s}} \tag{12}$$

Then the filtering rules in Appendix 1 are adopted to remove apparently wrong and overlapped candidate spans. Finally we keep top *k* as result spans for each paragraph, and gather them as result spans for the corresponding essay.

2.3 Word-Level Argumentation Tagging for Premise

Premises are at word level. We adopt word-level argumentation tagging through a BERT-CRF sequence tagging model to mine premises. For each essay, let $E = \{w_1, \cdots, w_{l_{es}}\}$ denotes the essay. The input sequence is:

$$\mathbb{E} = [CLS]\ E\ [SEP] \tag{13}$$

and the sequence is also encoded with BERT encoder for contextualized embedding:

$$H = \mathrm{BERT}(\mathbb{E}) \tag{14}$$

Then the embedding of each word is employed to score the word to be different tags through a linear layer:

$$[score_i^1,\ score_i^2,\ \cdots,\ score_i^k] = W_{tag} H_i \tag{15}$$

where *k* is the number of tag types, and $score_i^j$ ($j \in \{1, 2, \cdots, k\}$) is the score of $word_i$ to be marked as tag *j*. In our research, we adopt the same tag configuration as Chernodub et al. (2019), which is a compound of BIO label and argumentation types.

We also adopt a Conditional Random Field (CRF) model (Lample et al., 2016) as decoder. Specifically, for a predicted tag sequence *t*:

$$t = \{t_1,\ t_2,\ t_3,\ \cdots, t_{l_{es}}\}\ \ (t_i \in \{1, 2, \cdots, k\}) \tag{16}$$

t_i is the predicted tag of the word w_i, and the corresponding sequence score is:

$$seqscore_t = \sum_{i=1}^{l_{es}} score_i^{t_i} + \sum_{i=1}^{l_{es}-1} A_{t_i,\, t_{i+1}} \tag{17}$$

where *A* is trained one-step tag transition matrix. The final loss is defined as:

$$\mathcal{L}_{III} = -\sum_{t \in T} y_t \log(\mathrm{softmax}(seqscore_t)) \tag{18}$$

where y_t is tag sequence label, 1 for groundtruth tag sequence while 0 for others, and *T* is a set of all possible tag sequences.

During prediction, the *Viterbi* algorithm is adopted for decoding to obtain tag sequence with the highest sequence score, which will be considered as the submodule prediction.

2.4 Coarse-to-fine Argumentation Fusion

As mentioned above, we have obtained result spans of different argumentation types at corresponding levels respectively. However, the result spans of different argumentation types might be overlapped. Hence we propose a coarse-to-fine method for the fusion of them.

Specifically, let $priority_x$ denotes the priority of argumentation type x, where $x \in \{MC, C, P\}$. We follow the coarse-to-fine principle and set the highest priority for major claim, higher for claim, and the lowest for premise:

$$priority_{MC} > priority_C > priority_P \tag{19}$$

Then for each essay, we keep three sets, which respectively contain the result spans of major claim, claim and premise. For each essay, if a result span from one set is overlapped with another result span from another set according to Algorithm 1 in Appendix 1, then we reserve the span from the set with higher priority, and remove another span from its corresponding set. In this way, all sets will not share any overlapped spans and the fusion procedure is accomplished.

3 Experiments

In this section, at first we introduce the dataset we utilize and show our experiment setup. Then we introduce the evaluation metrics. Finally we list the baselines that we adopt for comparison.

3.1 Dataset

PE 2.0 dataset[1] (2017), which is based on PE 1.0 dataset (2014), is one of the most classical and widely used datasets in argumentation mining on essays. PE 2.0 annotates three kinds of argumentation components, namely **major claim** (MC), **claim** (C) and **premise** (P). Many previous researches (Persing and Ng, 2016; Stab and Gurevych, 2017; Eger et al., 2017; Chernodub et al., 2019; Petasis, 2019; Reimers et al., 2019; Spliethover et al., 2019) have adopted this dataset. Hence we also carry out our experiments on it.

Statistic information on text and argumentation components in PE 2.0 is shown in Table 1. Taking model generalization ability into account, we follow the dataset split[2] with 286 essays as the train set, 80 as the test set, and 36 as the development set.

Essay	**Paragraph**	**Word**
402	2235	148186
Major Claim	**Claim**	**Premise**
751	1506	3832

Table 1: Statistic information on PE 2.0 dataset

3.2 Experiment Setup

We implement our model with TensorFlow 1.14.0 and conduct our experiments on a computation node with a NIVIDIA RTX2080 GPU. In our experiments, pre-trained uncased BERT-base model[3] is adopted as encoder. We utilize BERTAdam optimizer with an initial learning rate of 5e-6, and choose a batch size of 4 to avoid out of memory problem, for BERT is extremely exhausting for memory. We also employ a hyper parameter optimization with dropout probability from $\{0.1, 0.2, 0.3\}$. In each case, we train 20 epochs, and choose model parameters with the best performance on the development set.

3.3 Evaluation Metrics

To accurately evaluate the performance of our model on mining all types of argumentation components, we employ following span-based evaluation metrics. For specific argumentation type, a prediction span

[1] https://www.informatik.tu-darmstadt.de/ukp/research_6

[2] https://github.com/UKPLab/acl2017-neural_end2end_am

[3] https://github.com/google-research/bert

of an essay is regarded as true only if it is exactly matched with a groundtruth span of the essay. We calculate mean precision $\overline{\mathbf{P}}$, mean recall $\overline{\mathbf{R}}$, as well as mean F1 score $\overline{\mathbf{F}}$ of each essay on the test set. Furthermore, we employ macro F score defined in Equation 20 as overall evaluation metric:

$$\mathbf{F}_{macro} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{F}_{MC_i} + \mathbf{F}_{C_i} + \mathbf{F}_{P_i}}{3} \tag{20}$$

where *n* is number of essays on the test set. Besides, as previous research (Eger et al., 2017; Chernodub et al., 2019), we also take the micro F score from Persing and Ng (2016) into account. The detailed definition of this metric is available in Appendix 2.

3.4 Baselines

We adopt TARGER from Chernodub et al. (2019) as a baseline. This model is a BiLSTM-CNN-CRF sequence tagging model, which has similar structure and shows similar performance with the STag$_{BLCC}$ model from Eger et al. (2017). In their experiments, Chernodub et al. (2019) adopted a 70/20/10 train/development/test dataset split. However, they did not report the detailed model performance on different argumentation types and their split is not available. Therefore we rerun their codes[4] with the dataset split which we adopt. Besides, we adopt the BERT-CRF sequence tagging model implemented by ourselves as another baseline.

4 Results

4.1 Overall Performance

Table 2 summarizes the overall performance[5]. Our multi-scale argumentation mining model obtains the best overall performance with the highest $\mathbf{F}_{micro}$ score of 66.93%, and the highest $\mathbf{F}_{macro}$ score of 64.03%. Besides, our model also shows the best performance on mining all types of argumentation components.

Input	Method	$\overline{\mathbf{F}}_{MC}$	$\overline{\mathbf{F}}_{C}$	$\overline{\mathbf{F}}_{P}$	$\mathbf{F}_{micro}$	$\mathbf{F}_{macro}$
PA	TARGER (2019)	51.04	41.71	65.78	61.16	52.84
	BERT-CRF	13.83	46.01	54.78	51.44	38.21
ES	TARGER (2019)	46.04	28.91	66.42	58.81	47.12
	BERT-CRF	51.11	45.27	68.94	62.02	55.11
Multi-scale	Our Model	**66.00**	**57.06**	**69.02**	**66.93**	**64.03**

Table 2: Overall performance ($\overline{\mathbf{F}}_i$ denotes mean F1 socre of argumentation type *i*. PA denotes paragraph. ES denotes essay.)

4.2 Submodule Performance

Experimental results on mining different argumentation types before fusion are summarized in Table 3. Our essay-level argumentation extraction submodule for major claim shows the best performance with the highest F1 score, as well as the highest precision and recall on mining major claims. As we have pointed out, major claims are at essay level. Thus BERT-CRF with essay as input performs better among all the sequence tagging models. However, it still conducts reasoning in a word-by-word way through CRF. As compared to the CRF, pointer network in our submodule can capture long-distance context information on essays. Therefore, the submodule significantly outperforms other word-level sequence tagging models.

Besides, our paragraph-level argumentation extraction submodule for claim obtains the best performance with the highest F1 score on mining claims. The submodule also obtains near the best precision and recall. Claims are at paragraph level. BERT-CRF with paragraph as input shows better performance among all the sequence tagging baselines. As compared to it, our submodule utilizes pointer network to

[4]https://github.com/achernodub/targer

[5]Error analysis and word-based sequence tagging results are respectively shown in Appendix 3 and Appendix 4.

conduct reasoning on paragraphs. Thus, the submodule shows apparent advantages on mining claims. However, its F1 score of 53.54%, though the highest among all models, is relatively low compared to other argumentation types. This might because the submodule ignoring the information from other paragraphs in the identical essay. Nevertheless, it is a challenging trade-off problem from the later ablation studies in Section 4.3.

Moreover, our word-level argumentation tagging submodule for premises has the best performance with the highest F1 score, as well as the highest precision on mining premises. It indicates that the pre-trained language model BERT is also powerful and adaptive to transfer in this task.

Input	Method	$\overline{\mathbf{P}}_{MC}$	$\overline{\mathbf{R}}_{MC}$	$\overline{\mathbf{F}}_{MC}$	$\overline{\mathbf{P}}_{C}$	$\overline{\mathbf{R}}_{C}$	$\overline{\mathbf{F}}_{C}$	$\overline{\mathbf{P}}_{P}$	$\overline{\mathbf{R}}_{P}$	$\overline{\mathbf{F}}_{P}$
PA	TARGER(2019)	58.33	48.75	51.04	43.92	45.54	41.71	68.49	65.37	65.78
	BERT-CRF	16.67	12.71	13.83	**54.78**	43.75	46.01	64.67	49.47	54.78
ES	TARGER(2019)	45.94	50.42	46.04	39.67	25.00	28.91	62.85	**72.05**	66.42
	BERT-CRF	55.52	52.08	51.11	41.46	**54.18**	45.27	76.17	64.11	68.94
*	Our Submodules	**65.00**	**69.17**	**66.00**	54.42	54.09	**53.54**	**76.17**	64.11	**68.94**

Table 3: Results on mining different argumentation types before fusion (PA denotes paragraph. ES denotes essay. * denotes our submodules take different inputs on mining different argumentation types.)

Fusion	$\overline{\mathbf{P}}_{MC}$	$\overline{\mathbf{R}}_{MC}$	$\overline{\mathbf{F}}_{MC}$	$\Delta_{\overline{\mathbf{F}}_{MC}}$	$\overline{\mathbf{P}}_{C}$	$\overline{\mathbf{R}}_{C}$	$\overline{\mathbf{F}}_{C}$	$\Delta_{\overline{\mathbf{F}}_{C}}$
before	65.00	69.17	66.00	-	54.42	54.09	53.54	3.52
after	65.00	69.17	66.00		66.13	51.65	57.06	
Fusion	$\overline{\mathbf{P}}_{P}$	$\overline{\mathbf{R}}_{P}$	$\overline{\mathbf{F}}_{P}$	$\Delta_{\overline{\mathbf{F}}_{P}}$	$\mathbf{F}_{micro}$	$\Delta_{\mathbf{F}_{micro}}$	$\mathbf{F}_{macro}$	$\Delta_{\mathbf{F}_{macro}}$
before	76.17	64.11	68.94	0.08	65.75	1.18	62.83	1.20
after	79.19	62.27	69.02		66.93		64.03	

Table 4: Efficiency of coarse-to-fine argumentation fusion mechanism

We also verify the efficiency of our coarse-to-fine argumentation fusion mechanism in Table 4. For major claim, the performance remains the same after fusion since we set the highest priority for major claim, and do not remove any such span. For claim, the performance is apparently improved with higher F1 score, which comes from the significant increase of precision and relatively slight decrease of recall. For premise, the performance also gets slightly promoted. Besides, the overall performance also gets promoted after fusion with respectively 1.18% and 1.20% absolute increase of $\mathbf{F}_{micro}$ and $\mathbf{F}_{macro}$ score. All these improvements indicate that our coarse-to-fine argumentation fusion mechanism is effective.

4.3 Ablation Study

4.3.1 Multi-span Extraction or Randomized Extraction

We respectively mine major claims and claims at different levels with different extraction strategy[6]. The results are summarized in Table 5. For major claims, under the same extraction strategy, extractions on essays significantly outperform extractions on paragraphs. However, the situation is exactly opposite for claims. Under the same extraction strategy, extractions on paragraphs obtain better performances. It shows that different type argumentations components should be handled at corresponding level.

Moreover, no matter what argumentation type, on paragraphs, randomized extractions outperform multi-span extractions. And on essays, multi-span extractions are better than randomized extractions. It may indicate that multi-span strategies is appropriate at essay-level extractions, and randomized strategies is appropriate at paragraph-level extractions. Actually, in usual, an essay contains more than one claim spans, where multi-span extraction is more appropriate. However, on most occasions, a paragraph has at most one claim span, or does not have any span, where randomized extraction is more appropriate.

[6]We also try to mine major claims and claims under the Machine Reading Comprehension (MRC) framework with essay title as query and guide information. The results are shown in Appendix 5.

The situation is similar for major claims. Therefore, the results show the effectiveness of our strategies chosen for different types of argumentation components.

Type	Input	ST	$\overline{P}$	$\overline{R}$	$\overline{F}$
MC	PA	ME	29.98	68.54	41.33
		RE	30.61	70.21	42.20
	ES	ME	65.00	69.17	66.00
		RE	63.12	65.83	63.58
C	PA	ME	48.92	49.12	48.40
		RE	54.42	54.09	53.54
	ES	ME	45.00	48.06	45.67
		RE	39.37	42.62	40.22

Table 5: Results of mining major claim and claim with different extraction strategy. (ST denotes extraction strategy. ME denotes multi-span extraction. RE denotes randomized extraction.)

Method	Input	ST	$\overline{P}$	$\overline{R}$	$\overline{F}$
AE	PA	ME	50.18	70.58	57.45
		RE	49.15	65.83	55.16
	ES	ME	60.88	61.83	59.25
		RE	53.63	54.44	52.22
AT	ES	-	**76.17**	**64.11**	**68.94**

Table 6: Results of argumentation extraction and argumentation tagging on mining premise (AE denotes argumentation extraction. AT denotes argumentation tagging, where the results come from our word-level argumentation tagging submodule before fusion. ST denotes extraction strategy. ME denotes multi-span extraction. RE denotes randomized extraction.)

4.3.2 Argumentation Extraction or Argumentation Tagging

We also try to mine premises with argumentation extraction method. The results are compared in Table 6. Our word-level argumentation tagging submodule for premise obtains the best performance with the highest F1 score as well as the highest precision and recall. This just indicates that premises are at word level, and argumentation tagging is more appropriate than argumentation extraction on mining them.

5 Related Work

Stab et al. (2014) modeled AM on essays as a sentence-level feature-based classification task, where each sentence is respectively classified through a set of linguistical features. Stab et al.(2017) firstly proposed a sequence tagging model to distinguish argumentation components and non-argumentation components, and employed a joint ILP (Integer Linear Programming) model to identify the types of argumentation components. However, they reported performance of different subtasks without overall performance. Potash et al. (2017) utilized pointer network to identify the types of argumentation components on the assumption that all argumentation components have already been identified, which means the exact boundaries of all the argumentation components are already available. Eger et al. (2017) further proposed a new end-to-end sequence tagging model, which firstly employs compound labels of BIO and argumentation types, and simultaneously identifies the types and exact locations of different argumentation components. Chernodub et al. (2019) tried to build application interface, which is called *TARGER* and is a BiLSTM-CNN-CRF sequence tagging model, for convenient argumentation mining on essays. Besides, latest research (Petasis, 2019; Spliethover et al., 2019) also aims to distinguish argumentation components from non-argumentation components with text segmentation based on sequence tagging models. Other work (Peldszus et al., 2015; Peldszus et al., 2016; Skeppstedt et al., 2018) focuses on arg-microtext corpus (Peldszus et al., 2015), which contains 112 independent short texts, where each can be considered as one paragraph and contains about 5 argumentation components on average.

6 Conclusion

We propose a multi-scale argumentation mining model for argumentation mining on essays. Our model respectively mines different types of argumentation components at corresponding levels. Moreover, a coarse-to-fine argumentation fusion mechanism is adopted to further improve the results. The experimental results on PE 2.0 dataset indicate that our model obtains the state-of-the-art performance, where the model obtains significantly improved performance on mining major claims and claims. The results reveal the importance of argumentation mining at different levels on different argumentation types. In the future, we will try to mine different argumentation types with multi-task learning method.

Acknowledgements

We would like to sincerely thank anonymous reviewers for their careful reviews, thoughtful comments and helpful suggestions. We also sincerely thank the committees for their effort to provide us such a good platform to show our study results. This work was supported by the Key Program of National Natural Science Foundation (No. 61732018).

References

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: state of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):10:1–10:25.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the 27th International Joint Conferences on Artificial Intelligence*, pages 5427-5433, Stockholm, Sweden.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual meeting of the Association for Computational Linguistics*, pages 1589–1599, Berlin, Germany.

Roy Bar-Haim and Indrajit Bhattacharya and Francesco Dinuzzo and Amrita Saha and Noam Slonim. 2017. Stance classification of context-dependent claim. In *Proceedings of the 54th Annual meeting of the Association for Computational Linguistics*, pages 251–261, Vancouver, Canada.

Johannes Daxenberger and Steffen Eger and Ivan Habernal and 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, Copenhagen, Denmark.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual meeting of the Association for Computational Linguistics*, pages 4568–4644, Florence, Italy.

Nils Reimers and Benjamin Schiller and Tilman Beck and Johannes Daxenberger and Christian Stab and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy.

Ran Levy and Yonatan Bilu and Daniel Hershcovich and Ehud Aharoni and Noam Slonim. 2014. Context Dependent Claim Detection. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland.

Ruty Rinott and Lena Dankin and Carlos'Alzate and Mitesh M. Khapra. 2015. Argument mining from speech: detecting claims in political debates. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal.

Ruty Rinott and Lena Dankin and Carlos Alzate and Mitesh M. Khapra and Ehud Aharoni and Noam Slonim. 2015. Show Me Your Evidence – an Automatic Method for Context Dependent Evidence Detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440-450, Lisbon, Portugal.

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

Ivan Habernal and Iryna Gurevych. 2016. Argumentation mining in user-generated web discourse. *Computational Linguistics*,43(1):125-179.

Steffen Eger and 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*, pages 11–22, Vancouver, Canada.

Artem Chernodub and Oleksiy Oliynyk and Philipp Heidenreich and Alexander Bondarenko and Matthias Hagen and Chris Biemann and Alexander Panchenko. 2019. TARGER: neural argument mining at your fingertips. In *Proceedings of the 59th Annual meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195-200, Florence, Italy.

Georgios Petasis. 2019. Segmentation of argumentation texts with contextualized word representations. In *Proceedings of the 6th Workshop on Argument Mining*, pages 1-10, Florence, Italy.

Jacob Devlin and Ming-Wei Chang and 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*, pages 4171–4186, Minneapolis, MN, USA.

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

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

Issac 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, San Diego California, USA.

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

Maximilian Spliethover and Jonas Klaff and Hendrik Heuer. 2019. Is it worth the attention? A comparative evaluation of attention layers for argument unit segmentation. In *Proceedings of the 6th Workshop on Argument Mining*, pages 74-82, Florence, Italy.

Minghao Hu and Yuxing Peng and Zhen Huang and Dongsheng Li. 2019. Retrieve, read, rerank: towards end-to-end multi-document reading comprehension. In *Proceedings of the 57th Annual meeting of the Association for Computational Linguistics*, pages 2285-2295, Florence, Italy.

Oriol Vinyals and Meire Fortunato and Navdeep Jaitly. 2015. Pointer Networks. In *Proceedings of 2015 Annual Conference on Neural Information Processing System*, pages 2285-2295, Montreal, Quebec, Canada.

Guillaume Lample and Miguel Ballesteros and Sandeep Subramanian and Kazuya Kawakami and Chris Dyer 2015. Neural architectures for named entity recognition. In *Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260-270, San Diego California, USA.

Xiaoya Li and Jingrong Feng and Yuxian Meng and Qinghong Han and Fei Wu and Jiwei Li 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual meeting of the Association for Computational Linguistics*, pages 5849-5859, Online.

Wei Wu and Fei Wang and Arianna Yuan and Fei Wu and Jiwei Li 2020. Coreference Resolution as Query-based Span Prediction. In *Proceedings of the 58th Annual meeting of the Association for Computational Linguistics*, pages 6953-6963, Online.

Peter Potash and 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, Copenhagen, Denmark.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation*, pages 801-816, Lisbon, Portuguese.

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.

Andreas Peldszus and Manfred Stede. 2016. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the 3rd Workshop on Argument Mining*, pages 103-112, Berlin, Germany.

Maria Skeppstedt and Andreas Peldszus and Manfred Stede. 2018. More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 155-163, Brussels, Belgium.

Appendix

A.1 Filtering Rules

Major claims and **claims** respectively serve as the core opinions and subopinions in an essay. To express an opinion, a sentence must consist of at least 3 parts, namely a subject, a predicate verb, and an object. Hence, we propose a set of reasonable filtering rules to remove apparently wrong candidate spans. Firstly, candidate spans with less than 3 words, e.g. *international tourism*, will be removed. Besides, we employ Stanford CoreNLP[1] to operate part-of-speech tagging on each candidate span. Candidate spans without at least one predicate verb, e.g. *cultural and environment values of the tourist countries*, will be also removed.

Furthermore, there are may exist some overlapped candidate spans, which have the same rough locations but different exact boundaries. To handle this issue, inspired by Hu et al.(2019), we judge overlapped candidate spans according to Algorithm 1. Then we reserve the one with the highest span probability, and remove the others.

Algorithm 1 IsOverlap

Input: $span_i$, $span_j$
Parameter: *threshold*

$words_i$ = {*word* for *word* in $span_i$}
$words_j$ = {*word* for *word* in $span_j$}
common_words = $words_i$ & $words_j$
minimun_length = minimun(length($words_i$), length($words_j$))
cover_rate = length(*common_words*) / *minimum_length*
if *cover_rate* >*threshold* **then**
    return True
else
    return False
end if

A.2 Definition of $\mathbf{F}_{micro}$

The micro F score $\mathbf{F}_{micro}$ from Persing and Ng(2016) is a span-based metric and is defined as:

$$AC^p = \sum_i \sum_{j=1}^{n} AC_{ij}^p \quad AC^g = \sum_i \sum_{j=1}^{n} AC_{ij}^g \quad AC^t = \sum_i \sum_{j=1}^{n} AC_{ij}^t \tag{A.1}$$

$$\mathbf{F}_{micro} = \frac{2AC^t}{AC^p + AC^g} \tag{A.2}$$

where *i* denotes the type of argumentation components and belongs to {major claim, claim, premise}, while *n* is number of essays on the test set. Besides, AC_{ij}^p, AC_{ij}^g and AC_{ij}^t respectively denote predicted, groundtruth and truly-predicted spans of argumentation type *i* in essay *j*.

A.3 Error Analysis

We adopt statistical methods to analyze the errors of our model. Totally, for single argumentation type, there exist 5 kinds of errors, which are:

- **Boundary** denotes the error on exact boundary for the argumentation component.
- **Type-in** denotes that the other type is falsely predicted as this type.
- **Type-out** denotes that this type is falsely predicted as other ones.
- **None-in** denotes that non-argumentation component is falsely predicted as this type.
- **None-out** denotes that this type is falsely predicted as non-argumentation component.

[1] https://stanfordnlp.github.io/CoreNLP/

The results are displayed in Figure A.1. Different argumentation types show diverse error modes. For all the types, None-out is the dominate error. This may because of the few shot of argumentation components as compared to the non-argumentation ones. For major claims, None-out, Type-in, and None-in errors are serious. It may be a bit difficult for the model to distinguish major claims from non-argumentation components. Claims come with pretty critical None-out, Type-in, and Type-out errors. This may indicate the model tends to mistake claims for non-argumentation components, as well as confuse claims with other argumentation types. As for premises, None-out, Boundary, and Type-out errors take dominant positions. The model may get into trouble in identifying exact boundaries of premises and distinguishing premises from non-argumentation components. Besides, the model also tends to mistake premises for other argumentation types.

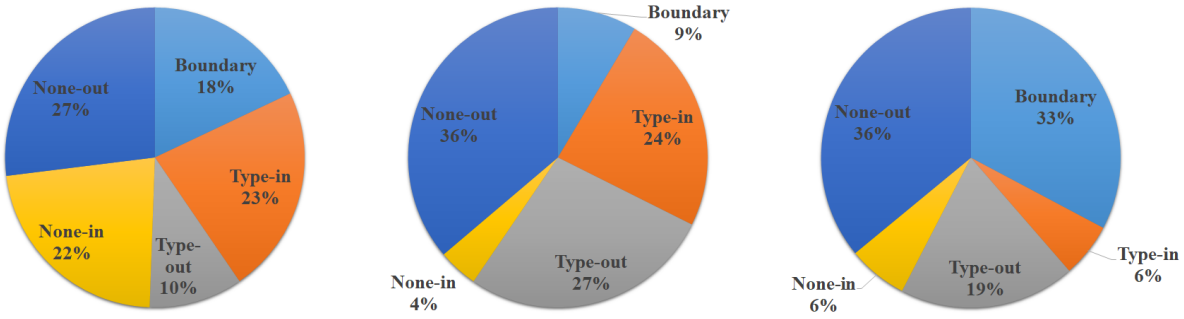

Figure A.1: Error statistics of different argumentation types on the test set (From left to right are respectively error statistics of major claim, claim, and premise.)

A.4 Word-based Sequence Tagging Results

Word-based sequence tagging results of different models are compared in Table A.1. Among all these models, BERT-CRF with essays as input shows the best word-based performance on all tag types. However, even this model, the F1 scores of major claim and claim are still low, where the F1 scores of B-MC and I-MC are both less than 70%, while the F1 scores of B-C and I-C are both less than 60%. Moreover, actually, for major claim, the minimum of the F1 scores of B-MC and I-MC can be considered as the upper boundary of corresponding span-based F1 score. The situation is similar for claim. That is to say, for these sequence tagging models, span-based F1 scores of major claim and claim will be respectively less than 64.58% and 58.51%. Therefore, sequence tagging models show extremely limited performance on mining major claims and claims.

input	method	metric	O	B-MC	I-MC	B-C	I-C	B-P	I-P
PA	TARGER	P	82.83	70.59	71.32	46.47	46.18	74.78	84.71
		R	85.17	54.90	59.87	47.70	52.87	72.56	81.64
		F1	83.98	61.76	65.09	47.08	49.30	73.65	83.15
	BERT-CRF	P	86.651	45.10	51.77	59.04	47.13	83.46	90.51
		R	87.36	15.03	80.57	48.36	64.67	66.13	74.27
		F1	87.00	22.55	63.03	53.17	54.52	73.79	81.59
ES	TARGER	P	95.87	54.12	51.32	41.80	42.09	69.32	77.36
		R	75.97	60.13	69.91	25.99	27.81	79.60	92.65
		F1	84.77	56.97	59.19	32.05	33.49	74.11	84.32
	BERT-CRF	P	92.81	68.89	60.44	52.08	55.14	85.39	88.99
		R	84.61	60.78	79.76	70.07	62.33	74.41	86.96
		F1	**88.52**	**64.58**	**68.77**	**59.75**	**58.51**	**79.52**	**87.96**

Table A.1: Word-based sequence tagging result (PA denotes paragraph. ES denotes essay.)

A.5 Machine Reading Comprehension Framework

Inspired by Li et al. (2020) and Wu et al. (2020), we try to handle the task under the Machine Reading Comprehension (MRC) framework to further improve the performance on mining major claims and claims. As shown in Figure 1 in our paper, the title of an essay is a condensed summary of the essay, which explicitly points out the topic and even directly proposes the core opinion. Hence, we adopt essay title as query and guide information.

We respectively employ new MRC inputs for our submodules in Section 2.1 and Section 2.2 to mining major claims and claims. More specifically, to mine major claims, for each essay, let $T = \{w_1, w_2, \cdots, w_{l_t}\}$ denotes the title, and $E = \{w_1, w_2, \cdots, w_{l_{es}}\}$ denotes the essay. We concatenate the title and the essay text as MRC input:

$$\mathbb{C}_{T \oplus E} = [CLS]\ T\ [SEP]\ E\ [SEP] \tag{A.3}$$

And the concatenation is encoded with BERT encoder:

$$H = \mathrm{BERT}(\mathbb{C}_{T \oplus E}) \tag{A.4}$$

Similarly, to mine claims, for each paragraph, let $T = \{w_1, w_2, \cdots, w_{l_t}\}$ denotes the essay title, and $P = \{w_1, w_2, \cdots, w_{l_{pa}}\}$ denotes the paragraph. These two are also concatenated as MRC input:

$$\mathbb{C}_{T \oplus P} = [CLS]\ T\ [SEP]\ P\ [SEP] \tag{A.5}$$

The concatenation is also encoded with BERT encoder:

$$H = \mathrm{BERT}(\mathbb{C}_{T \oplus P}) \tag{A.6}$$

Then the subsequent argumentation extractions remain the same.

The results are compared in Table A.2. MRC framework with essay title as query leads to worse performance. Actually, essay titles are diverse. They can be a statement, e.g. *International tourism is now more common than ever before*, a question, e.g. *Can technology alone solve the world's environmental problems*?, or a phrase, e.g. *Living and studying overseas*. It may be pretty difficult for the model to understand the role of the essay title as query. The essay title query will act as disturbing factor rather than guide information for argumentation mining. Hence, MRC framework with essay title as query fails to show promoted performance.

Type	Title Query	$\overline{\mathbf{P}}$	$\overline{\mathbf{R}}$	$\overline{\mathbf{F}}$
MC	+	63.12	65.83	63.58
	−	**65.00**	**69.17**	**66.00**
C	+	50.79	49.84	49.69
	−	**54.42**	**54.09**	**53.54**

Table A.2: Results of MRC framework (+) and our submodules (−)