Have my arguments been replied to? Argument Pair Extraction as Machine Reading Comprehension

Jianzhu Bao[1,2*], **Jingyi Sun**[1,2*], **Qinglin Zhu**[1,2], **Ruifeng Xu**[1,3†]
[1]Harbin Institute of Technology (Shenzhen), China
[2]Joint Lab of China Merchants Securities and HITSZ
[3]Peng Cheng Laboratory, Shenzhen, China
{jianzhubao, sunjingyihit}@gmail.com
zhuqinglin@stu.hit.edu.cn, xuruifeng@hit.edu.cn

Abstract

Argument pair extraction (APE) aims to automatically mine argument pairs from two interrelated argumentative documents. Existing studies typically identify argument pairs indirectly by predicting sentence-level relations between two documents, neglecting the modeling of the holistic argument-level interactions. Towards this issue, we propose to address APE via a machine reading comprehension (MRC) framework with two phases. The first phase employs an argument mining (AM) query to identify all arguments in two documents. The second phase considers each identified argument as an APE query to extract its paired arguments from another document, allowing to better capture the argument-level interactions. Also, this framework enables these two phases to be jointly trained in a single MRC model, thereby maximizing the mutual benefits of them. Experimental results demonstrate that our approach achieves the best performance, outperforming the state-of-the-art method by 7.11% in F_1 score.

1 Introduction

As a salient part of argument mining (AM), the analysis of dialogical argumentation has received increasing research attention (Morio and Fujita, 2018; Chakrabarty et al., 2019; Ji et al., 2021; Cheng et al., 2021; Yuan et al., 2021). Argument pair extraction (APE), proposed by Cheng et al. (2020), is a new task within this field that focuses on extracting interactive argument pairs from two interrelated documents (e.g., peer reviewer and rebuttal). Figure 1 presents an example of APE where two interrelated documents are segmented into arguments and non-arguments at sentence level. Two arguments from different documents that discuss the same issues are regarded as an argument pair.
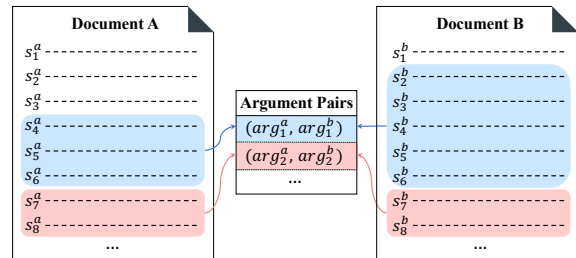

Figure 1: A simplified example of APE task, where each dashed line in the two documents denotes a sentence. s_j^i is the j-th sentence in document i, and arg_j^i is an argument in the j-th argument pair from document i. Sentences without colors indicate non-arguments, while sentences covered by colors can form arguments. Two arguments with the same color are regarded as an argument pair.

Previous works (Cheng et al., 2020, 2021) commonly address APE by decomposing it into two sentence-level subtasks, i.e., a sequence labeling task and a sentence relation classification task. These methods identify arguments by sentence-level sequence labeling and determine whether two sentences belong to the same argument pair by sentence relation classification. Afterwards, the argument pairs are inferred indirectly by certain rules combining the results of the two subtasks. However, such a paradigm only considers sentence-level relations, while the holistic argument-level relations can not be well modeled.

In this paper, we argue that APE can be considered as a multi-turn machine reading comprehension (MRC) task with two phases, i.e., an AM phase and an APE phase. Specifically, in the first turn, a special AM query is employed to identify all the arguments in the first document (AM phase). Afterwards, in each subsequent turn, every identified argument is treated as an APE query to extract its paired arguments from the second document (APE phase). Similarly, this process can also be performed in another direction, that is, using the

*Equal Contribution
†Corresponding Author

arguments identified in the second document as queries to extract the paired arguments from the first document. We train these two phases jointly in a single MRC model, allowing them to benefit each other. By considering arguments as queries, our proposed MRC framework can better capture the interactions between each query argument and the queried document, thus extracting the argument pairs at the argument level. In addition, considering the long length of the documents, we utilize Longformer (Beltagy et al., 2020) to model longer contexts.

We evaluate our method on the large benchmark dataset (Cheng et al., 2020). Results show that our proposed method significantly outperforms the current state-of-the-art method by 7.11% in F_1 score.

2 Related Work

2.1 Argument Mining

Argument mining aims to analyze the structure of argumentation, and it contains various subtasks, such as argument component identification (Moens et al., 2007; Goudas et al., 2015; Ajjour et al., 2017; Jo et al., 2019), argument relation prediction (Nguyen and Litman, 2016; Cocarascu et al., 2020; Jo et al., 2021), argumentation structure parsing (Stab and Gurevych, 2017; Kuribayashi et al., 2019; Morio et al., 2020; Bao et al., 2021), argumentation strategy analysis (Khatib et al., 2018; Morio et al., 2019), etc.

Most previous works mainly focus on monological argumentation, while dialogical argumentation (Morio and Fujita, 2018; Chakrabarty et al., 2019) is relatively less emphasized. Recently, the analysis of dialogical argumentation has attracted increasing attention in the field of argument mining. Cheng et al. (2020) propose the APE task which involves identifying arguments and extracting argument pairs in peer review and rebuttal. Ji et al. (2021) identify interactive argument pairs in online debate forums based on the discrete variational autoencoders. Cheng et al. (2021) address the APE task based on a table-filling approach. Yuan et al. (2021) construct a dialogical argumentation knowledge graph for identifying argument pairs.

2.2 Machine Reading Comprehension

Machine reading comprehension (MRC) aims to extract answer spans from a passage according to a given query (Seo et al., 2017; Chen et al., 2017; Devlin et al., 2019; Wen et al., 2021). Formulating NLP tasks as MRC tasks has been a rising trend in recent years, such as dependency parsing (Gan et al., 2021), relation extraction (Levy et al., 2017), named entity recognition (Li et al., 2020), sentiment analysis (Chen et al., 2021; Mao et al., 2021). Unlike previous studies above, we employ a MRC framework to analyze the complex argumentative relations between two documents with excessively long length.

3 Methodology

3.1 Task Formulation

We assume that two interrelated documents $D_a = (s_1^a, s_2^a, ..., s_{n^a}^a)$ and $D_b = (s_1^b, s_2^b, ..., s_{n^b}^b)$ are given, where s_j^i denotes the j-th sentence in document i. We need to extract the collection of argument pairs $P = \{(arg_i^a, arg_i^b)\}_{i=1}^{|P|}$, where arg_i^a and arg_i^b respectively represent the arguments in document D_a and D_b, and they compose the i-th argument pair. Note that each argument consists of one or more consecutive sentences. For example, $arg_i^a = (s_{start}^{a,i}, s_{start+1}^{a,i}, ..., s_{end}^{a,i})$ where $start$ and end denote the start and end sentence index.

To frame APE as a multi-turn MRC task, two types of queries are constructed, i.e., the argument mining (AM) query and the argument pair extraction (APE) query. Intuitively, we could consider the process of extracting argument pairs from the perspective of two directions, i.e., $D_a \rightarrow D_b$ and $D_b \rightarrow D_a$. For the $D_a \rightarrow D_b$ direction, we first construct an AM query using a special token whose corresponding answers are all the arguments in document D_a. After recognizing all arguments through the AM query, each recognized argument is considered as an APE query whose corresponding answers are its paired arguments in document D_b. Similarly, for the $D_b \rightarrow D_a$ direction, we first query document D_b with the AM query, and then generate the APE queries for document D_a. Finally, the argument pairs can be derived by fusing the answer results of all APE queries.

3.2 MRC Framework

3.2.1 Encoder

Since APE is a document-level task with excessively long text, we adopt Longformer to capture contextual information over longer distances. For brevity, we only describe the MRC process in the $D_a \rightarrow D_b$ direction below, and the $D_b \rightarrow D_a$ direction can be performed similarly.

Formally, we use a special token "[AM]" to represent the AM query q^{am}, which aims to identify all the arguments $A^a = \{arg_k^a\}_{k=1}^{|A^a|}$ in document D_a where arg_k^a indicates the k-th argument in D_a. Then, each identified argument arg_k^a is considered as an APE query $q_k^{a,ape}$, i.e., $q_k^{a,ape} = arg_k^a = (s_{start}^{a,k}, ..., s_{end}^{a,k})$. Note that we use gold arguments as APE queries during training.

With these queries, we first concatenate the AM query q^{am} and the document D_a as an input sequence for AM:

$$I^{am} = ([\text{s}], q^{am}, [/\text{s}], [\text{s}], s_1^a, s_2^a, ..., s_{n^a}^a, [/\text{s}]) \quad (1)$$

Also, we concatenate each APE query $q_k^{a,ape}$ and the document D_b to obtain multiple input sequences for APE:

$$I_k^{ape} = ([\text{s}], q_k^{a,ape}, [/\text{s}], [\text{s}], s_1^b, s_2^b, ..., s_{n^b}^b, [/\text{s}]) \quad (2)$$

where [s] and [/s] are special tokens of Longformer.

Subsequently, for each sequence above, we feed it into Longformer to get the hidden representation of each token in the input document. Specifically, to enable Longformer to better learn argument-specific representations, we add global attention to the tokens of the query. Afterwards, we derive the hidden representation of each sentence through mean pooling on token representations in this sentence. Further, to better model the long-term dependency among sentences, the hidden representations of sentences are fed into LSTM to derive the contextual sentence representation matrix $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n)$.

3.2.2 Answer Span Prediction

For each turn, one or more answer spans will be extracted as arguments. Note that, in each direction, the first turn aims to extract all arguments, while the following turns aim to extract arguments that can form pairs with the query argument.

Specifically, inspired by Li et al. (2020), we fed $\mathbf{H}$ into two binary classifiers to predict the start and end sentence positions of arguments. After obtaining all start and end positions, we further employ another binary classifier to determine whether each start and end position pair (matched by Cartesian product) forms an answer span. Note that the input of this span classifier is the concatenation of the start and end sentence representations from $\mathbf{H}$.

3.2.3 Training

During training, the three classifiers described in Section 3.2.2 yield three cross-entropy losses, i.e., a start loss, an end loss, and a span loss. We simply sum these losses up as the training objective of our model. In addition, the AM phrase and the APE phrase are trained jointly in a single MRC model.

3.2.4 Inference

During inference, the $D_a \rightarrow D_b$ direction uses the trained MRC model to first identify all the arguments in D_a by the AM query and then extract all the argument pairs in D_b by the APE queries. Similarly, the $D_b \rightarrow D_a$ direction can be performed in the same manner by simply exchanging the order of D_a and D_b. Each APE query in both directions yields one or more argument pairs, where each argument pair contains the query argument and one extracted argument. We simply merge all argument pairs extracted by all APE queries into a union set to obtain the final inference results.

4 Experiments

4.1 Experimental setup

4.1.1 Dataset

Our experiments are conducted on the large APE benchmark dataset, namely the Review-Rebuttal (RR) dataset (Cheng et al., 2020), which contains 4,764 pairs of review-rebuttal passages of ICLR. Following the setup of (Cheng et al., 2021), we also evaluate our method on two versions of the train/dev/test (8:1:1) split, i.e., RR-Passage-v1 and RR-Submission-v2. Note that in our method, we view review passage and rebuttal passage as document D_a and document D_b, respectively.

4.1.2 Implementation Details

We adopt Longformer-base-4096 [1] as base encoder, and we use sliding window attention with the window size of 512. We train our model 6 epochs with a batch size of 4. AdamW (Kingma and Ba, 2015) is used as the optimizer, and the learning rates for Longformer and other layers are 1e-5 and 1e-3.[2]

The evaluation metrics contain two aspects, namely AM and APE. Different from (Cheng et al., 2021, 2020), sentence pairing is not included as a metric because we extract argument pairs directly.

---

[1] https://huggingface.co/allenai/longformer-base-4096
[2] Our source code is available at https://github.com/HLT-HITSZ/MRC_APE

Data	Methods	Argument Mining			Argument Pair Extraction		
		Pre.	**Rec.**	**F₁**	**Pre.**	**Rec.**	**F₁**
RR-Submission-v2	PL-H-LSTM-CRF	67.02	68.49	67.75	19.74	19.13	19.43
	MT-H-LSTM-CRF	70.74	69.46	70.09	27.24	26.00	26.61
	MLMC	69.53	**73.27**	71.35	37.15	29.38	32.81
	MRC-APE-Bert	**73.36**	68.35	70.77	42.26	34.06	37.72
	MRC-APE-Sep.	72.45	71.58	72.01	41.09	36.99	38.93
	MRC-APE (Ours)	71.83	73.05	**72.43**	41.83	**38.17**	**39.92**
RR-Passage-v1	PL-H-LSTM-CRF	73.10	67.65	70.27	21.24	19.30	20.22
	MT-H-LSTM-CRF	71.85	71.01	71.43	30.08	29.55	29.81
	MLMC	66.79	**72.17**	69.38	**40.27**	29.53	34.07
	MRC-APE-Bert	66.81	69.84	68.29	34.70	35.53	35.11
	MRC-APE-Sep.	75.27	67.90	71.39	36.63	40.05	38.26
	MRC-APE (Ours)	**76.39**	70.62	**73.39**	37.70	**44.00**	**40.61**

Table 1: Main results on RR-Submission-v2 and RR-Passage-v1 (%). The best scores are in bold.

We select the best parameters based on the performance (i.e., average F_1 scores of AM and APE) on the dev set. All scores are averaged across 5 distinct trials using different random seeds.

4.1.3 Baselines

We compare our model with several baselines. **PL-H-LSTM-CRF** (Cheng et al., 2020) independently trains an argument mining task and a sentence pairing task, while **MT-H-LSTM-CRF** (Cheng et al., 2020) trains two subtasks in a multi-task framework. **MLMC** (Cheng et al., 2021) is an attention-guided model based on a table-filling approach, which is the current state-of-the-art method.

Furthermore, we implement two additional baselines. For a fair comparison with MLMC, **MRC-APE-Bert** replaces Longformer with Bert, where documents with excessively long length are splited into several segments. Instead of jointly training AM and APE phases, **MRC-APE-Sep.** trains the two phases separately.

4.2 Results and Analysis

4.2.1 Main Results

As shown in Table 1, our model achieves the best performance on both versions of the RR dataset. Concretely, on RR-Submission-v2, our model significantly outperforms the current state-of-the-art model MLMC by at least 7.11% in APE F_1 score. On RR-Passage-v1, our model obtains at least a 6.54% higher APE F_1 score than the MLMC. Also, our model achieves the best performance on AM. Furthermore, without applying Longformer as the base encoder, MRC-APE-Bert still outperforms MLMC in APE F_1 score, demonstrating that our improvement is not only brought by Longformer. However, for the AM task, MAC-APE-Bert

Method	APE			
	Pre.	**Rec.**	**F₁**	$\Delta(\mathbf{F_1})$
MRC-APE (Ours)	41.83	**38.17**	**39.92**	-
w/o $D_b \rightarrow D_a$	**49.47**	31.33	38.36	1.56
w/o $D_a \rightarrow D_b$	46.68	26.02	33.41	6.51
w/o LSTM	44.98	34.51	39.06	0.86
w/o GA	38.20	30.66	34.02	5.90

Table 2: The results of ablation experiments on RR-Submission-v2 (%). The best scores are in bold. *w/o* GA indicates that the global attention is not included in Longformer.

achieves slightly lower F_1 score than MLMC. The reason may be that, in MLMC, the predictions of the AM task are influenced by the APE task through a complex attention interaction mechanism. However, our model does not require such a complex design and can achieve much better results on the APE task. Besides, our MRC-APE achieves better results than MRC-APE-Sep. on both AM and APE tasks, indicating that jointly training two phases in a single MRC model could maximize the mutual benefits of the two phases.

In addition, to analyze the error propagation from the first phase to the second phase, we use the true label of AM task to predict APE task. Under this setting, our model can achieve around 59.44% F_1 score for APE task, showing effectiveness in identifying argument pairs.

4.2.2 Ablation Study

The ablation study results are shown in Table 2. It can be observed that using two directions contributes greatly to our method. Also, using the arguments recognized in D_a to extract the paired arguments in D_b is more critical in the RR dataset, removing it causes a 6.51% decrease in APE F_1 score. Without the LSTM to capture the long-

term dependency among sentences, the APE F_1 score decreases by 0.86%. Furthermore, the performance drops heavily without the global attention, because it enables more interactions between the query argument and the queried document, thus better argument-specific representations could be learned.

5 Conclusion

In this paper, we propose to frame the argument pair extraction (APE) task as a machine reading comprehension (MRC) task. Our MRC framework addresses APE through two phases with two types of queries, that is, argument mining (AM) query and argument pair extraction (APE) query. Our proposed method can better model the argument-level interactions, thus facilitating the extraction of argument pairs. Experimental results on a large benchmark dataset demonstrate that our proposed method achieves state-of-the-art performance.

Acknowledgments

References

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, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 118–128. Association for Computational Linguistics.

Jianzhu Bao, Chuang Fan, Jipeng Wu, Yixue Dang, Jiachen Du, and Ruifeng Xu. 2021. A neural transition-based model for argumentation mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6354–6364. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: argument mining for persuasive online discussions. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2933–2943. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1870–1879. Association for Computational Linguistics.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12666–12674. AAAI Press.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7000–7011. Association for Computational Linguistics.

Liying Cheng, Tianyu Wu, Lidong Bing, and Luo Si. 2021. Argument pair extraction via attention-guided multi-layer multi-cross encoding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6341–6353. Association for Computational Linguistics.

Oana Cocarascu, Elena Cabrio, Serena Villata, and Francesca Toni. 2020. Dataset independent baselines for relation prediction in argument mining. In *Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 45–52. IOS Press.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Leilei Gan, Yuxian Meng, Kun Kuang, Xiaofei Sun, Chun Fan, Fei Wu, and Jiwei Li. 2021. Dependency parsing as mrc-based span-span prediction. *CoRR*, abs/2105.07654.

Theodosios Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news, blogs, and the social web. *Int. J. Artif. Intell. Tools*, 24(5):1540024:1–1540024:22.

Lu Ji, Zhongyu Wei, Jing Li, Qi Zhang, and Xuanjing Huang. 2021. Discrete argument representation learning for interactive argument pair identification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5467–5478. Association for Computational Linguistics.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard H. Hovy. 2021. Classifying argumentative relations using logical mechanisms and argumentation schemes. *Trans. Assoc. Comput. Linguistics*, 9:721–739.

Yohan Jo, Jacky Visser, Chris Reed, and Eduard H. Hovy. 2019. A cascade model for proposition extraction in argumentation. In *Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019, Florence, Italy, August 1, 2019*, pages 11–24. Association for Computational Linguistics.

Khalid Al Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. 2018. Modeling deliberative argumentation strategies on wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2545–2555. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tatsuki Kuribayashi, Hiroki Ouchi, Naoya Inoue, Paul Reisert, Toshinori Miyoshi, Jun Suzuki, and Kentaro Inui. 2019. An empirical study of span representations in argumentation structure parsing. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4691–4698. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 333–342. Association for Computational Linguistics.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, 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, ACL 2020, Online, July 5-10, 2020*, pages 5849–5859. Association for Computational Linguistics.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13543–13551. AAAI Press.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *The Eleventh International Conference on Artificial Intelligence and Law, Proceedings of the Conference, June 4-8, 2007, Stanford Law School, Stanford, California, USA*, pages 225–230. ACM.

Gaku Morio, Ryo Egawa, and Katsuhide Fujita. 2019. Revealing and predicting online persuasion strategy with elementary units. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6273–6278. Association for Computational Linguistics.

Gaku Morio and Katsuhide Fujita. 2018. End-to-end argument mining for discussion threads based on parallel constrained pointer architecture. In *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 11–21. Association for Computational Linguistics.

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, Yuta Koreeda, and Kohsuke Yanai. 2020. Towards better non-tree argument mining: Proposition-level biaffine parsing with task-specific parameterization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3259–3266. Association for Computational Linguistics.

Huy Nguyen and Diane J. Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

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

Haoyang Wen, Anthony Ferritto, Heng Ji, Radu Florian, and Avi Sil. 2021. VAULT: variable unified long text representation for machine reading comprehension. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 1035–1042. Association for Computational Linguistics.

Jian Yuan, Zhongyu Wei, Donghua Zhao, Qi Zhang, and Changjian Jiang. 2021. Leveraging argumentation knowledge graph for interactive argument pair identification. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2310–2319. Association for Computational Linguistics.