A Structure-Aware Argument Encoder for Literature Discourse Analysis

Anonymous ACL submission

Abstract

Existing research for argument representation learning mainly treats tokens in the sentence equally and ignores the implied structure information of argumentative context. In this paper, we propose to separate tokens into two groups, namely framing tokens and topic ones, to capture structural information of arguments. In addition, we consider high-level structure by incorporating paragraph-level position information. A novel structure-aware argument encoder is proposed for literature discourse analysis. Experimental results on both a self-constructed corpus and a public corpus show the effectiveness of our model.

1 Introduction

With the growing amount of scientific literature, researchers pay increasing attention to developing computational methods for analyzing scientific literature (Kirschner et al., 2015; Stab and Gurevych, 2017; Green, 2018; Lauscher et al., 2018; Accuosto and Saggion, 2019), aiming to identify various components of arguments automatically (Abend et al., 2009; Judea and Strube, 2017; Lukin et al., 2017; Durmus and Cardie, 2018; Lugini and Litman, 2020). Existing research focuses on constructing annotated corpus and learning representation of sentences for literature discourse analysis. They tend to treat tokens in a sentence equally and ignore the implied structure information of argumentative context (Stab and Gurevych, 2014; Wachsmuth et al., 2016; Zhang et al., 2016; Lawrence and Reed, 2017).

Figure 1 shows two annotated abstracts of scientific literature, in which sentences are classified into four types, namely *background*, *method*, *result* and *conclusion*. We have some findings. First, tokens in sentences can be divided into two groups, i.e., topic words and framing words. Topic words provide the fundamental knowledge of this argument while framing words organize the expression. Second, the same argument components often use similar framing structure in discourses across topics. For example, structures like '... *is employed / investigated to* ...' usually appear in the *method* section. Third, argument components are sensitive to their positions. For example, *background* almost always comes before *method* part and *conclusion* usually locates at the end. Motivated by these findings, we propose a structure-aware argument encoder (SAE) based on the transformer to enhance the literature discourse analysis. Experimental results show the effectiveness of our proposed model both on a self-constructed corpus and a public corpus.

Abstract A:

Background : [This study investigates how does the Stern Review affect the Economics of Climate Change, …]
Method: [An approach to support climate policy evaluation, consisting in the Threshold 21 (T21) model, … is employed to analyze Ecuador's energy, …] …
Result: [Results indicate ... would reduce GHG emissions in the electric power sector, ...] …
Conclusion: [Finally, authors find …positive economic increases energy consumption, ...]

Abstract B:

Background : [Combustion is a source of atmospheric aerosols, including organic carbon (OC) and black carbon (BC) .]
Method: [Pilot-scale coal combustor was used to investigate the OC and BC aerosol formation under combustion conditions.]
Result: [It was found that BC aerosol formation was sensitive to the fuel-oxidizer equivalence ratio.] ……
Conclusion: [The BC and OC aerosol formation indicate that the formation pathways of OC aerosol ...]

Figure 1: Two samples of annotated abstracts. Framing tokens are highlighted in blue font and blue dotted lines and the rest are topic tokens. The division rule of tokens can be referred in section 3.

Our contributions are two-fold: (1) we propose a novel transformer encoder that considers topic tokens and framing tokens separately to incorporate the structure of an argument for its representation learning; (2) we construct a large scale annotated corpus of scientific literature across different topics as a new benchmark.

Corpus	Area	Content	Size	Type	IAA
DiGAT (Kirschner et al., 2015)	Education	Full-text	24	-	0.50 (F1)
Gold Standard (Sateli and Witte, 2015)	Computer Science	Full-text	30	2	-
Dr. Inventor (Ronzano and Saggion, 2015)	Computer Science	Full-text	40	5	0.66 (Kappa)
PubMed-SciDT (Dasigi et al., 2017)	Medical	Experiment	75	7	-
Biomedical-Claims (Achakulvisut et al., 2019)	Medical	Abstract	1,500	2	0.63 (Kappa)
CCSA (ours)	Climate Science	Abstract	**2,018**	4	**0.68** (Kappa)

Table 1: Comparison between the CCSA corpus and other human-annotated corpora for scientific literature.

	Bg.	Meth.	Res.	Con.
Number	3,939	4,306	5,962	4,625
Proportion	20.9%	22.9%	31.7%	24.5%

Table 2: Distribution of different argument component types. Bg., Meth., Res., Con. are the abbreviations of background, method, result and conclusion.

2 CCSA Corpus

There are several public annotated corpora for scientific literature analysis (Liakata et al., 2010; Kirschner et al., 2015; Sateli and Witte, 2015; Ronzano and Saggion, 2015; Dasigi et al., 2017; Accuosto and Saggion, 2019; Achakulvisut et al., 2019), most of which focus on medicine and computer science. However, as a highly controversial research area, climate science is less explored. To bridge the gap, we create the **C**limate **C**hange **S**cientific **A**rgumentation (CCSA) corpus. Table 1 shows a comparison between the CCSA corpus and several annotated corpora for scientific literature, and our CCSA corpus has the advantages of corpus size and inter-annotator agreement.

Data Source We search for *climate change* in the ISI Web of Science[1] 2020 and collect all the retrieved papers published from 2000 to 2020 as the source. The domain of climate change covers a wide range of topics. In order to balance various sub-focus, we use Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to cluster all papers into different topics and choose similar number of publications from each group for annotation.

Annotation Scheme We treat each sentence in the abstract as an argument component and classify them into four types. C1) ***Background*** explains the motivation and background. C2) ***Method*** presents experimental procedures. C3) ***Result*** includes data, facts, and descriptions of outcomes, without any subjective speculations or judgements. C4) ***Conclusion*** gives opinions of the author. Invalid sentences, such as copyright information, are labeled as *other* types.

Annotation Process Undergraduate students are hired for the annotation, about half of them are majored in environmental sciences. We develop a web-based annotation platform and each abstract is annotated by three annotators. The inter-annotator agreement for argument type annotation is 0.68 in terms of Fleiss's Kappa coefficient (Falotico and Quatto, 2015), which shows a moderate consistency. The final result is determined by majority votes. If there is a disagreement, the label will be determined by the annotator with the greatest confidence [2]. There were 2,018 abstracts and 18,832 valid argument components in CCSA corpus. Table 2 depicts the distribution of the argument type.

3 Structure-aware Argument Encoder

In order to incorporate the structure information of an argument, we propose a novel structure for argument representation learning, named **S**tructure-aware **A**rgument **E**ncoder (SAE). The main component of SAE is a transformer structure with multiple attention mechanisms to capture interactions between different groups of tokens. The overall architecture is shown in Figure 2.

Argument Structure In scientific discourse, some technical terms may introduce some noise to the identification of the argument structure. In SAE, we divide the tokens in each sentence into framing tokens and topic tokens. ***Framing Token*** contains the structural information in the argument component. ***Topic Token*** contains the topic information in the argument component, such as technical terms in the research field.

The sentence is tokenized and tagged with POS (Part-of-speech) using NLTK (Hardeniya et al.,

[1] http://isiknowledge.com/

[2] We calculate the divergence between each annotator and the determined results as the confidence.

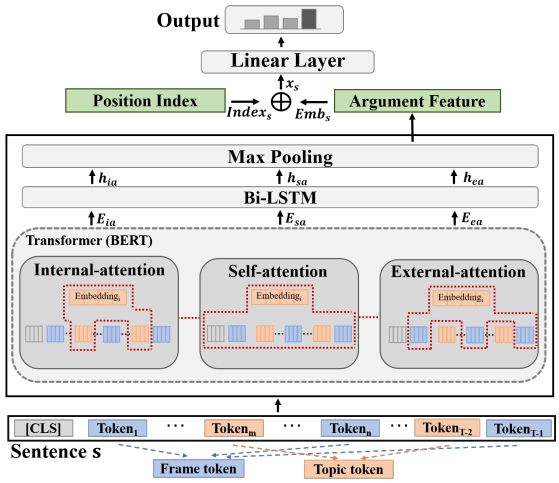

Figure 2: The overall architecture of our Structure-aware Argument Encoder (SAE).

2016). We regard *Singular Noun (NN)*, *Plural Noun (NNS)*, *Singular Proper Noun (NNP)* and *Plural Proper Noun (NNPS)* as topic tokens and others as framing tokens. Any method can be adopted for token division, not just POS tagging.

Argumentative Attention Mechanism To utilize the information of the token types, in addition to ***self-attention***, our argumentative attention mechanism contains two extra attention patterns. ***Internal-attention*** takes effect among tokens of the same type, i.e., framing tokens attend to framing tokens, and so do topic tokens. Internal-attention is utilized to explore the internal influence of tokens of the same type. ***External-attention*** takes effect among tokens with different types, i.e., framing tokens attend to topic tokens, and topic tokens attend to framing tokens. External-attention is expected to explore the influence between tokens with different types.

Argument Representation Suppose the **input** s is a sentence with T tokens $s = [t_0, t_1, ..., t_{T-1}]$, the structure-aware argument encoder is first adopted to obtain the contextual **token embeddings** E based on argumentative attention:

$$E = [e^0, ..., e^{T-1}] = F(t_0, ..., t_{T-1}) \quad (1)$$

where $F(\cdot)$ is transformer encoder. We can obtain E_{ia}, E_{ea} and E_{sa} through $F_{ia}(\cdot)$, $F_{ea}(\cdot)$ and $F_{sa}(\cdot)$, which are transformer encoders with internal-attention, external-attention and self-attention. The parameters of the three transformer encoders are shared, but due to their different attention mechanisms, different features can be extracted. The token embeddings E are then fed into a token-level bidirectional LSTM layer, and the last hidden states from both directions are concatenated as the **sentence embedding** h:

$$\begin{aligned}[[\overrightarrow{h^0}; \overleftarrow{h^0}], ...[\overrightarrow{h^{T-1}}; \overleftarrow{h^{T-1}}]] &= \text{Bi-LSTM}(E) \\ h &= [\overrightarrow{h^{T-1}}; \overleftarrow{h^{T-1}}]\end{aligned} \quad (2)$$

We obtain h_{ia}, h_{ea} and h_{sa} with E_{ia}, E_{ea} and E_{sa} respectively, and further use a max-pooling layer to extract the **argument feature** Emb_s of sentence s:

$$Emb_s = \text{max-pooling}(h_{ia}, h_{ea}, h_{sa}) \quad (3)$$

Argument components are sensitive to their positions and the position information is an important feature for its type. We use the standardized index of the sentence in the abstract as an additional **position feature** concatenated to argument feature as the final **argument representation**:

$$x_s = [Emb_s; Index_s] \quad (4)$$

The predicted probability distribution $p(y|s)$ of argument categories is obtained after x_s is fed into a multilayer perceptron (MLP) layer.

4 Experiment

Experimental Setup We focus on the task of **argument component identification** which aims to predict the argument type of argument component (sentences). We conduct experiments on our CCSA corpus. To demonstrate that our SAE is domain-independent, we also conduct experiments on another scientific publication abstract corpus *biomedical-claims*[3] (Achakulvisut et al., 2019). It annotates whether a sentence is a *claim*, whose setting is similar to CCSA.

For CCSA corpus, we take the macro F1 as the evaluation metric of this multi-classification problem, and the F1 score of each sentence type on the test set is also reported. For biomedical-claims corpus, we report precision, recall and F1 score on the test set. The experiment configuration details are shown in A.1.

To prove our argumentative attention mechanism has the advantage of modeling topic tokens and framing tokens, we also implement a variant of our SAE model that utilizes token types

[3] https://github.com/titipata/detecting-scientific-claim

in a simpler way, namely ***parameterized SAE (p-SAE)***. Specifically, we initialize a learnable embedding layer for framing tokens and topic tokens instead of argumentative attention mechanism, and add them to token embeddings as input, similar to the segment embedding in BERT.

Overall Performance For CCSA corpus, we compare our SAE and p-SAE with following baselines: BERT (Devlin et al., 2019), bidirectional LSTM (Bi-LSTM) (Graves et al., 2013) and Sentence Encoder (SE), which contains a BERT layer and on top of it, a Bi-LSTM layer. Compared with SAE and p-SAE, SE is a combination of BERT and Bi-LSTM without the information of token types. For biomedical-claims corpus, we present the state-of-the-art model based on transfer learning (TL-CRF) in the original paper as baseline (Achakulvisut et al., 2019).

Table 3 shows main results of CCSA corpus, which indicate that our SAE achieves competitive macro F1 score on the argument component identification task. It is worth noting that SAE improves the identification of *conclusion* part most, because the *conclusion* is the most argumentative part, which shows that our model has excellent effect in exploring argumentative structure. Similarly, results of scientific publication corpus are shown in Table 4 indicating that the model has better performance in identifying scientific claims.

Ablation Study Table 3 shows the results of ablation study. Internal-attention affects *conclusion* part most and external-attention affects *method* part most, which shows that argumentative texts, such as *conclusion* part focus more on the organization of structure. However, the structure of *method* part needs to be combined with some professional terms through external-attention. The macro F1 score of *conclusion* part drops down most without internal-attention, which shows the effectiveness of modeling topic tokens and framing tokens separately in argumentative structure.

Domain Adaptation We apply the model trained with CCSA on the test set of biomedical-claims to evaluate the ability of generalization of SAE. Since the sentence types of the two corpora are different, we do label mapping as follows: the predicted *conclusion* label is converted to claim, and the others are converted to non-claim. We migrate three models, namely SE, p-SAE and SAE, and the results are shown in Table 4.

Model	Bg.	Meth.	Res.	Con.	Macro F1
Bi-LSTM	59.6	79.4	59.6	55.0	59.6
BERT	69.6	83.3	78.0	57.5	72.1
SE	69.1	84.3	78.2	57.9	72.4
p-SAE	**72.9**	85.0	**78.6**	63.1	74.9
SAE	72.3	**86.2**	77.9	**65.7**	**75.5**
Ablation study					
SAE w/o Ia	-1.3	-1.5	-0.9	**-2.3**	-1.5
SAE w/o Ea	-0.2	**-3.5**	+1.1	-1.1	-0.8

Table 3: Performance on test set of CCSA corpus. Bg., Meth., Res., Con. are the abbreviations of *background*, *method*, *result* and *conclusion*. Ia and Ea represents *Internal-attention* and *External-attention*.

Model	Precision	Recall	F1
TL-CRF	86.6	72.7	79.0
Bi-LSTM	82.8	63.4	66.6
BERT	84.7	80.8	82.5
SE	84.8	81.9	83.2
p-SAE	84.5	83.1	83.8
SAE	**86.6**	**83.6**	**85.0**
Domain adaptation			
SE (CCSA)	80.2	78.2	79.1
p-SAE (CCSA)	**83.3**	77.0	79.5
SAE (CCSA)	81.8	**78.9**	**80.2**

Table 4: Performance on test set of biomedical-claims corpus (Achakulvisut et al., 2019). TL-CRF is the SOTA result in the original paper.

Although the research fields and categories involved in the two scientific literature corpora are different, our model still shows strong transfer capability without any training. Among them, both p-SAE and SAE that consider the argument structure outperform SE. SAE with multiple attention mechanisms performs better than p-SAE, which also illustrates the advantages of our proposed SAE in terms of domain adaptation.

5 Conclusion

In this paper, we propose a structure-aware argument encoder (SAE) that considers token types in the sentence and separate tokens into two groups, namely topic tokens and framing tokens. Multiple argumentative attention mechanisms are utilized to capture internal and external interactions among different groups of tokens. Experimental results on a self-constructed corpus and another publicly corpus of scientific literature show the effectiveness of our model.

Ethical Statement

In this paper, different ethical restrictions deserve discussion.

All data in our self-constructed corpus are available online and other corpora in this paper are publicly available sources. We strictly followed the platform's policies and rules when crawling data from web platforms. We did not employ any author-specific information in our research.

The reward for annotating an article is determined by the number of sentences in the abstract. We pay \$0.03 for each sentence, averaging about \$0.24 per article. All annotators are people who are willing to participate and over the age of 18. We have an online chat group for making announcements and answering questions.

Our corpus may includes some bias, such as political bias and social bias, and our model might have inherited some forms of these bias. In order to limit these bias as much as possible, we filter controversial articles and removed data with offensive information when possible.

References

Omri Abend, Roi Reichart, and Ari Rappoport. 2009. Unsupervised argument identification for semantic role labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 28–36, Suntec, Singapore. Association for Computational Linguistics.

Pablo Accuosto and Horacio Saggion. 2019. Discourse-driven argument mining in scientific abstracts. In *Natural Language Processing and Information Systems*, pages 182–194.

Titipat Achakulvisut, Chandra Bhagavatula, Daniel Acuna, and Konrad Kording. 2019. Claim extraction in biomedical publications using deep discourse model and transfer learning.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 601–608, Cambridge, MA, USA. MIT Press.

Pradeep Dasigi, Gully APC Burns, Eduard Hovy, and Anita de Waard. 2017. Experiment segmentation in scientific discourse as clause-level structured prediction using recurrent neural networks. *arXiv preprint arXiv:1702.05398*.

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.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045, New Orleans, Louisiana. Association for Computational Linguistics.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49(2):463–470.

A. Graves, N. Jaitly, and A. Mohamed. 2013. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 273–278.

Nancy Green. 2018. Proposed method for annotation of scientific arguments in terms of semantic relations and argument schemes. In *Proceedings of the 5th Workshop on Argument Mining*, pages 105–110, Brussels, Belgium. Association for Computational Linguistics.

Nitin Hardeniya, Jacob Perkinsand Deepti Chopra, Nisheeth Joshi, and Iti Mathur. 2016. *Natural Language Processing: Python and NLTK*. Packt Publishing.

Alex Judea and Michael Strube. 2017. Event argument identification on dependency graphs with bidirectional LSTMs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 822–831, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. pages 1–11.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. 2018. Investigating the role of argumentation in the rhetorical analysis of scientific publications with neural multi-task learning models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3326–3338, Brussels, Belgium. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2017. Mining argumentative structure from natural language text using automatically generated premise-conclusion topic

models. In *Proceedings of the 4th Workshop on Argument Mining*, pages 39–48, Copenhagen, Denmark. Association for Computational Linguistics.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers.

Luca Lugini and Diane Litman. 2020. Contextual argument component classification for class discussions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1475–1480, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain. Association for Computational Linguistics.

Francesco Ronzano and Horacio Saggion. 2015. Dr. inventor framework: Extracting structured information from scientific publications. In *International conference on discovery science*, pages 209–220. Springer.

Bahar Sateli and René Witte. 2015. Semantic representation of scientific literature: bringing claims, contributions and named entities onto the linked open data cloud. *PeerJ Computer Science*, 1:e37.

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, Doha, Qatar. Association for Computational Linguistics.

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

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.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

A Appendix

A.1 Experiment Details

We use BERT-base model (bert-base-uncased) to initialize the parameters of the transformer encoder, and the parameters of bidirectional LSTM (Bi-LSTM) are randomly initialized. All models are trained on 4 Nvidia GeForce RTX 2080 Ti GPUs with the same random seed. The batch size is 32, the dropout rate is 0.1, the learning rate is 1e-5, the hidden size for the Bi-LSTM layers is 200, the max length of a sentence is 100. We split our CCSA corpora and biomedical-claims corpus into training, validation and test sets with the proportion of $6:2:2$ respectively. The best performing model on the validation set are evaluated on the test set.