Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6162–6169
Marseille, 20-25 June 2022

Incorporating Zoning Information into Argument Mining from Biomedical Literature

Boyang Liu, Viktor Schlegel, Riza Batista-Navarro, Sophia Ananiadou
Department of Computer Science, The University of Manchester
Manchester, United Kingdom
{boyang.liu-2@postgrad., viktor.schlegel@, riza.batista@, sophia.ananiadou@}manchester.ac.uk

Abstract

The goal of text zoning is to segment a text into zones (i.e., Background, Conclusion) that serve distinct functions. Argumentative zoning, a specific text zoning scheme for the scientific domain, is even considered as the antecedent for argument mining by many researchers. Surprisingly, however, little work is concerned with exploiting zoning information to improve the performance of argument mining models, despite the relatedness of the two tasks. In this paper, we propose two transformer-based models to incorporate zoning information into argumentative component identification and classification tasks. One model is for the sentence-level argument mining task and the other is for the token-level task. In particular, we add the zoning labels predicted by an off-the-shelf model to the beginning of each sentence, inspired by the convention commonly used biomedical abstracts. Moreover, we employ multi-head attention to transfer the sentence-level zoning information to each token in a sentence. Based on experiment results, we find a significant improvement in F1-scores for both sentence- and token-level tasks. It is worth mentioning that these zoning labels can be obtained with high accuracy by utilising readily available automated methods. Thus, existing argument mining models can be improved by incorporating zoning information without any additional annotation cost.

Keywords: Argument Mining, Text Zoning, Biomedical Literature

1. Introduction

Argument mining (also known as argumentation mining) is a task that aims at analyzing the argumentative structure of discourse. This task can be divided into four sub-tasks (Eger et al., 2017): argumentative component recognition (separating argumentative units from non-argumentative units), argumentative component classification (i.e., the distinction between claims and premises), argumentative relation recognition (finding relations between argumentative components) and argumentative relation classification (i.e., whether argumentative components support or attack each other). Based on granularity, argument mining can be divided into sentence-level and token-level tasks. The former means that the boundary of an argumentative component is the same as the sentence, while the latter means that the length of argumentative components can range from less than a clause to several sentences.

In scientific literature, arguments (see examples in Figure 1) are fundamental building blocks whose main purpose is to persuade others to accept the academic opinion (i.e., *claims*) proposed by the authors. This means that the key to understanding a scientific paper is to find its argumentative structure. However, with the growth of literature (Wang and Lo, 2021), it is extremely time-consuming to manually analyse the latest research and understand the argumentative structure of related literature on such a large scale. Automatic identification of the argumentative structure in scientific literature is important for researchers and practitioners to follow the literature and find statements that corroborate or contradict each other. In this paper, we focus on argumentative component identification and classification from biomedical literature abstracts.

> ***Background***: *We have recently suggested that bolus 5-fluorouracil (5-FU) may work via a RNA directed mechanism while ...*
>
> ***Patients and methods***: *Two hundred fourteen patients from nineteen Italian centers were randomized to the control arm ...*
>
> ***Results***: {*Nine CR and twenty-seven PR were obtained on one hundred eleven evaluable patients treated in experimental arm (RR = 32%, 95% confidence interval (95% CI): 24%-42%), while two CR and eleven PR were observed among one hundred three evaluable patients in control arm (RR = 13%, 95% CI: 7%-21%)* }$_{premise1}$. ... {*Eighty percent of patients receiving second-line chemotherapy in control arm were treated with continuous infusion 5-FU* }$_{premise5}$.
>
> ***Conclusions***: *Alternating,* [*schedule-specific biochemical modulation of FU is more active than ...*]$_{claim1}$. [*However, the overall survival was similar suggesting that alternating bolus and infusional 5-FU upfront may be as effective as giving them in sequence as first- and second-line treatment*]$_{claim2}$.

Figure 1: An abstract from PubMed 11142481. We remove several sentences for brevity. The sequences in curly brackets are premises (pieces of evidence supporting or attacking claims) and in square brackets are claims.

Text zoning aims at segmenting a text into zones (Gnehm, 2018). Here, each zone differs from others and consists of text parts in terms of a particular func-

tion. For example, email zoning (Repke and Krestel, 2018) segments an email into five zones including *body*, *header*, *signoff*, *signature* and *greetings*. A job advertisement can be divided into eight zones (i.e., *company description*, *reason of vacancy*...)(Gnehm and Clematide, 2020). As for scientific literature, there are several zoning schemes which have been proposed (Teufel and others, 1999; Liakata et al., 2010; Kim et al., 2011; Dernoncourt and Lee, 2017). Among them, argumentative zoning (Teufel and others, 1999) is considered as the antecedent for argument mining in scientific literature in previous research (Lawrence and Reed, 2020; Accuosto and Saggion, 2020). It is a sentence-level scheme used in the classification of sentences by their functions within a scientific paper. For example, a sentence belongs to the *Background* zone if it is used as a description of generally accepted background knowledge and it belongs to the *Aim* zone if it is a statement of a research goal. Even though zoning information is highly related to argumentative components, only Achakulvisut et al. (2020) use this information to support sentence-level claim identification from biomedical abstracts. To the best of our knowledge, there exists no work which investigated the effect of zoning information on the tasks of argumentative component identification and classification.

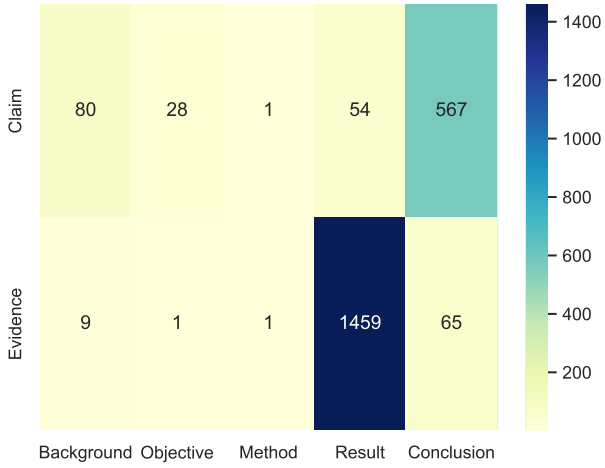

Figure 2: Distribution of argumentative components and zoning information within the training subset of PubMedRCT dataset. Zoning labels are predicted labels using a tool named HSLN (Jin and Szolovits, 2018)

In this paper, we work towards closing this gap by performing a fine-grained analysis of the impact of zoning information on the tasks of argumentative component identification and classification. We choose the PubMedRCT (Dernoncourt and Lee, 2017) as the zoning scheme used in our paper. This scheme consists of five zones, namely *Background*, *Objective*, *Method*, *Result* and *Conclusion* (see Figure 1 for some examples). By doing our own frequency analysis on AbstRCT (Mayer et al., 2020) dataset, we find that argumentative components mainly exist in the *Result* and *Conclusion* zones, as shown in Figure 2. Specifically, premises are more likely to occur in the *Result* zone and claims are more likely to occur in the *Conclusion* zone. Based on these findings, our hypothesis is that relying on zoning information, a model can mine argumentative components more accurately. We investigate the impact of zoning information on both token-level and sentence-level argument mining tasks.

Our contribution is divided into three parts: Firstly, to the best of our knowledge, this is the first study to utilise zoning information in the tasks of argumentative component identification and classification. Secondly, we propose a direct yet effective method for exploiting regularities in the writing style of biomedical abstracts, to verify the effectiveness of zoning information and minimise the impact of changes in model complexity. Thirdly, experimental evaluation shows that zoning information is helpful in both token-level and sentence-level argument mining tasks.

2. Related work

Text Zoning for Scientific Literature There are different zoning schemes for different domains, including, argumentative zoning (Teufel and others, 1999) for computational linguistics, CoreSC (Liakata et al., 2010) for chemistry, MAZEA (Dayrell et al., 2012) for physical sciences and engineering, and life and health sciences, and PIBOSO (Kim et al., 2011), GENIA-MK (Thompson et al., 2011; Shardlow et al., 2018) and PubMedRCT (Dernoncourt and Lee, 2017) for biomedicine. Argumentative zoning is the earliest schema that includes seven categories of zones, such as *Aim, Background, Contrast*. The idea of argumentative zoning is to follow the knowledge claims made by authors. For example, sentences for the description of new knowledge claims belong to *OWN* zone, while for the description of existing knowledge claims belong to *OTHER* zone. CoreSC, meanwhile, is a concept-driven scheme. It seeks to retrieve the structure of research components from a paper as generic high-level Core Scientific Concepts and thus obtains humanly-readable representations of the research process, including categories such as *Model* (to describe a theoretical model or framework) or *Conclusion* (to describe statements inferred from research results). A detailed comparison between these two schemes can be found in (Liakata et al., 2012). MAZEA and PIBOSO both consider six classes, the former includes *Background, Gap, Purpose, Method, Result and Conclusion*, while the latter includes *Background, Population, Intervention, Outcome, Study Design and Other*. GENIA-MK classifies sentences that describe bio-event into different categories based on their knowledge types (i.e., *Investigation, Observation*). All these five schemes were used for manually annotated datasets. In contrast, annotations in the PubMedRCT200k (Dernoncourt and Lee,

2017) dataset were obtained automatically based on PubMedRCT scheme designed for biomedicine. Observing that in biomedical literature, there exist zoning labels provided by publication authors themselves, Dernoncourt and Lee (2017) selected abstracts with zoning labels as the documents for their dataset. They then used a rule-based method to map author-provided labels to the 5 categories in the scheme and annotated each sentence. Adding this type of information is less laborious than adding other information such as PICO (Stylianou and Vlahavas, 2021) or discourse relations (Accuosto and Saggion, 2020), where labels are not readily available from publications. Although these schemes are not directly designed for argument mining, they are helpful in locating important arguments in scientific literature.

Argument Mining from Scientific Literature Recently, argument mining from scientific literature has received more attention, in part due to the challenges brought about by the inherent complexity of the structure and language used in specialised domains (Kirschner et al., 2015). Transformer-based models have been dominant in the approaches used in this domain. Mayer et al. (2020) compared different transformer-based models, demonstrating that SciBERT model (Beltagy et al., 2019) performs best on biomedical literature. Accuosto et al. (2021) employed cased SciBERT to mine argumentative structures from both computational linguistics and biomedical literature. Other researchers also investigated other models. For example, Galassi et al. (2021a) designed a logic tensor network for neuro-symbolic argument mining. Galassi et al. (2021b) proposed a multi-task attentive residual network for the argument mining task in different scientific domains; they made the assumption that the boundary of each argumentative component has already been detected correctly and focussed only on the classification task.

In this paper, we propose to use additional external knowledge, similar to previous work. For instance, Stylianou and Vlahavas (2021) combined PICO with argument mining and obtained significant improvement. Accuosto and Saggion (2019) found that incorporating discourse information significantly contributes to the identification of the argumentative function. With regard to zoning more specifically, Lauscher et al. (2018) designed a tool for analysing argument and rhetorical aspects in scientific writing. This tool can be used for both argumentative component identification and discourse role (similar to zoning labels) classification tasks. However, it does not consider the relation between these two tasks. A similar work to ours is that of Achakulvisut et al. (2020), which employed a transfer learning-based model to transfer knowledge from zoning to solve the claim identification task from biomedical abstracts. However, they only considered the sentence-level task and ignored the identification and classification of premises in biomedical literature. Differently from other work, we incorporate zoning information into whole argumentative component identification and classification tasks.

3. Data

We evaluate our model on two datasets on medical scientific abstracts. One dataset is used for the token-level task and the other for the sentence-level task.

AbstRCT dataset (Mayer et al., 2020).This is a token-level dataset. It consists of three types of argumentative components, namely *major claim*, *claim*, and *evidence* [1]. This dataset has three parts. The biggest part is the neoplasm corpus, which is split into the training set, development set, and test set. Additionally, there are two other test sets. The glaucoma test set includes only abstracts concerning glaucoma, whereas the second one is a mixed set with 20 abstracts concerning each disease in the dataset (neoplasm, glaucoma, hypertension, hepatitis and diabetes), respectively.

SciARG dataset (Accuosto et al., 2021).This is a sentence-level dataset, which means that the annotators consider sentences as annotation units. They propose a fine-grained scheme that contains eleven types of argumentative components (i.e., proposal, observation).

4. Methodology

We propose two models (depicted in Figure 3): the sentence-level argument mining model (SLAM) and the token-level argument mining model (TLAM), for the sentence-level and token-level tasks, respectively. SLAM is based on Accuosto et al. (2021) while TLAM is based on Mayer et al. (2020). The main difference between their models and ours is the utilisation of zoning information in such a way that changes to the models are minimal, and that they directly assess the effect of zoning information. In the following subsections, we describe how we combined zoning information with argument mining and present the details of our two models.

4.1. Utilisation of Zoning Information

To the best of our knowledge, there exists no dataset annotated with both zoning and argumentative component labels, so we need to predict the zoning labels for both the AbstRCT and SciARG datasets. As for zoning scheme, we selected PubMedRCT (Dernoncourt and Lee, 2017) for two reasons. Firstly, this scheme is used to annotate the biggest dataset PubMedRCT200k and there exists an off-the-shelf tool named HSLN (a hierarchical sequential labelling network) [2] (Jin and Szolovits, 2018) with high accuracy. Secondly, the PubMedRCT200k dataset is automatically annotated

[1] Here they call premises as evidences. To be in line with them, we use evidence instead of premise when mentioning this dataset.

[2] https://github.com/jind11/HSLN-Joint-Sentence-Classification

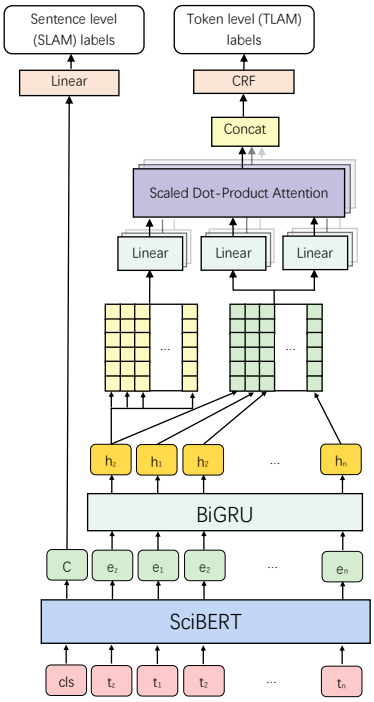

Figure 3: Overview of our model. t_z represents zoning labels, and cls is the special token [CLS] in SciBERT

based on the already existing labels in abstracts provided by the authors.

Given an abstract X that consists of m sentences $X = (x_0, x_1, ..., x_m)$, we first apply the HSLN model on each sentence to obtain the zoning label.

$$z_i = HSLN(x_i)$$

As discussed above and shown in Figure 1, the convention typically used in biomedical abstracts is to explicitly place the zoning labels at the beginning of each zone (e.g. Background, Method, etc.). Here, each zone consists of one or more sentences. Inspired by this, we placed the corresponding zoning label in front of each sentence in the abstract (since zoning is formally a sentence classification task). Afterwards, we used these sentences enriched with zoning information as the input to SLAM and TLAM.

$$input = concatenate(z_i, x_i)$$

We employ this direct methodology to empirically verify that the improvement in fact stems from utilising zoning information rather than the design of a more complex model.

4.2. Sentence-Level Argument Mining (SLAM)

For the sentence-level argument mining task, we do not need to identify boundaries of argumentative components, so we treat it as a sentence classification task. We employed the pre-trained SciBERT model (Beltagy et al., 2019) to obtain sentence embeddings, drawing upon the results of Mayer et al. (2020) who showed that SciBERT yields the best results in biomedical literature argument mining. We used a linear layer as the sentence classifier.

Specifically, we followed the work of Accuosto et al. (2021) and directly used the conventionally used `[CLS]` token in BERT-based models as the representation of the class of each sentence in an abstract. The `[CLS]` token is then passed to a linear layer. Finally we employed a Softmax function to obtain the probability distribution of argumentative component types.

$$y_i = Softmax(W[ClS] + b)$$

In line with Accuosto et al. (2021), we chose cross entropy loss as the loss function for the sentence-level model.

4.3. Token-Level Argument Mining (TLAM)

We treat the token-level argument mining task as a sequence tagging problem, incorporating both the argumentative component identification and classification tasks. Similar to Mayer et al. (2020), we used SciBERT to obtain token embeddings and passed them to a BiGRU (Cho et al., 2014) sequence encoder. Finally we employed a conditional random field (CRF) layer to capture label dependencies. Furthermore, we added a multi-head attention operation to transfer the sentence-level zoning labels into token-level labels.

In particular, we used the embeddings of each token rather than the `[CLS]` token as the output of SciBERT:

$$e_z, e_1, e_2, ..., e_n = SciBERT(input)$$

BiLSTM has proven to be effective for the task of sequence tagging. However, Mayer et al. (2020) found that BiGRU performs better than BiLSTM on the AbstRCT dataset. Therefore, we selected BiGRU as the sequence encoder. We concatenate both forward and backward hidden state vectors $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ to obtain the encoding h_t for each token in a sequence:

$$\overrightarrow{h_t} = GRU_{forward}(e_z, e_1, e_2, ..., e_n)$$

$$\overleftarrow{h_t} = GRU_{backward}(e_n, e_{n-1}, ..., e_z)$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$$

Considering that the sequence labelling task is a token-level task whereas adding zoning labels in front of each sentence only provides sentence-level information, we employed multi-head attention, which is is similar to the method used in the transformer architecture (Vaswani et al., 2017), to transform sentence-level into token-level information.

Given the representation of a sequence after the BiGRU encoder $h = (h_z, h_1, h_2, ..., h_n)$, we used a duplicate of the zoning token h_z as query Q to add the zoning

information into each token, while the key K and value V are the same as in the transformer:

$$Q = [h_z, h_{z1}, h_{z2}, ..., h_{zn}]$$

$$K = V = [h_z, h_1, h_2, ..., h_n]$$

where h_{zi} is identical with h_z. We then used scaled-dot attention to obtain the representation of each head:

$$head_i = Softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i$$

$$Q_i = QW_i^Q, K_i = KW_i^K, V_i = VW_i^V$$

where W_i^Q,W_i^K and W_i^V are trainable parameters. The final output of the multi-head attention $O_{attention}$ is calculated by concatenating all representations of each head.

$$O_{attention} = Concatenate(head_1, ..., head_h)$$

Finally, a CRF layer is used to learn the dependency between labels:

$$Y = CRF(O_{attention})$$

where Y is a series of predicted labels. As with the sentence-level task, we also employed cross entropy loss as the loss function for the token-level task.

5. Empirical Study

In this section, we describe the experiments we conducted and the analysis of experiment results.

5.1. Baselines

In this work, we designed a rule-based heuristic method and chose two existing transformer-based models as baselines. The reason for choosing the latter two is that they are similar to our models. In this way, we can maintain comparability by minimising changes to their model architecture, thus directly testing the effect of incorporating zoning information.

Heuristic method. As depicted in Figure 2, zoning and argumentative components are strongly related. To directly assess the extent to which zoning information can help in identifying and classifying argumentative components, we designed a heuristic method for the token-level task, which applies the following rules: sentences labelled as *Background*, *Objective* and *Method* are classified as non-argumentative sentences, and all the tokens of non-argumentative sentences are all labelled as O (Outside). Sentences labelled as *Result* are considered evidences and labelled as *Conclusion* are classified as claims. The first token in *Result* and *Conclusion* sentences are labelled as B-evidence and B-claim respectively, while succeeding tokens are labelled as I-evidence and I-claim.

Mayer et al. (2020) employed a fine-tuned SciBERT model with a BiGRU network and a CRF layer, which is a common method for sequence tagging tasks. We use it as a baseline for the token-level task.

Accuosto et al. (2021) used the cased version of SciBERT as a base model and feed the representation of the `[CLS]` token into a linear classifier followed by a Softmax function. We utilised this as a baseline for the sentence-level task.

5.2. Experimental Settings

The token-level task is a BIO sequence labelling task. Like Mayer et al. (2020), we merged major claims and claims into claims considering the negligible occurrences of major claims. Finally, the token-level task was cast as a five labels (i.e., B-Claim, I-Claim, B-Evidence, I-Evidence and Outside) sequence tagging task. For this task, we used the uncased SciBERT model, and fine-tuned it with Adam optimizer (Kingma and Ba, 2014) for three epochs. The hidden dimension of a single GRU for each direction in the BiGRU sequence encoder was set to 768. We set the learning rate to 5×10^{-5}. For the sentence-level task, we used the cased SciBERT model. We used the Adam optimizer with a learning rate of 2×10^{-5}. The number of training epoch was set to 15. Both uncased and cased SciBERT were downloaded from Huggingface (Wolf et al., 2020).

5.3. Results and Discussion

We report macro-averaged (F1) and micro-averaged (f1) scores for the token-level task as Mayer et al. (2020) did, and macro-averaged F1-scores weighted by class cardinality for the sentence-level task as Accuosto et al. (2021) did. All these scores are a mean across ten different runs of the model training with different random seeds. In the sentence-level task, we also report the results of a specific task named *main unit identification* proposed by Accuosto et al. (2021), which aims at finding the sentence describing the most significant contribution of a research paper. The results of token-level and sentence-level argument mining are shown in Table 1 and Table 2, respectively. In Table 1, C-F1 stands for macro-averaged F1-score for claims and E-F1 corresponds to macro-averaged F1-scores for evidence.

We find that the heuristic method obtains very high macro- and micro-averaged F1-scores, despite its simplicity. In the neoplasm test set of AbstRCT dataset, it even achieves the highest C-F1 compared to other BERT-based models. This result is in line with our finding in Figure 2. Both of them suggest that zoning information is indeed useful for argumentative component identification and classification, even without the help of additional semantic information.

[1]We downloaded their code from https://gitlab.com/tomaye/ecai2020-transformer_based_am to reproduce these results. We also directly employed their code in our evaluation.

Models	Neoplasm				Glaucoma				Mixed			
	f1	C-F1	E-F1	F1	f1	C-F1	E-F1	F1	f1	C-F1	E-F1	F1
Heuristic method	87.23	**73.88**	81.96	80.20	87.32	80.04	80.93	82.17	88.00	73.71	83.97	80.97
Mayer et al. (2020) [1]	90.05	69.65	84.17	83.48	91.50	76.52	83.53	85.67	90.88	72.50	83.62	84.27
TLAM	90.78	73.80	86.17	**85.59**	92.18	80.45	85.99	87.82	91.63	75.58	**85.76**	86.05
TLAM_without_Att	90.82	72.94	85.65	85.09	91.93	79.47	84.68	87.03	91.68	74.69	85.33	85.70
TLAM_Single_B	90.12	71.45	85.85	84.64	91.74	80.45	85.62	87.66	90.99	74.63	84.86	85.41
TLAM_Single_O	90.15	72.03	85.79	84.79	91.74	80.38	85.87	87.73	90.94	74.28	84.75	85.25
TLAM_Single_M	90.05	71.89	85.88	84.77	91.79	**80.90**	**86.18**	**87.99**	91.33	74.87	85.37	85.72
TLAM_Single_R	90.68	72.84	86.35	85.29	**92.19**	80.19	85.29	87.52	**91.69**	75.56	85.50	85.98
TLAM_Single_C	**90.94**	73.39	**86.39**	85.51	91.86	80.45	85.45	87.60	91.61	**76.31**	85.38	**86.16**

Table 1: Results for token-level argument mining. All the reported results are statistically significant. Best results are highlighted in bold. TLAM is our token-level argument mining model. F1 and f1 stand for macro- and micro-averaged F1-scores, respectively. C-F1 and E-F1 stand for macro-averaged F1-scores for claim and evidence, respectively. TLAM_Single_{B,O,M,R,C} means the model only exploits a single zoning label, i.e., *Background*, *Objective*, *Method*, *Result* and *Conclusion* respectively.

Regarding the token-level experiment, from the obtained results we observe that the overall macro-averaged F1-score improves by 2.11, 2.15 and 1.78 percentage points in the neoplasm, glaucoma and mixed test sets, respectively. This improvement mainly comes from C-F1. This is consistent with the results of the heuristic method, whose C-F1 is comparable with transformer-based models in both neoplasm and glaucoma test sets. All improvements in micro-averaged F1-score are less than one percentage point, which is mainly due to the dominance of the 'O' label.

From the results reported in Table 2, we find that the sentence-level task benefits from zoning information as well, not only for classifying argumentative components, but also for the identification of main units. Interestingly, even though the number of argumentative component types is higher (eleven) than the number of zoning types (five), the fine-grained component type classification task can still benefit from the coarser zoning labels.

Method	Component	Main Unit
Accuosto et al. (2021) [2]	67.38	86.76
SLAM	**69.08**	**88.79**

Table 2: Results for sentence-level argument mining. Best results are highlighted in bold. SLAM is our sentence-level argument mining model.

5.4. Ablation Study

To understand the influence of multi-head attention, we ran both TLAM and TLAM_without_Att models. The difference between them is that the latter does not include multi-head attention and directly uses the output of BiGRU as the input of the CRF layer. It is evident in Table 1 that multi-head attention improves the macro-average F1-score by roughly 0.5 percentage points. Furthermore, we conducted experiments to investigate the contribution of each type of zoning label. Unlike other experiments that test the contribution of one label type by removing this type to detect the degradation in the model's performance, we conducted experiments to test the results of incorporating only one type of zoning label. For instance, TLAM Single_B only adds *Background* label before the sentences that belong to *Background* zone, while the sentences that belong to other zones are sent to the SciBERT directly without any processing. The results are shown in Table 1.

From this table, we observe that even when using only one type of zoning label, the five models perform better than the model developed by Mayer et al. (2020), which does not include zoning information. Among the five labels, models using the *Result* and *Conclusion* labels alone can obtain results comparable with the model using all five types of labels, when considering all three test sets. It is noticeable that the model incorporating only the *Conclusion* label outperforms the model that uses all five types of labels in four different F1-score. It is also worth noting that using the *Method* label alone achieves the best performance in the glaucoma test set. We posit that this is due to the high proportion of *Method* sentences (30%) in abstracts, especially in the glaucoma test set (35%). Even though this type of zoning label is least relevant to argumentative components, it helps to effectively exclude these non-argumentative sentences given the high frequency of occurrences.

It is clear that the information provided by *Background* and *Objective* leads to the least improvement in model performance. One possible reason is that these two labels have little correlation with the appearance of argumentative components, as shown in Figure 2. Another reason could be that the accuracy of the predictions of these two types of labels is relatively low (75.6 F1-score for *Background* and 70.7 for *Objective*), and these two labels tend to be confused by the HSLN model(Jin and Szolovits, 2018). Improvements might be obtained when using gold-standard zoning labels rather than predicted labels.

[2]Results taken directly from their paper

6. Conclusion

In this paper, we propose a method to leverage zoning information for the argumentative component identification and classification tasks in the biomedical domain. Specifically, we added the predicted zoning label in front of each sentence, which is then given as input to the encoding layer. We propose a sentence-level model and a token-level model for the sentence-level and the token-level argument mining task, respectively. Experiment results performed at these two different levels demonstrate the effectiveness of utilising zoning information for the task of argument mining. Considering that we only focus on the biomedical domain, one possible research direction is to test whether the zoning information is useful in other scientific domains.

7. Bibliographical References

Accuosto, P. and Saggion, H. (2019). Discourse-driven argument mining in scientific abstracts. In *International Conference on Applications of Natural Language to Information Systems*, pages 182–194. Springer.

Accuosto, P. and Saggion, H. (2020). Mining arguments in scientific abstracts with discourse-level embeddings. *Data & Knowledge Engineering*, 129:101840.

Accuosto, P., Neves, M., and Saggion, H. (2021). Argumentation mining in scientific literature: from computational linguistics to biomedicine. In *Frommholz I, Mayr P, Cabanac G, Verberne S, editors. BIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval; 2021 Apr 1; Lucca, Italy. Aachen: CEUR; 2021. p. 20-36*. CEUR Workshop Proceedings.

Achakulvisut, T., Bhagavatula, C., Acuna, D., and Kording, K. (2020). Claim extraction in biomedical publications using deep discourse model and transfer learning.

Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November. Association for Computational Linguistics.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Dayrell, C., Candido Jr, A., Lima, G., Machado Jr, D., Copestake, A., Feltrim, V. D., Tagnin, S., and Aluisio, S. (2012). Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1604–1609.

Dernoncourt, F. and Lee, J. Y. (2017). PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

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

Galassi, A., Lippi, M., and Torroni, P. (2021a). Investigating logic tensor networks for neural-symbolic argument mining.

Galassi, A., Lippi, M., and Torroni, P. (2021b). Multi-task attentive residual networks for argument mining. *arXiv preprint arXiv:2102.12227*.

Gnehm, A.-S. and Clematide, S. (2020). Text zoning and classification for job advertisements in german, french and english. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 83–93.

Gnehm, A.-S. (2018). Text zoning for job advertisements with bidirectional lstms. In *Proceedings of the 3rd Swiss Text Analytics Conference*, pages 66–74.

Jin, D. and Szolovits, P. (2018). Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3100–3109, Brussels, Belgium, October-November. Association for Computational Linguistics.

Kim, S. N., Martinez, D., Cavedon, L., and Yencken, L. (2011). Automatic classification of sentences to support evidence based medicine. *BMC bioinformatics*, 12(2):1–10.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirschner, C., Eckle-Kohler, J., and Gurevych, I. (2015). Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11.

Lauscher, A., Glavaš, G., and Eckert, K. (2018). ArguminSci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 22–28, Brussels, Belgium, November. Association for Computational Linguistics.

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

Liakata, M., Teufel, S., Siddharthan, A., and Batchelor, C. (2010). Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2054–2061.

Liakata, M., Thompson, P., de Waard, A., Nawaz, R., Pander Maat, H., and Ananiadou, S. (2012). A three-way perspective on scientific discourse annotation for knowledge extraction. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 37–46. The Association for Computational Linguistics (ACL).

Mayer, T., Cabrio, E., and Villata, S. (2020). Transformer-based argument mining for healthcare applications. In *ECAI 2020*, pages 2108–2115. IOS Press.

Repke, T. and Krestel, R. (2018). Bringing back structure to free text email conversations with recurrent neural networks. In *European Conference on Information Retrieval*, pages 114–126. Springer.

Shardlow, M., Batista-Navarro, R., Thompson, P., Nawaz, R., McNaught, J., and Ananiadou, S. (2018). Identification of research hypotheses and new knowledge from scientific literature. *BMC medical informatics and decision making*, 18(1):1–13.

Stylianou, N. and Vlahavas, I. (2021). Transformed: End-to-end transformers for evidence-based medicine and argument mining in medical literature. *Journal of Biomedical Informatics*, 117:103767.

Teufel, S. et al. (1999). *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, Citeseer.

Thompson, P., Nawaz, R., McNaught, J., and Ananiadou, S. (2011). Enriching a biomedical event corpus with meta-knowledge annotation. *BMC bioinformatics*, 12(1):1–18.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, L. L. and Lo, K. (2021). Text mining approaches for dealing with the rapidly expanding literature on covid-19. *Briefings in Bioinformatics*, 22(2):781–799.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.