DMON: A Simple yet Effective Approach for Argument Structure Learning

Wei Sun, Mingxiao Li, Jingyuan Sun, Jesse Davis, Marie-Francine Moens
Department of Computer Science, KU Leuven
Celestijnenlaan 200A 3001 Heverlee, Belgium
Sun.Wei, Mingxiao.Li, Jingyuan.Sun, Jesse.Davis,Sien.Moens@kuleuven.be

Abstract

Argument structure learning (ASL) entails predicting relations between arguments. Because it can structure a document to facilitate its understanding, it has been widely applied in many fields (medical, commercial, and scientific domains). Despite its broad utilization, ASL remains a challenging task because it involves examining the complex relationships between the sentences in a potentially unstructured discourse. To resolve this problem, we have developed a simple yet effective approach called Dual-tower Multi-scale cOnvolution neural Network (DMON) for the ASL task. Specifically, we organize arguments into a relationship matrix that together with the argument embeddings forms a relationship tensor and design a mechanism to capture relations with contextual arguments. Experimental results on three different-domain argument mining datasets demonstrate that our framework outperforms state-of-the-art models. The code is available at `https://github.com/VRCMF/DMON.git`.

Keywords: argument structure learning, argument mining

1. Introduction

Argument structure learning (ASL) (Moens, 2013; Lawrence and Reed, 2020) involves detecting and tagging relationships between argumentative components in a text. Figure 1 shows an illustrative example of an argumentative structure for a medical report where pairs of sentences are annotated with whether there is a supportive or attacking relationship between them. This problem is a cornerstone in the semantic analysis of natural language text because it helps to elucidate the relational structure. Consequently, it helps facilitate more accurate and deeper comprehension of text and hence plays a critical role in various NLP applications such as patient-generated content analysis (Mayer et al., 2020; Stylianou and Vlahavas, 2021), legal reasoning (Poudyal et al., 2020), and opinion mining (Niculae et al., 2017).

Despite of its broad application, solving the ASL task is still challenging due to the complexity of text structures and diversity of relationships. Moreover, real-world data often contains inconsistencies and are largely unstructured. Fully understanding the relationship between two arguments often require contextual knowledge from other arguments, or even their relationships.

A key challenge posed by ASL is that fully understanding the relationship between two arguments often requires capturing contextual knowledge about other arguments and their relationships. In Figure 1, to classify the relationship **C** $\rightarrow$ **D**, examples of contextual argument relationships are **A** $\rightarrow$ **D** and **B** $\rightarrow$ **D**. Mayer et al. (2020), Stylianou and Vlahavas (2021) and Galassi et al. (2021) tried to conduct pairwise relation classification for ASL without contextual information, yielding sub-optimal classification performance. A more recent attempt by Hua and Wang (2022) encodes the contextual arguments with a transformer architecture. This helped to improve its accuracy, but they still ignored the relationships between contextual arguments.

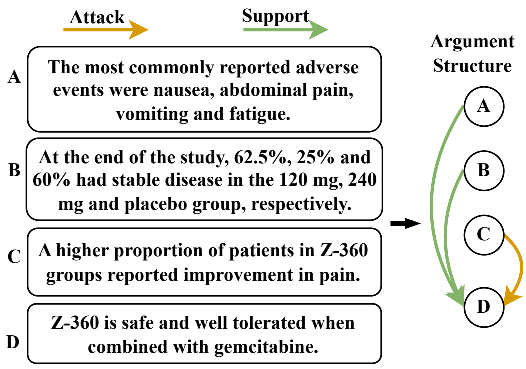


Figure 1: A simple example of an argumentative structure showing attack (orange arrow) and support (green arrow) relationships.

In this paper, we for the first time propose to exploit contextual argument relationships to solve the ASL task. As shown in Figure 2, we represent the argument structure as a relationship tensor to capture the contextual information about argument relationships. This also allows us to naturally model the relationships between pairs of arguments that can be bidirectional and asymmetric. We propose a bidirectional learning approach that uses a separate model for each direction. Moreover, training is also hampered by the fact that there is limited labeled data for ASL problems due to the high an-

LREC-COLING 2024, pages 5109–5118
20-25 May, 2024.

notation costs. Therefore, we propose a cropping strategy that randomly samples a subtensor that maintains the ordering of the selected relationships.

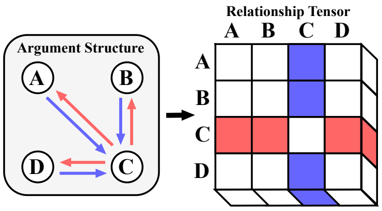


Figure 2: We select argument C as the observation object. This example shows the correlation between the head (red) and tail (blue) relationship information and a relationship tensor. Each element in this relationship tensor is the concatenation of two arguments.

The main contributions of this paper are the following:

- We propose a novel approach called DMON to encode contextual arguments and their relationships by connecting the argument structure with a relationship tensor.
- We propose a bidirectional learning mechanism that allows distinguishing head and tail arguments in a relationship.
- We design a cropping strategy to handle the scarcity of training data.
- Experimental results on three different-domain argument mining datasets show that our method outperforms state-of-the-art models for the ASL task.

First, we discuss related works in section 2. Next, in section 3, we provide a detailed description of the DMON framework. Then, in section 4, we present experimental results on two argument mining datasets from different domains and conduct an ablation study to analyze the proposed framework and its outcomes. Finally, in section 5, we summarize this paper.

2. Related Work

2.1. Argument Structure Learning

Argument Structure Learning is a challenging but essential task in text mining (Moens, 2013; Lawrence and Reed, 2020). Most papers tackle argument relationship classification by performing pairwise ASL and classifying two concatenated sentences due to it simplicity and effectiveness (Mayer et al., 2020; Stylianou and Vlahavas, 2021; Galassi et al., 2021).

In Hua and Wang (2022), argument components and their contextual sentences were encoded together using a RoBERTa encoder to obtain neighboring arguments information. The experimental results demonstrate that incorporating contextual information enhances performance. This method primarily relies on self-attention to capture relationships between the concatenated sentences. However, self-attention can be computationally expensive as it involves computing a pairwise similarity matrix for every token in all argument components. In this paper we explicitly model contextual argumentative relationships and, given that a convolution operation can effectively capture input-data features (Alzubaidi et al., 2021; Andreoli, 2019; Dumoulin and Visin, 2016), we utilize convolutional modules to represent a pair of arguments and contextual pairs of arguments.

2.2. Structured Learning in NLP

Structured learning (SL) also called structured prediction plays a crucial role in many NLP tasks. It models complex relationships and dependencies within text data to improve the performance of various discourse-related applications, such as sentiment analysis (Ein-Dor et al., 2022), and summarization (Xu et al., 2020). In argument mining, Niculae et al. (2017) use a structured support vector machine, while Bao et al. (2021) implement transition-based dependency parsing to reveal the argumentative structure.

3. Method

Problem Setting: Given a document containing n sentences, we treat each one as a potential argument and use A to denote the argument set. We convert A into the set of ordered pairs $C_A = \{(a_i, a_j)|a_j \in A \text{ and } a_i \in A\} \in \mathbb{R}^{n^2}$. The goal of ASL is to classify all relations in the set C_A with the domain-specific labels contained in the given dataset.

Method Overview: We introduce a Dual-tower Multi-scale cOnvolution neural Network (DMON) for the ASL task. The model has four components (Figure 3). Firstly, we use an encoder to extract pairwise argument representations. Given that an argumentative graph is a directed acyclic graph (DAG), potential argumentative relationships can be represented as an asymmetric relationship matrix, or as a relationship tensor when it includes the pairwise argument embeddings. Secondly, during training, a cropping strategy selects sub-tensors from the relationship tensor. Thirdly, a bidirectional learning mechanism is applied to the cropped relationship tensors to capture contextual arguments and their relationships. Finally, we employ label fusion to

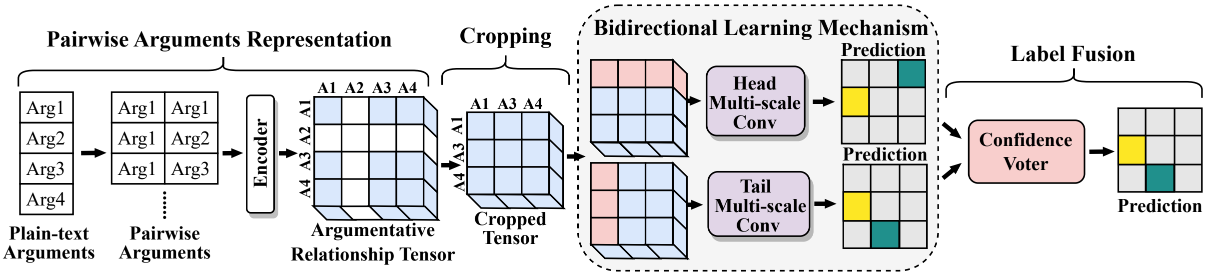


Figure 3: The structure of Multi-scale Residual Convolution Neural Network (DMON) during training. During testing, no cropping mechanism is used. In this example, the yellow and green cells in a prediction matrix correspond to attack and support relations.)

merge two predicted label matrices into one label matrix. During testing, we feed full relationship tensors instead of cropped tensors into the model.

3.1. Pairwise Arguments Representation

Following the literature (Mayer et al., 2020; Hua and Wang, 2022; Stylianou and Vlahavas, 2021), we treat each sentence as a potential argument. To capture pairwise interactions, we create all possible pairs (a_i, a_j) and concatenate them with a special token placed between them. Next, we use a linkBERT model to encode the paired arguments into an average pooled embedding with dimensionality d. We organize these into a tensor $\mathbf{H} \in \mathbb{R}^{n \times n \times d}$. Figure 2 illustrates how to transform the sentences of a discourse into a relationship tensor. Each cell in the relationship tensor represents the concatenation of a coupled argument, which typically consists of two elements: the first element is known as the head argument, while the second element is referred to as the tail argument.

3.2. Cropping Strategy

We use a cropping strategy to mitigate scarcity of labeled training data. During each training iteration, we sample a new sub-tensor $\mathbf{H}' \in \mathbb{R}^{m \times m \times d}$ from $\mathbf{H}$. Concretely, we sample m indices $\{i_0, i_1, \ldots, i_{m-1}\}$ without replacement from $\{0, 1, \ldots, n-1\}$ where $m < n$ and m is called the window size. Now we describe in mathematical notation how these indices induce a sub-tensor. $\mathbf{H}^{'}_{j,k}$ represents an element of sub-tensor $\mathbf{H}'$ whose row and column indexes are j and k, and $\mathbf{H}'_{j,k} = \mathbf{H}_{i_j, i_k}$. The cropping strategy keeps the discourse order of arguments and relations of the full relationship tensor in its rows and columns. Because the cropping strategy looks at the sub-graph induced by the selected vertices, it maintains the alignment of contextual arguments (graph's vertices) and relations (graph's edges). The cropped relationship tensors are resampled in each training iteration, which can be viewed as a form of data augmentation.

3.3. Bidirectional Learning Mechanism

We have developed a bidirectional learning mechanism to predict the labels of a relationship between its head and tail. Head and tail relationship information is captured by applying a multi-scale (1D) convolution on the relationship tensor both horizontally and vertically, respectively.

Because the relationship tensor can take into account both short- and long-distance relationships between the argument sentences in a discourse, we leverage a multi-scale residual module (MSRM). Prior work (Li and Yu, 2020) reveals that a multi-filter convolutional layer (similar to an Inception block) can capture varied relationships. However, (Li et al., 2018) stated that the Inception architecture leads to the underutilization of local features. Therefore, we choose the MSRM (Li et al., 2018) as our base module and use different kernel sizes for the convolutional filters to capture the short- and long distance relationships. Because of the asymmetry of the relationships, we apply the MSRM both horizontally and vertically on the relationship tensor to generate predictions while taking into account the contextual argument structure of the discourse. Figure 4 illustrates the structure of the MSRM.

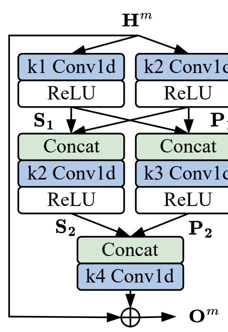


Figure 4: The structure of the multi-scale residual module (MSRM). k1, k2, k3 and k4 represent the kernel size of the respective 1D convolution layer.

We illustrate here how to capture the contextual arguments and head relationship information using the horizontal branch during trainng with a cropped tensor. Firstly, we obtain the representations of the first row of the relationship tensor, $\mathbf{h}_1 = \mathbf{H}^{'}[1,:,:] \in \mathbb{R}^{1\times m\times d}$ and pass them into the MSRM. We denote the 1D convolutional layer with kernel size k_x as $\mathcal{F}_{k_x}(\cdot)$. We omit the bias item in the equations and the output $\mathbf{o}_1 \in \mathbb{R}^{1\times m\times d}$ is:

$$\mathbf{S}_1 = \mathrm{ReLU}(\mathcal{F}_{k_1}(\mathbf{h_1})), \tag{1}$$
$$\mathbf{P}_1 = \mathrm{ReLU}(\mathcal{F}_{k_2}(\mathbf{h_1})), \tag{2}$$
$$\mathbf{S}_2 = \mathrm{ReLU}(\mathcal{F}_{k_2}(\mathrm{Concat}(\mathbf{S}_1, \mathbf{P}_1))), \tag{3}$$
$$\mathbf{P}_2 = \mathrm{ReLU}(\mathcal{F}_{k_3}(\mathrm{Concat}(\mathbf{S}_1, \mathbf{P}_1))), \tag{4}$$
$$\mathbf{o}_1 = \mathcal{F}_{k_4}(\mathrm{Concat}(\mathbf{S}_2, \mathbf{P}_2)) + \mathbf{h_1}, \tag{5}$$

where $\mathrm{ReLU}(\cdot)$ represents the ReLU activation function and $\mathrm{Concat}(\cdot)$ is the feature concatenation operation. We repeat the above calculations for m times and we concatenate all output features to get the resulting tensor of the head convolution $\mathbf{O}_h^m = \mathrm{Concat}(o_1, \cdots, o_m) \in \mathbb{R}^{m\times m\times d}$. Thirdly, a linear classifier layer consisting of flattened logits are applied to produce the predictions. On the horizontal branch, this is $\hat{\mathbf{y}}^h \in \mathbb{R}^l$ where l represents the label space dimension. The vertical branch uses the same operations as above but the logits are denoted as $\hat{\mathbf{y}}^t \in \mathbb{R}^l$.[1]

The cross-entropy losses of the horizontal and vertical branches are calculated as follows:

$$\mathcal{L}_h = -\frac{1}{m^2}\sum_{i=1}^{m^2}\sum_{j=1}^{l} y_{i,j}\log(\hat{y}_{i,j}^h), \tag{6}$$

$$\mathcal{L}_{td} = -\frac{1}{m^2}\sum_{i=1}^{m^2}\sum_{j=1}^{l} y_{i,j}\log(\hat{y}_{i,j}^t), \tag{7}$$

where $\mathbf{y} \in \mathbb{R}^l$ are one-hot encoded vectors representing the ground-truth labels. We adopt joint training for the two branch losses and the joint training loss is defined as:

$$\mathcal{L} = \lambda_h\mathcal{L}_h + \lambda_t\mathcal{L}_t, \tag{8}$$

where λ_h and λ_t are scaling factors for the horizontal and vertical branches, respectively.

3.4. Label Fusion

During testing, a confidence voter fuses the label predictions from the horizontal and vertical branches. Inspired by Vyas et al. (2018) and Weng et al. (2023), we measure the confidence scores of logits by the difference between top-1 and top-2 probabilities. Assume that we have logits $\widetilde{\mathbf{y}^h} \in \mathbb{R}^l$ and $\widetilde{\mathbf{y}^t} \in \mathbb{R}^l$, the difference between the top-1 and top-2 probabilities for $\widetilde{\mathbf{y}^h}$ is $\widetilde{\mathbf{c}^h}$:

$$\widetilde{\mathbf{s}^h} = \mathrm{Softmax}(\widetilde{\mathbf{y}^h}), \tag{9}$$
$$\widetilde{\mathbf{c}^h} = \mathrm{Topk}(\widetilde{\mathbf{s}^h}, k_1) - \mathrm{Topk}(\widetilde{\mathbf{s}^h}, k_2), \tag{10}$$

where $\mathrm{Topk}(\cdot, k)$ returns the $k-$th largest elements from the given input and $\mathrm{Softmax}(\cdot)$ is the softmax function.[2] Secondly, we get the confidence score for the vertical branch $\widetilde{\mathbf{c}^t}$ by following the above computations. Thridly, we applied the argmax operation to get the predictions of two branches, $\overline{\mathbf{y}^h} \in \mathbb{R}^{n\times n}$ and $\overline{\mathbf{y}^t} \in \mathbb{R}^{n\times n}$. We merge two matrices into $\overline{\mathbf{y}} \in \mathbb{R}^{n\times n}$ based on confidence scores $\widetilde{\mathbf{c}^h}$ and $\widetilde{\mathbf{c}^t}$. The merging process for the $\overline{y}_{i,j}$ is:

$$\overline{y}_{i,j} = \begin{cases} \overline{y^h}_{i,j}, & \text{if } \widetilde{s^h}_{i,j} \geq \widetilde{s^t}_{i,j} \\ \overline{y^t}_{i,j}, & \text{otherwise} \end{cases} \tag{11}$$

4. Experiments

The goal is to evaluate the Macro F1 score (it returns objective results on imbalanced ASL datasets) of the detection and correct classification of argumentative relationships in a discourse, to compare the results with the results of state-of-the-art baselines, and to assess the influence of the proposed components of the DMON model.

4.1. Datasets

We conduct experiments in the medical, legal and scientific domains, represented by the AbstRCT (Mayer et al., 2020), the Cornell eRulemaking Corpus (CDCP) (Niculae et al., 2017), and the SciDTB (Yang and Li, 2018) datasets.

AbstRCT: The AbstRCT corpus consists 659 medical documents about the treatment for specific disease (neoplasm, glaucoma, hypertension, hepatitis-b, diabetes). Following Mayer et al. (2020), the corpus is divided into three datasets based on disease category: neoplasm, glaucoma, and mixed. The neoplasm (neo) dataset contains 350 documents for training, 50 for validation and 100 notes for testing. The neoplasm train set is used as train set for the glaucoma (gla) and mixed (mix) dataset, which each have 100 instances for testing. The labels of the relationships of the three AbstRCT datasets are 'attack', 'support', and 'un-related'.

SciDTB: The SciDTB (SCI) dataset includes 1049 scientific abstracts collected from the ACL Anthology. It consists of 743 examples for training, 154 samples for validation and 152 for testing. The

[1]Note that during testing (inference stage) in the above computations m is replaced by n (number of paired arguments of the full text) and $\mathbf{H}^{'}$ by $\mathbf{H}$.

[2]Results with other normalization functions (e.g., L1, L2, min-max) did not improve results.

Models	Neoplasm				Glaucoma				Mixed			
	F1	S-F1	A-F1	U-F1	F1	S-F1	A-F1	U-F1	F1	S-F1	A-F1	U-F1
AMPERE++	63.73	-	-	-	-	-	-	-	-	-	-	-
Roberta	67.00	-	-	-	66.00	-	-	-	69.00	-	-	-
AMCT-Sci	68.16	59.99	49.12	95.45	62.28	64.71	24.78	95.24	69.43	55.31	58.00	94.76
TransforMED	69.96	58.72	55.65	95.51	69.72	65.32	47.00	96.88	71.82	57.14	**63.41**	94.90
RESATTARG	70.92	52.77	**65.38**	94.54	68.40	54.73	**56.00**	94.36	67.66	49.62	59.09	94.21
DMON	**76.30** $\pm$0.71	**68.25** $\pm$0.98	64.13 $\pm$0.47	**97.33** $\pm$0.08	**74.16** $\pm$1.10	**73.16** $\pm$1.17	53.41 $\pm$0.63	**97.27** $\pm$0.06	**74.07** $\pm$1.11	**68.35** $\pm$1.42	54.05 $\pm$0.74	**97.08** $\pm$0.09

Table 1: Experimental results of argument structure learning in terms of macro-F1 scores on three AbstRCT medical datasets. S-F1, A-F1, U-F1 and F1 refer to the average macro-F1 score of the support relation, of the attack relation, of no-relation, and of their average, respectively. For DMON the mean results over 5 runs with variance are shown.

dataset has more fine-grained discourse relationship categories while the number of labels is 27.
CDCP: The CDCP dataset contains 731 user comments about consumer debt collection practices from an eRulemaking website and include 581 examples for training and 150 for testing. Labels of the CDCP dataset are 'related' and 'un-related'.

4.2. Experimental Set-up

Model Settings: The maximum sequence length is 128. The combination of kernel sizes for the multi-scale convolution module is set to $\{k_1, k_2, k_3, k_4, k_5\} = \{7, 5, 5, 3, 1\}$. The window size for the cropping is 13. λ_h and λ_t are set to 0.5. k_1 and k_2 is set as 1 and 2 in the label fusion part.
Training Details: We fine-tune the BioLinkBERT (Yasunaga et al., 2022) for the AbstRCT dataset and LinkBERT (Yasunaga et al., 2022) for the SciDTB and CDCP dataset. For the training, all baseline models and our framework are trained with FP16. We train all models on a *NVIDIA GeForce RTX 3090* GPU. All models use the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning scheduler initialized at $2e^{-5}$ and linearly decreased to 0.

4.3. Baselines

We consider models that classify the argument relationships given a representation of the pairs of sentences obtained with a pretrained encoder.
AMPERE++ (Hua and Wang, 2022) uses a Roberta model to concatenate 20 neighbouring sentences with an argument and only predicts whether this argument is a head or tail argument. The authors did not name this model so we call it AMPERE++.
BERT (Devlin et al., 2019) fine-tunes a pretrained BERT model to encode an argument pair and predict its relationship.
Roberta (Liu et al., 2019) fine-tunes a pretrained Roberta model to represent a pair of sentences and classifies the relationships.
AMCT-Sci (Stylianou and Vlahavas, 2021) is similar to Roberta, but encodes argument pairs with a domain-specific BERT model.[3]
TransforMED (Stylianou and Vlahavas, 2021) integrates a medical knowledge system to extract medical entities from arguments. The authors inject medical knowledge into their model by concatenating features of arguments and medical entities.
We also consider models that use attention mechanisms to model relationships between arguments.
RESARG (Galassi et al., 2021) used a BiLSTM to extract textual feature and then applied a residual neural network to deal with the ASL task.
RESATTARG (Galassi et al., 2021) extended RESARG with a coarse-grained parallel co-attention mechanism to predict argumentative relations.
TSP-PLBA (Morio et al., 2020) consists of task-

Models	Full-F1	F1	R-F1	U-F1
RESARG	-	67.60	38.99	96.20
RESATTARG	-	73.64	50.00	97.28
BERT	32.83	73.89	50.47	97.30
AMCT-Sci	34.58	75.89	54.61	97.17
Roberta	35.96	75.88	54.25	97.51
DMON	**48.37** $\pm$0.54	**87.36** $\pm$0.25	**77.03** $\pm$0.36	**98.70** $\pm$0.02

Table 2: Experimental results of argument structure learning in terms of macro-F1 scores on the SciDTB datasets. For DMON the mean results over 5 runs with variance are shown. Full-F1, R-F1, U-F1 and F1 refer to the average macro-F1 score of the full label space, the related relation, of no-relation, and of related and no-relation, respectively. For DMON the mean results over 5 runs with variance are shown.

[3] https://huggingface.co/allenai/scibert_scivocab_uncased

specific parameterization (TSP) and proposition-level biaffine attention (PLBA) to capture argument structure from documents. TSP encodes the arguments while PLBA predicts argument relations by using a biaffine scoring function.
We then consider models that train a transition-based dependency parser.
BERT-Trans (Bao et al., 2021) leverage the BERT language model to obtain representation and propose a neural transition-based model to generate a sequence of actions (shift, delete-delay, delete, right-arc, right-arc-delay, and left-arc) that build an argument structure (predicted nodes and relations).

Models	F1	R-F1	U-F1
TSP-PLBA	-	34.00	-
AMPERE++	63.10	-	-
RESATTARG	64.40	30.60	98.30
BERT-Trans	67.80	37.30	98.30
DMON	**68.14** $\pm$0.45	**38.26** $\pm$0.74	**98.37** $\pm$0.06

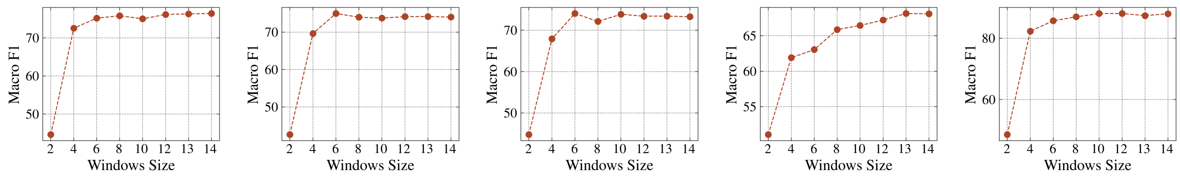
Table 3: Experimental results of argument structure learning in terms of macro-F1 scores on the CDCP datasets. For DMON the mean results over 5 runs with variance are shown. R-F1, U-F1 and F1 refer to the average macro-F1 score of the related relation, of no-relation, and of their average, respectively.

4.4. Results and Discussion

We run each model five times with different seeds and report the average macro-F1 scores and their variance obtained on the three absRCT datasets the CDCP dataset and the SciDTB dataset.
AbsRCT: Table 1 shows that the DMON outperforms all baselines on all average F1 scores. Compared with the state-of-the-art model TransforMED, our model improves the average macro-F1 scores by 6.34, 5.76, and 2.25 percentage points on the Neoplasm, Glaucoma, and Mixed datasets, respectively. Even though TransforMED explicitly injects external medical knowledge, our approach still performs better.
SciDTB: Table 2 shows that the DMON outperforms baseline models by a large margin when evaluated on the SciDTB dataset. Compared to the Roberta model, our model improves the average macro-F1 scores by 12.41, 11.48, and 22.78 percentage points on Full-F1, F1, and R-F1 scores.
CDCP: Table 3 shows that DMON also achieves the best performance in terms of average macro-F1 score on the CDCP dataset. Compared to the BERT-Trans model our method improves the averaged macro-F1 and R-F1 by 0.79 and 1.7 percentage points, respectively. Compared with the BERT-Trans, our model is simple (can be applied to other pairwise classification models) and can achieve better performance. Additionally, BERT is the transformer encoder but the proposed neural transition-based model is to generate a sequence of actions (shift, delete-delay, delete, right-arc, right-arc-delay, and left-arc) that build an argument structure (predicted nodes and relations). To train this transition system, they need to convert argument structure learning data into the transition-based structure data. This preprocessing adds complexity.

Discussion: We observe that the performance gains of DMON oompared to state-of-the-art baselines are different when analyzing the results obtained on the three domains. For instance, when comparing with baseline RESATTARG, DMON improves the macro-F1 scores by 13.92, 5.38, 3.74 percentage points on SciDTB, Neoplasm, and CDCP datasets, respectively. The AbstRCT (of which Neoplasm is a part) and CDCP are imbalanced and have a high ratio of sentence pairs that exhibit no argumentative relationship ("unrelated" relationship). This could be a reason why DMON has somewhat lower performance gains. On the other hand, AbstRCT and CDCP have few relationship types and it might be that the baselines already deal with these in a satisfactory way, while they have more difficulties with the 27 relationship types of SciDTB. For this more difficult case of argumentative structure learning, DMON has the highest gains in performamce and improves the average macro-F1 scores by 12.41, 11.48, and 22.78 percentage points on all 27 relationship type, on the "unrelated" class and on the other 26 relationships, respectively.

Large Language Models: As large language models capture the attention of the NLP communities, we conduct a comparison of LLMs with the proposed algorithm. We pick the GPT 3.5 turbo (gpt-3.5-turbo-0613), a widely used LLM, and validate it on the ASL task. We use the in-context prompt learning (ICL) (Brown et al., 2020) to validate the ASL task. We found the zero-shot ICL is much worse than the few-shot ICL, so we reported the results of the GPT 3.5 turbo by using the few-shot ICL method. Experimental results show the number of the demonstration samples is 2. We report the Macro-F1 scores of the GPT 3.5 on the AbstRCT (gla), AbstRCT (neo), AbstRCT (mix), CDCP and SciDTB datasets, which are 12.03, 13.84, 11.19, 11.89, and 17.24, respectively. Results of the GPT 3.5 turbo are much lower than the fine-tuned models, i.e., our DMON model, whose scores are 73.16, 76.30, 74.07, 87.36, and 68.14. We think the reason why the GPT 3.5 performs badly on the ASL datasets is because LLMs fail to deal with tasks which require complex reasoning ability.



(a) Neoplasm Dataset (b) Glaucoma Dataset (c) Mixed Dataset (d) CDCP Dataset (e) SciDTB Dataset

Figure 5: Macro-F1 scores when changing window size of contextual arguments.

4.5. Ablation Study and Analysis of Results

We conduct ablation experiments to study DMON's components and analyze them.

4.5.1. Bidirectional Learning Mechanism and Confidence Voter

Table 4 shows results when ablating the bidirectional learning mechanism and confidence voter. The F1 scores obtained on all datasets are largely reduced if we remove the bidirectional learning mechanism and confidence voter. The macro-F1 scores on all three datasets (Neoplasm, Glaucoma, and CDCP) decrease when we just use the predictions of one of the two branches (i.e., head or tail relationships) and use no confidence voter. The same pattern holds when we directly remove one of two branches when training the model. The macro-F1 score is reduced on the Mixed dataset when we remove the confidence voter and use tail prediction. The macro-F1 shows the same pattern when we only remove the head branch.

Model	Neo	Gla	Mix	CDCP	SCI
DMON	**76.30**	**74.16**	74.07	**68.14**	**87.36**
w/o Voter (h)	74.31	72.04	73.42	67.50	87.08
w/o Voter (t)	75.74	73.83	74.84	67.50	86.58
w/o T	73.74	70.10	71.40	65.39	85.75
w/o H	75.62	71.48	**74.88**	66.55	87.08
w/o H+T	69.10	71.28	70.35	56.24	74.27

Table 4: Experimental results of argument structure learning in terms of average macro-F1 scores on the Neoplasm (Neo), Glaucoma (Gla), Mixed (Mix), and CDCP datasets. *w/o* Voter (h) or Voter (t) removes the confidence voter and leverages head or tail prediction, respectively. *w/o* T or H removes either the head branch or tail branch when training the model. *w/o* H+T completely removes the bidirectional learning mechanism.

4.5.2. Cropping Strategy

We analyze several aspects of the cropping strategy.

Training with Different Cropped Tensors: Figure 5 compares results of DMON using cropped tensors with the results of using the full relationship tensor as input. We observe that DMON using cropped tensors outperforms DMON with the full tensor on the four datasets. Using the cropping strategy improves the macro-F1 score of our framework by 0.2, 0.95, 4.25, 1.38, and 0.18 percentage points on Neoplasm, Glaucoma, Mixed and SciDTB datasets, respectively. The cropped tensors offer more variation in the training data, and its representations contribute to the generalization capabilities of the model.

Model	Neo	Gla	Mix	CDCP	SCI
DMON$_{\text{win=13}}$	**76.30**	**74.16**	**74.07**	**68.14**	**87.36**
DMON$_{full}$	76.10	73.21	69.82	66.76	87.18

Table 5: Results of the cropping strategy with window size of 13 (**DMON$_{\text{win=13}}$**) and of using the complete relationship tensor (**DMON$_{full}$**) during training.

Information Alignment: To study the information alignment of the cropping with the original relationship tensor, we develop two shuffle approaches called order shuffle (*ord*) and random shuffle (*rad*). *ord* scatters the order of the arguments, that is, horizontal and vertical indexes of the cropped relationship tensor are shuffled. *rad* randomly chooses pairwise samples to fill the cropped relationship tensor. Figure 6 shows an example illustrating the *ord* and *rad* approaches. Table 6 shows that the

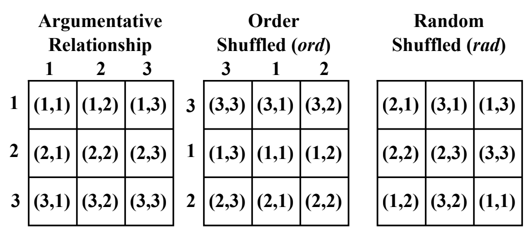


Figure 6: An example that illustrates the order shuffle and random shuffle. For simplicity we only show the relationship matrix.

macro-F1 score decreases by using the *rad* method.

Therefore, it is important to keep head and tail relationships aligned, that is, correctly representing the asymmetric relationships between arguments. Table 6 also reveals that the order of the argumentative sentence pairs in the relationship tensor is important as it implicitly captures the coherence of the discourse.

original	*ord*	*rad*	Neo	Gla	Mix	CDCP	SCI
✓	✗	✗	**76.30**	**74.16**	**74.07**	**68.14**	**87.36**
✗	✓	✗	75.20	68.68	72.44	61.07	75.88
✗	✗	✓	71.43	74.31	72.85	65.86	86.69
✗	✓	✓	59.98	55.08	60.02	50.78	66.27

Table 6: An experiment for exploring the information alignment of the cropping strategy. *original* represents DMON without any shuffling strategies.

Contextual Windows Size: The results demonstrate that encoding contextual arguments and their relationships is beneficial. Figure 5 shows the macro-F1 scores by changing the window size of the cropped relationship tensor for the evaluated datasets. The macro-F1 score curve shows an upward trend by encoding more neighboring arguments and relationships, but slightly decreases when considering the full discourse (Table 5).

5. Conclusion

In this paper, we have proposed a novel framework called the Dual-tower Multi-scale cOnvolution neural Network (DMON) to deal with the ASL task that in a flexible way can learn the argumentative DAG structure taking into account contextual argumentative relationships. A sentence or clause on its own seldom fulfils an argumentative role in a discourse, it is only when paired with another sentence and in the context of other sentences that its argumentative role becomes apparent. In an argumentative DAG structure a sentence can have multiple parents and children, and our model can deal with this flexibility. We conduct experiments on four datasets covering the medical, legal and scientific domains, namely abstRCT, CDCP SciDTB and achieve new state-of-the-art performance, when compared to several strong baselines. Furthermore, we perform ablations and in-depth analyses to prove the effectiveness of each component of our model.

Limitations

The limitation of our paper are reflected as follows: **1)** Computation limitations prevented us from fully exploring the effectiveness of the self-attention mechanism in argument structure learning. In future work, we might leverage more powerful GPUs when implementing a multi-head self-attention mechanism and thoroughly investigate its impact. Nevertheless the performance of our model is better than previous state-of-the-art models that use attention and are computationally more expensive. **2)** We will test whether our bidirectional learning mechanism can be embedded to other pairwise classification models in NLP to improve their performance. In the future, we will leverage balancing approaches, such as focal loss and resampling, to alleviate this problem.

6. Acknowledgements

This research was funded by the CHIST-ERA projects ANTIDOTE (ERA-NET CHIST-ERA IV FET PROACT JTC 2019) and iTRUST (ERA-NET CHIST-ERA JTC 2021).

7. Bibliographical References

Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. 2021. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74.

Jean-Marc Andreoli. 2019. Convolution, attention and structure embedding. *arXiv preprint arXiv:1905.01289*.

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 (Volume 1: Long Papers)*, pages 6354–6364.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. 2022. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11963–11975.

Vincent Dumoulin and Francesco Visin. 2016. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. *arXiv preprint arXiv:1704.06104*.

Liat Ein-Dor, Ilya Shnayderman, Artem Spector, Lena Dankin, Ranit Aharonov, and Noam Slonim. 2022. Fortunately, discourse markers can enhance language models for sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10608–10617.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2018. Argumentative link prediction using residual networks and multi-objective learning. In *Proceedings of the 5th Workshop on Argument Mining*, pages 1–10.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2020. Attention in natural language processing. *IEEE transactions on neural networks and learning systems*, 32(10):4291–4308.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2021. Multi-task attentive residual networks for argument mining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Xinyu Hua and Lu Wang. 2022. Efficient argument structure extraction with transfer learning and active learning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 423–437, Dublin, Ireland. Association for Computational Linguistics.

Wilfried Imrich, Sandi Klavzar, and Douglas F Rall. 2008. *Topics in graph theory: Graphs and their Cartesian product*. CRC Press.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse centric evaluation of machine translation with a densely annotated parallel corpus. *arXiv preprint arXiv:2305.11142*.

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

Fei Li and Hong Yu. 2020. Icd coding from clinical text using multi-filter residual convolutional neural network. In *proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8180–8187.

Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. 2018. Multi-scale residual network for image super-resolution. In *Proceedings of the European conference on computer vision (ECCV)*, pages 517–532.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. 2016. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020*, pages 2108–2115. IOS Press.

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

Marie-Francine Moens. 2013. Argumentation mining: Where are we now, where do we want to be and how do we get there? In *Proceedings of the 5th 2013 Forum on Information Retrieval Evaluation, FIRE '13, New Delhi, India, December 4-6, 2013*, pages 2:1–2:6. ACM.

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*, pages 3259–3266.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, Vancouver, Canada. Association for Computational Linguistics.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

Prakash Poudyal, Jaromír Šavelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. Echr: legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75.

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

Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L Willke. 2018. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 550–564.

Juanjuan Weng, Zhiming Luo, Dazhen Lin, Shaozi Li, and Zhun Zhong. 2023. Boosting adversarial transferability via fusing logits of top-1 decomposed feature. *arXiv preprint arXiv:2305.01361*.

Florian Wolf and Edward Gibson. 2006. *Coherence in natural language: data structures and applications*. MIT Press.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7338–7345.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

An Yang and Sujian Li. 2018. SciDTB: Discourse dependency TreeBank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Haokui Zhang, Wenze Hu, and Xiaoyu Wang. 2022. Parc-net: Position aware circular convolution with merits from convnets and transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 613–630. Springer.