CIARAM: Class Imbalance Aware Generative Framework for Relational Argument Mining

Anonymous ACL submission

Abstract

Relational Argument Mining (RAM) is a key task of computational argumentation, which aims to classify the relationships such as *Support* or *Attack* between argument component (AC) pairs. Traditional approaches primarily rely on graph-based modelling with external knowledge sources, which are complex in nature. Also, these approaches struggle with RAM datasets when relation classes are imbalanced, as they are not designed for class-imbalanced scenarios. In this work, we propose ***CIARAM*** framework to reformulate RAM as a text-to-text generation problem to generate relational labels in a flattened text format. To address the class imbalance, we employ a data augmentation strategy using a decoder-only Large Language Model (LLM) to balance the underrepresented relation classes. Across five standard RAM benchmarks, CIARAM achieves State-of-the-Art (SoTA) results, with ***Macro-F1 score gains ranging from 5.05% to 12.88%***[1], demonstrating the strong potential of our approach.

1 Introduction

Relational Argument Mining (RAM) is a specialized task within computational argumentation that focuses on identifying the relationships between pairs of arguments, as shown in Fig. 1. Specifically, given two arguments, the goal is to determine whether $Arg2$ ***Supports*** $Arg1$ or $Arg2$ ***Attacks*** $Arg1$. Unlike traditional Argument Mining tasks, which primarily extract argumentative components and relations (Lawrence and Reed, 2019), RAM seeks to understand the interplay between arguments. RAM has various potential applications, including online debate (Slonim et al., 2021), legal document interpretation (Habernal et al., 2023), opinion aggregation (Cocarascu and Toni, 2017), scientific literature analysis (Fergadis et al., 2021), etc.

[1]Our code is available here.

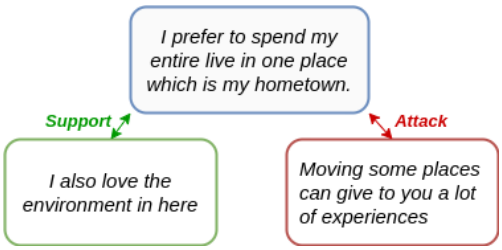

Figure 1: Examples of related argument component pairs taken from Student Essay corpus (Opitz and Frank, 2019) highlighting the *Support* and *Attack* relations.

The primary challenge of RAM is that the relationship between the arguments is often implicit (Saadat-Yazdi et al., 2023), requiring contextual inference. The diversity in linguistic expressions and domain dependency makes generalization difficult (Cabrio and Villata, 2018). On top of that, the standard RAM corpora have a class imbalance, where examples of a certain class are significantly greater than examples of the other class, leading to biased models (Henning et al., 2023). Recent work in (Sun et al., 2022) handles these complexities using graph-based approaches with fine-grained phrase-level similarities (similar words/phrases). Though effective, it overlooks the whole argument-level interaction, where multiple phrase-level interactions are present. More recently, (Saadat-Yazdi et al., 2023) uses the culture-specific (domain-dependent) knowledge from external sources to model the discourse dynamics. However, this external knowledge might not be useful for out-of-distribution data where the culture-specific constraints are different. Additionally, none of the existing RAM methods in the literature has addressed the class imbalance problem. As a result, they often exhibit sub-optimal performance in RAM tasks in class-imbalanced datasets.

With the rise of the generative paradigm, several NLP tasks have been reformulated as text-to-text generation problems, where the input is given as

plain text, and the expected output is structured with a flattened representation of target labels. For example, (Athiwaratkun et al., 2020) solved NER and intent classification problems in a unified target sequence. Specifically, *"((AddToPlaylist)) Add [Kent James | artist] to the [Disney | playlist] soundtrack."* is the target sequence of the original input text *"Add Kent James to the Disney soundtrack."*, where the intent is *"AddToPlaylist"* and the named entities are *"Kent James"* and *"Disney"* of type *"artist"* and *"playlist"* respectively. A similar methodology is applied in (Kawarada et al., 2024) to solve traditional argument mining tasks such as argument component classification and relation classification, which showed improved performances. However, the usability of this flattened representation is unexplored when solving RAM tasks.

In this paper, we propose ***C**lass **I**mbalance **A**ware **R**elational **A**rgument **M**ining*, i.e., **CIARAM**, a simple, yet effective text-to-text generation framework for RAM. The input and output of CIARAM is based on the flattened text representation. It also takes care of the minority classes of the class-imbalanced datasets using a decoder-only LLM. For the class-imbalanced datasets, we take the instances of the majority class and, using LLM, we apply a data augmentation strategy to balance the minority class with the same count as the majority class for that dataset. Thus, CIARAM has three steps: (i) Balancing the minority classes with data augmentation strategy for the class imbalance datasets; (ii) Preparation of flattened representations for both input and output sequences; and (iii) Fine-tuning an encoder-decoder model for the proposed text-to-text generation task with the flattened sequences.

Upon experimentation on five standard diverse-domain RAM datasets including the class-imbalanced ones, CIARAM produces SoTA results with substantial improvements over the existing baselines. In summary:

1. We propose a simple yet effective framework for RAM called, **CIARAM** based on the text-to-text generation paradigm, where both input and output are represented as flattened sequences.

2. By utilizing the data augmentation strategy using an LLM, we mitigate the class-imbalanced problem of the imbalanced datasets to improve the CIARAM performance.

3. Upon performing extensive experiments on diverse domain datasets and an ablation study, we demonstrate the improved performance of CIARAM, producing substantial improvements over the current SoTA.

2 Related Work

Previous works in RAM used transformer-based models like BERT and RoBERTa models to learn contextual representations better (Ruiz-Dolz et al., 2021). Multi-task learning frameworks further enhanced RAM by jointly addressing multiple tasks (Tran and Litman, 2021; Liu et al., 2023). Some models integrated external commonsense knowledge. Examples include ARK and KE-RoBERTa, which leveraged resources such as ConceptNet and WordNet (Paul et al., 2020; Saadat-Yazdi et al., 2023). Graph-based neural networks now capture structural dependencies from PLMs (Sun et al., 2022). Recently, DISARM has used adversarial training and discourse marker detection to further push RAM boundaries despite challenges in cross-domain applicability and small datasets (Contalbo et al., 2024).

3 Methodology

Our methodology consists of three key steps as shown in Fig. 2: (i) Data augmentation to handle minority classes of class-imbalanced datasets, (ii) Preparation of flattened representations for both input & output, and (iii) Fine-tuning an encoder-decoder model for the proposed text-to-text generation task.

Data Augmentation: To address class imbalance in RAM datasets, we use *Llama-3.1-instruct* to generate additional instances for underrepresented relation classes. Given $Arg1$ and $Arg2$ from the majority class, we prompt the model to generate an opposing argument of $Arg2$, to which we call it $Arg3$. As a result, a new minority-class relation is created between $Arg1$ and $Arg3$, holding the exact opposite relation of $Arg1$ and $Arg2$. This process continues until class distribution is balanced. Example instances are shown in Step 1 of Fig. 2. Notably, only the *Support* and *Attack* classes are augmented in imbalanced datasets, while the *No Relation* class remains unchanged. Different relation classes of the datasets, including the augmented ones, are shown in Fig. 3.

Flattened Representation: We propose a structured approach to input and output representations

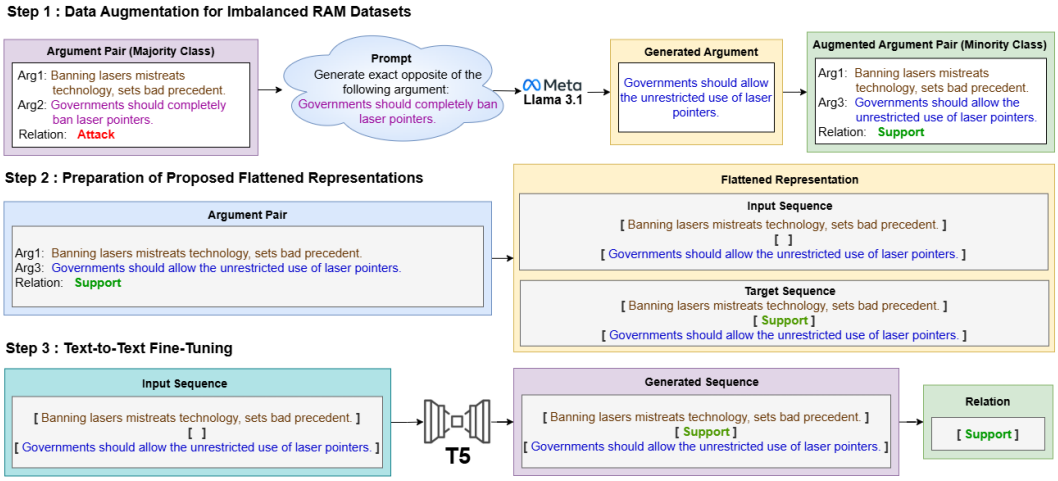

Figure 2: Step-by-step illustration of the **CIARAM** framework.

to solve the RAM task. The input is formatted as [$Arg1$][ ][$Arg2$], while the target output is structured as [$Arg1$][$Relation$][$Arg2$], where [$Relation$] represents the relationship between the $Arg1$ & $Arg2$. Notably, for augmented examples, $Arg3$ is applicable instead of $Arg2$. An illustrative example is given in Step 2 of Fig. 2. This flattened representation gives the model a rich context by presenting $Arg1$ and $Arg2$ in both the input and output. During the generation, the model fills the empty slot of the input "[ ]" with the relation classes in the output sequence.

Text-to-Text Fine-Tuning: Using the flattened input sequence, we fine-tune an encoder-decoder model to generate the flattened output sequence as shown in Step 3 of Fig. 2. During the inference, we post-process the flattened output sequence to extract the corresponding relation class of the related arguments.

4 Experimental Setup

Datasets: We evaluate CIARAM on five publicly available standard RAM datasets: **Student Essay (Essay)** (Opitz and Frank, 2019), **Debatepedia (Debate)** (Paul et al., 2020), **Presidential Debates (M-Arg)** (Mestre et al., 2021), and **Debatepedia-Normative (Normative)** and **Debatepedia-Causal (Causal)** (Jo et al., 2021). Among these, **M-Arg** and **Essay** exhibit class imbalance. Therefore, data augmentation is applied only to these two datasets, while the others remain unchanged. Further details of the datasets are provided in Appendix B.

Implementation Details: We fine-tuned the *Flan-T5-XL* model using the QLoRA adapter for parameter-efficient text-to-text generation with flattened input-output representations. Training was conducted on NVIDIA A100 GPU upon five datasets with a learning rate of 0.0005 and a maximum sequence length of 128 tokens. We used a batch size of 64 for both training and inference, running for 10,000 steps while evaluating every 200 steps to select the best model. Results are averaged over three runs. For all experiments, we consider **Macro-F1 score** as the evaluation metric. Further details on QLoRA hyperparameters are provided in Appendix A.

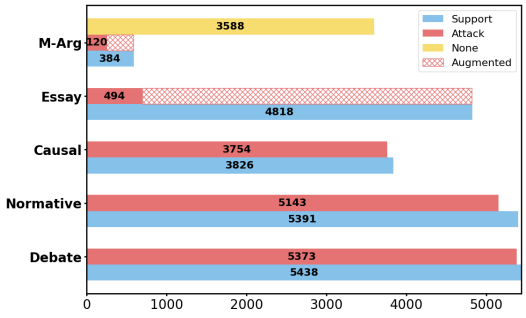

Figure 3: Distribution of different relation classes across the five datasets.

Baselines: We consider the following SoTA models as baselines: **BiLSTM** (Cocarascu and Toni, 2017), **LSTM-ATT** (Ma et al., 2017), **Hybrid-Net** (Chen et al., 2018), **BERT** (Sun et al., 2022), **BERT+LX** (Jo et al., 2021), **BERT+MT** (Jo et al., 2021), **LogBERT** (Jo et al., 2021), **ARK** (Paul et al., 2020), **KE-RoBERTa** (Saadat-Yazdi et al., 2023), **DPGNN** (Sun et al., 2022), **DISARM** (Contalbo et al., 2024). Details of these baselines are described in Appendix C.

Model	Essay	Debate	M-Arg
ARK	60	64	-
KE RoBERTa	70	75	49
RoBERTa+	65.15	74.7	50.37
RoBERTa+ INJ	65.83	74.97	49.35
DISARM (MTL)	69.74	76.14	50.88
DISARM	70.1	76.22	51.34
CIARAM (Ours)	**75.15** (+5.05)	**89.1** (+12.88)	**57.26** (+5.92)

Table 1: Comparison of Macro-F1 scores of CIARAM with existing baselines. Best scores are in **bold**, and improvements over SoTA are marked in **Green**.

Model	Normative	Causal
BiLSTM	71	68.3
LSTM + Att	71.5	70.3
Hybrid Net	67.2	58.8
BERT	79.4	80.7
BERT-LX	78.4	81.5
BERT-MT	79.6	77.5
Log BERT	80.7	80.8
DPGNN	82.9	84.1
CIARAM (Ours)	**93.3** (+10.4)	**94.5** (+10.4)

Table 2: Comparison of Macro-F1 scores of CIARAM with existing baselines. Best scores are in **bold**, and improvements over SoTA are marked in **Green**.

5 Results and Discussion

Main Results: Table 1 presents a performance comparison of CIARAM against existing baselines on the Debate, Essay, and M-Arg datasets, with the latter two being class-imbalanced. Additionally, Table 2 reports results for the Normative and Causal datasets. Across all five datasets, CIARAM achieves SoTA performance, highlighting the advantages of a flattened text-to-text generation approach over traditional methods. The improvements are substantial for class-imbalanced datasets. This demonstrates the effectiveness of the data augmentation strategy in mitigating class imbalance for minority classes.

Verification of Augmented Data: To evaluate the quality of augmented data, we manually verified 10% of the generated arguments from the Essay and M-Arg datasets. The assessment checked contextual validity with the intended relationship (Support or Attack). As shown in Table 3, 87% of Essay and 87.5% of M-Arg arguments generated were contextually valid, indicating high reliability. This reinforces the effectiveness of augmentation in mitigating class imbalance in RAM tasks. Thus, CIARAM achieves SoTA performance even with slightly noisy opposing arguments, showcasing robustness in real-world conditions.

Dataset	Total	Valid	Percentage (%)
Essay (10%)	341	210	87.0
M-Arg (10%)	24	21	87.5

Table 3: Manual verification of augmented opposite arguments generated using *Llama-3.1-instruct*.

Method	Essay	M-Arg
CIARAM *(with Aug)*	**75.1**	**57.2**
CIARAM *(w/o Aug)*	69.7 **(-5.4)**	53.2 **(-4.0)**

Table 4: Ablation Study of CIARAM: *with* and *without* data augmentation on class-imbalanced datasets.

Model	Debate	Essay	M-Arg	Normative	Causal
Flan-T5-XL (Fine-Tuned)	**89.1**	**75.15**	**57.26**	**93.3**	**94.5**
Llama-3.1 (Zero-Shot)	59.10	45.50	25.00	74.40	69.20
(5-shot)	77.29	37.92	34.80	71.99	74.79
(10-shot)	78.83	39.44	36.03	75.28	77.90
(20-shot)	79.29	42.02	34.15	69.88	82.79

Table 5: Performance Comparison of the RAM Task: Fine-Tuned *Flan-T5-XL* vs. Zero/Few-Shot *Llama-3.1*.

Ablation Study: To evaluate the impact of data augmentation, we compare CIARAM's performance *with* and *without* augmented data on the imbalanced Essay and M-Arg datasets. As shown in Table 4, removing augmented data causes a significant drop in Macro-F1 by 5.4 and 4.0 points for Essay and M-Arg, respectively. This highlights the positive effect of data augmentation in addressing class imbalance to enhance performance.

Zero/Few-shot vs Fine-tuning: According to Table 5, although zero/few-shot performance of RAM task using *Llama-3.1-instruct* improves with more examples, it consistently falls short of fine-tuned *Flan-T5-XL*, which outperforms it by a significant margin across all datasets. Details of *Llama-3.1-instruct* prompts are given in Appendix D.

6 Conclusion

This paper presents **CIARAM**, a simple yet efficient framework for RAM that leverages the text-to-text generation paradigm, representing both input and output as flattened sequences. To tackle class imbalance in standard RAM datasets, we incorporate a data augmentation strategy using an LLM, boosting CIARAM's performance. Through extensive experiments and an ablation study, we show that CIARAM delivers strong performance, significantly outperforming the current SoTA.

7 Limitations and Future Scope

One key challenge is the potential for generative models to introduce hallucinations, generating arguments that do not accurately reflect the original stance. Additionally, while Llama-based augmentation improves class balance, it may introduce artifacts that do not fully capture natural argumentation patterns. We used *Flan-T5-XL* as our base model. Exploring other encoder-decoder models, such as BART, could provide insights into their performance within the current setup.

References

Ben Athiwaratkun, Cicero Nogueira dos Santos, Jason Krone, and Bing Xiang. 2020. Augmented natural language for generative sequence labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5427–5433. International Joint Conferences on Artificial Intelligence Organization.

Di Chen, Jiachen Du, Lidong Bing, and Ruifeng Xu. 2018. Hybrid neural attention for agreement/disagreement inference in online debates. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 665–670, Brussels, Belgium. Association for Computational Linguistics.

Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379, Copenhagen, Denmark. Association for Computational Linguistics.

Michele Luca Contalbo, Francesco Guerra, and Matteo Paganelli. 2024. Argument relation classification through discourse markers and adversarial training. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18949–18954, Miami, Florida, USA. Association for Computational Linguistics.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2023. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, 32(3):1–38.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540, Dubrovnik, Croatia. Association for Computational Linguistics.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. 2021. Classifying argumentative relations using logical mechanisms and argumentation schemes. *Transactions of the Association for Computational Linguistics*, 9:721–739.

Masayuki Kawarada, Tsutomu Hirao, Wataru Uchida, and Masaaki Nagata. 2024. Argument mining as a text-to-text generation task. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, St. Julian's, Malta. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Boyang Liu, Viktor Schlegel, Riza Batista-Navarro, and Sophia Ananiadou. 2023. Argument mining as a multi-hop generative machine reading comprehension task. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10846–10858, Singapore. Association for Computational Linguistics.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. *Preprint*, arXiv:1709.00893.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Juri Opitz and Anette Frank. 2019. Dissecting content and context in argumentative relation analysis. In *Proceedings of the 6th Workshop on Argument Mining*, pages 25–34, Florence, Italy. Association for Computational Linguistics.

Debjit Paul, Juri Opitz, Maria Becker, Jonathan Kobbe, Graeme Hirst, and Anette Frank. 2020. Argumentative relation classification with background knowledge. In *Comma*.

Ramon Ruiz-Dolz, Jose Alemany, Stella M. Heras Barbera, and Ana Garcia-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(6):62–70.

Ameer Saadat-Yazdi, Jeff Z. Pan, and Nadin Kokciyan. 2023. Uncovering implicit inferences for improved relational argument mining. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2484–2495, Dubrovnik, Croatia. Association for Computational Linguistics.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen, Lena Dankin, Lilach Edelstein, Liat Ein Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, and Ranit Aharonov. 2021. An autonomous debating system. *Nature*, 591:379–384.

Yang Sun, Bin Liang, Jianzhu Bao, Min Yang, and Ruifeng Xu. 2022. Probing structural knowledge from pre-trained language model for argumentation relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3605–3615, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nhat Tran and Diane Litman. 2021. Multi-task learning in argument mining for persuasive online discussions. In *Proceedings of the 8th Workshop on Argument Mining*, pages 148–153, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Hyperparameters

We use the following hyperparameters setting for fine-tuning with QLoRA:

Parameter	Value
r	16
lora alpha	32
lora dropout	0.05
bias	none
task type	SEQ_2_SEQ_LM
target modules	$q, v, k, o, wo, wi_0, wi_1$
load_in_4bit	True
bnb_4bit_quant_type	nf4
bnb_4bit_use_double_quant	True
bnb_4bit_compute_dtype	torch.bfloat16

Table 6: Hyperparameter setting of QLoRA

B Dataset Description

The description of the five publicly available standard RAM datasets are given as follows:

- **Student Essay (Essay)** (Opitz and Frank, 2019): A corpus of argumentative essays written by second-language speakers, annotated with *attack/support* relations.

Dataset	Train	Dev	Test
Essay	3,070	1,142	1,100
Debate	6,486	2,163	2,162
M-Arg	3,283	410	411
Normative	11,098	472	707
Causal	6,581	496	330

Table 7: Dataset statistics.

- **Debatepedia (Debate)** (Paul et al., 2020): A dataset of structured arguments extracted from Debatepedia, containing pro/con arguments on controversial topics, following a *binary classification scheme (attack/support)*.

- **Presidential Debates (M-Arg)** (Mestre et al., 2021): Transcripts from U.S. presidential debates, annotated with three classes: *support, attack, and neutral*.

- **Debatepedia-Normative (Normative) and Debatepedia-Causal (Causal)** (Jo et al., 2021): Two subcorpora derived from Debatepedia, containing argument pairs categorized based on normative and causal reasoning. These datasets follow a *binary classification scheme (support/attack)*.

C Details of Baselines

We compare our proposed approach with several SoTA models, including both traditional machine learning and deep learning-based methods:

- **BiLSTM** (Cocarascu and Toni, 2017): A dual BiLSTM architecture to encode argument component (AC) pairs independently.

- **LSTM-ATT** (Ma et al., 2017): An LSTM with interaction-based attention to enhance AC pair representations.

- **Hybrid-Net** (Chen et al., 2018): A BiLSTM-based model incorporating self- and cross-attention for better argument pair modeling.

- **BERT** (Sun et al., 2022): A vanilla BERT model that uses the `[CLS]` token representation for classification.

- **BERT+LX** (Jo et al., 2021): A BERT-based model that incorporates external linguistic features such as *factual consistency and sentiment coherence*.

- **BERT+MT** (Jo et al., 2021): A multitask learning-based approach using ARC jointly

with textual entailment and sentiment classification.

- **LogBERT** (Jo et al., 2021): A variation of BERT pre-trained on logical reasoning tasks before fine-tuning on ARC.
- **ARK** (Paul et al., 2020): A method that employs a cross-attention mechanism with BiLSTMs and integrates external commonsense knowledge from ConceptNet and WordNet for enhanced argument relation classification.
- **KE-RoBERTa** (Saadat-Yazdi et al., 2023): A knowledge-enhanced RoBERTa model that incorporates *commonsense reasoning from external knowledge graphs*.
- **DPGNN** (Sun et al., 2022): A *dual prior graph neural network* that integrates syntactic dependencies and probing knowledge from *pre-trained language models (PLMs)* for fine-grained argument relation classification.
- **DISARM** (Contalbo et al., 2024): A RoBERTa-based approach that combines *multi-task learning and adversarial training* by aligning *argument relation classification (ARC) and discourse marker detection (DMD)* into a unified latent space. DISARM utilizes the *Discovery dataset* to learn discourse marker-based representations that improve ARC performance.

D Zero/Few-Shot Prompt Details

We did not perform an extensive search for the optimal prompt, as finding the most effective prompt is challenging. Instead, we used the same input-output format as the text-to-text generation model to construct the zero/few-shot prompts. Details are shown below:

D.1 Zero-Shot Prompt

For the zero-shot setting, the prompt consisted only of the input format without any example demonstrations:

```
  Classify the relationship between the arguments in the format [Arg1][Rel][Arg2].
Use only one of these labels: Support, Attack.

Example Format:
Input: [Arg1][][Arg2]
Output: [Arg1][Rel][Arg2]

Real Example
Input: [without the cooperation , there would be no victory of competition][][we should attach more importance to cooperation during primary education]
Output:

NOTE: Only give the output in the same format. No unnecessary texts or explanations please.
```

D.2 Few-Shot Prompts

For the few-shot settings, we included k examples ($k = 5, 10, 20$) demonstrating the relationship between arguments before providing the test instance. The format for the few-shot prompts was as follows:

D.2.1 Two-Class Task (Support/Attack)

Each prompt contained k examples drawn from the training data, formatted as follows:

```
  Classify the relationship between the arguments in the format [Arg1][Rel][Arg2].
Use only one of these labels: Support, Attack.

Example Format:
Input: [Arg1][][Arg2]
Output: [Arg1][Rel][Arg2]

Here are some examples:

Example 1:
Input: [Now he says I should have closed it earlier.][][I didn't say either of those things.]
Output: [Now he says I should have closed it earlier.][Attack][I didn't say either of those things.]

Example 2:
Input: [Leadership crisis is only worsened by not passing $700b plan][][The $700 billion bailout plan for the 2008 US financial crisis is a good idea.]
Output: [Leadership crisis is only worsened by not passing $700b
```

plan][Support][The $700 billion bailout plan for the 2008 US financial crisis is a good idea.]

...

Example 19:
Input: [Right of return jeopardizes Israeli welfare, so invalid][][The Palestinians have the right to return.]
Output: [Right of return jeopardizes Israeli welfare, so invalid][Attack][The Palestinians have the right to return.]

Example 20:
Input: [More of a right to leave than right to return.][][The Palestinians have the right to return.]
Output: [More of a right to leave than right to return.][Attack][The Palestinians have the right to return.]

Real Example

Input: [without the cooperation , there would be no victory of competition][][we should attach more importance to cooperation during primary education]
Output:

NOTE: Only give the output in the same format. No unnecessary texts or explanations please.

The number of examples varied based on the setting ($k = 5, 10, 20$).

D.2.2 Three-Class Task (Support/Attack/None)

For the three-class task, the format remained the same, with an additional class label None:

Classify the relationship between the arguments in the format [Arg1][Rel][Arg2].
Use only one of these labels: Support, Attack, None.

Example Format:
Input: [Arg1][][Arg2]
Output: [Arg1][Rel][Arg2]

Here are few examples:

Example 1:
Input: [It's a fact.][][It's been totally discredited.]
Output: [It's a fact.][Attack][It's been totally discredited.]

Example 2:
Input: [We have an election coming up.][][You think she would rule for you?]
Output: [We have an election coming up.][None][You think she would rule for you?]

...

Example 19:
Input: [it is necessary to make sure that people can live a long life][][animal experiments have negative impact on the natural balance]
Output: [it is necessary to make sure that people can live a long life][Attack][animal experiments have negative impact on the natural balance]

Example 20:
Input: [Now he says I should have closed it earlier.][][I didn't say either of those things.]
Output: [Now he says I should have closed it earlier.][Attack][I didn't say either of those things.]

Real Example

Input: [students learn far more from other sources , such as the Internet and television][][students learn far more from their teachers than from other source]
Output:

NOTE: Only give the output in the same format. No unnecessary texts or explanations please.

As before, k varied between 5, 10, and 20 based on the setting.

These prompts were used to evaluate the impact of few-shot learning on classification performance.