KNOWCOMP POKEMON Team at DialAM-2024: A Two-Stage Pipeline for Detecting Relations in Dialogical Argument Mining

Zihao Zheng[1], Zhaowei Wang[2], Qing Zong[2], Yangqiu Song[2],
[1]Harbin Institute of Technology(Shenzhen), Guangdong, China
[2]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
{melfeszheng, zongqing0068}@gmail.com, {zwanggy, yqsong}@cse.ust.hk

Abstract

Dialogical Argument Mining (**DialAM**) is an important branch of Argument Mining (**AM**). DialAM-2024 is a shared task focusing on dialogical argument mining, which requires us to identify argumentative relations and illocutionary relations among proposition nodes and locution nodes. To accomplish this, we propose a two-stage pipeline[1], which includes the Two-Step S-Node Prediction Model in Stage 1 and the YA-Node Prediction Model in Stage 2. We also augment the training data in both stages and introduce context in Stage 2. We successfully completed the task and achieved good results. Our team **KNOWCOMP POKEMON** ranked **1*st*** in the ARI Focused score and **4*th*** in the Global Focused score.

1 Introduction

Dialogues contain a wealth of information about arguments and their relationships, but the structure and content of dialogues are casual, which poses challenges for extracting argument structures. To handle it, Budzynska et al. (2014) provides a method for analyzing dialogue and argument structures, as well as the relations between them, using Inference Anchoring Theory (IAT) (Budzynska and Reed, 2011). In dialogues, the content of the discussions serves as locution nodes, while their propositional content serves as proposition nodes. Among these nodes, three types of relation nodes are used for connection: argumentative relations between propositions, illocutionary relations between locutions and propositions, and transitional relations between locutions. This method helps extract argument structures from dialogues, enabling further argument mining and analysis. By employing this approach, Hautli-Janisz et al. (2022) has introduced QT30, an English corpus of meticulously analyzed dialogical argumentation. This corpus encompasses the argumentative structure derived from 30 debates from the BBC television program Question Time.

The DialAM task in ACL2024 (Ruiz-Dolz et al., 2024) is the first shared task focused on dialogical argument mining. It consists of two tasks. The first task is to identify Propositional Relations, aiming to detect argumentative relations between the identified and segmented propositions in the argumentative dialogue. The second task is the Identification of Illocutionary Relations, which aims to detect the illocutionary relations between the locutions uttered in the dialogue and the argumentative propositions associated with them.

To address the two tasks proposed by DialAM-2024, we introduce a two-stage pipeline. Based on initial locutions and propositional contents, we utilize data augmentation by adding data that does not fit any relation in the relation set to increase the gap between data within and outside the relation set. Thus, we can predict the relationships between propositional contents using our proposed two-step S-node prediction model to address the first task. Building upon this, we further tackle the task of identifying illocutionary relations by bringing context to prediction and employing a multi-classification YA-node prediction model. Adopting this method, our team **Pokemon** ranked **1*st*** in the ARI Focused score and **4*th*** in the Global Focused score.

Our paper is structured as follows: Section 2 presents related work on argument mining. Section 3 describes the details of our proposed method, a two-stage pipeline. Section 4 outlines the experiments we conducted, including the models and methods used in each stage, as well as the overall pipeline experiments. Section 5 makes a conclusion and provides further discussion.

[1]Codes are avilable at https://github.com/HKUST-KnowComp/KnowComp-DialAM2024-ACL2024

Proceedings of the 11th Workshop on Argument Mining, pages 110–118

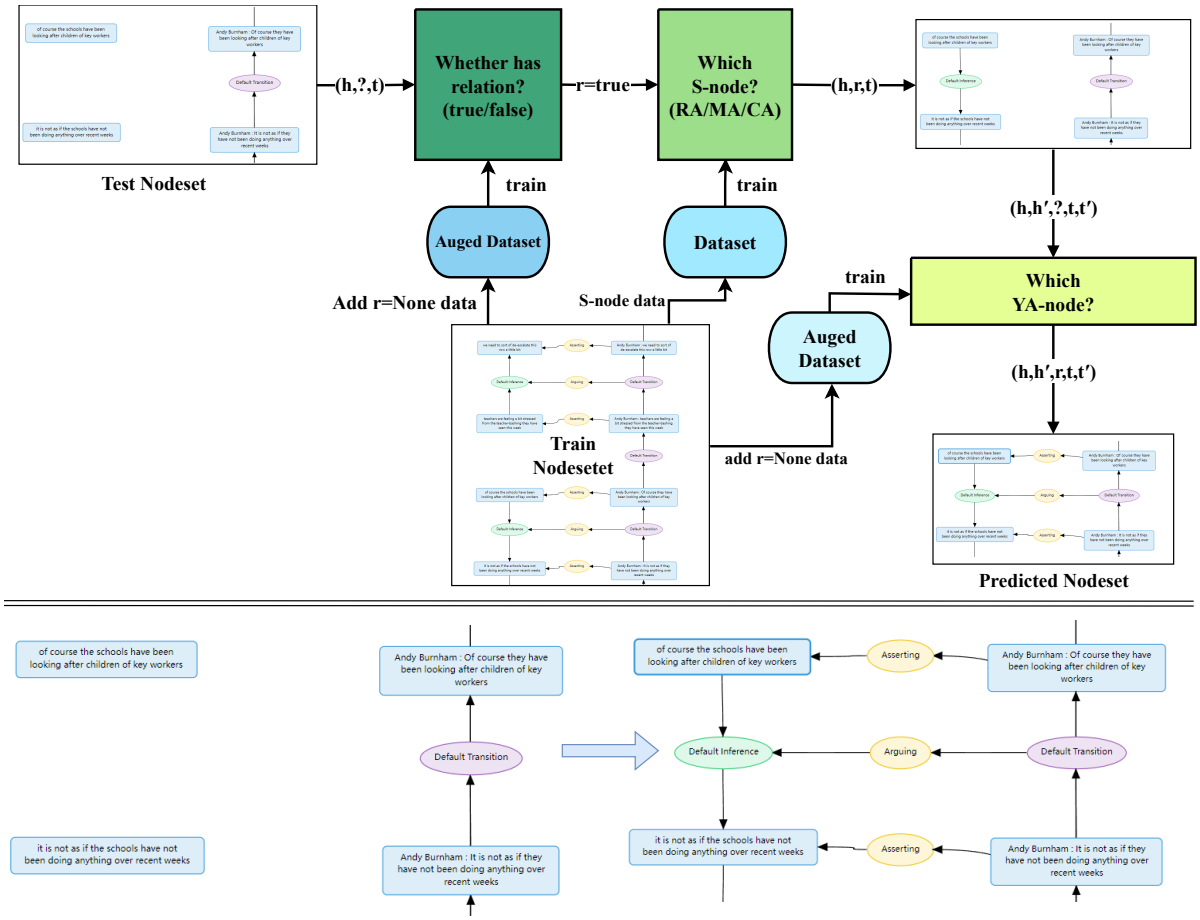

Figure 1: The 2-stage Pipeline.

2 Related Work

Argument Mining: Argument Mining involves the automatic extraction and analysis of arguments from various sources, such as texts, debates, and social media discussions (Stab and Gurevych, 2014; Habernal and Gurevych, 2017; Carlile et al., 2018; Lawrence and Reed, 2019). Some recent works study the stance and persuasiveness of the arguments in multi-modal data like tweets on Twitter (Liu et al., 2022; Zong et al., 2023b). Other works focus on dialogical argumentation, exploring how arguments are put forward, supported, and attacked through dialogue (Haddadan et al., 2019; Visser et al., 2020). QT30 corpus (Hautli-Janisz et al., 2022), which is built on Inference Anchoring Theory (IAT) (Budzynska and Reed, 2011), a prominent framework in manual argument analysis, is the largest dialogical argumentation corpus in English.

3 Method

We have developed a pipeline (Fig. 1) to address the challenge of dialogical argument mining. This pipeline consists of two stages designed to address the task of identifying propositional relations and illocutionary relations, respectively.

3.1 Two-Step S-node Prediction Model

Our primary objective in the first stage is to detect argumentative relations between propositions (I-node). According to QT30 (Hautli-Janisz et al., 2022), This kind of relation (S-node) consists of Inference (RA-node), Rephrase (MA-node), and Conflict (CA-node). However, it is worth noting that not all I-node pairs have relations. Consequently, an initial determination should be made regarding the presence of a relation between two given I-nodes, followed by a secondary prediction of the specific scheme of the relation. This binary step-wise approach forms the foundation of our two-step prediction model.

Inspired by the approach proposed by Parikh et al. (2016), we adopt a similar representation

using pairs to denote our problems. Specifically, for any two distinct I-nodes denoted as h and t, wherein h represents the head node and t the tail node, the task is to predict the relation r between h and t given the tuple (h, t) and subsequently deriving the final triple (h, r, t).

The first step of determining relation existence is framed as a binary classification task, given the pair (h, t), with the relation set $R = \{true, false\}$. The principle of cross-entropy loss shapes the loss function of the model.

Similarly, the second step of ascertaining the specific relation between the I-nodes is structured as a ternary classification task, with the relation set $R = \{RA, CA, MA\}$.

3.2 YA-node Prediction Model

The illocutionary relations (YA-node) include (11 distinct types in total): 1) Asserting, Challenging, Pure Questioning, Assertive Questioning, Rhetorical Questioning between I-nodes and L-nodes, 2) Arguing, Disagreeing, Default Illocuting, Restating between TA-nodes and S-nodes, and 3) Agreeing, Challenging, Disagreeing between TA-nodes and I-nodes (Hautli-Janisz et al., 2022). The relationship between L-node and I-node is relatively direct, indicating an illocutionary relation between locutions and their propositional content. However, for the occasion where YA-nodes are connected to TA-nodes or S-nodes, since TA-nodes and S-nodes themselves do not have much meaning when considered alone, we take the context into account, that is, considering two L-nodes connected by TA-nodes and two or more I-nodes connected by S-nodes.

Our task still remains to predict the relation r between the given head node h and tail node t. Additionally, the head and tail nodes may be followed by their respective contexts h' and t'.

This is also a multi-classification task to predict the illocutionary relation r given (h, h', t, t'). The relation set $R = \{r_0, r_1, r_2, ..., r_{11}\}$, where r_0 indicates there's no illocutionary relation between the node pairs. The model's loss function is cross-entropy loss.

3.3 Data Augmentation

While we have discussed the pipeline of our framework in the above two sections (i.e., Section 3.1 and Section 3.2), we also introduced data augmentation techniques to further improve the performance of fine-tuned models in our framework.

Within the training dataset of the first step of the first stage, I-node pairs already connected by S-nodes are categorized as $r = true$. It becomes imperative to introduce $r = false$ data manually. To this end, a set number of I-node pairs without S-node connections are randomly selected to represent the training data for $r = false$. Specifically, in each nodeset within our training set, we randomly select some node pairs from all possible I-node pairs. These selected I-node pairs must satisfy the condition that there is no S-node connecting them. We think that there are no significant argumentative relations between these selected I-node pairs. Meanwhile, the training dataset for the second step is solely comprised of I-node pairs with established S-node connections, but the connections are further categorized into RA, MA, and CA.

In the training set of the YA-node prediction model of the second stage, in addition to the tuples (h, h', r_{1-11}, t, t') that already have YA-node connections as training data, a certain number of tuples (h, h', r_0, t, t') need to be extracted from node pairs that do not have YA node connections, artificially created as training data with $r = r_0$, i.e., $r = None$.

4 Experiments

4.1 Setup

The baseline models we employed include DeBERTa-base (He et al., 2021), DeBERTa-large, DeBERTa-MNLI, RoBERTa-MNLI (Liu et al., 2019). We also tried LLaMa-3-8B (AI@Meta, 2024) with LoRa (Hu et al., 2022).

The learning rate during training is 1e-5, the weight decay is 0.01, and fp16 is enabled during the training process. When utilizing Lora, the parameter r is set to 64, and alpha is set to 16. Due to time constraints, the testing of other LoRa parameters was not completed.

Our dataset comprises a total of 1,478 nodesets. We randomly selected 78 nodesets as the evaluation set, leaving the remaining 1,400 nodesets for the training set. A more detailed data description is in appendix D.

4.2 Experimental Results of S-node Prediction

First, we artificially generated a certain amount of $r = false$ data in this step and evaluated the impact of this additional data volume. Therefore, we performed experiments by controlling the ratio

Model	General Metrics			Focused Metrics		
	precision	recall	f1	precision	recall	f1
RoBERTa-MNLI	0.114	0.369	0.046	0.494	0.533	0.488
DeBERTa-large	0.099	0.376	0.050	0.511	0.548	0.503
LLaMa-3-8B-LoRa	0.100	0.289	0.018	0.261	0.432	0.315
DeBERTa-large	0.351	0.443	0.322	0.351	0.266	0.282
RoBERTa-MNLI	0.317	0.470	0.306	0.449	0.334	0.355

Table 1: Experiments on different methods of the first stage of S-node prediction. The two models in the lower part of the table are the 2nd-step models, while the four models in the upper part are four-label classification models.

Model	General Metrics			Focused Metrics		
	precision	recall	f1	precision	recall	f1
DeBERTa-large	0.746	0.862	0.784	0.757	0.760	0.753
RoBERTa-MNLI	0.650	0.772	0.691	0.834	0.842	0.834
DeBERTa-MNLI	0.627	0.744	0.667	0.823	0.830	0.823

Table 2: Experiments on the second stage of YA-node prediction.

Type	General Metrics			Focused Metrics		
	precision	recall	f1	precision	recall	f1
ARI	0.463	0.324	0.359	0.320	0.466	0.306
ILO	0.542	0.499	0.514	0.564	0.646	0.594

Table 3: The result of our submitted system

of the amount of $r = false$ data to the amount of $r = true$ data to observe the results.

Moreover, we experimented with a four-label direct classification model and compared the results with those of the two-step model we ultimately employed.

The results of the first experiment are shown in the appendix A. Based on the experimental results, the 1:1 data ratio produced the best outcome. We believe that the 1st-step model only needs to determine whether a relationship exists without considering factors such as the distribution of various relationships that the 2nd-step model should concern. Therefore, the 1:1 data ratio makes it easier for the model to distinguish the differences between $r = true$ and $r = false$ data.

The results of the second experiment are shown in Table 1. Our two-step model framework uses the *DeBERTa-base-1* model, which had the best performance in the first experiment, as the 1st-step model. It can be observed that the models trained directly for four-class classification achieve higher focused scores but have very low general scores. On the other hand, our two-step model achieves a significant improvement in general scores at the expense of sacrificing some focused scores. Overall, the two-step method yields better results.

4.3 Experimental Results of Y-node Prediction

We tested the performance of different models in Stage 2. In the experiments of this stage, we trained 12-label classification models. In addition to the training data for the 11 labels extracted from the nodesets, inspired by the experiments in the previous stage, we also included an equal amount of $r = None$ data in training.

The experimental results are shown in Table 2. Most of the models had higher Focused scores than General scores. Among them, DeBERTa-large received the highest General score, whereas RoBERTa-MNLI achieved the highest Focused score.

4.4 Experimental Results of the Pipelines

The composition of the pipeline submitted by us in DialAM-2024 is as follows: DeBERTa-base + RoBERTa-MNLI as the first stage model, and DeBERTa-large as the second stage model. The result is shown in Table 3. Our pipeline achieved first place in the ARI Focused score and fourth place in the Global Focused score.

We also modified the models in stage 1 and stage 2 and tested these different pipelines on the test dataset, which was finally released by DialAM-2024. The results are presented in appendix C, and we found that we have achieved a much higher score, with the ILO-focused scores surpassing 0.87.

5 Conclusion

We propose a two-stage pipeline that predicts argumentative relations and illocutionary relations based on the initial locutions and propositions. This method utilizes data augmentation to optimize the training data and employs a two-step model to predict the relations, incorporating contextual information during prediction. Ultimately, our method achieves good performance in the DialAM24 shared task.

However, due to time constraints and limited computational resources, there are still many aspects of our method that have not been fully optimized. For example, we could appropriately incorporate additional information in locutions to assist the prediction process. It is also worth exploring the possibility of first determining the correspondence between locutions and propositions before predicting the remaining relations. These areas can be further explored and researched.

Limitations

In this paper, we design a pipeline that utilizes knowledge of language models, like T5 and DeBERTa, to solve this argument mining problem. For LLMs, we only tested Llama3 (8B) (AI@Meta, 2024) by fine-tuning a small fraction of parameters. For future works, we can try more LLMs, like Llama2 (Touvron et al., 2023) and Mistral (Jiang et al., 2023) with more sizes (e.g., 13B, 70B). Meanwhile, we can augment our argument-mining pipeline with various external knowledge, including commonsense knowledge (Sap et al., 2019; Do et al., 2024; Deng et al., 2023; Wang et al., 2024a; Wu et al., 2023) event-centric knowledge (Wang et al., 2022, 2023; Fang et al., 2024; Wang et al., 2024c,b; Fan et al., 2023) and factual knowledge (Choi et al., 2023). More importantly, we can also add more modalities like images for relation detection in dialogical argument mining (Zong et al., 2023a; Shen et al., 2024).

Acknowledgement

The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20), and the GRF (16211520 and 16205322) from RGC of Hong Kong. This paper was also supported by the Tencent AI Lab Rhino-bird Focused Research Program. We also thank the UGC Research Matching Grants (RMGS20EG01-D, RMGS20CR11, RMGS20CR12, RMGS20EG19, RMGS20EG21, RMGS23CR05, RMGS23EG08).

References

AI@Meta. 2024. Llama 3 model card.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska-Shah. 2014. Towards argument mining from dialogue. *Frontiers in Artificial Intelligence and Applications*, 266:185–196.

Katarzyna Budzynska and Chris Reed. 2011. Speech acts of argumentation: inference anchors and peripheral cues in dialogue. In *Proceedings of the 10th AAAI Conference on Computational Models of Natural Argument*, AAAIWS'11-10, page 3–10. AAAI Press.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. 2023. Kcts: Knowledge-constrained tree search decoding with token-level hallucination detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14035–14053.

Zheye Deng, Weiqi Wang, Zhaowei Wang, Xin Liu, and Yangqiu Song. 2023. Gold: A global and local-aware denoising framework for commonsense knowledge graph noise detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3591–3608.

Quyet V Do, Tianqing Fang, Shizhe Diao, Zhaowei Wang, and Yangqiu Song. 2024. Constraintchecker: A plugin for large language models to reason on commonsense knowledge bases. *arXiv preprint arXiv:2401.14003*.

Wei Fan, Weijia Zhang, Weiqi Wang, Yangqiu Song, and Hao Liu. 2023. Chain-of-choice hierarchical policy learning for conversational recommendation. *arXiv preprint arXiv:2310.17922*.

Tianqing Fang, Zhaowei Wang, Wenxuan Zhou, Hongming Zhang, Yangqiu Song, and Muhao Chen. 2024. Getting sick after seeing a doctor? diagnosing and mitigating knowledge conflicts in event temporal reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3846–3868.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4684–4690. Association for Computational Linguistics.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. QT30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 3291–3300. European Language Resources Association.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

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*.

Zhexiong Liu, Meiqi Guo, Yue Dai, and Diane Litman. 2022. ImageArg: A multi-modal tweet dataset for image persuasiveness mining. In *Proceedings of the 9th Workshop on Argument Mining*, pages 1–18, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

Ramon Ruiz-Dolz, John Lawrence, Ella Schad, and Chris Reed. 2024. Overview of DialAM-2024: Argument Mining in Natural Language Dialogues. In *Proceedings of the 11th Workshop on Argument Mining*, Thailand. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Xiangqing Shen, Yurun Song, Siwei Wu, and Rui Xia. 2024. Vcd: Knowledge base guided visual commonsense discovery in images. *arXiv preprint arXiv:2402.17213*.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020. Argumentation in the 2016 US presidential elections: annotated corpora of television debates and social media reaction. *Lang. Resour. Evaluation*, 54(1):123–154.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Jiayang Cheng, et al. 2024a.

Candle: iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. *arXiv preprint arXiv:2401.07286*.

Zhaowei Wang, Quyet V Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Y Wong, and Simon See. 2023. Cola: contextualized commonsense causal reasoning from the causal inference perspective. *arXiv preprint arXiv:2305.05191*.

Zhaowei Wang, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Y Wong, and Simon See. 2024b. Absinstruct: Eliciting abstraction ability from llms through explanation tuning with plausibility estimation. *arXiv preprint arXiv:2402.10646*.

Zhaowei Wang, Haochen Shi, Weiqi Wang, Tianqing Fang, Hongming Zhang, Sehyun Choi, Xin Liu, and Yangqiu Song. 2024c. Abspyramid: Benchmarking the abstraction ability of language models with a unified entailment graph. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3991–4010.

Zhaowei Wang, Hongming Zhang, Tianqing Fang, Yangqiu Song, Ginny Wong, and Simon See. 2022. Subeventwriter: Iterative sub-event sequence generation with coherence controller. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1590–1604.

Siwei Wu, Xiangqing Shen, and Rui Xia. 2023. Commonsense knowledge graph completion via contrastive pretraining and node clustering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13977–13989.

Qing Zong, Zhaowei Wang, Baixuan Xu, Tianshi Zheng, Haochen Shi, Weiqi Wang, Yangqiu Song, Ginny Wong, and Simon See. 2023a. Tilfa: A unified framework for text, image, and layout fusion in argument mining. In *Proceedings of the 10th Workshop on Argument Mining*, pages 139–147.

Qing Zong, Zhaowei Wang, Baixuan Xu, Tianshi Zheng, Haochen Shi, Weiqi Wang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023b. TILFA: A unified framework for text, image, and layout fusion in argument mining. In *Proceedings of the 10th Workshop on Argument Mining, ArgMining 2023, Singapore, December 7, 2023*, pages 139–147. Association for Computational Linguistics.

A Experiments on different data ratios

We conducted experiments using DeBERTa models, with the numbers following the model name indicating the data ratio, i.e., the ratio between the amounts of $r = false$ and $r = true$ data. The results are showin in Table 4.

B Full Experiments of Y-node Prediction

The results are showin in Table 5. Except for the *LLaMa-3-8B* model trained with LoRa, which performed significantly worse, the other models achieved high scores. We speculate that *LLaMa-3-8B* model may not be well-suited for this multi-classification task compared to these smaller models specifically designed for this. Most of the models had higher Focused scores than General scores. Among them, DeBERTa-large received the highest General score, whereas RoBERTa-MNLI achieved the highest Focused score.

C Experiments of the pipelines

We modified the model in the second step of stage 1, as well as the model in stage 2, and tested the performance of these different pipelines. The results are shown in the Table 6.

To our surprise, the second pipeline, DeBERTa-base + RoBERTa-MNLI + RoBERTa-MNLI, which performed slightly worse on the evaluation set, obtained the highest score in the test set. Its ILO score was significantly higher than the score of the pipeline we submitted.

We speculate that this might be because our evaluation set consisted of only 78 randomly selected nodesets from the training dataset, which could have significant differences in data distribution and relationship distribution compared to the final test set. As a result, the pipeline that performed best on the validation set may have had poorer performance on the test set, while some pipelines that performed slightly worse on the validation set happened to achieve better scores on the test set.

D Additional Data Description

Our dataset comprises a total of 1,478 nodesets. We randomly selected 78 nodesets as the evaluation set, leaving the remaining 1,400 nodesets for the training set.

The training set contains 5,365 RA data samples, 1,181 CA data samples, 5,596 MA data samples, and 32,626 YA data samples. In the evaluation set, there are 268 RA data samples, 59 CA data samples, 279 MA data samples, and 1,631 YA data samples.

The selected 78 nodesets are: 'nodeset18321', 'nodeset21402', 'nodeset21463', 'nodeset23939', 'nodeset18455', 'nodeset19912', 'nodeset23828', 'nodeset21575', 'nodeset17918', 'nodeset23771', 'nodeset21041', 'nodeset18846', 'nodeset18850', 'nodeset23887', 'nodeset18775', 'nodeset21044', 'nodeset18877', 'nodeset23794', 'nodeset23512', 'nodeset25524', 'nodeset21390', 'nodeset23605', 'nodeset23769', 'nodeset23526', 'nodeset17938', 'nodeset19911', 'nodeset20342', 'nodeset21438', 'nodeset18311', 'nodeset19159', 'nodeset19742', 'nodeset23547', 'nodeset18764', 'nodeset21384', 'nodeset21294', 'nodeset19153', 'nodeset20755', 'nodeset23869', 'nodeset17923', 'nodeset20303', 'nodeset23894', 'nodeset23715', 'nodeset23484', 'nodeset20332', 'nodeset23505', 'nodeset21577', 'nodeset21595', 'nodeset19341', 'nodeset21023', 'nodeset23746', 'nodeset20871', 'nodeset25400', 'nodeset18271', 'nodeset20343', 'nodeset21473', 'nodeset21571', 'nodeset25691', 'nodeset21452', 'nodeset18848', 'nodeset23721', 'nodeset18794', 'nodeset25522', 'nodeset25499', 'nodeset21393', 'nodeset17940', 'nodeset23876', 'nodeset23927', 'nodeset23498', 'nodeset23900', 'nodeset19095', 'nodeset20981', 'nodeset21603', 'nodeset21451', 'nodeset18266', 'nodeset25754', 'nodeset19091', 'nodeset23859', 'nodeset23834'

Model	General Metrics			Focused Metrics		
	precision	**recall**	**f1**	**precision**	**recall**	**f1**
DeBERTa-base-2	0.548	0.674	0.530	0.539	0.324	0.389
DeBERTa-base-1.5	0.550	0.672	0.536	0.506	0.290	0.358
DeBERTa-base-1	0.541	0.671	0.507	0.526	0.332	0.397

Table 4: Experiments on three different data ratios.

Model	General Metrics			Focused Metrics		
	precision	**recall**	**f1**	**precision**	**recall**	**f1**
DeBERTa-large	0.746	0.862	0.784	0.757	0.760	0.753
LLaMa-3-8B-LoRa	0.252	0.213	0.105	0.491	0.517	0.502
XLM-RoBERTa-large	0.557	0.855	0.622	0.799	0.808	0.799
DeBERTa-base	0.549	0.795	0.607	0.791	0.802	0.792
RoBERTa-MNLI	0.650	0.772	0.691	0.834	0.842	0.834
DeBERTa-MNLI	0.627	0.744	0.667	0.823	0.830	0.823

Table 5: Full Experiments on the second stage of YA-node prediction.

Model	Type	General Metrics			Focused Metrics		
		precision	**recall**	**f1**	**precision**	**recall**	**f1**
DeBERTa-base + RoBERTa-MNLI	ARI	**0.463**	0.324	**0.359**	0.320	**0.466**	**0.306**
+ DeBERTa-large (**submitted**)	ILO	0.542	0.499	0.514	0.564	0.646	0.594
DeBERTa-base + RoBERTa-MNLI	ARI	0.463	0.324	0.359	0.320	0.466	0.306
+ RoBERTa-MNLI	ILO	0.660	**0.796**	**0.705**	**0.873**	**0.902**	**0.883**
DeBERTa-base + DeBERTa-large	ARI	0.366	**0.469**	0.331	**0.393**	0.261	0.285
+ DeBERTa-large	ILO	**0.676**	0.763	0.703	0.662	0.648	0.652

Table 6: Experiments of different pipelines.