Overcome Noise and Bias: Segmentation-Aided Multi-Granularity Denoising and Debiasing for Enhanced Quarduples Extraction in Dialogue

Xianlong Luo[1,2] **Meng Yang** [1,2*] **Yihao Wang**[1,2]
[1]School of Computer Science and Engineering, Sun Yat-Sen University
[2]Key Laboratory of Machine Intelligence and Advanced Computing (SYSU), Ministry of Education, China
luoxlong@mail2.sysu.edu.cn, yangm6@mail.sysu.edu.cn, wangyh357@mail2.sysu.edu.cn

Abstract

Dialogue Aspect-based Sentiment Quadruple analysis (DiaASQ) extends ABSA to more complex real-world scenarios (i.e., dialogues), which makes existing generation methods encounter heightened noise and order bias challenges, leading to decreased robustness and accuracy. To address these, we propose the Segmentation-Aided multi-grained Denoising and Debiasing (SADD) method. For noise, we propose the Multi-Granularity Denoising Generation model (MGDG), achieving word-level denoising via sequence labeling and utterance-level denoising via topic-aware dialogue segmentation. Denoised Attention in MGDG integrates multi-grained denoising information to help generate denoised output. For order bias, we first theoretically analyze its direct cause as the gap between ideal and actual training objectives and propose a distribution-based solution. Since this solution introduces a one-to-many learning challenge, our proposed Segmentation-aided Order Bias Mitigation (SOBM) method utilizes dialogue segmentation to supplement order diversity, concurrently mitigating this challenge and order bias. Experiments demonstrate SADD's effectiveness, achieving state-of-the-art results with a 6.52% F1 improvement.

1 Introduction

Dialogue Aspect-based Sentiment Quadruple Extraction task (DiaASQ) (Li et al., 2023a) is a subtask of Aspect-based Sentiment Analysis (ABSA), aiming to extract sentiment quadruples in dialogues, i.e., Target: mentioned objects, Aspect: components of targets, Opinion: expressions conveying comments, and Sentiment: polarity of targets. Recently, Li et al. (2023a) proposed a discriminative model to control the information fusion among utterances, ultimately classifying different elements separately. However, this method fails to utilize the connections between tuple elements fully. Generative methods (Zhang et al., 2021a,b; Mao et al., 2022; Gou et al., 2023) succeeded in framing ABSA as a text-to-text task with robust generalization capabilities and fully leverage element connections, which inspired us.

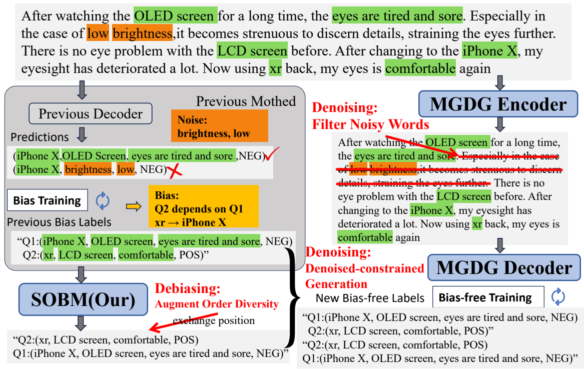

Figure 1: **Noise** refers to irrelevant words in dialogue (highlighted in orange), which lead the model to generate incorrect quadruples. **Order Bias** occurs when the model erroneously learns non-existent tuple order dependencies (highlighted in yellow boxes). Through denoising and debiasing, our SADD method enhances the performance of quadruple extraction.

However, generative methods still face two significant challenges: Noise and Order Bias, as illustrated in Fig 1. **1. Noise** is extraneous words in dialogues that interfere with the quadruple generation process, as illustrated by the orange words in Fig. 1. These extraneous words often disrupt the predicted quadruples; for instance, the terms 'brightness' and 'low' interfere with previous methods, leading to an incorrect quadruple. **2. Order Bias** is an irrational causal relationship caused by the fixed order of quadruple labels, like the yellow relationships in Fig. 1. As shown in Fig. 1, we formulate Diaasq task as a text-to-text problem: Input text → "Q1, Q2" (just like Fig. 1) , where the label is a sequence of tuples. However, the order between the tuples does not inherently exist, and the generation of Q2 should not be conditioned on

* Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 839–856
November 12-16, 2024 ©2024 Association for Computational Linguistics

Q1. This labeling scheme compels previous models to establish an order dependency from Q2 to Q1 ('xr->iPhone') and a causal relationship between the input and the order of tuples. However, such order dependency and causal relationships do not actually exist. These incorrect constraints hinder the model's generalization. A further explanation of noise and bias is shown in Appendix A.1. To address these, we propose a novel **S**egmentation-**A**ided multi-granularity **D**enoising and **D**ebiasing (**SADD**) method, including the following modules.

Denoising: Specifically, we first propose a novel **M**ulti-**G**ranularity **D**enoising **G**eneration (**MGDG**) module to reduce noise at the word and utterance levels. As shown in Fig. 1, our MEDG module identifies and eliminates the noise "Especially in ... before.", thereby achieving denoising. At the word level, we employ sequence labeling to label tuple elements. At the utterance level, we adopt topic-aware dialogue segmentation to achieve topic-centric utterance clustering, followed by generating topic masks based on clusters. Finally, we merge probability from the sequence labeling task and topic masks from the segmentation task into the decoder's denoised attention to generate denoised output. By emphasizing in-tuple and topic-related elements, denoised attention effectively makes the model more accurate and robust in tuple extraction tasks.

Our **T**opic-**A**ware **D**ialog **S**egmentation (**TADS**) differs from previous segmentation methods by explicitly introducing fine-grained topics information. Unlike existing methods (Wu et al., 2020a; Xie et al., 2021) that directly analyze complex contexts between utterances, we establish fine-grained relations between topic words and utterance sentences by cross-attention interaction, ultimately indirectly analyzing relationships between sentences. These improve models' robustness and accuracy in the segmentation of complex dialogue.

Debiasing: For the second challenge, we begin with theoretically analyzing the direct cause of order bias: the gap between the ideal and actual training objectives. By further analyzing the gap and the Maximum Likelihood Estimation (MLE) from a distribution perspective, we find a solution to augment order diversity at the data level, yet this poses a one-to-many learning problem. To solve these challenges, we propose a **S**egmentation-aided **O**rder **B**ias **M**itigation (**SOBM**) method to tackle order bias as shown in the lower part of Fig. 1. We leverage dialogue segmentation to generate multiple inputs that meet a specific criterion. We then pair these inputs with various feasible labels to create new samples, thereby increasing the diversity of tuple orders. The SOBM narrows the gap between ideal and actual training objectives, thereby mitigating order bias in the generation method.

In summary, our contributions are as follows:

1. We introduce a novel multi-granularity denoising generation model to mitigate interference noise through word-level sequence labeling and utterance-level topic masks.

2. We propose a topic-aware dialogue segmentation model to streamline context analysis and establish fine-grained relationships between utterances by introducing topic words as a bridge.

3. We uncover the direct cause of order bias and mitigate its impact by enhancing the data distribution through dialogue segmentation.

4. Our SADD method is validated on the widely used dataset and achieves state-of-the-art performance with a 6.52% F1 improvement.

2 Related Works

Aspect-Based Sentiment Analysis (Thet et al., 2010) primarily focuses on short texts (i.e., 1 or 2 sentences text) like reviews and emphasizes sentiment interpretability. ABSA methods analyze elements such as target (Li et al., 2019a,b), target categories (Zhang et al., 2021a), specific aspects, direct opinions (Peng et al., 2020) and so on. Quadruple extraction, involving four key elements, is a more comprehensive sentiment analysis task. Mainstream ABSA methods include sequence labeling (Wu et al., 2020b; Chen et al., 2022; Liang et al., 2023) and generative methods (Gao et al., 2022; Yu et al., 2023; Gou et al., 2023), with the latter known for robustness and generalization. However, existing ABSA models face challenges in dealing with complex textual content and structures when applied to dialogue texts, highlighting the need for advancements in this domain.

Dialogue Segmentation aims to segment a dialogue into pieces based on topics discussed, enhancing comprehension for downstream tasks (Zhong et al., 2022). Existing unsupervised Deep Learning(DL) methods use a pre-trained model without fine-tuning for segmentation (Xu et al., 2021b; Devlin et al., 2019; Xing and Carenini, 2021). DL-based methods directly analyze the context of two utterances and predict their relationships with fine-tuned CLS tokens, like TOD-BERT (Wu et al., 2020a) and RetroTS-T5 (Xie et al., 2021). However, analyzing two utterances directly can be challenging, especially with complex contexts involving multiple topics or lacking explicit topics.

Previous Methods for Addressing Tuple Order Bias mainly focused on addressing the order bias by modifying the model. They used non-autoregressive transformers (Sui et al., 2021; Tan et al., 2021) or set up multiple output heads (Ye et al., 2021) to generate results in an unordered manner. However, these methods have limited the generality of the model. "Set" (Li et al., 2023b) adjusts the loss function to force the model to minimize overall loss for all feasible labels globally. However, this approach actually forces models to learn a one-to-many mapping, hindering them from converging to optimal performance.

3 Task Definition

The input of the DiaASQ task is a n-utterance and N-word dialogue D=$\{u_1, \ldots, u_n\}$, where u_i represents the i-th utterance. DiaASQ aims to extract all quadruples $(target, aspect, opinion, sentiment)$ from the dialogue, where the target, aspect, and opinion are sub-strings of D, and $sentiment \in \{pos\ , neg, other\}$. In the example "I didn't buy it since my friend said the **Xiaomi 11** has **poor battery life**," the corresponding quadruple is (**Xiaomi 11**, **battery life**, **poor**, neg).

4 Method

In the DiaASQ task, generation models face two significant challenges: noise and order bias. To mitigate noise, we propose a novel Multi-Granularity Denoising Generation approach involving sequence labeling, topic-aware dialogue segmenting, and denoising generation, as shown in Fig. 2. By employing sequence labeling and topic-aware dialogue segmentation, we acquire denoising information at both the utterance and word levels. Then, we integrate this multi-grained denoising information to guide the model in generating quadruples more accurately and robustly. For order bias, we uncover its cause as the gap between the actual and the ideal training objective. We propose a novel Segmentation-aided Order Bias Mitigation (SOBM) method to narrow the gap with dialog segmentation. This method simultaneously addresses both the one-to-many training challenge and the order bias.

4.1 Multi-Granularity Denoising Generation

Due to the extensive content and intricate structure of dialogues, the model is susceptible to noise. To address noise, we propose a novel Multi-Granularity Denoising Generation method to reduce the noise at the word and utterance levels. Specifically, we leverage sequence labeling to mitigate noise at the word levels, and employ topic-aware dialogue segmentation to cluster sentences with the same topics, thereby eliminating noise from irrelevant sentences. We generate denoised outputs with the decoder's denoised attention which combines multi-grained information.

4.1.1 Labeling for Word-level Denoising

Word-level denoising identifies and emphasizes quadruple elements to reduce noise. For a dialogue D, we concatenate all utterances and encode them using the generation model's encoder: e=Encoder$([u_1; \ldots; u_n])$. Then, we employ a classification layer to label the quadruple elements in e with a loss $\mathcal{L}_{labeling}$. Each word e_i in e is classified into one of four categories (None, Target, Aspect, Opinion) using p_i=Softmax$(W_1 * e_i + b_1)$, where $p_i \in \mathbb{R}^4$. This process classifies all words in e to generate $P \in \mathbb{R}^{N\times 4}$.

4.1.2 Topic-aware Dialogue Segmentation for Utterance-level Denoising

Existing dialog segmentation methods directly analyze the complex context between utterances to determine their relationship, i.e., whether they belong to the same topic. However, these methods can struggle with complex utterance contexts, especially those involving multiple topics or lacking explicit topic mentions. To simplify the context analysis, we indirectly establish fine-grained relationships between utterances by examining their relationships with the same topic. This employs topics as bridges, streamlining the contextual analysis and enhancing the model's robustness in complex contexts. Moreover, we utilize cross-attention for fine-grained information fusion between topics and utterances, which helps resolve semantic-level coreferences for topics (Experiment 5.3.3).

Fine-grained Interaction We designate those words labeled as "Target" (in section 4.1.1) as the primary "topics" of the utterances because the target words are the cores of the quadruples and are highly relevant to the utterance topics. The topic embedding t_i for i-th topic (i.e., target) is selected from e according to its posi-

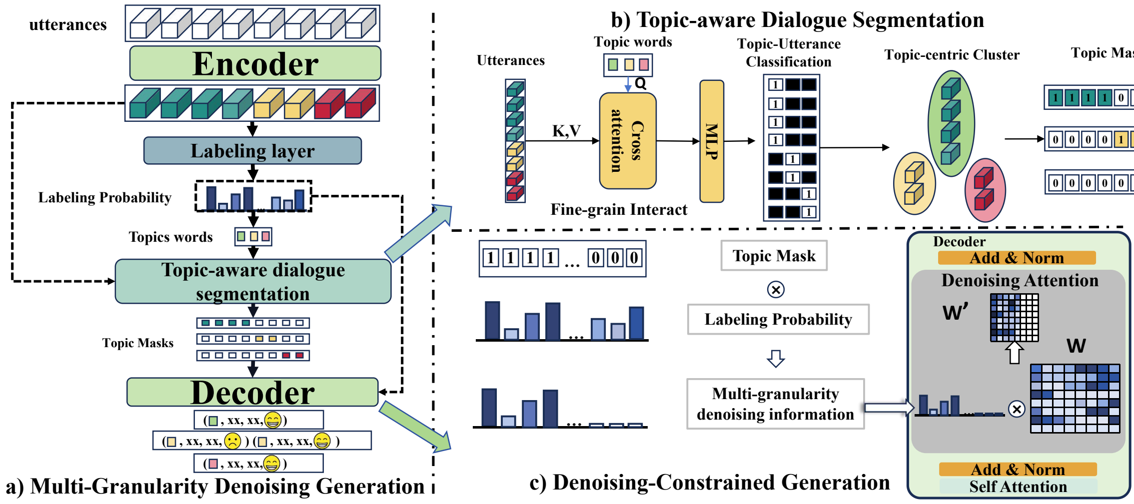

Figure 2: (a) Overview of the MGDG model. (b) Topic-aware Dialogue Segmentation module utilizes cross-attention to explore fine-grained correlations between utterances and topics, facilitating topic-centric clustering of utterances. Subsequently, we create a topic mask for each cluster. (c) The Denoising-Constrained Generation module integrates the denoising information into cross-attention to guide generation, resulting in denoised outputs.

tion. All topic embeddings are concatenated into $\mathbf{T}_{tp}$=$[t_1;\ldots;t_k]$$\in$$\mathbb{R}^{k\times dim}$. The utterance embedding e_{u_i} for i-th utterance u_i is directly extracted from e without pooling. $e_{u_i} \in \mathbb{R}^{|u_i|\times dim}$, where $|u_i|$ means the number of words in u_i. Feed them to cross-attention layers ($\mathbf{T}_{tp}$ as Query, u_i as Key and Value): O=softmax$\left(\frac{\mathbf{T}_{tp}\ (e_{u_i})'}{\sqrt{dim}}\right) e_{u_i}$, where $O$$\in$$\mathbb{R}^{k\times dim}$. Pass O to a classification layer to predict whether u_i has fine-grained associations (e.g. discussing relations) with $\{t_1,\ldots,t_k\}$ concurrently, with loss $\mathcal{L}_{topic}$. During training, the positions of the "Target" words are determined by the ground truth; during testing, they are determined by the predictions of the preceding module.

Topic Mask Applying these steps to all utterances $\{u_1,\ldots,u_n\}$, we predict the relationships between all utterances and $\{t_1,\ldots,t_k\}$. If both u_i and u_j discuss t_v, these two utterances can be grouped into the same v-th cluster. In this way, we establish fine-grained relations between utterances and aggregate utterances with the same topic into topic-centric clusters. Based on these clusters, we generate topic masks. Each topic mask $m^{(i)} \in \mathbb{R}^N$ masks out all utterances not in the i-th cluster.

4.1.3 Denoising-Constrained Generation

Denoised Attention Learning irrelevant context can lead attention mechanisms to focus on harmful information. To mitigate this, we restrict the attention scope and adjust its weight to maintain global interaction features while minimizing interaction with harmful data. When generating quadruples related to k-th topic, we incorporate its corresponding topic mask $m^{(k)} \in \mathbb{R}^N$ and the probabilities $P \in \mathbb{R}^{N\times 4}$ from section 4.1.1 into decoder's cross-attention:

$$\hat{P}_j = 1 - P_{j,0}\ ;\ r_j = \left(1+\hat{P}_j\right)\cdot m_j^{(k)} \quad (1)$$

$$w'_i = \frac{r_j \cdot \exp\left(w_{i,j}\right)}{\sum_j r_j \cdot \exp\left(w_{i,j}\right)} \quad (2)$$

where $P_{j,0}$ denotes probabilities of the input dialogue's j-th word belonging to the "None" category, $\hat{P}_j \in \mathbb{R}^N$ denotes probabilities of j-th input word being quadruple elements, $m_j^{(k)} \in \{0,1\}$ indicates whether the j-th word is masked $r \in \mathbb{R}^N$ is multi-granularity denoising information, $w \in \mathbb{R}^{N\times N}$ is the original cross-attention weights, $w_{i,j}$ signifies the weight of the i-th generated token relative to the j-th input token, and w'_i is the weights after adjusted to incorporate the multi-granularity denoising information. During training, the topic masks are replaced by ground truth masks; during testing, we employ the predicted topic masks.

Multi-granularity denoising To ensure the compatibility of our method with pre-trained models, we can directly replace the cross-attention in pre-trained generation models' decoders with the Denoised Attention. Train this generation task with a loss $\mathcal{L}_{generation}$. The topic mask m_i enables utterance-level denoising by constraining cross-attention scope to utterances within i-th topic cluster. This diminishes noisy utterances that do not mention the potential 'targets'. The probabilities

P facilitate word-level denoising by guiding the model to prioritize words identified as quadruple elements by the sequence labeling module. This effectively reduces noise from non-quadruple words. This multi-granularity denoising approach controls attention scope and adjusts attention weight to reduce noise, thereby enhancing extraction accuracy and robustness.

Overall loss $\mathcal{L}$=$\mathcal{L}_{labeling}$+$\mathcal{L}_{topic}$+$\mathcal{L}_{generation}$.

4.2 Order Bias Mitigation

Although previous works have shown the effectiveness of generative extraction methods, they often overlooked the accompanying issue of order bias, as shown in Figure 1. Existing solutions for order bias exhibit poor generalizability and scalability. To address order bias and ensure strong generalizability, we begin with a theoretical analysis revealing that the gap between practical and ideal training objectives leads to order bias. By further analyzing the gap and MLE from a distribution perspective, we find a data-driven solution to narrow the gap. However, this solution faces a one-to-many training challenge. To address this, we leverage dialog segmentation to enrich the order diversity within the data distribution, thereby mitigating the one-to-many training issue and order bias.

4.2.1 Ideal-Actual Training Gap

Ideal Training Objective According to Appendix B.2.1, the MLE loss for generative methods is :

$$\min_{\theta} -\mathrm{E}_{x\sim p(x)}\left[\mathrm{E}_{y\sim p(y|x)}\left[\log p_{\theta}(y|x)\right]\right] \quad (3)$$

where p represents the data distribution, and $p(x)$ denotes the probability of x occurring in the natural language context. When training a generative model for DiaASQ, for each input x, the associated ideal goal $\mathbb{S}$ is an unordered set of quadruples. By concatenating the quadruples in $\mathbb{S}$ in all possible permutation orders Π, we get a set of all feasible labels ($\Pi(\mathbb{S})$=$\{\pi_1(\mathbb{S}), \pi_2(\mathbb{S}), \dots\}$). According to Appendix B.2.1, for each sample with input as x, the ideal training loss (MLE) needs learning all feasible labels:

$$\min_{\theta}\left[-p(x)\sum_{y\in\Pi(\mathbb{S})} p(y|x)\log\left(p_{\theta}(y|x)\right)\right] \quad (4)$$

Actual Training Objective Neural network systems often struggle with learning one-to-many mappings (Vargas et al., 2017; Berner et al., 2021; Mukhamediev et al., 2022; Taye, 2023) because multiple labels imply multiple descending gradients, making it difficult for the model to adjust parameters and converge to optimal performance. Consequently, when constructing a training dataset, only one label $\pi_k(\mathbb{S}) \in \Pi(\mathbb{S})$ corresponds to each input x. Thus, the actual training objective is:

$$\min_{\theta}\left[-p(x)p(\pi_k(\mathbb{S})|x)\log\left(p_{\theta}(\pi_k(\mathbb{S})|x)\right)\right] \quad (5)$$

Following the calculations in Appendix B.3, the **Ideal-Actual Training Gap** Δ between the ideal training loss(MLE$_{ideal}$) and the actual training loss(MLE$_{actual}$) is:

$$\Delta = \mathrm{MLE}_{ideal} - \mathrm{MLE}_{actual} \quad (6)$$

$$= \frac{-p(x)}{|\mathbb{S}|}\left[\sum_{y\in(\Pi(\mathbb{S})-\{\pi_k(\mathbb{S})\})} \log p_{\theta}(y|x)\right] \neq 0 \quad (7)$$

where $|\mathbb{S}|$ is the number of elements in $\mathbb{S}$. The difference in Eq. (7) cannot be approximated to 0, indicating a **gap between the actual and ideal training objectives**. Clearly, the ideal training objective needs learning all feasible labels $\Pi(\mathbb{S})$ to capture the unordered nature of quadruples. However, in practice, the model is trained on only one feasible label $\pi(\mathbb{S})$, neglecting training with other feasible labels. This may lead the model to learn non-existent order biases and spurious causal relationships between input and order.

4.2.2 Segmentation-aided Order Bias Mitigation

Idea and Challenge Inspired by the MLE insights from the distribution perspective in Appendix B.2.1, a straightforward idea to narrow the gap is to augment the dataset with feasible label samples, allowing the model to learn more feasible labels to approximate the ideal training objective:

$$\min_{\theta} -\left[p_{\text{aug}}(x)\sum_{y\in\Pi(\mathbb{S})} p_{\text{aug}}(y|x)\log\left(\frac{p_{\text{aug}}(y|x)}{p_{\theta}(y|x)}\right)\right] \quad (8)$$

where p_{aug} represents the data distribution after augmenting with feasible labels. However, as mentioned earlier, it's challenging for a model to learn multiple outputs y for a single input x.

Order Diversity Augmentation: To address this issue, we propose constructing an input set $Ag(x)$ for x ($x{\in}Ag(x)$). **Each $\hat{x}{\in}Ag(x)$ shares the same quadruples and similar semantics with x.** Then we pair $\hat{x}{\in}Ag(x)$ with feasible labels $y \in \Pi(\mathbb{S})$ in a one-to-one manner to create new samples $(\hat{x}, y)$. For the original sample with input x, the objective in this augmented distribution is:

$$\min_\theta - \left[\sum_{(\hat{x},y)\in(Ag(x),\Pi(\mathbb{S}))} p_{\text{aug}}(\hat{x}) p_{\text{aug}}(y|\hat{x}) \log \left(\frac{p_{\text{aug}}(y|\hat{x})}{p_\theta(y|\hat{x})}\right)\right] \quad (9)$$

Clearly, in this augmented dataset, the training objective can approximate the ideal objective, as demonstrated in Appendix B.3.1.

AI rewriting tools (such as ChatGPT) and traditional data augmentation methods struggle to generate dialogue inputs with the same quadruples and similar semantics without human intervention, as shown in Appendix B.1.1 and Experiment C.6. We propose a cost-effective solution based on dialogue segmentation to address this problem, which divides the dialogue into segments based on their semantic topics, ensuring they are semantically isolated. These segments are then rearranged and concatenated in all possible orders to form an augmented dialogue input set like $Ag(x)$. Each input in this set shares similar semantics because rearranging semantically independent segments does not affect the overall semantics. Each input in this set contains the same quadruples, as all the words remain unchanged. We then pair these inputs with multiple feasible labels to create new samples, thereby increasing order diversity and enhancing the data distribution. In this augmented dataset, as mentioned earlier, the actual training objective closely approximates the ideal training objective, thus alleviating order bias. For simplicity, our dialog segmentation scheme is based on the inherent reply thread structure (shown in section 5.1) within the dataset. It works because utterances connected by reply relationships often share similar semantic topics, making them inseparable, while others are separable.

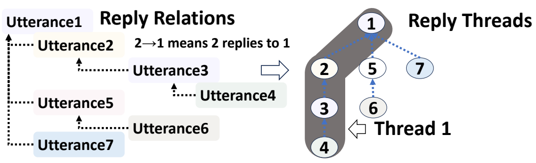

Figure 3: Example of **Reply thread** in a dialog.

5 Experiments

5.1 Experimental Settings

Dataset The Diaasq dataset (Li et al., 2023a) comprises both English(EN) and Chinese(ZH) datasets and provides dialogue texts with reply threads. A reply thread is a collection of utterances linked by reply relationships, as shown in Fig. 3. More detail is in Appendix C.1.

Metrics We use micro F1 for the pair extraction task and both Micro F1 and Identification F1 (Barnes et al., 2021) for the quadruple extraction task, following the dataset creators' recommendations. Micro F1 considers tuples with all words correct as TP and any incorrect word as FP. Identification F1 is similar but ignores sentiment elements.

Baselines We compared with generative models like **ParaPhrase** (Zhang et al., 2021a) and discriminative models like **CRF-ExtractClassify** (CEC)(Cai et al., 2021), **SpERT**(Eberts and Ulges, 2020), **Span-ASTE**(Xu et al., 2021a), and **MvI**(Li et al., 2023a). **ParaPhrase**(Zhang et al., 2021a) introduces a novel paraphrase modeling paradigm to frame the ASQP task as a paraphrase generation process. **MvI** (Li et al., 2023a) uses multi-view information to control information fusion and then extracts quadruples by decoding Tagging Grid.

Settings We use BART (Lewis et al., 2020) (440M) for both EN and ZH datasets. We train the model for 10 epochs (2 hours) on 4 3090 GPUs with a batch size of 5 and a learning rate of 5e-5. The ratio of the three losses is 1:1:1. The number of cross-attention layers is 3(Appendix C.8). More detail is in Appendix C.3. All reported results are averaged over multiple runs.

5.2 Main Result

The results are presented in Table 1. In the quadruple extraction task, our SADD method achieves a maximum improvement of 5.56% micro F1 and 6.52% Iden F1 in the EN dataset compared to the previous best model(MvI), demonstrating the effectiveness of our method. Because discriminative models are not influenced by bias, our method's major advantage over them lies in denoising. With the multi-granularity denoising generation module, we achieve up to a 6.52% Iden F1 improvement compared to the best discriminative method(MvI) on the EN dataset. Compared to generative models, our method's greatest strength lies in order bias mitigation. With the segmentation-aided order bias mitigation module, we achieve up to a 16.56% Iden F1 improvement compared to the ParaPhrase in the EN dataset. Further insights into the impact of order bias on the results can be found in the Appendix C.5. In the Pair Extraction task, our model achieved an average 3.19% micro F1 improvement in all datasets over the previous best approaches. This underscores the effectiveness of our method in

Table 1: Main Results. 'D' denotes discriminative methods, while 'G' indicates generation methods. T-A means the target-aspect pair extraction task, T-O refers to target-opinion, and A-O refers to aspect-opinion.

Type	Method	EN					ZH				
		Pair Extraction(F1)			Quadruple(F1)		Pair Extraction(F1)			Quadruple(F1)	
		T-A	T-O	A-O	Micro	Iden	T-A	T-O	A-O	Micro	Iden
D.	CEC	34.31	20.94	19.21	11.59	12.80	32.47	26.78	18.90	8.81	9.25
	SpERT	28.33	21.39	23.64	13.07	13.38	38.05	31.28	21.89	13.00	14.19
	Span-ASTE	42.19	30.44	45.90	26.99	28.34	44.13	34.46	32.21	27.42	30.85
	MvI	47.91	45.58	44.27	33.31	36.80	48.61	43.31	45.44	34.94	37.51
G.	ParaPhrase	37.22	32.19	30.78	24.54	26.76	37.81	34.32	27.76	23.27	27.98
	SADD (Ours)	**50.82**	**49.64**	**49.70**	**38.87**	**43.32**	**51.13**	**46.72**	**47.87**	**37.80**	**41.05**

enhancing extraction performance across various tasks, indicating its generalizability. By employing topic-aware dialogue segmentation to form target-centric clusters, our model effectively diminishes noise from quadruples with different targets during aspect and opinion extraction tasks associated with a specific target (TA, TO task). Furthermore, in aspect-opinion pair extraction (AO task), our model primarily benefits from the sequence labeling probability, which diminishes non-quadruple noise.

5.3 Analysis

5.3.1 Ablation Study

Table 2: Ablation studies of MGDG and SOBE components on DiaASQ Dataset.

Method	Components		EN		ZH	
	MGDG	SOBM	Micro	Iden	Micro	Iden
Baseline			29.31	32.30	30.45	33.64
+MGDG	✓		36.35	40.64	35.76	39.21
+SOBM		✓	34.96	37.86	35.70	38.39
SADD (Ours)	✓	✓	**38.36**	**42.94**	**37.80**	**41.05**

We conducted an ablation study to validate the effectiveness of our Multi-Granularity Denoising Generation (MGDG) and Segmentation-aided Order Bias Mitigation (SOBM) components, detailed in Table 2. Compared to the baseline, integrating the MGDG module brings a maximum 8.34% Iden F1 improvement in the EN dataset. It indicates that the MGDG module significantly enhances tuple extraction accuracy and robustness by reducing noise. We also compared our MGDG module with existing segmentation methods in Section 5.3.3. Furthermore, the integrated SOBM module brings a maximum 5.65% micro F1 improvement in the EN dataset compared to the baseline. It demonstrates the effectiveness of SOBM in mitigating order bias. We also compared our SOBM module with existing debias methods in Section 5.3.4 and investigated the effects of different data augmentation strategies in the SOBM in Appendix C.6.

Table 3: The proportion of errors attributed to noise

	MvI	SADD(our)	Δ
Proportion	79.88	48.67	-31.21

5.3.2 Statistics and Case Studies

We conducted a comparative analysis between our proposed method and the SOTA method (MvI) regarding the proportion of errors attributed to noise, as shown in Table 3. The significant proportion of errors, amounting to 79.88%, underscores the inadequacy of previous methods in handling noise effectively, thereby highlighting the necessity for denoising techniques. Furthermore, our denoising approach resulted in a notable reduction of 31.21% in the proportion of errors attributed to noise, affirming our method's effectiveness. Figure 4 presents several case studies where the previous SOTA method (MvI) failed to provide good predictions, whereas our model demonstrated superior performance. The two examples primarily illustrate how noise leads to an increase in irrelevant quadruples and a decline in quadruple quality. In the first example, due to the interference of noisy words like "Meizu 18", "machine," "backup," and "main," MvI produced several erroneous and irrelevant quadruples. In the second example, the MvI model's prediction of the quadruple "(mate series, appearance, much better, pos)" is compromised by the noise word "p series," leading to the erroneous generation of "(p series, appearance, much better, pos)" instead. Noise detrimentally affects the quality of predicted quadruples. In contrast, our model remains unaffected by such disturbances.

5.3.3 Further Ablation Study on TADS

To assess the effectiveness of the **Topic-aware Dialogue Segmentation (TADS)** method, we com-

Dilogues	MvI	SADD(our)	Label	Explanation
speaker0:... speaker1:"After watching your 5 - minute long test , I bought the p40pro . It 's really good [hee hee] . Taking photo is stable and the workmanship is excellent . 90hz is well optimized .", speaker0:... speaker1:... speaker2:"Why is P40Pro and not mate40Pro ? Meizu 18 is just a backup machine , what about the main machine ?", speaker3:"The mate40p feels too bad , not suitable for holding it all the time , but it has full functions and is more suitable for the main machine in life", speaker2:... speaker4:...	(mate40p, feels, too bad, neg) (p40pro, Taking photo, stable, pos) (p40pro, 90hz, well optimized, pos) (mate40p, functions, full, pos) (p40pro, workmanship, excellent, pos) (Meizu 18, machine, backup, neg) (mate40p, machine, main, pos)	(mate40p, feels, too bad, neg) (p40pro, Taking photo, stable, pos) (p40pro, 90hz, well optimized, pos) (mate40p, functions, full, pos) (p40pro, workmanship, excellent, pos)	(mate40p, feels, too bad, neg) (p40pro, Taking photo, stable, pos) (p40pro, 90hz, well optimized, pos) (mate40p, functions, full, pos) (p40pro, workmanship, excellent, pos)	"Machine" is not an aspect of Meizu or Mate40p; instead, it refers to their entities. Therefore, the two additional quadruples predicted are incorrect.
speaker0:... speaker1:... speaker2:"When the Android phone of Dimensity 9000 comes out , such as OPPO 's , it will definitely be good . And Huawei 's flagship is really no better than Oppo 's flagship . Oppo 's flagship machine has good quality control and texture . But it is very cheap , much cheaper than Huawei .", speaker1:... speaker3:... speaker0:... speaker4:"Honestly , I personally think the appearance of the mate series is much better than the p series ." speaker5: ...	(Oppo, quality control, good, pos) (p series, appearance, much better, pos) (Oppo, texture, good, pos)	(Oppo, quality control, good, pos) (mate series, appearance, much better, pos) (Oppo, texture, good, pos)	(Oppo, quality control, good, pos) (mate series, appearance, much better, pos) (Oppo, texture, good, pos)	In the dialogue, the phrase "much better" describes the "mate series" rather than the "p series".

Figure 4: Case Study. The orange words represent the noise that causes errors in the MvI model.

pare it with existing methods detailed in Appendix C.7, as shown in Table 4. Compared to **TOD-BERT**(Wu et al., 2020a) , our methods achieved a maximum of 6.23 % Iden F1 improvement in the ZH dataset. This underscores the effectiveness of incorporating topic information to simplify contextual analysis, enhancing segmentation accuracy and robustness by avoiding the direct analysis of complex utterances. Compared to **TSP**, our methods achieved a maximum of 3.74% Iden F1 improvement in the ZH dataset. This demonstrates that utilizing cross-attention to mine fine-grained associations can enhance the model's robustness in complex situations, such as utterances with multiple topics and implicit topics. Compared to **SMGD**, our methods achieved a maximum 12.1% Iden F1 improvement in the EN dataset. This highlights that the pre-labeling topic words are necessary for the topic-aware dialogue segmentation module. The SMGD method, which segments dialogues without pre-labeling topics, struggles to analyze complex context interactions between utterances. In contrast, our method benefits from pre-labeling topics, which simplifies contextual analysis by focusing only on interactions between topics and utterances. Compared to **RT**, our methods achieved a maximum 3.96% Iden F1 improvement in the EN dataset. This indicates that our method can handle utterances related to multiple topics, thereby performing more accurate dialogue segmentation and denoising without removing any topic-related information. Compared to **TWM**, our methods achieved a maximum 6.74% Iden F1 improvement in the EN dataset. This demonstrates that utilizing cross-attention to mine fine-grained associations can help resolve topic-level coreferences.

Table 4: Result of various dialogue segmentation methods combined with SOBM and MGDG. NN means the method is totally a Neural Network method.

Method	NN	Components		EN		ZH	
		Topic	Fine-grain	Micro	Iden	Micro	Iden
TOD-BERT	✓			34.76	38.42	32.12	34.82
TSP	✓	✓		36.30	40.18	34.30	37.31
SMGD	✓		✓	27.67	31.22	28.39	30.87
RT				35.78	39.36	35.36	38.51
TWM		✓		32.85	36.58	31.78	36.00
TADS (Ours)	✓	✓	✓	**38.87**	**43.32**	**37.80**	**41.05**

Table 5: Results of Methods Addressing Order Bias.

Method	EN		ZH	
	Micro	Iden	Micro	Iden
Set	31.83	35.26	29.81	33.52
SOBM(Ours)	**34.96**	**37.86**	**35.70**	**38.39**

5.3.4 Further Ablation Study on SOBM

To evaluate the effectiveness of our debiasing solution, we compare it with an existing method called Set (Li et al., 2023b) introduced in Section 2 , as shown in Table 5. Our method outperforms Set by a maximum of 5.89% micro F1 in the ZH dataset, highlighting its effectiveness in mitigating order bias. In contrast to Set's struggle with one-to-many learning at the loss level, our approach augments inputs to avoid learning one-to-many mappings and mitigate order bias at the data level, thereby improving performance and generalizability.

6 Conclusion

This paper introduces a novel Segmentation-Aided multi-grained Denoising and Debiasing (SADD) model for denoising and debiasing in the DiaASQ task. For noise, we propose a Multi-Granularity Denoising Generation(MGDG) model to denoise at both word and utterance levels with denoised attention. For order bias, we analyze its direct causes and propose a distribution-based solution.

We then introduce the Segmentation-aided Order Bias Mitigation (SOBM) method, which utilizes dialogue segmentation to increase order diversity, thereby simultaneously alleviating the challenges of one-to-many learning and order bias. Extensive experiments show SADD's SOTA performance.

7 Limitations

1. A limitation we encountered is the increased training time due to the augmented dataset.

2. The BART model encounters challenges when processing long-text inputs, particularly in dialogue scenarios, due to the increasing time complexity of attention mechanisms as the input length grows. This results in higher time overhead compared to short-text ABSA. More efficient attention mechanisms tailored for long textual inputs in dialogue contexts need to be developed to mitigate this issue.

3. We didn't fully utilize the inherent information in the DiaASQ dataset, such as speaker information or reply relationships, which could improve the model's comprehension of dialogue content.

8 Ethics Statement

In all our experiments, we utilized pre-existing datasets widely used in previous research. While analyzing experimental results, We made diligent efforts to maintain fairness and honesty, ensuring that our work did not cause harm to any individuals.

Regarding broader impacts, this work can contribute to further research in sentiment analysis and the utilization of generative methods for simplifying and automating the extraction of user opinions in real-world applications. However, it's noteworthy that this work utilizes fine-tuning large-scale pre-trained language models for generating sentiment triplets. Since the large-scale pre-training corpora originate from the internet, predicted sentiment polarity may be subject to unintended biases associated with gender, race, and intersectional identities (Tan and Celis, 2019). Large pre-trained language models often inherit biases present in their training data, potentially leading to biased sentiment analysis results, particularly when evaluating texts from underrepresented or marginalized groups, thereby perpetuating and amplifying societal prejudices. It is crucial for the natural language processing community to consider these biases more extensively. Fortunately, these issues are actively being addressed within the research community, including efforts to standardize datasets and methodologies.

We obtained licenses for all artifacts used in our study, and our data was obtained from open-source repositories. Our use of existing artifacts is consistent with their intended use. Our method's specific intended use is to extract quadruples from dialogues and is compatible with the original access conditions. We read and checked each sample to ensure that the data used does not contain information that names or uniquely identifies individual people or offensive content.

9 Acknowledgment

This work was supported by the National Natural Science Foundation of China No.62176271) and Guangdong Basic and Applied Basic Research Foundation (Grant no. 2024A1515011692).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. 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 3387–3402, Online. Association for Computational Linguistics.

Julius Berner, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. 2021. The modern mathematics of deep learning. *CoRR*, abs/2105.04026.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. 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 340–350, Online. Association for Computational Linguistics.

Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2974–2985, Dublin, Ireland. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

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.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2006–2013. IOS Press.

Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. 2022. LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7002–7012, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. Mvp: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4380–4397. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023a. Diaasq: A benchmark of conversational aspect-based sentiment quadruple analysis. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13449–13467. Association for Computational Linguistics.

Jiangnan Li, Yice Zhang, Bin Liang, Kam-Fai Wong, and Ruifeng Xu. 2023b. Set learning for generative information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13043–13052. Association for Computational Linguistics.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A unified model for opinion target extraction and target sentiment prediction. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6714–6721. AAAI Press.

Zheng Li, Xin Li, Ying Wei, Lidong Bing, Yu Zhang, and Qiang Yang. 2019b. Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4589–4599. Association for Computational Linguistics.

Shuo Liang, Wei Wei, Xian-Ling Mao, Yuanyuan Fu, Rui Fang, and Dangyang Chen. 2023. STAGE: span tagging and greedy inference scheme for aspect sentiment triplet extraction. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13174–13182. AAAI Press.

Yue Mao, Yi Shen, Jingchao Yang, Xiaoying Zhu, and Longjun Cai. 2022. Seq2path: Generating sentiment tuples as paths of a tree. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2215–2225. Association for Computational Linguistics.

Ravil I. Mukhamediev, Yelena Popova, Yan Kuchin, Elena Zaitseva, Almas Kalimoldayev, Adilkhan Symagulov, Vitaly Levashenko, Farida Abdoldina, Viktors Gopejenko, Kirill Yakunin, Elena Muhamedijeva, and Marina Yelis. 2022. Review of artificial intelligence and machine learning technologies: Classification, restrictions, opportunities and challenges. *Mathematics*, 10(15).

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The*

Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 8600–8607. AAAI Press.

Dianbo Sui, Chenhao Wang, Yubo Chen, Kang Liu, Jun Zhao, and Wei Bi. 2021. Set generation networks for end-to-end knowledge base population. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9650–9660, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13209–13220.

Zeqi Tan, Yongliang Shen, Shuai Zhang, Weiming Lu, and Yueting Zhuang. 2021. A sequence-to-set network for nested named entity recognition. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3936–3942. ijcai.org.

Mohammad Mustafa Taye. 2023. Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions. *Comput.*, 12(5):91.

Tun Thura Thet, Jin-Cheon Na, and Christopher S. G. Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *J. Inf. Sci.*, 36(6):823–848.

Rocio Vargas, Amir Mosavi, and Ramon Ruiz. 2017. Deep learning: a review.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020a. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020b. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585, Online. Association for Computational Linguistics.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. 2021. TIAGE: A benchmark for topic-shift aware dialog modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1684–1690, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linzi Xing and Giuseppe Carenini. 2021. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 167–177, Singapore and Online. Association for Computational Linguistics.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021a. Learning span-level interactions for aspect sentiment triplet extraction. 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 4755–4766, Online. Association for Computational Linguistics.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021b. Topic-aware multi-turn dialogue modeling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14176–14184. AAAI Press.

Jiacheng Ye, Tao Gui, Yichao Luo, Yige Xu, and Qi Zhang. 2021. One2Set: Generating diverse keyphrases as a set. 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 4598–4608, Online. Association for Computational Linguistics.

Chengze Yu, Taiqiang Wu, Jiayi Li, Xingyu Bai, and Yujiu Yang. 2023. Syngen: A syntactic plug-and-play module for generative aspect-based sentiment analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 504–510. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Confer-*

ence on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pages 11765–11773. AAAI Press.

A Appedix for Introduction

A.1 Definition and Example of Noise and Bias

Noise is words that interfere with the generation process when the model generates a certain quadruple.

Order Bias: Due to the constraints of seq2seq tasks, the model learns a nonexistent causal relationship from the input to the order of quadruples. The model ends up overfitting to a specific order we've arbitrarily defined, which affects its generalization ability. We term this as "Order Bias."

Example:

Input ⋆: Utterance 1: ...The battery of the iPhone was quite good and the system was smooth ...

Utterance 2: ... The battery of Samsung phones is worse. ... I also bought a Samsung phone for my girlfriend...

Utterance 3: ... Xiaomi can also be considered, mainly because the price is very low...

Output ♡:

"iPhone" Quads: (iPhone, battery, quite good, POS), (iPhone, system, smooth, POS)

"Samsung" Quads: (Samsung, battery, worse, NEG)

"Xiaomi" Quads: (Xiaomi, price, very low, POS).

Example of Noise: Specifically, when the model generates the quadruple ♣" (iPhone, battery, quite good, POS)", it selects words from the input ⋆. Words in the input ⋆ but not in the quadruple are the words that interfere with the generation process of the quadruple ♣. So, these words are the noise (The definition of Noise). For instance, words such as "bought" and "considered" can introduce significant noise, potentially leading the model to generate incorrect quadruples.

Example of Bias: When we transform the quadruple extraction task into a text-to-text generation task, we need to design a sentence as the label. Considering the quadruples (Samsung, battery, worse, NEG) and (Xiaomi, price, very low, POS), we have to decide the order between them when constructing labels for the seq2seq task. Whether it's "(Samsung, battery, worse, NEG) (Xiaomi, price, very low, POS)" or "(Xiaomi, price, very low, POS) (Samsung, battery, worse, NEG)", the model is forced to learn the corresponding order and move away from the other orders. But in fact, any order is correct. This confusion leads the model to seek semantic clues from the input to find out why this order. The model attempts to find a nonexistent causal relationship between the input and the order of quadruples to find why this order exists. As a result, the model overfits to our arbitrarily defined order, impacting its generalization ability, thus leading to a bias.

B Appedix for Method

B.1 Appendix for section 4.1

B.1.1 Augment with Chatgpt

We aim to keep the quadruple elements unchanged while constructing semantically similar inputs. However, AI paraphrasing tools like ChatGPT 3.5 and ChatGPT 4 often fail to preserve the quadruple elements and may alter the original semantics. Firstly, there's the issue of maintaining quadruple elements. ChatGPT often modifies the opinion part of the quadruples. Changing the quadruple elements renders the original labels incompatible with the input, resulting in a failed input construction. It is nearly impossible to determine whether the original quadruple elements remain unchanged through code analysis, as the appearance of characters in the text does not necessarily imply their association with the same quadruple or a quadruple relationship between them. Additionally, manually verifying whether quadruple elements have changed would require significant effort. Secondly, there's the issue of preserving the original input's semantics. ChatGPT also frequently alters semantics, disregarding certain parts of the content or even producing dialogues with entirely opposite meanings. We demonstrate some examples where ChatGPT rewriting resulted in changes to quadruple elements and altered semantics, as shown in Figure 8 and Figure 9. Therefore, AI rewriting tools like ChatGPT may not be suitable for our augmentation task.

B.2 Appendix for section 4.2

B.2.1 From a Data Distribution Perspective: MLE Loss

Currently, generative extraction models are primarily trained using Maximum Likelihood Estimation (MLE). Given the data distribution p and a parametric model with parameters θ, Maximum likelihood estimation (MLE) minimizes:

$$L_{MLE}(\theta) = -\mathrm{E}_{x\sim p(x)}[\mathrm{E}_{y\sim p(y|x)}[logp_\theta(y|x)]] \tag{10}$$

where x represents the input context, y represents the generation label.

It is well known that MLE can be seen as minimizing the Kullback-Leibler (KL) divergence between the data distribution p and the model-estimated distribution p_θ. The equation below shows the relationship between MLE loss and KL divergence:

$$D_{KL}(p \parallel p_\theta) \tag{11}$$

$$= \sum_X p(x) \sum_Y p(y|x) \log\left(\frac{p(y|x)}{p_\theta(y|x)}\right) \tag{12}$$

$$= \sum_X p(x) \sum_Y p(y|x) \log\left(p(y|x)\right) \tag{13}$$

$$- \sum_X p(x) \sum_Y p(y|x) \log\left(p_\theta(y|x)\right) \tag{14}$$

$$= -H + \mathrm{MLE}(\theta) \tag{15}$$

where H is the "Entropy" and is independent of model parameters θ, it can be disregarded in the training loss. Hence, MLE loss and KL divergence share the same minimum. By minimizing the MLE loss, we encourage the predicted distribution p_θ to align with the data distribution p closely.

Learning All Feasible Lables It is worth noting that Equation 12 indicated that **we need to learn all feasible labels for the input.** In many specific tasks, only one label y corresponds to a given input, with a probability $p(y|x) = 1$, and the probabilities $p(other|x)$ for other texts are all 0. However, in some tasks, there may be multiple labels $\{y_1, y_2, \ldots\}$ that match a given input, with probabilities $p(y_1|x), p(y_2|x), \ldots$ all non-zero. Unfortunately, these probabilities are often immeasurable, which has led to prior research overlooking multiple feasible labels and instead focusing only on one label. **Failing to learn all feasible labels fully, and instead focusing on just one, increases the risk of introducing bias into the model.**

B.3 Proof of Ideal-Actual Training Gap

We prove that, for each sample, the Ideal-Actual Training Gap Δ, i.e., the difference between the ideal MLE loss and the actual MLE loss is not zero, thereby demonstrating a disparity between the ideal training objective and the actual training objective.

Given one sample with input as x and model parameter θ, the difference Δ between the ideal MLE loss and the actual MLE loss is as follows:

$$\begin{aligned}\Delta =& MLE_{ideal} - MLE_{actual}\\ =& -\left[p(x) \sum_{y\in\Pi(\mathbb{S})} p(y|x)\log p_\theta(y|x)\right]\\ &+ [p(x)p(\pi_k(\mathbb{S})|x)\log p_\theta(\pi_k(\mathbb{S})|x)]\\ =& -p(x)\left[\sum_{y\in\Pi(\mathbb{S})} p(y|x)\log p_\theta(y|x) - p(\pi_k(\mathbb{S})|x)\log p_\theta(\pi_k(\mathbb{S})|x)\right]\end{aligned}$$

In this task, all feasible labels contain the same quadruples but in different orders. Moreover, all permutation orders are equivalent. Therefore, all labels are equivalent, resulting in equal probabilities for each label. That is, for $p(y|x)$, the probability of each feasible label $y \in \Pi(\mathbb{S})$ is the same, so $p(y|x) = \frac{1}{|\Pi(\mathbb{S})|}$, where $|\Pi(\mathbb{S})|$ represents the number of elements in $\mathbb{S}$. Of course, $p(\pi_k(\mathbb{S})|x) = \frac{1}{|\Pi(\mathbb{S})|}$. Consequently, we can further simplify the above expression:

$$\begin{aligned}\Delta =& MLE_{ideal} - MLE_{actual}\\ =& -\frac{p(x)}{|\mathbb{S}|}\left[\sum_{y\in\Pi(\mathbb{S})} \log p_\theta(y|x) - \log p_\theta(\pi_k(\mathbb{S})|x)\right]\\ =& -\frac{p(x)}{|\mathbb{S}|}\left[\sum_{y\in\{\Pi(\mathbb{S})\text{-}\{\pi_k(\mathbb{S})\}\}} \log p_\theta(y|x)\right]\\ \not\approx& \; 0\end{aligned}$$

In dialogue datasets, each sample contains more than one quadruple, so $\Pi(\mathbb{S})$-$\{\pi_k(\mathbb{S})\} \neq \varnothing$. Therefore, in this scenario, the Ideal-Actual training gap Δ between the ideal MLE loss and the actual MLE loss cannot approximate 0. This indicates a gap between the ideal training objective and the actual training objective.

B.3.1 Objective Approximation

Our approach to supplementing necessary samples with various feasible labels involves the following steps: Firstly, construct an input set $Ag(x)$ where each input shares the same quadruple elements and exhibits similar semantics. Then, combine these inputs with multiple feasible labels to create samples. **Within the augmented dataset**, we will illustrate that the Ideal-Actual training gap Δ between the ideal Maximum Likelihood Estimation (MLE) loss and the actual MLE loss for any given sample is **approximately** 0. This demonstration serves to indicate that the training objective on this augmented dataset can closely approximate the ideal training objective.

Given one sample with input as x and model parameter θ, the gap Δ between the ideal MLE loss and the actual MLE loss is as follows:

$$\begin{aligned}\Delta =& MLE_{ideal} - MLE_{actual} \\ =& -\left[p(x) \sum_{y\in\Pi(\mathbb{S})} p(y|x)\log p_\theta(y|x) \right] \\ &+ \left[\sum_{(\hat{x},y)\in(Ag(x),\Pi(\mathbb{S}))} p_{\text{aug}}(\hat{x}) p_{\text{aug}}(y|\hat{x}) \log p_\theta(y|\hat{x}) \right]\end{aligned} \tag{16}$$

Because x and each $\hat{x} \in Ag(x)$ are semantically similar, they occur with the same probability in natural language contexts. With a sufficiently large sample size in the dataset, under the guarantee of the "Law of the Large Numbers," we can assert that $p(x) \approx p_{aug}(\hat{x})$. Thus, we can simplify the above formula to:

$$\begin{aligned}\Delta =& MLE_{ideal} - MLE_{actual} \\ \approx& -p(x) \left[\sum_{y\in\Pi(\mathbb{S})} p(y|x)\log p_\theta(y|x) \right. \\ &\left. - \sum_{(\hat{x},y)\in(Ag(x),\Pi(\mathbb{S}))} p_{\text{aug}}(y|\hat{x}) \log p_\theta(y|\hat{x}) \right]\end{aligned} \tag{17}$$

In this task, all feasible labels contain the same quadruples but in different orders. Moreover, all permutation orders are equivalent. Therefore, all labels are equivalent, resulting in equal probabilities for each label. That is, for $p(y|x)$, the probability of each feasible label $y \in \Pi(\mathbb{S})$ is the same, so $p(y|x) = \frac{1}{|\Pi(\mathbb{S})|}$. In the augmented dataset, $\hat{x}$ and x share the same quadruple elements, implying that all feasible labels associated with them are the same. Furthermore, as our augmented dataset encompasses all feasible labels, we have $p_{\text{aug}}(y|\hat{x}) = \frac{1}{|\Pi(\mathbb{S})|}$. Hence, $p(y|x) = p_{\text{aug}}(y|\hat{x}) = \frac{1}{|\Pi(\mathbb{S})|}$. This allows for further simplification of the above expression:

$$\begin{aligned}\Delta =& MLE_{ideal} - MLE_{actual} \\ \approx& -\frac{p(x)}{|\mathbb{S}|} \left[\sum_{y\in\Pi(\mathbb{S})} \log p_\theta(y|x) \right. \\ &\left. - \sum_{(\hat{x},y)\in(Ag(x),\Pi(\mathbb{S}))} \log(p_\theta(y|\hat{x}) \right]\end{aligned} \tag{18}$$

An input x consists of two components: the quadruple elements x_q and the non-quadruple context x_c. Therefore, we can decompose x in the above equation as follows:

$$\begin{aligned}\Delta =& MLE_{ideal} - MLE_{actual} \\ \approx& -\frac{p(x)}{|\mathbb{S}|} \left[\sum_{y\in\Pi(\mathbb{S})} \log\left[p_\theta(y|x_q) p_\theta(y|x_o)\right] \right. \\ &\left. - \sum_{(\hat{x},y)\in(Ag(x),\Pi(\mathbb{S}))} \log\left[p_\theta(y|\hat{x}_q) p_\theta(y|\hat{x}_o)\right] \right]\end{aligned} \tag{19}$$

Here, x_q is correlated with the label y, while x_c is independent of the label y. So, we have

$$\begin{aligned}\Delta =& MLE_{ideal} - MLE_{actual} \\ \approx& \frac{p(x)\log p_\theta(y)}{|\mathbb{S}|} \left[-\sum_{y\in\Pi(\mathbb{S})} \log p_\theta(y|x_q) + \sum_{(\hat{x},y)\in(Ag(x),\Pi(\mathbb{S}))} \log p_\theta(y|\hat{x}_q) \right]\end{aligned} \tag{20}$$

When constructing $\hat{x}$, we ensure that it shares the same quadruple elements as x, hence $\hat{x}_q = x_q$. Consequently, $\log(p_\theta(y|\hat{x}_q)) = \log(p_\theta(y|x_q))$. Hence, we can simplify the above expression to:

$$\begin{aligned}\Delta =& MLE_{ideal} - MLE_{actual} \\ \approx& \frac{p(x)}{|\mathbb{S}|} \left[-\sum_{y\in\Pi(\mathbb{S})} \log p_\theta(y|x_q) + \sum_{(\hat{x},y)\in(Ag(x),\Pi(\mathbb{S}))} \log(p_\theta(y|x_q) \right] \\ \approx& \ 0\end{aligned} \tag{21}$$

The above equation can be approximated to 0 because the variable y in Equation 21 can cover all feasible labels during actual training, aligning it with the ideal scenario. Therefore, in this case, the difference between the ideal MLE loss and the actual MLE loss can be approximated to 0. **This indicates that when training the model in the augmented dataset, the actual training objective can closely approximate the ideal training objective.**

C Appendix for Experiment

C.1 Dataset Detail

The dataset used is called Diaasq, including both a Chinese and an English dataset. The dataset is divided into train/test/dev sets in an 8:1:1 ratio. Aside from the dialogue text, the dataset also includes important details such as the speaker for each utterance, dialogue reply relationships, and reply thread relationships. Every dialogue originates from a root utterance, and multiple speakers take part in responding to preceding utterances. Multi-threaded and multi-turn dialogues form a tree structure based on reply relationships. In other words, dialogues are structured like trees, following reply relationships. **Each reply thread consists of all the utterances along the path from a leaf node**

to the root node, as illustrated in Figure 5. The dataset labels consist of ground truth tuples and the positions of their element. The statistical information of the dataset is shown in Table 6.

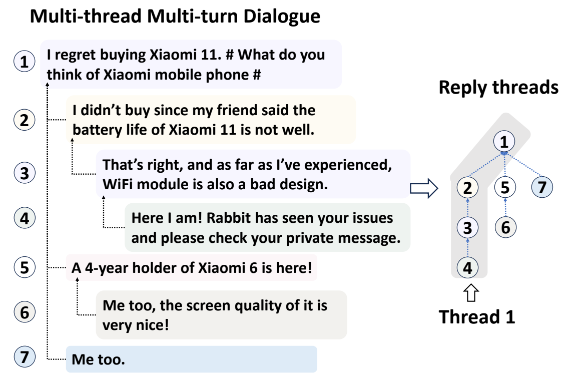

Figure 5: Reply threads. "2" →"1" means utterance "2" replies to utterance "1".

C.2 Detail of Metrics

We use micro F1 for the pair extraction task and both micro F1 and identification F1 for the quadruple extraction task, as stated in (Barnes et al., 2021). In micro F1, predicted tuples with all correct words are considered true positives (TP), while tuples with any incorrect word are considered false positives (FP). Tuples that were not predicted correctly are considered false negatives (FN). On the other hand, Identification F1 is similar to Micro F1, but it does not take sentiment elements into account.

C.3 Detail of Experiment Setting

We use BART (Lewis et al., 2020) (440M) for both EN and ZH datasets. We train the model for ten epochs (2 hours) on 4 3090 GPUS with a batch size of 5 and a learning rate of 5e-5, while other layers employ a learning rate of 8e-5. We use 3 cross-attentithreen layers. During testing, the beam search size is set to 2. All reported results are averaged over multiple runs.

C.4 Detail of Compared Baseline

CRF-ExtractClassify(CEC)(Cai et al., 2021) is a two-stage model that initially extracts aspect-opinion pairs and then predicts category-sentiment based on the extracted aspect-opinion pairs. **SpERT**(Eberts and Ulges, 2020) is a span-based transformer model for joint entity and relation extraction, initially extracting spans, filtering them, and finally classifying relationships among the spans. Modify this model to support quadruple extraction classification. **Span-ASTE**(Xu et al., 2021a) is a span-based model that explicitly considers interactions between the entire span of targets and opinions when predicting sentiment relations. Modify the final stage of SpanASTE to enumerate triplets, aligning it with the DiaASQ task. **ParaPhrase**(Zhang et al., 2021a), an end-to-end generation approach, introduces a novel paraphrase modeling paradigm to frame the ASQP task as a paraphrase generation process. **MvI**(Li et al., 2023a) method leverages speaker information, reply relationships, and thread information in dialogues to control information fusion between dialogues. Finally, it extracts quadruples based on the decoding output of Grid Tagging.

C.5 More analysis for main Experiment Results

ParaPhrase is a generative model that outperforms the discriminative model Span-ASTE on short text datasets but falls short on dialogue datasets. This is because the dialogue dataset has an increasing number of tuples, which widens the gap between the actual and ideal training objectives, i.e., increasing gap in $\Pi(\mathbb{S})$ and $\pi_k(\mathbb{S})$ as indicated by Equation 4 and 5. This amplifies the order bias interference in ParaPhrase. In contrast, Span-ASTE remains unaffected by tuple order bias, resulting in a reversal of performance on dialogue datasets shown in Table 1.

C.6 Augmentation Strategies in SOBM

To investigate the effectiveness of the augmentation strategy in SOBM, we compared it with other augmentation methods. By determining whether to shuffle the tuples in labels and the segmented fragments in inputs, we get various augmented datasets. We also compared SOBM with traditional data augmentation methods: synonyms, replacement, and deletion (**SRD**). The results are presented in Table 8. Compared to row 1, our method surpasses the first method by a maximum of 2.52 %(micro F1) on the EN dataset. The first method creates biased samples, while our method helps alleviate biases, improving the model's robustness and generalizability. The second method is actually a type of standard data augmentation method. So, it outperforms the first method by a maximum of 1.04%(micro F1) in the ZH dataset. However, the comparison between rows 5 and 2 shows that our method outperforms the second method by a maximum of

Table 6: Statistical information of the Diaasq dataset.

	Pairs			Quadruples			Utterance Length			Dialogue Length		
	Pairt-a	Pairt-o	Paira-o	Quad	Intra	Cross	Avg	Min	Max	Avg	Min	Max
EN	5894	7432	4994	5514	4287	1227	31	3	156	231	85	481
ZH	6041	7587	5358	5742	4467	1275	29	3	142	219	76	462

Table 8: Result of different augmentation methods. TDA - Traditional Data Augment.

Method	TDA	Shuffle		EN		ZH	
		Input	OutPut	Micro	Iden	Micro	Iden
w/o				36.35	41.64	35.76	39.21
In only		✓		37.31	41.68	36.80	39.78
Out only			✓	36.44	41.35	36.64	39.53
SRD	✓		✓	37.44	41.95	36.29	39.12
SOBM (Our)		✓	✓	**38.87**	**43.32**	**37.80**	**41.05**

1.64%(Iden F1) in the EN dataset. This emphasizes that our approach isn't merely an optional data augmentation technique but rather a necessary debiasing technique. Compared to row 3, our method outperforms the third method by a maximum of 2.43%(micro F1) in the EN dataset. The second method introduces a one-to-many learning challenge, while our method avoids this by pairing feasible labels with newly constructed inputs, facilitating models to converge to optimal performance. Compared to row 4, our method outperforms the fourth method by a maximum of 1.93%(Iden F1) in the EN dataset. This highlights the superiority of our augmentation technique over traditional methods in dialogue processing.

C.7 Compared Dialogue Segmentation Methods

Here is the detail of the compared dialogue segmentation methods:

1. TOD-BERT (Wu et al., 2020a): This method directly classifies the utterance relationships without introducing topic information. The method interacts with the contextual information between two utterances, and the classification is performed on the fused contextual information to achieve dialogue segmentation. Pass the fused contextual information through an MLP layer and then classify to determine whether the two utterances share the same topic or whether they need to be segmented. This method is the most commonly used approach in existing works.
2. Topic-Sentence Pair: This approach introduces "topics" and then performs classification on topic-utterance pairs, similar to our method. However, instead of using cross-attention for fine-grained information fusion, **it uses a concatenation operation to pool information.** Firstly, it performs average pooling on a topic word and on an utterance. Then, it concatenates the two pooled embeddings and passes them through an MLP layer for classification to determine whether the utterance belongs to the given topic. Apply this process to all the topics and utterances to finish the segmentation.
3. Simultaneously Multi-Granularity Denoising: This method incorporates sequence labeling and dialogue segmentation into the dialogue segmentation module. It doesn't need to **pre-label topics** topics. Instead, it views each word in an utterance as a potential topic and categorizes the connection between each word and the utterance. Based on the classification results, the method identifies words linked to any sentence as topics, while those without connections are not considered topics. This approach achieves both topic-centric clustering and topic labeling simultaneously. **However, this means that when we perform dialogue segmentation, there is no explicit guidance from topic information. Consequently, it must deal with the intricate contextual interactions between utterances.**
4. Reply Thread (RT): This method doesn't employ neural networks. Instead, it directly uses the inherent reply thread structures (as shown in Section C.1) in the dataset as the final dialogue segmentation scheme. In this segmentation scheme, every utterance, except for the initial dialogue, is assigned to a single topic-centric cluster. However, this approach lacks finer segmentation granularity, as seen in cases where an utterance may relate to multiple topics, as illustrated by the black utterances in Fig. 2.
5. Topic Word Match(TWM): This technique begins by labeling topic words within the utter-

ances. Then, it utilizes string-matching algorithms to determine whether an utterance belongs to a specific topic. Specifically, it checks if an utterance contains the topic string at the string level. If the topic is found in the utterance, it's considered to belong to that topic; otherwise, it's not. However, this method is limited to establishing connections between an utterance and a topic only when the utterance explicitly mentions the topic at the string text level. When an utterance indirectly references a topic or discusses related content, such as using pronouns, this approach proves ineffective.

C.8 Hyparameter Experiment

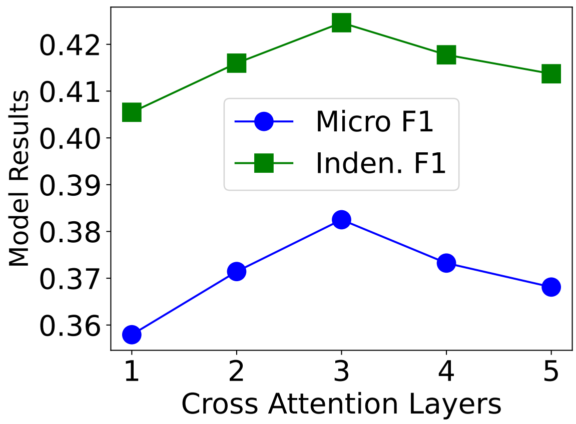

Figure 6: Results for the different number of cross attention layers.

We also investigated the impact of the number of cross-attention layers on model performance, keeping the batch size constant at four due to GPU memory limitations. The results are shown in Figure 6. The figure illustrates that increasing the number of cross-attention layers initially enhances model performance but then diminishes it. When there are fewer cross-attention layers, the model lacks sufficient interaction between topic and utterance information, limiting the exploration of their relationship. Conversely, an excessive number of cross-attention layers leads to overfitting due to a surplus of parameters and limited data, resulting in the incorporation of non-topic-related information during interaction.

C.9 LLM's performance

We experimented with various fine-tuning methods, fine-tuning the Qwen1.5(7B) (Bai et al., 2023) model on the English dataset. Fine-tuning methods include full parameters fine-tuning, Lora (Hu et al., 2021) fine-tuning, and Qlora (Dettmers et al., 2024) fine-tuning. The results are depicted in Fig. 7. Clearly, even with a smaller parameter setting, our approach outperforms the results of larger models.

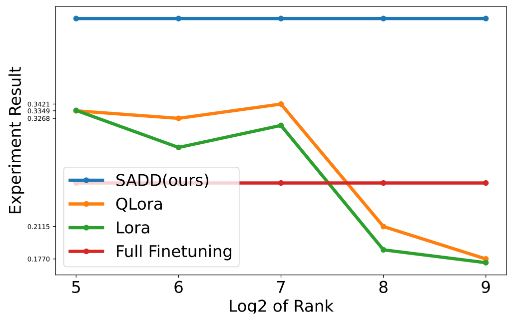

Figure 7: LLM's Performance. The x-axis represents the logarithmic values of Lora's rank number, and the y-axis represents the Iden F1 results.

Please rewrite the following dialogues in their original language so that change the content, but keep the general semantics and the elements in the quadruples unchanged:
Quadruples:["neg", "k50", "take photos", "a little better"], ["pos", "pro", "photo", "far worse than"], ["neg", "12", "photo", "far worse than"], ["neg", "K40", "game", "always 460"], ["pos", "Xiaomi", "taking pictures", "better"], ["neg", "Redmi", "taking pictures", "Don't consider"]
Here is the dialogue:
["I hope k50 can take photos a little better [School Season][School Season][School Season]", "Taking photos is really important !", "No matter how good the camera parameters are , it is useless without technology [doge]", "I went to the store to try the 12 and 12pro , the 12 's photo is far worse than the pro 's", "I agree with the photo too ! The top - level configuration version of the camera can also be enhanced", "The light is dark at night , and it is inconvenient to scan the code when going out .", "The telephoto scan code is even more fragrant .", "Do n't consider taking pictures using Redmi , Xiaomi is better [DOGE] .", "[Doge] [Doge] [Doge] Yes , I do n't know if you have encountered it , my K40 game is always 460 [Tears] [Tears] [Tears] ."]

["I hope the K50 can improve its photo-taking capabilities just a bit. [School Season][School Season][School Season]", "Taking photos really matters!", "No matter how impressive the camera specs are, without the right technology, they're useless. [doge]", "I visited the store to test out the 12 and 12 Pro. The 12's camera performance is far worse than the Pro's.", "I totally agree about the camera! Even the highest-tier camera setup can be further enhanced.", "It gets really dark at night, making it inconvenient for scanning codes when out and about.", "Scanning codes with the telephoto lens is even more satisfying.", "If you're considering photo quality, go for Xiaomi over Redmi. Xiaomi performs better. [DOGE]", "[Doge] [Doge] [Doge] You know what, I've been experiencing this issue too. My K40's gaming performance is consistently at 460. [Tears] [Tears] [Tears]."

Unexpected Changed Quaruples:
["neg", "k50", "take photos", "a little better"], ["neg", "K40", "game", "always 460"], ["pos", "Xiaomi", "taking pictures", "better"], ["neg", "Redmi", "taking pictures", "Don't consider"]

Figure 8: Chatgpt example 1.

Please rewrite the following dialogues in their original language so that change the content, but keep the general semantics and the elements in the quadruples unchanged:
Quadruples:["pos", "vivo", "rear camera module", "better"], ["neg", "11U", "rear camera module", "stands out too much"], ["pos", "11U", "design", "recognizable"], ["pos", "11U", "design", "very beautiful"],
Here is the dialogue:
["The price is not bad", "11U said (all are little brothers [doge] refers to the back design)", "The 11U rear camera module stands out too much , which vivo has done better . The next generation of Ultra has changed again , and Xiaomi really has no inheritance . . . In fact , the design of 11U is very beautiful , and it is also recognizable .", "The recognition is too great", "iQOO can be regarded as a firm high - end , Xiaomi has not hurry to learn [doge] ?", "Now I really do n't understand which of the X series and iQOO is the son of vivo ?"]

["The price is quite reasonable.", "Regarding the 11U, all the younger models (referred to as 'little brothers' [doge]) have similar rear designs.", "The rear camera module of the 11U stands out too much, which contrasts with vivo's more subtle approach. The next generation of Ultra has made further changes, highlighting Xiaomi's lack of design continuity. However, the design of the 11U is undeniably beautiful and distinctive.", "The design is truly remarkable.", "iQOO can definitely be considered a solid high-end option. Is Xiaomi in any hurry to catch up? [doge]", "Honestly, I'm confused about whether the X series or iQOO is the true heir to vivo."]

Unexpected Changed Quaruples:
["pos", "vivo", "rear camera module", "better"], ["pos", "11U", "design", "recognizable"], ["pos", "11U", "design", "very beautiful"]

Figure 9: Chatgpt example 2.