DMIN: A Discourse-specific Multi-granularity Integration Network for Conversational Aspect-based Sentiment Quadruple Analysis

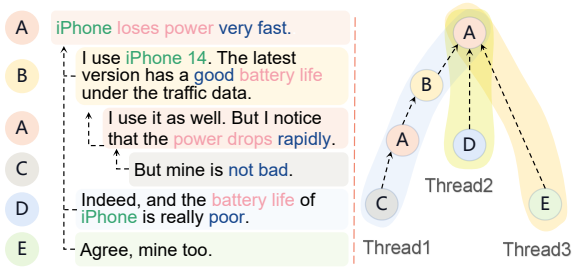
Peijie Huang[†], **Xisheng Xiao**[†], **Yuhong Xu**[*], **Jiawei Chen**
College of Mathematics and Informatics, South China Agricultural University, China
pjhuang@scau.edu.cn, xishengxiao.mail@gmail.com, xuyuhong@scau.edu.cn,
jw_chen@stu.scau.edu.cn

Abstract

Conversational Aspect-based Sentiment Quadruple Analysis (DiaASQ) aims to extract fine-grained sentiment quadruples from dialogues. Previous research has primarily concentrated on enhancing token-level interactions, still lacking in sufficient modeling of the discourse structure information in dialogue. Firstly, it does not incorporate interactions among different utterances in the encoding stage, resulting in a limited token-level context understanding for subsequent modules. Secondly, it ignores the critical fact that discourse information is naturally organized at the utterance level and learning it solely at the token level is incomplete. In this work, we strengthen the token-level encoder by utilizing a discourse structure called "thread" and graph convolutional networks to enhance the token interaction among different utterances. Moreover, we propose an utterance-level encoder to learn the structured speaker and reply information, providing a macro understanding of dialogue discourse. Furthermore, we introduce a novel Multi-granularities Integrator to integrate token-level and utterance-level representations, resulting in a comprehensive and cohesive dialogue contextual understanding. Experiments on two datasets demonstrate that our model achieves state-of-the-art performance. Our codes are publicly available at https://github.com/SIGSDSscau/DMIN.

1 Introduction

Conversational Aspect-based Sentiment Quadruple Analysis (DiaASQ) (Li et al., 2023) is a new compound subtask of Aspect-based Sentiment Analysis (ABSA). As shown in Figure 1, DiaASQ aims to extract all (**t**, **a**, **o**, **s**) sentiment quadruples present in a conversation, where the target **t** (the subject of discussion), aspect **a** (specific attribute of target) and opinion **o** (attitude or evaluation towards aspect) represent substrings of dialogue, and sentiment **s** represents one of the categories of positive, negative, or neutral. Given an example sentence, "iPhone loses power very fast.", the corresponding elements are "iPhone", "loses power", "very fast", and "negative", respectively. Each quadruple can be considered as a fine-grained opinion that conveys sentiment towards a specific aspect of the target.



Type	Target	Aspect	Opinion	Sentiment
Intra-utt	iPhone	loses power	very fast	negative
Intra-utt	iPhone 14	battery life	good	positive
Cross-utt	iPhone 14	power drops	rapidly	negative
Cross-utt	iPhone 14	power drops	not bad	neutral
Intra-utt	iPhone	battery life	poor	negative

Figure 1: An example dialogue (top-left), along with the corresponding tree-like reply structure (top-right) and sentiment quadruples (bottom), where different sentiment elements are highlighted in various colors, dashed lines represent reply relationships, and letters inside circles denote different speakers.

As shown in Figure 1, the dialogue has a structural discourse, including multiple utterances and corresponding speakers, and except for the root utterance, each speaker's utterance has a reply object. Compared to aspect-based sentiment quad prediction (ASQP) (Zhang et al., 2021; Mao et al., 2022), another subtask of ABSA that extracts quadruples from plain text, DiaASQ faces two additional challenges. On the one hand, the relatively lengthy nature of the dialogue makes it more challenging to establish long-term dependency relationships be-

[†] Equal contribution.
[*] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 16326–16338

tween targets, aspects, and opinions. On the other hand, the dialogue encompasses crucial reply relationships and structured discourse information, making it a challenging task to model the discourse information of the dialogue accurately. Specifically, the elements of the quadruples may come from different utterances. Taking Figure 1 as an example, the opinion term "not bad" in Speaker-C's utterance is related to the target term "iPhone 14" in Speaker-B's utterance and the aspect term "power drops" in Speaker-A's utterance. Compared with intra-utterance quadruples, such cross-utterance quadruples require a more comprehensive understanding of the interaction of the utterances and speakers.

In DiaASQ, Li et al. (2023) designed a new labeling scheme of grid tagging (Wu et al., 2020) and proposed the MVQPN network, which models the discourse structure of conversations at the token level through three kinds of mask multi-head attention mechanisms (Vaswani et al., 2017). However, the model didn't capture the interaction of utterances in the encoding stage. Cai et al. (2023) tried to solve this issue by encoding the whole dialogue in the pretrain language models (PLMs) layer. Lai et al. (2023) changed the parallel attentions to continuous, deepening the network structure.

Though achieving a promising performance, the model's context modeling for structured conversation is incomplete and inadequate: (1) It does not consider the influence of other utterances when encoding each sentence, resulting in an incomplete contextual understanding for subsequent modules; (2) They neglect the critical fact that discourse information is naturally organized at the utterance level, and learning structured discourse information solely from a token-level perspective is inadequate for achieving optimal results. For example, the expression "utterance-A replies to utterance-B" is more natural and efficient than a lot of "word-A replies to word-B". The same goes for speaker information.

In this work, we propose the **D**iscourse-specific **M**ulti-granularity **I**ntegration **N**etwork, DMIN, to provide a more complete contextual understanding for structured conversation, enhancing the extraction of intra-utterance quads and cross-utterance quads. For the first issue mentioned above, we employ the Concrete Knowledge Encoder (CKEncoder) to capture syntactic and semantic information at the token level, which leverages the discourse structure called "thread" to enhance the interaction between different utterances. As shown in the top-right corner of Figure 1, a dialogue can be structured as a tree based on the reply relationships, and the so-called "thread" refers to the subtree derived from the root node of the conversation tree. For the second issue, we learn the speaker and reply relationships at the utterance level through a Global Discourse Encoder (GDEncoder). The GDEncoder offers a more natural and comprehensive structural discourse information from a macro perspective. Furthermore, as the two modules' information granularity and representation dimensions are different, direct fusion is not feasible. Therefore, we propose a novel Multi-Granularity Integrator to effectively combine the token-level and utterance-level information.

In essence, DiaASQ requires the extraction of substrings from the text, necessitating that the model's minimum granularity must be at the token level. However, the natural organization of dialogue discourse structure is based on the utterance level, generating representations at the utterance level. Our model provides these two kinds of complementary information and fuses them effectively, yielding a more comprehensive and cohesive contextual understanding.

Our contributions can be summarized as follows:

(1) We strengthened the token-level encoder and introduced a global perspective to learn the structural discourse information at the utterance level, providing a more comprehensive approach for modeling the context of structured conversations.

(2) We proposed an original multi-granularity fusion module, addressing the fusion challenge of two different-dimensional information.

(3) Our experimental results on two datasets demonstrated that our model achieves state-of-the-art performance, showing a 6.72% and 3.85% improvement in Micro F1, respectively.

2 Related Work

2.1 Aspect-based Sentiment Analysis

DiaASQ is one of the new subtasks of ABSA. The early research on ABSA primarily focused on plain text with short lengths and without any structures. Initially, studies concentrated on single-element extraction tasks such as extracting aspect terms **a** (Li et al., 2018) and analyzing sentiment polarity **s** (Li et al., 2021). Subsequent tasks involved the analysis of composite sentiment elements, for example, the output of (**a, s**) for Aspect-Opinion Pair Extraction (AOPE) (Wu et al., 2021), the output of (**a, o,**

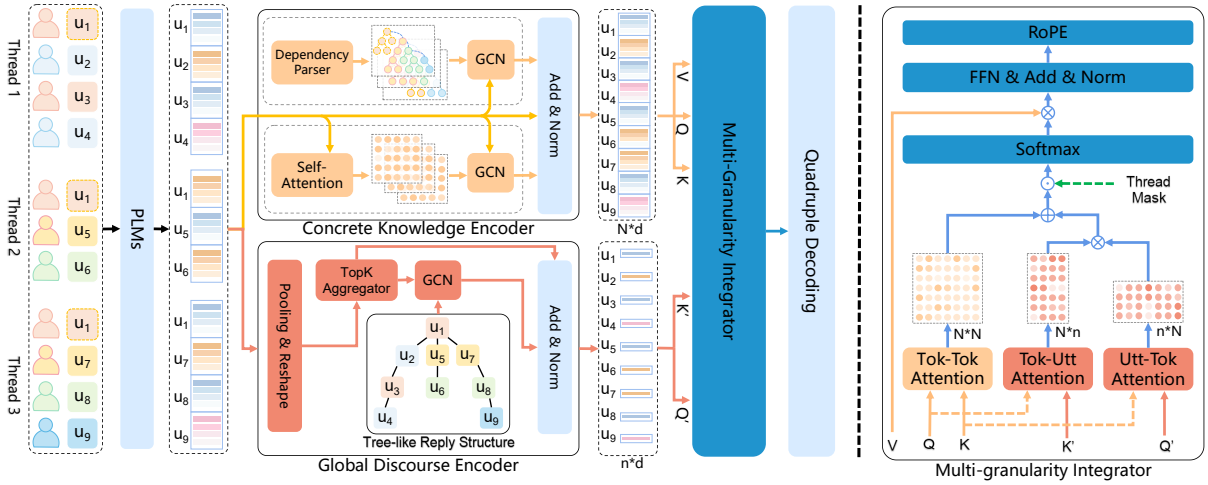


Figure 2: The overall architecture of our proposed DMIN.

s) for Aspect-Sentiment-Term Extraction (ASTE) (Chen et al., 2022), the output of sentiment quadruples (**a, c, o, s**) for Aspect Sentiment Quad Prediction (ASQP) (Zhang et al., 2021; Mao et al., 2022) and so on. Here, '**c**' refers to predefined categories, with each aspect associated with a specific category. In ASQP, recent works have utilized a generation-based approach (Zhang et al., 2021) and specific templates (Mao et al., 2022; Hu et al., 2022) to extract quadruples, thereby mitigating the potential error propagation. However, these approaches struggle to incorporate structured discourse information (speakers and reply relationships) into the generation-based framework naturally.

2.2 Conversational Aspect-based Sentiment Quadruple Analysis

Unlike other conversational understanding tasks (Cheng et al., 2023a,b), DiaASQ requires handling explicit reply relationships. In DiaASQ, Li et al. (2023) re-designed the labeling scheme of the grid-tagging method (Wu et al., 2020), decomposing the original task into Entity Boundary Prediction, Entity Pair Prediction, and Polarity Prediction. Additionally, they designed the speaker mask, reply mask, and thread mask, together with the mask multi-head attention and Rotary Position Embedding (RoPE) (Su et al., 2021) to strengthen the awareness of the dialogue discourse. Overall-QPN (Cai et al., 2023) is a model based on MVQPN that proposes a method for encoding the entire dialogue at the PLMs layer. However, dialogues are often very long and may exceed the acceptable range of PLMs. Lai et al. (2023) modified three parallel attentions of MVQPN to be consecutive and trained the model relied on the k-fold strategy and manual rules. They also utilized weights trained on a Chinese dataset to initialize the model for training on English datasets.

3 METHODOLOGY

The overall architecture of our proposed DMIN is shown in Figure 2. In DiaASQ, each dialog is represented as a training sample denoted as $D = \{u_1, ..., u_n\}$ with the corresponding relying record $r = \{l_1, ..., l_n\}$ and speakers $s = \{s_1, ..., s_n\}$, where relying record l_i denotes the *i-th* utterance reply to the l_i*-th* one. Each utterance $u_i = \{w_1, ..., w_{m_i}\}$ where m_i is the length of u_i.

Following the labeling scheme of grid tagging proposed by Li et al. (2023), we decompose the original DiaASQ task into three joint jobs, and the model aims to predict the entity boundary labels $y^{ent} \in \{tgt, asp, opi, other\}$, the entity pair labels $y^{pair} \in \{h2h, t2t, other\}$ and the polarity labels $y^{pol} \in \{pos, neg, neu, other\}$, where *tgt*, *asp*, and *opi* denote the token-level relations between the head and tail of a target, aspect, and opinion term, respectively (e.g., the label tgt between head token "iPhone" and tail token "14" denotes a target term "iPhone 14"). The labels *h2h* (head-to-head) and *t2t* (tail-to-tail) are used to align the head and tail tokens between a pair of entities in two types (e.g., the target's head token "iPhone" and aspect's head token "battery" is connected by h2h, while the tail token "14" and "life" is connected with t2t).

3.1 Textual Features

We observed that utterances within the same thread generally exhibit topical solid relevance. Therefore,

we leverage the discourse units, thread, to enhance the contextual extraction capabilities of PLMs (Devlin et al., 2019). This approach strikes a balance by staying within the maximum acceptable text length of PLMs while effectively enhancing the interaction between utterances. Additionally, in order to obtain the speaker's representation information, we add the speaker's ID after the corresponding utterance $u'_i = \{[cls], u_i, s_i\}$, where $[cls]$ is the special token of PLMs.

To ensure generality, we designate u'_1 as the root utterance and include it at the beginning of each thread. Assuming that *k-th* thread $t_k = \{u'_1, u'_i, u'_{i+1}, ..., u'_j\}$, the representations of it can be defined as follows:

$$\boldsymbol{H}^t_k = \{\boldsymbol{H}^{u'}_1, \boldsymbol{H}^{u'}_i, ..., \boldsymbol{H}^{u'}_j\} = \text{PLMs}(t_k), \quad (1)$$

$$\boldsymbol{H}^{u'}_i = \{\boldsymbol{h}^{cls}_i, \boldsymbol{H}^u_i, \boldsymbol{h}^s_i\}, \quad (2)$$

where each utterance feature $\boldsymbol{H}^u_i \in \mathbb{R}^{m_i \times d}$ consists of token-level representations.

3.2 Concrete Knowledge Encoder

DiaASQ requires extracting specific entities from dialogue, necessitating a profound understanding of token-level knowledge and information. We thus propose a CKEncoder to enhance the token features for the subsequence module by incorporating two dedicated modules to learn syntactic knowledge and semantic information, respectively. Both modules are based on Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017; Chen et al., 2022), which is capable of effectively modeling relationships and dependencies between nodes. Assuming that the graph contains n nodes and the activation function is denoted by σ, the representation of the i^{th} node in the l^{th} layer of GCN can be formulated as follows:

$$\boldsymbol{h}^l_i = \sigma\left(\sum_{j=1}^{n} \boldsymbol{A}_{ij}\boldsymbol{W}^l\boldsymbol{h}^{l-1}_j + \boldsymbol{b}^l\right), \quad (3)$$

where $\boldsymbol{A}$ represents the adjacency matrix, $\boldsymbol{W}$ and $\boldsymbol{b}$ is learnable parameters and bias.

Syntactic GCN. In many previous works of ABSA (Zhang et al., 2022; Chen et al., 2022), dependency parsing trees have been proven to have the ability to establish dependency relationships between aspect words and opinion words. We thus use GCNs to learn syntactic information from the dependency tree[1]. However, dependency parsers are generally only applicable to short texts and cannot be directly applied to dialogue texts. For this, we have made some adaptive improvements by establishing dependency relationships for each utterance and then linking their root token nodes to each other based on the utterance reply relationships within the same thread. Specifically, we construct a syntactic adjacency matrix for *k-th* thread as follows:

$$\boldsymbol{A}^{syn}_{k,ij} = \begin{cases} 1, & \text{if words } w_i, w_j \text{ contain} \\ & \text{dependency relationship,} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

The input for the first layer of GCNs is $\boldsymbol{A}^{syn}$ and thread text feature $\boldsymbol{H}^t$, and the syntactic representation $\boldsymbol{H}^{syn} = \text{GCNs}(\boldsymbol{A}^{syn}, \boldsymbol{H}^t)$ is obtained from GCNs using Eq.(3).

Semantic GCN. We further learn the semantic information from the self-attention mechanism (Vaswani et al., 2017), enhancing utterance interaction in thread range. The semantic adjacency matrix can be formulated as:

$$\boldsymbol{A}^{sem} = \text{Atten}(\boldsymbol{H}^t\boldsymbol{W}^Q, \boldsymbol{H}^t\boldsymbol{W}^K), \quad (5)$$

$$\text{Atten}(\boldsymbol{Q}, \boldsymbol{K}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}). \quad (6)$$

where d is the dimension of thread text features $\boldsymbol{H}^t$. Similar to Syntactic GCN, the semantic representation $\boldsymbol{H}^{sem} = \text{GCNs}(\boldsymbol{A}^{sem}, \boldsymbol{H}^t)$ is obtained by Eq.(3) using $\boldsymbol{A}^{sem}$ and $\boldsymbol{H}^t$ as input.

Feature Fusion. By employing residual connections (He et al., 2016) and layer normalization operations, denoted as LN, we combine $\boldsymbol{H}^{syn}$ and $\boldsymbol{H}^{sem}$ with text feature $\boldsymbol{H}^t$ to obtain the token-level concrete representation for each thread:

$$\boldsymbol{H}^{tok} = \text{LN}(\boldsymbol{H}^t + \boldsymbol{H}^{syn} + \boldsymbol{H}^{sem}). \quad (7)$$

3.3 Global Discourse Encoder

Discourse information such as speaker and reply relationships is naturally organized at the utterance level. We first designed the TopK Aggregator to obtain sentence representation and speaker information and then learned structured reply relationships through GCNs. Particularly, root utterance $\boldsymbol{H}^u_1$ is obtained by avg-pooling at the first as it is repeated in each thread.

TopK Aggregator. Here we calculate each utterance's token score $\boldsymbol{S}^u = \{\boldsymbol{s}^u_1, \ldots, \boldsymbol{s}^u_n\}$ and then get the index of maximum κ tokens:

$$\boldsymbol{s}^u_i = \boldsymbol{H}^u_i \cdot \boldsymbol{W}^s + b^s, \quad (8)$$

[1]We use the spaCy as dependency parser: https://spacy.io

$$\boldsymbol{idx}_i^u = \text{argmax}(\boldsymbol{S}_i^u, \kappa), \quad (9)$$

where parameter $\boldsymbol{W}^s \in \mathbb{R}^{d\times 1}$, *i-th* utterance's token scores $\boldsymbol{s}_i^u \in \mathbb{R}^{m_i\times 1}$, $\kappa = m_i * \lambda$ and $\lambda \in (0, 1]$ is a hyperparameter. Then the weighted representation can be obtained by taking the element-wise product of the corresponding tokens and scores:

$$\boldsymbol{H}_i^{wu} = \text{softmax}(\boldsymbol{S}_i^u[\boldsymbol{idx}_i^u]) \odot \boldsymbol{H}_i^u[\boldsymbol{idx}_i^u]. \quad (10)$$

We use the concatenation operator "||" and a linear layer to combine the max features, average features, and speaker information, getting the overall representation $\boldsymbol{h}_i^o \in \mathbb{R}^{1\times d}$ of utterance :

$$\boldsymbol{h}_i^o = \text{MLP}(\max(\boldsymbol{H}_i^{wu})||\,\text{avg}(\boldsymbol{H}_i^{wu})||\boldsymbol{h}_i^s). \quad (11)$$

Discourse GCN. To acquire the reply relationships of the entire dialogue, we construct a discourse adjacency matrix $\boldsymbol{A}^{dsc}$:

$$\boldsymbol{A}_{ij}^{dsc} = \begin{cases} 1, & \text{if utterances } u_i, u_j \text{ contain} \\ & \text{replying relationship,} \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

We learn the utterance-level structured context and speaker information from the GCNs and finally get the discourse features:

$$\boldsymbol{H}^{dsc} = \text{LN}(\text{GCNs}(\boldsymbol{A}^{dsc}, \boldsymbol{H}^o) + \boldsymbol{H}^o), \quad (13)$$

where $\boldsymbol{H}^o = \{\boldsymbol{h}_1^o, ..., \boldsymbol{h}_n^o\}$ is the utterance-level representations.

3.4 Multi-Granularity Integrator

We have acquired the token-level representation from CKEncoder and learned the utterance-level discourse structure information from GDEncoder. However, these two representations have different dimensions and cannot be directly fused. Therefore, we propose a Multi-Granularity Integrator to solve this issue, yielding a complete contextual representation.

Multi-Granularity Attention. Specifically, the integrator addresses the challenge through unique attention score matrix (Vaswani et al., 2017), namely Token to Token Attention Score $\boldsymbol{S}^{tok\text{-}tok} \in \mathbb{R}^{N\times N}$, Token to Utterance Attention Score $\boldsymbol{S}^{tok\text{-}utt} \in \mathbb{R}^{N\times n}$ Utterance to Token Attention Score $\boldsymbol{S}^{utt\text{-}tok} \in \mathbb{R}^{n\times N}$:

$$\boldsymbol{S}^{tok\text{-}tok} = \text{Atten}(\boldsymbol{W}^1\boldsymbol{Q}^{tok}, \boldsymbol{W}^2\boldsymbol{K}^{tok}), \quad (14)$$

$$\boldsymbol{S}^{tok\text{-}utt} = \text{Atten}(\boldsymbol{W}^1\boldsymbol{Q}^{tok}, \boldsymbol{W}^3\boldsymbol{K}^{utt}), \quad (15)$$

$$\boldsymbol{S}^{utt\text{-}tok} = \text{Atten}(\boldsymbol{W}^4\boldsymbol{Q}^{utt}, \boldsymbol{W}^2\boldsymbol{K}^{tok}), \quad (16)$$

where n and N respectively represent the overall number of utterances and tokens in the dialogue, $\boldsymbol{Q}^{tok}$, $\boldsymbol{K}^{tok}$, and $\boldsymbol{V}^{tok}$ in the Eq.(14) $\sim$ Eq.(18) are representations obtained by concatenating all the token features of $\boldsymbol{H}^{tok}$ in the order they appear in the conversation, $\boldsymbol{Q}^{utt}$ and $\boldsymbol{K}^{utt}$ represent the utterance-level discourse feature learned from the GDEncoder.

By combining these two different granularities of attention, together with token level feature, we obtain the integrated representation $\boldsymbol{H}^{itg} \in \mathbb{R}^{N\times d}$:

$$\boldsymbol{A}^{itg} = \boldsymbol{S}^{tok\text{-}utt} \cdot \boldsymbol{S}^{utt\text{-}tok} + \boldsymbol{S}^{tok\text{-}tok}, \quad (17)$$

$$\boldsymbol{H}^{itg} = \text{softmax}(\boldsymbol{A}^{itg} \odot \boldsymbol{M}^{th}) \cdot \boldsymbol{V}^{tok}, \quad (18)$$

where $\boldsymbol{M}^{th}$ refers to the thread mask proposed by Li et al. (2023), $\boldsymbol{M}_{ij}^{th} = 1$ if *i-th* and *j-th* token within the same thread.

Finally, we conduct a Feedforward layer and a LayerNorm over the representations, followed by a tag-wise MLP layer to yield the feature representation $\boldsymbol{v}_i^\gamma$ for each token:

$$\boldsymbol{H}^f = \text{LN}(\text{FFN}(\boldsymbol{H}^{itg}) + \boldsymbol{H}^{itg}), \quad (19)$$

$$\boldsymbol{v}_i^\gamma = \text{MLP}(\boldsymbol{h}_i^f), \quad (20)$$

where $\gamma \in \{y^{ent} \cup y^{pair} \cup y^{pol}\}$ indicates a specific label.

Rotary Position Embedding (RoPE). RoPE (Su et al., 2021) can guide a better understanding of dialogue context. Following Li et al. (2023), the model fuses RoPE into token representations $\boldsymbol{u}_i^\gamma = \boldsymbol{R}(\theta, i)\boldsymbol{v}_i^\gamma$, where $\boldsymbol{R}(\theta, i)$ is a positioning matrix parameterized by θ and the absolute index i of $\boldsymbol{v}_i^\gamma$.

3.5 Quadruple Decoding and Learning

According to the grid tagging method (Li et al., 2023), the score s_{ij}^γ indicating the probability of relation label γ between w_i and w_j can be calculated as $s_{ij}^\gamma = (\boldsymbol{u}_i^\gamma)^T\boldsymbol{u}_j^\gamma$. Then we put a softmax layer over all elements to determine the relation label γ.

The training loss $\mathcal{L}$ of all subtasks can be defined as:

$$\mathcal{L}_\epsilon = -\frac{1}{G\cdot N^2}\sum_{g=1}^{G}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\alpha}^\epsilon y_{ij}^\epsilon \log(p_{ij}^\epsilon), \quad (21)$$

$$\mathcal{L} = \mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{pair}} + \mathcal{L}_{\text{pol}}, \quad (22)$$

where $\epsilon \in \{ent, pair, pol\}$ indicates the subtask, N is the total token length in a dialogue, G is total training instances, y_{ij}^ϵ is ground-truth label, p_{ij}^ϵ is the prediction, $\boldsymbol{\alpha}^\epsilon$ are weighting hyperparameters.

4 EXPERIMENT

4.1 Datasets

We evaluate our proposed model on Chinese dataset **ZH** (Li et al., 2023) and English dataset **EN** (Li et al., 2023), which are closely related to electronic products collected from social media. The dataset ZH contains 1000 dialogues and 5742 sentiment quadruples in total, while the dataset EN contains 5514 quadruples. The ratio of training, validation, and testing sets for two datasets is 8:1:1. In the trainsets, there are 1013 quadruples in ZH (about 21.99%) and 972 quadruples in EN (about 22.02%) are cross-utterance quadruples, where the elements (aspect, target, and opinion) in one quadruple are extracted from more than one utterance. More details about datasets can be found in Appendix A.

4.2 Baselines

As few prior methods are designed for DiaASQ, we compare with several strong-performing systems closely related to the task, which have been adjusted to support DiaASQ by Li et al. (2023). Additionally, considering the powerful zero-shot and few-shot capabilities of the current popular large-scale language models (LLMs), we conducted some experiments on ChatGPT to validate its performance on DiaASQ:

- **Three-stage model.** CRF-Extract-Classify (Cai et al., 2021) is an end-to-end system with extraction, filter, and combination stages for quadruple ABSA.
- **Span-based models.** Span-ASTE (Xu et al., 2021) and SpERT (Eberts and Ulges, 2020) are span-based approach for entities and relations joint extraction.
- **Generative model.** ParaPhrase (Zhang et al., 2021) is a generative seq-to-seq model for the quadruple ABSA.
- **LLM.** ChatGPT-3.5-turbo[2] is a large-scale language model based on GPT3 (Brown et al., 2020). We set the temperature parameter of ChatGPT to 0 to obtain stable outputs.

We also compared some models specifically designed for DiaASQ:

- **MVQPN** is a grid-tagging-based model (Li et al., 2023) for DiaASQ and it strengthens the discourse awareness of dialogues at the token level. Our DMIN is based on it.
- **Overall-QPN**[3] (Cai et al., 2023) is based on MVQPN, and models the overall dialogue with PLMs at the encoding stage. Overall-QPN was the second-place solution in the DiaASQ competition organized by NLPCC 2023 [4]. To ensure a fair comparison, we followed Li et al. (2023) and standardized the use of chinese-roberta-wwm-ext-base (102 million parameters) (Cui et al., 2021) for the ZH dataset, instead of Erlangshen-DeBERTa-v2-320M-Chinese (320 million parameters) (He et al., 2020).

4.3 Implementation Details

Settings. For dataset EN and ZH, we use Roberta-Large (Liu et al., 2019) and Chinese-Roberta-wwm-ext-base (Cui et al., 2021) as the PLMs layer and set the ratio λ of top-k as 0.5 and 0.8, respectively. The layer numbers of Syntatic GCN and Semantic GCN are set to 3, while Discourse GCN is 2. The batch size and the dropout rate are set to 2 and 0.1. We set the learning rate as 1e-4, except for the PLMs, which is 1e-5. The results of our implemented models are based on an average of 5 random runs on the test set. More experimental details can be found in Table 4.

Metrics. Following Li et al. (2023), we use the *exact F1* as the metric. We adopt Micro F1 and Identification F1 (Barnes et al., 2021) to measure the performance of quadruple extraction, which is the most important and challenging task of DiaASQ. Micro F1 measures the whole quadruple, while identification F1 does not distinguish the polarity and is more suitable for evaluating the model's boundary prediction and element-matching ability. To further analyze the performance, we detect the F1 scores of the span pair, i.e., Target-Aspect, Aspect-Opinion, and Target-Opinion, denoted as T-A, T-O, and A-O, respectively.

4.4 Main Results

The main experimental result is shown in Table 1. There are some notable observations:

(1) As our strong baseline method, we observe MVQPN already surpasses previous models by a large margin. Thanks to our token-level and utterance-level encoders, in terms of the most important metric, Micro F1, DMIN can effectively

[2]https://chat.openai.com/chat

[3]https://github.com/terence1023/NLPCC2023-DiaASQ.

[4]We did not compare with the first-place solution (Lai et al., 2023) because it relied on manual rules.

Dataset	Model	Pair Extraction(F1)			Quadruple(F1)	
		T-A	T-O	A-O	Micro	Ident.
ZH	ChatGPT$_{\text{zero-shot}}$	23.86	10.55	15.81	13.77	18.15
	ChatGPT$_{\text{one-shot}}$	29.90	17.48	25.59	18.26	20.56
	CRF-Extract-Classify (Cai et al., 2021)	32.47	26.78	18.90	8.81	9.25
	SpERT (Eberts and Ulges, 2020)	38.05	31.28	21.89	13.00	14.19
	ParaPhrase (Zhang et al., 2021)	37.81	34.32	27.76	23.27	27.98
	Span-ASTE (Xu et al., 2021)	44.13	34.46	32.21	27.42	30.85
	Overall-QPN (Cai et al., 2023)	52.86	50.98	53.33	37.77	43.56
	MVQPN (Li et al., 2023)	48.61	43.31	45.44	34.94	37.51
	Ours DMIN	**57.62**	**51.65**	**56.16**	**44.49**	**47.50**
EN	ChatGPT$_{\text{zero-shot}}$	23.26	16.07	14.34	10.98	12.99
	ChatGPT$_{\text{one-shot}}$	26.18	20.33	21.20	13.20	14.67
	CRF-Extract-Classify (Cai et al., 2021)	34.31	20.94	19.21	11.59	12.80
	SpERT (Eberts and Ulges, 2020)	28.33	21.39	23.64	13.07	13.28
	ParaPhrase (Zhang et al., 2021)	37.22	32.19	30.78	24.54	26.76
	Span-ASTE (Xu et al., 2021)	42.19	30.44	45.90	26.99	28.34
	Overall-QPN (Cai et al., 2023)	50.70	49.46	50.31	35.37	39.73
	MVQPN (Li et al., 2023)	47.91	45.58	44.27	33.31	36.80
	Ours DMIN	**53.49**	**52.66**	**52.09**	**39.22**	**42.31**

Table 1: The overall performance of different baseline models and our proposed DMIN, where 'T/A/O' represents Target/Aspect/Opinion, respectively. All the scores are averaged values over five runs under different random seeds.

improve the performance of MVQPN by 9.55% on the ZH dataset and 5.91% on the EN dataset. As for Overall-QPN, the current best baseline, DMIN also achieves a notable improvement of 6.72% and 3.85%. This showcases our better performance in extracting complete quadruples.

(2) Regarding identification F1, DMIN outperforms Overall-QPN by 3.94% and 2.58% on the ZH and EN, respectively. This demonstrates the superior performance of DMIN in entity extraction and target-aspect-opinion relationship matching.

(3) DMIN achieves improvements on all metrics in Pair Extraction compared with Overall-QPN, indicating that it has excellent ability in pairing binary relationships. Taking the EN dataset as an example, DMIN achieved the largest improvement in the T-O metric, 3.2%, while obtaining a relatively smaller improvement in the A-O metric, 1.78%.

(4) Regarding the LLM experiments, we observed that the ChatGPT-3.5-turbo model did not perform well in the zero-shot and one-shot settings on the DiaASQ task. When compared to another supervised generative model called ParaPhrase, it showed a difference of approximately 5% in performance. This could be because ChatGPT-3.5-turbo lacks an understanding of the structural information within the dialogue. To further investigate this, we conducted an experiment where we did not provide any information about the reply relationships or speaker identities. As a result, ChatGPT's predicted F1 scores exhibited minimal fluctuations. The prompts used in the experiment are presented in the Appendix C.

4.5 Ablation Study

As DMIN is focused on structured context understanding, we additionally observed the performance of Micro F1 on cross-utterance cases as one of the indicator to explore each module's contribution toward the structured context understanding.

Effects of Token-level Concrete Knowledge. From Table 2, it can be observed that removing the CKEncoder (w/o CKEncoder) results in a decline on all of metrics for both datasets. Taking dataset ZH as an example, the model decrease with approximately 1.28% on Micro F1, 0.72% on identification F1, and 3.17% on cross-utterance F1, respectively. The significant decrease of cross-utterance

Model	ZH			EN		
	Micro F1	Ident. F1	Cross-Utt.	Micro F1	Ident. F1	Cross-Utt.
Ours DMIN	44.49	47.50	31.23	39.22	42.31	25.56
w/o CKEncoder	43.21 (↓1.28)	46.78 (↓0.72)	28.06 (↓3.17)	38.54 (↓0.68)	42.07 (↓0.24)	23.75 (↓1.81)
w/o SynGCN	42.30 (↓2.19)	45.99 (↓1.51)	27.64 (↓3.59)	36.97 (↓2.25)	40.62 (↓1.69)	25.35 (↓0.21)
w/o SemGCN	42.73 (↓1.76)	46.58 (↓0.92)	25.83 (↓5.4)	38.25 (↓0.97)	40.92 (↓1.39)	22.93 (↓2.63)
w/o Utt-Discourse	42.94 (↓1.55)	46.26 (↓1.24)	28.32 (↓2.91)	38.54 (↓0.68)	42.08 (↓0.23)	23.03 (↓2.53)
w/o DscGCN	43.33 (↓1.16)	46.10 (↓1.40)	29.26 (↓1.97)	38.12 (↓1.10)	41.20 (↓1.11)	21.65 (↓3.91)
w/o Speaker	42.69 (↓1.80)	46.08 (↓1.42)	28.31 (↓2.92)	38.98 (↓0.24)	42.19 (↓0.12)	22.78 (↓2.78)
w/o Thread	42.11 (↓2.38)	44.69 (↓2.81)	26.87 (↓4.36)	36.36 (↓2.86)	39.46 (↓2.85)	16.74 (↓8.82)

Table 2: The results of ablation study on datasets. All of the metrics are Micro F1 scores, where "Cross-Utt." refers to the Micro F1 scores of the model on cross-utterance quadruples.

F1 demonstrates that syntactic knowledge and semantic information provide auxiliary support for model's structured context comprehension. Additionally, we conducted ablative experiments on two sub-modules separately (w/o SemGCN and w/o SynGCN), and the results showed a varying degree of decrease as expected.

Effects of Utterance-level Discourse. We initially removed all speakers, GDEncoder, and the corresponding integration mechanism (w/o Utt-Discourse) to assess the impact on structured context understanding. Consequently, the model's performance has a decrease of approximately 1.55% and 0.68% in Micro F1 scores on the two datasets. Furthermore, we conducted additional experiments by individually removing the reply relationship (w/o DscGCN) and speaker information (w/o Speaker) to evaluate their respective contributions. The results showed that their removal led to varying performance drops ranging from approximately 0.24% to 1.80% Micro F1, confirming their importance. Notably, we observed that the performance of cross-utterance quad extraction on the EN dataset was more influenced by replying information,while the ZH dataset was more influenced by speaker information.These experiments confirm the importance of discourse information, such as speaker and reply relationships, and also validate the effectiveness of our proposed Multi-Granularity Integrator.

Effects of Thread. The inclusion of the "thread" in the encoding stage allows for a wider range of token interactions. As demonstrated in Table 2, we conducted an experiment where we replaced the thread-range embedding and GCNs with utterance-range counterparts (w/o thread). This resulted in a decrease in overall micro scores of more than 2.38% on both datasets. Particularly, the cross-utterance quadruples experienced a larger drop, with a decrease of 8.82% on the EN dataset. This highlights the importance of the thread in achieving optimal performance in the model.

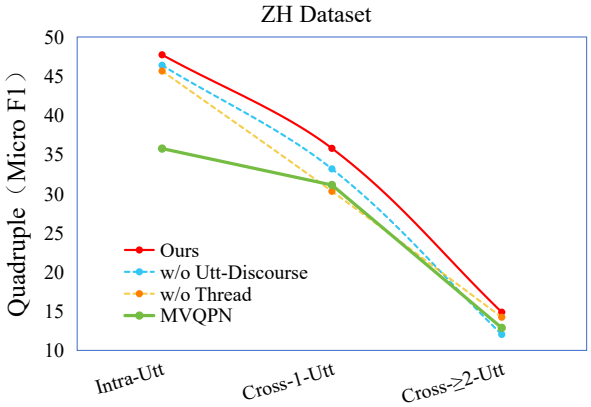


Figure 3: Performances on different cross-utterance levels.

4.6 Improvement Analysis

In Section 4.5, we conducted ablation studies on various modules to validate their contributions to structured dialogue understanding and additionally observed the changes in F1 scores of the model on more challenging cross-utterance cases to explore the capabilities of each module in addressing such difficult examples. In this section, we further performed an in-depth analysis by examining the performance of different levels of cross-utterance quad extraction. As shown in Figure 3, as the cross-utterance level increases, the performance of all models gradually decreases. Our DMIN achieved a

significant improvement of approximately 12% and 4% in short-range (intra-utterance) and medium-range (cross-1 utterance), respectively. Even in the most challenging long-range cases (cross-$\geq$2), our performance was approximately 2% higher than the baseline. Furthermore, we studied the performance of utterance-level discourse information and thread range encoding method. The results indicated that the thread range encoding method was more beneficial for short to medium-range quad extraction, while utterance discourse information played a more significant role in understanding the dialogue structure over longer distances.

4.7 Case Study

In this section, we present a close look at our DMIN and the baseline MVQPN via case studies. In the complex dialogue shown in Figure 4, there are a total of 3 threads, 4 cross-utterance quadruples, and 4 intra-utterance quadruples. The baseline model correctly predicts only 2 of the quadruples. Specifically, it mistakenly pairs the opinion "not good" from the sixth utterance with the target "Pro" from the fifth utterance, indicating a misunderstanding of the dialogue structure. In contrast, DMIN performs better and accurately extracts 3 cross-utterance quadruples and 1 intra-utterance quadruple, including the cross-2 utterance quadruple (id=1). Although DMIN incorrectly predicts the boundary of the opinion phrase "joint names is useless" as "useless" (id=4), it can be observed that DMIN correctly matches the dependencies between elements across utterances.

A complex dialogue shown in Figure 6(a) mostly consists of inter-utterance quads. DMIN predicts a significantly higher number of accurate quads compared to the baseline. In the case of simple dialogue shown in Figure 6(b) that mostly contains intra-utterance quads, DMIN performs slightly better than the baseline. In both cases, DMIN extracts fewer false positive cases, revealing another reason for the score improvement achieved by DMIN.

Speaker	Utterance
0	OnePlus 1, 3, 7pro users, see how bad OnePlus 9, 9Pro, 9R, 9RT are this year. What? don't let me say
1	9pro and Kazakh Soviet co-branded is good for taking pictures [doge]
0	The mode updated later is a bit interesting. In the early stage, I feel that it is not as good as X3pro [allow sad]
1	Harzu is good [like]
2	Pro is ok, good workmanship
0	Aesthetics are not good, and taking pictures with joint names is useless. If the workmanship is not good, it is not a OnePlus. How can the workmanship be bad with the green factory production line.
3	We both use the same phone.

Id	Type	Gold Label [T, A, O, P]	MVQPN's Pred	Ours DMIN's Pred
1	cross	['9pro', 'Harzu', 'good', 'pos']	✖	✔
2	cross	['9pro', 'mode', 'a bit interesting', 'pos']	✖	✔
3	cross	['Pro', 'Aesthetics', 'not good', 'neg']	✔	✔
4	cross	['Pro', 'taking pictures', 'joint names is useless', 'pos']	✖	['Pro', 'taking pictures', 'useless', 'pos']
5	intra	['Pro', 'workmanship', 'good', 'pos']	['Pro', 'workmanship', 'not good', 'neg']	✖
6	intra	['green factory', 'workmanship', 'How can the workmanship be bad', 'pos']	✖	✖
7	intra	['9pro', 'taking pictures', 'good', 'pos']	✔	✔
8	intra	['OnePlus', 'workmanship', 'not good', 'neg']	✖	['OnePlus', 'workmanship', 'bad', 'neu']
9	error	/	/	['9pro', 'mode', 'not as good as', 'neg']

Figure 4: Case study. The major target, aspect, and opinion in dialog are colored differently, and the incorrectly predicted quads by the model are marked in red.

5 CONCLUSION

This paper addresses previous work's limitations and proposes a fresh viewpoint for DiaASQ to better model the discourse structure information in dialogue. Specifically, we enhance the utterance interactions at the token-level granularity on the thread scale and then capture global discourse information at the utterance-level granularity on the dialogue scale, which is more efficient and macro. Furthermore, we introduce a novel integrator to tackle the challenge of integrating data across diverse granularities, yielding a comprehensive and cohesive contextual understanding. The experimental results demonstrate the effectiveness of our proposed DMIN.

6 Acknowledgements

This work was supported by the National Natural Science Foundation of China (71472068 and 62306119) and the Natural Science Foundation of Guangdong Province (2021A1515011864).

7 Limitations

In this paper, we considered utilizing both syntactic knowledge and semantic information to enhance token-level representations. However, we just adopted a simple additive fusion method to combine the representations from the two modules. In the future, more efficient fusion methods can be explored for better integration. Additionally, as an emerging task, the DiaASQ task focuses on structured dialogue text and holds significant research value. However, it poses challenges in terms of annotation difficulty, and the availability of suitable datasets is limited. In the future, we plan to apply our proposed model framework to a wider range of datasets and domains to further validate its effectiveness and generalizability.

References

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, pages 1877–1901.

Chenran Cai, Qin Zhao, Ruifeng Xu, and Bing Qin. 2023. Improving conversational aspect-based sentiment quadruple analysis with overall modeling. In *Proceedings of the 12th National CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part III*, volume 14304 of *Lecture Notes in Computer Science*, pages 149–161. Springer.

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.

Xuxin Cheng, Bowen Cao, Qichen Ye, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023a. ML-LMCL: Mutual learning and large-margin contrastive learning for improving ASR robustness in spoken language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6492–6505, Toronto, Canada. Association for Computational Linguistics.

Xuxin Cheng, Zhihong Zhu, Bowen Cao, Qichen Ye, and Yuexian Zou. 2023b. MRRL: Modifying the reference via reinforcement learning for non-autoregressive joint multiple intent detection and slot filling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10495–10505, Singapore. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

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 *Proceedings of the 24th European Conference on Artificial Intelligence, ECAI 2020, Santiago de Compostela, Spain, Aug 29 - Sep 8*, pages 2006–2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022. Improving aspect sentiment quad prediction via template-order data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7900, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017*. OpenReview.net.

Yongquan Lai, Shixuan Fan, Zeliang Tong, Weiran Pan, and Wei Wei. 2023. Conversational aspect-based sentiment quadruple analysis with consecutive multi-view interaction. In *Proceedings of the 12th National CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part III*, volume 14304 of *Lecture Notes in Computer Science*, pages 162–173. Springer.

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

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021. Dual graph convolutional networks for 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 (Volume 1: Long Papers)*, pages 6319–6329, Online. Association for Computational Linguistics.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4194–4200.

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*.

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*, pages 2215–2225, Dublin, Ireland. Association for Computational Linguistics.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfenn Liu. 2021. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. 2021. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3957–3963.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. 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.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. 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.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. 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.

Zheng Zhang, Zili Zhou, and Yanna Wang. 2022. SSEGCN: Syntactic and semantic enhanced graph convolutional network for aspect-based sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4916–4925, Seattle, United States. Association for Computational Linguistics.

A Datasets

The statistics of experimental datasets is shown in Table 3. The Chinese version of the dataset contains a total of 1000 dialogues, 7,452 utterances, and 5,742 sentiment quadruples, while the English version contains 5,514 quadruples Each dialog has around five speakers on average, and the dataset contains 1,275 (22.2%, in Chinese) and 1,227 (22.3%, in English) cross-utterance quadruples, respectively. As for pair labels, in the training set of dataset ZH, there are 4699, 5931, and 3989 instances for T-A, T-O, and A-O, respectively. In the training set of dataset EN, there are 4823, 6062, and 4297 instances for T-A, T-O, and A-O, respectively.

Dataset		Pair			Quadruple		
		T-A	T-O	A-O	Intra.	Cross.	Total.
EN	train	4,699	5,931	3,989	3,442	972	4,414
	val	603	750	509	423	132	555
	test	592	751	496	422	123	545
ZH	train	4,823	6,062	4,297	3,594	1,013	4,607
	val	621	758	538	440	137	577
	test	597	767	523	433	125	558

Table 3: The statistics of experimental datasets. "Intra." represents intra-utterance quadruples, while "Cross." represents cross-utterance quadruples.

Prompt - Quad Extraction

Schema and Regulations

[Task Description]
Conversational aspect-based quadruple analysis aims to predict the (target, aspect, opinoin, polarity) sentiment element from the dialogue, where target refers to the mobile phone model or brand (such as iPhone 14, Xiaomi, mate30, etc.), aspect refers to a certain aspect of the target (such as battery life, photography, price, etc.), opinion is the evaluation word of the aspect (such as good, very expensive, really unbearable, etc.), polarity is the corresponding sentiment polarity (including positive, negative, neutral). target, aspect, opinion are entities extracted from the dialogue, and polarity is one of {pos, neg, neu}.

[Regulations]
Input format: I will give you a dialogue, which contains multiple utterances U, each utterance U is composed of [uid][sid][rid][utterance], where uid and sid respectively represent the unique identifiers of utterance U and speaker S, and rid represents the uid of the object to which the current utterance replies.
Output specification: You need to find all the sentiment quadruples (target, aspect, opinoin, polarity). Do not output anything other than the quadruple. Do not output duplicate quadruples. Connect multiple quadruples with "#", the output format is: (t1, a1, o1, p1)#(t2, a2, o2, p2)#...#(tn, an, on, pn)

Example

Input=
[u0][s0][reply to u0][So who is better than 11u in taking pictures ?]
[u1][s1][reply to u0][vivo X70 Pro+ is the strongest]
[u2][s2][reply to u1][Telephoto 11ultra is better , portrait vivo is better]
....
Output=('11u', 'taking pictures', 'better', 'pos')#('vivo X70 Pro+', 'taking pictures' ...

Sentence input

Input=
[u0][s0][reply to u0][I use iPhone 14 . The latest version has a good battery life under the ...]
....
Ouput=

GhatGPT response

(iPhone 14, battery life, good, pos)#(Android, battery life, good, pos)#....

Prompt - Pair Extraction

Schema and Regulations

[Task Description]
Conversational aspect-based pair analysis aims to predict the (target-aspect), (target-opinion) and (aspect-opinion) sentiment element pairs from the dialogue, where target refers to the mobile phone model or brand (such as iPhone 14, Xiaomi, mate30, etc.), aspect refers to a certain aspect of the target (such as battery life, photography, price, etc.), opinion is the evaluation word of the aspect (such as good, very expensive, really unbearable, etc.). target, aspect, opinion are entities extracted from the dialogue.

[Regulations]
Input format: I will give you a dialogue, which contains multiple utterances U, each utterance U is composed of [uid][sid][rid][utterance], where uid and sid respectively represent the unique identifiers of utterance U and speaker S, and rid represents the uid of the object to which the current utterance replies.
Output specification: You need to find all the sentiment pairs (target-aspect), (target-opinion) and (aspect-opinion). Do not output anything other than the pair. Do not output duplicate pairs.
The output format is: <target, aspect>=(t1, a1)#(t2, a2)#...#(tn, an)
<target, opinion>=(t1, o1)#(t2, o2)#...#(tn, on)
<aspect, opinion>=(a1, o1)#(a2, o2)#...#(an, on)

Example

Input=
[u0][s0][reply to u0][So who is better than 11u in taking pictures ?]
[u1][s1][reply to u0][vivo X70 Pro+ is the strongest]
....
Output=
<target, aspect>=(Xiaomi, algorithm)#(x70Pro+, color reproduction)#...
<target, opinion>=(OV, good)#(x70Pro+, good)#(vivo X70 Pro+, strongest)#(11ultra, better)# ...
<aspect, opinion>=(Telephoto, better)#(taking pictures, strongest)#(telephoto, yyds)#(takin...

Sentence input

Input=[u0][s0][reply to u0][I use iPhone 14 . The latest version has a good battery life under ...]
....
Ouput=

GhatGPT response

<target, aspect>=(vivo, photography)#...
<target, opinion>=(X70Pro+, better)#...
<aspect, opinion>=(photography, strong)#...

Figure 5: The prompt of ChatGPT for quadruple extraction (left) and pair extraction (right).

B Experiment Details

As shown in the Table 4, our parameter size is only about 6% higher than the baseline. After completing dependency parsing, the time required to train one epoch is very close to that of MVQPN. This indicates that the computational complexity of DMIN is not as high as it may appear.

Attribute	Value
Optimizer	AdamW
$\boldsymbol{\alpha}^{ent}, \boldsymbol{\alpha}^{rel}, \boldsymbol{\alpha}^{pol}$	2, 9, 6
Learning rate(BERT)	1e-5
Learning rate(Other)	1e-4
Max grad norm	1.0
Weight decay	0.01
Max Epoch	40
Early Stop	10
Batch size	2 (dialogues)
θ	10,000
Parameter scale (DMIN)	128M
Parameter scale (MVQPN)	120M
Memory Consumption (DMIN)	11G
Memory Consumption (MVQPN)	8G
Training time/epoch (DMIN)	1min58s
Training time/epoch (MVQPN)	1min49s

Table 4: The details of main experiment.

C ChatGPT Experiments

We conducted preliminary tests on the performance of ChatGPT-3.5-turbo under 0-shot and 1-shot conditions. Initially, we set the temperature parameter of ChatGPT to 0 to obtain stable outputs. Considering the length limitation of ChatGPT-3.5-turbo, we employed quad extraction and pair extraction tasks for the large model separately. Furthermore, we requested the large model to output quads and pairs in the form of entity words, rather than precise word indices. Figure 5 illustrates the prompts designed for our experiment, consisting of three parts: "Schema and Regulations" provides the definition of the task, the input data format, and the output specifications; "Example" presents a sample provided to the large model for reference under the 1-shot condition, while it is not provided under the 0-shot condition; "Sentence Input" represents the dialogue that the large model needs to perform sentiment analysis on. It can be observed from Table 1 that the large model achieves a significant improvement under the one-shot condition compared to the zero-shot condition. However, it still falls short of the performance of the supervised training-based generative model ParaPhrase.

D Case study

Additional case study results are shown in Figure 6. The target, aspect, and opinion in the dialogue are colored differently, and the incorrectly predicted quads by the model are marked in red.

Speaker	Utterance
0	11U is a lot of stacking materials , and the actual experience is particularly poor
1	Let 's talk about it when you have a 11u [two ha]
2	The photoshoot is terrible , not at all up to the gn2 real level , and the photography is equally terrible . If you do n't believe me , look at my space , I 'm already out , the performance is bad , the charging speed is not bad , the screen is good , the r corner is too ugly , oh , by the way , do n't brag about the new system , I used it for half a year without optimization
1	You are really interesting , blahblahblah , other people use 11u to take good pictures , and the photography is also good , although not as good as Apple Samsung . Oh , by the way , whether taking pictures is good or not also depends on the individual [two ha] .
3	Indeed , the system is not good
0	right ?
3	It 's very bad to use . The first few systems often have a black screen and no response . The camera and zoom will also freeze . The game has not been optimized until now . I use it now for playing games and taking pictures , and I still use the iPhone for daily use
4	Your Hasselblad is worthless in front of me !
5	Is this why Lei Jun pursues Leica ? [question]

Id	Type	Gold Label [T, A, O, P]	MVQPN's Pred	Ours DMIN's Pred
1	cross	['11U', 'Hasselblad', 'worthless', 'neg']	✖	✖
2	cross	['11U', 'camera', 'will also freeze', 'neg']	✖	['11U', 'camera', 'freeze', 'neg']
3	cross	['11U', 'game', 'not been optimized', 'neg']	✖	✔
4	cross	['11U', 'system', 'not good', 'neg']	✔	✔
5	cross	['11U', 'systems', 'often have a black screen and no response', 'neg']	✖	✖
6	cross	['11U', 'zoom', 'will also freeze', 'neg']	✖	✖
7	cross	['11u', 'charging speed', 'not bad', 'pos']	['11U', 'charging speed', 'not bad', 'pos']	✖
8	cross	['11u', 'performance', 'bad', 'neg']	✖	✖
9	cross	['11u', 'photography', 'equally terrible', 'neg']	['11U', 'photography', 'equally terrible', 'neg']	✔
10	cross	['11u', 'photoshoot', 'terrible', 'neg']	✔	✔
11	cross	['11u', 'r corner', 'too ugly', 'neg']	['11U', 'r corner', 'too ugly', 'neg']	✔
12	cross	['11u', 'screen', 'good', 'pos']	✖	✖
13	cross	['11u', 'system', 'without optimization', 'neg']	✖	✖
14	intra	['11U', 'actual experience', 'particularly poor', 'neg']	['11U', 'experience', 'particularly poor', 'neg']	✔
15	intra	['11U', 'materials', 'a lot of stacking', 'pos']	✖	✖
16	intra	['11u', 'photography', 'good', 'pos']	✔	✔
17	intra	['11u', 'photography', 'not as good as', 'neg']	✔	✔
18	intra	['11u', 'pictures', 'good', 'pos']	['11u', 'take good pictures', 'good', 'pos']	✔
19	intra	['Apple', 'photography', 'not as good as', 'pos']	✔	['Apple Samsung', 'photography', 'not as good as', 'neg']
20	intra	['Samsung', 'photography', 'not as good as', 'pos']	✔	✔

(a)

Speaker	Utterance
0	4000 + buying Apple and Huawei is always the best solution
1	No more than 6,000 , Apple , Huawei , not as good as dog [laughing but not speaking] [laughing and not speaking]
2	Smart people [cool][cool][cool][cool] ! Huawei must buy high - end ! Low - end just look at them ! 4000 price Xiaomi is the first choice ! Do whatever you want if you have enough money ! The optimal solution for Huawei Apple is 5000 to 6000 + ! Around 4000 , you can only consider Xiaomi , O V , Meizu ! Do not ask me why ! Because that 1000 to 2000 is that called some tax [cool][cool][cool][cool][cool][cool]
3	The configuration of Huawei with more than 4000 may be worse than that of Xiaomi with more than 4000 , but the actual experience must be far better than Xiaomi
4	But Huawei 's latest model does n't have 5 G
1	Do n't buy it now , no need
4	It 's really not necessary , I really want to use emui , and glory is fine too
5	Huawei 's p40 with more than 4000 really suck [allow sad]
6	Taking pictures is really hard to say . .

Id	Type	Gold Label [T, A, O, P]	MVQPN's Pred	Ours DMIN's Pred
1	cross	['p40', 'Taking pictures', 'hard to say', 'neg']	✔	✔
2	intra	['Huawei', 'actual experience', 'far better', 'pos']	✔	✔
3	intra	['Huawei', 'configuration', 'worse', 'neg']	✔	✔
4	intra	['Xiaomi', 'actual experience', 'far better', 'neg']	['Xiaomi', 'actual experience', 'far better', 'pos']]	✔
5	intra	['Xiaomi', 'configuration', 'worse', 'pos']	✖	✖
6	error	/	['Huawei', 'Taking pictures', 'hard to say', 'neg']	['Huawei', 'Taking pictures', 'hard to say', 'neg']
7	error	/	['Xiaomi', 'experience', 'far better', 'neg']	/
8	error	/	['Huawei', 'experience', 'far better', 'neg']	/

(b)

Figure 6: Additional case study results presentation.

Figure 6(a) presents a complex dialogue primarily composed of inter-utterance quads. In this case, the main cross-utterance quads are densely distributed in the first and second threads, with the terms "11U" in the first utterance and "11u" in the second utterance being the main target terms. These two target terms are semantically similar and closely located, making them easily confused. In quads with IDs 7, 9, and 11, the baseline model incorrectly predicts "11U" instead of "11u", while DMIN correctly identifies the dialogue structure and identifies the correct target. As a result, DMIN achieves a Micro F1 score of 60.61% on this case, significantly surpassing the baseline's 36.36%. In the case of a simple dialogue shown in Figure 6(b), which mainly consists of intra-utterance quads, DMIN performs slightly better than the baseline. However, in both cases, due to the correct understanding of the structured dialogue context, DMIN extracts fewer incorrect quads, meaning it has fewer false positive cases, thus revealing another reason for the score improvement achieved by DMIN.