DiaASQ: A Benchmark of Conversational Aspect-based Sentiment Quadruple Analysis

Bobo Li[1], **Hao Fei**[2], **Fei Li**[1], **Yuhan Wu**[1], **Jinsong Zhang**[1], **Shengqiong Wu**[2],
Jingye Li[1], **Yijiang Liu**[1], **Lizi Liao**[3], **Tat-Seng Chua**[2] **and Donghong Ji**[1*]

[1] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University

[2] Sea-NExT Joint Lab, National University of Singapore [3] Singapore Management University

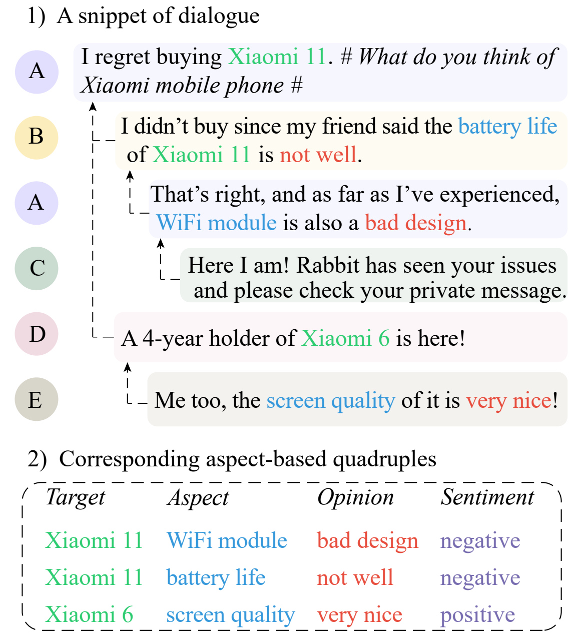
{boboli,lifei_csnlp,yuhanwu,jinsongzhang,theodorelee,cslyj,dhji}@whu.edu.cn
{haofei37,dcscts}@nus.edu.sg swu@u.nus.edu lzliao@smu.edu.sg

Abstract

The rapid development of aspect-based sentiment analysis (ABSA) within recent decades shows great potential for real-world society. The current ABSA works, however, are mostly limited to the scenario of a single text piece, leaving the study in dialogue contexts unexplored. To bridge the gap between fine-grained sentiment analysis and conversational opinion mining, in this work, we introduce a novel task of conversational aspect-based sentiment quadruple analysis, namely DiaASQ, aiming to detect the quadruple of *target-aspect-opinion-sentiment* in a dialogue. We manually construct a large-scale high-quality DiaASQ dataset in both Chinese and English languages. We deliberately develop a neural model to benchmark the task, which advances in effectively performing end-to-end quadruple prediction, and manages to incorporate rich dialogue-specific and discourse feature representations for better cross-utterance quadruple extraction. We hope the new benchmark will spur more advancements in the sentiment analysis community. Our data and code are open at https://github.com/unikcc/DiaASQ.

1 Introduction

It is meaningful to empower machines to understand human opinion and sentiment, which motivates the study of sentiment analysis (Pang and Lee, 2007; McDonald et al., 2007; Ren et al., 2016; Cambria, 2016). ABSA is an important branch of sentiment analysis aiming to detect the sentiment trends towards the fine-grained aspects of targets, which has received consistent research attention within last few years (Li et al., 2018; Fan et al., 2019; Chen et al., 2020; Wu et al., 2021; Chen et al., 2022a). The initial ABSA revolves around the study of *aspect terms* and *sentiment polarities* (Tang et al., 2016; Fan et al., 2018; Li et al., 2019). Later, the extraction of *opinion terms* is considered, resulting in a triplet analysis (i.e., *aspect-opinion-sentiment*) of ABSA (Peng et al., 2020; Chen et al., 2021). The latest trend of ABSA has been upgraded into the quadruple form by adding the *category* element into the triplet ABSA (Cai et al., 2021; Zhang et al., 2021a). The quadruple ABSA promisingly completes the ABSA definition and helps the comprehensive understanding of the opinion picture.



1) A snippet of dialogue

A: I regret buying Xiaomi 11. # *What do you think of Xiaomi mobile phone* #

B: I didn't buy since my friend said the battery life of Xiaomi 11 is not well.

A: That's right, and as far as I've experienced, WiFi module is also a bad design.

C: Here I am! Rabbit has seen your issues and please check your private message.

D: A 4-year holder of Xiaomi 6 is here!

E: Me too, the screen quality of it is very nice!

2) Corresponding aspect-based quadruples

Target	*Aspect*	*Opinion*	*Sentiment*
Xiaomi 11	WiFi module	bad design	negative
Xiaomi 11	battery life	not well	negative
Xiaomi 6	screen quality	very nice	positive

Figure 1: Illustration of the conversational aspect-based sentiment quadruple analysis (DiaASQ). The dialogue utterances produced by the corresponding speakers (marked at left) are organized into replying structure.

Yet we notice that all the current ABSA research is confined to the scenario of a single piece of text (i.e., sentence or document). For example, currently the most popular ABSA benchmark, SemEval (Pontiki et al., 2014, 2015, 2016), comes with only sentence-level annotations. This may limit the application of ABSA. Essentially, in the

[*]Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 13449–13467
July 9-14, 2023 ©2023 Association for Computational Linguistics

	ASTE	TOWE	MAMS	CASA	DiaASQ
Target	✗	✗	✗	✓	✓
Aspect	✓	✓	✓	✗	✓
Opinion	✓	✓	✓	✓	✓
Polarity	✓	✗	✓	✓	✓
Dialogue-level	✗	✗	✗	✓	✓
Multi-sentiment	✗	✗	✓	✗	✓
Multilingual	✗	✗	✗	✗	✓

Table 1: A comparison between our DiaASQ dataset and existing popular ABSA datasets, including: ASTE (Peng et al., 2020), TOWE (Fan et al., 2019), MAMS (Jiang et al., 2019), and CASA (Song et al., 2022).

real-world environment ABSA has a broader application under dialogue contexts. For example, people are more likely to discuss certain products, services, or politics on social media (e.g., Twitter, Facebook, Weibo) in the form of multi-turn and multi-party conversations. Also, it is practically meaningful to develop sentiment-support dialog systems to facilitate the clinical diagnosis, and treatment (Liu et al., 2021a). Unfortunately, no effort has been dedicated to the research of a holistic dialog-level ABSA.

In this paper, we consider filling the gap of dialogue-level ABSA. We follow the line of recent quadruple ABSA and present a task of conversational aspect-based sentiment quadruple analysis, namely **DiaASQ**. DiaASQ sets the goal to detect the fine-grained sentiment quadruple of *target-aspect-opinion-sentiment* given a conversation text, i.e., an opinion of sentiment polarity has been expressed toward the target with respect to the aspect. As exemplified in Fig. 1, multiple users (speakers) on social media discuss different angles of a product (i.e., '*Xiaomi*' brand cellphone) in dialogue threads of multiple turns. The task aims to extract three quadruples over the dialog: (*'Xiaomi 11', 'WiFi module', 'bad design', 'negative'*), (*'Xiaomi 11', 'battery life', 'not well', 'negative'*) and (*'Xiaomi 6', 'screen quality', 'very nice', 'positive'*).

To benchmark the task, we manually annotate a large-scale DiaASQ dataset. We collect millions of conversational corpus of source comments and discussions closely related to electronic products from Chinese social media. We hire well-trained workers to explicitly label the DiaASQ data (i.e., the elements of quadruples, targets, aspects, opinions, and sentiments) based on the crowd-sourcing technique, which ensures a high quality of annotations. Finally, we yield the dataset with 1,000 dialogue snippets in total with 7,452 utterances. To facilitate the multilinguality of the benchmark, we further translate and project the annotations into English. Data statistics show that each dialog involves around 5 speakers, and 22.2% of the quadruples are in the cross-utterance format.

Compared with previous single-text-based ABSA, DiaASQ challenges in two main aspects. First, DiaASQ includes four subtasks. Directly applying the existing best-performing graph-based ABSA model to enumerate all possible target, aspect, and opinion terms could cause a combinatorial explosion. Second, the elements of a quadruple are scattered around the whole conversation due to the complex replying structure, which requires the model to do cross-utterance extraction. To solve these challenges, we present an end-to-end DiaASQ framework. Specifically, based on the grid-filling method (Wu et al., 2020), we redesign the tagging scheme to fulfill the four subtasks in one shot effectively. Moreover, during the dialogue text encoding, we additionally model the dialogue-specific representations for utterance interaction and meanwhile encode the relative distance as cross-utterance features. Experiments on the DiaASQ data indicate that our model shows significant superiority than several strong baselines.

To sum up, this work contributes in threefold:

- We pioneer the research of dialogue-level aspect-based sentiment analysis. Specifically, we introduce a conversational aspect-based sentiment quadruple analysis (DiaASQ) task.
- We release a dataset for the DiaASQ task in both Chinese and English languages, which is of high quality and at a large scale.
- We introduce a model to benchmark the DiaASQ task. Our method solves the task end-to-end and meanwhile effectively learns the dialogue-specific features for better cross-utterance sentiment quadruple extraction.

2 Related Work

2.1 Fine-grained Sentiment Analysis

All the existing ABSA tasks and their derivations revolve around predicting several elements or combinations: *aspect term*, *sentiment polarity*, *opinion term*, *aspect category*[1], *target*. The initial ABSA task aims to classify the sentiment polarities given aspects (Tang et al., 2016; Fan et al., 2018; Li et al., 2019). Later, a wide range of new compound ABSA-related tasks is proposed, such

[1] For example, the aspect '*WiFi module*' in Fig. 1 belongs to the *hardware* category).

as aspect-opinion paired extraction (Zhao et al., 2020; Wu et al., 2021), aspect-category prediction (Wang et al., 2019; Jiang et al., 2019; Dai et al., 2020), triplet extraction (Peng et al., 2020; Chen et al., 2021, 2022b), and structured opinion extraction (Shi et al., 2022; Wu et al., 2022), etc.

The latest attention has been placed on the quadruple or quintuple ABSA, where the *aspect category* element is added into the triplet extraction (Cai et al., 2021; Zhang et al., 2021a; Liu et al., 2021b; Fei et al., 2022a). Compared to all prior ABSA tasks, the sentiment quadruples provide much more complete opinion details that can facilitate downstream applications better. In this work, we follow this line, while our work differs in three aspects. First, we consider adding the element of *target* instead of *category*. Second, current quadruple and quintuple ABSA datasets all are incrementally annotated based on the existing SemEval data (Pontiki et al., 2014, 2015, 2016); while we newly craft our data from real-world environment. Third, this work mainly focuses on the conversation contexts instead of sentence pieces.

2.2 Dialogue Opinion Mining

In NLP community, dialogue applications show increasing impacts to real-world environments (Liao et al., 2021; Ni et al., 2022; Liao et al., 2022). The emotion and sentiment analysis in conversation scenarios is an essential branch of opinion mining. Previous dialogue-level opinion mining has been limited to the coarse granularity, where the representative task is dialogue emotion detection (Li et al., 2020; Hu et al., 2021; Li et al., 2022). Yet as we indicated earlier, sentiment analysis in conversation at a fine-grained level has practical value. In this paper, we pioneer the research of dialogue-level ABSA, presenting the conversational aspect-based sentiment quadruple analysis task.

In Table 1 we compare our DiaASQ data with existing popular ASBA benchmarks. It is worth noticing that, although CASA (Song et al., 2022) is a dialogue-level sentiment analysis dataset, it may fail to provide a comprehensive understanding of opinion status due to the absence of key elements (e.g., *aspect*). In contrast, our DiaASQ dataset covers *target*, *aspect*, *opinion* and *sentiment*, which is by now the most comprehensive ABSA benchmark among all the other corpus. In addition, sentiment understanding in DiaASQ is more complex and thus more challenging. For example, one aspect term could correspond to multiple sentiments. Besides, DiaASQ contains both Chinese and English versions, which will facilitate the research community to different languages.

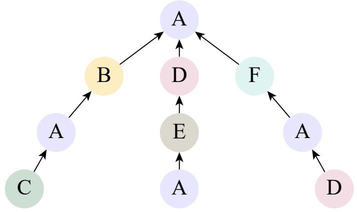


Figure 2: The tree-like dialogue replying structure.

3 Data Construction

We construct a new dataset to facilitate the DiaASQ task. The raw corpus is collected from the largest Chinese social media, Weibo[2]. We crawl nine million posts and comments from the tweets history of 100 verified digital bloggers. Each conversation is derived from a root post, and multiple users (i.e., multiple speakers) are attended to reply to a predecessor post. The multi-thread and multi-turn dialogue forms a tree structure, as illustrated in Fig. 2. We preprocess the raw dialogues to make the contexts integrated. First, we filter the topic-related conversations by a manually created keyword dictionary in the mobile phone field, which includes hundreds of hot words, like phone band names, aspects words to describe a mobile phone, etc. Then, we normalize the tweet language expressions (e.g., abusive language, hate speech) by human examination or consulting lexicons; we prune away those meaningless replying branches that deviate too much from the main topic. We also limit the maximum number of utterances to ten for better controllable modeling. After a strict cleaning procedure, we obtain the final 1,000 dialogues.

During the annotation stage, all the conversation texts are labeled with a team of crowd-workers who are pre-trained under the SemEval ABSA (Pontiki et al., 2014) annotation guideline[3]. Also, the linguistic and computer science experts inspect the labeling schema. After annotating, annotators are required to cross-examine the labels. Also, some automatic rules are applied to verify the labeling consistency. Finally, Cohen's Kappa score of quadruples is 0.86, which indicates our annotation corpus has reached a high-level agreement.

Data Insights. We randomly split the conversation snippets into train/valid/test sets, in the ra-

[2] https://weibo.com/

[3] Appendix § C details the data acquisition and annotation.

		Dialogue			Items			Pairs			Quadruples		
		Dia.	Utt.	Spk.	Tgt.	Asp.	Opi.	Pair$_{t\text{-}a}$	Pair$_{t\text{-}o}$	Pair$_{a\text{-}o}$	Quad.	Intra.	Cross.
ZH	Total	1,000	7,452	4,991	8,308	6,572	7,051	6,041	7,587	5,358	5,742	4,467	1,275
	Train	800	5,947	3,986	6,652	5,220	5,622	4,823	6,062	4,297	4,607	3,594	1,013
	Valid	100	748	502	823	662	724	621	758	538	577	440	137
	Test	100	757	503	833	690	705	597	767	523	558	433	125
EN	Total	1,000	7,452	4,991	8,264	6,434	6,933	5,894	7,432	4,994	5,514	4,287	1,227
	Train	800	5,947	3,986	6,613	5,109	5,523	4,699	5,931	3,989	4,414	3,442	972
	Valid	100	748	502	822	644	719	603	750	509	555	423	132
	Test	100	757	503	829	681	691	592	751	496	545	422	123

Table 2: Data statistics. 'Dia.', 'Utt.', and 'Spk.' refer to dialogue, utterance, and speaker, respectively. 'Tgt', 'Asp', and 'Opi' refer to target, aspect, and opinion terms, respectively. 'Intra' and 'Cross' refer to the intra-/cross-utterance quadruples. A quadruple is cross-utterance if any two elements of the (target, aspect, and opinion) in one quadruple distribute in different utterances. Since some words will be added, dropped, or merged during the translating process, the numbers of annotation items in Chinese and English versions of datasets are somewhat different.

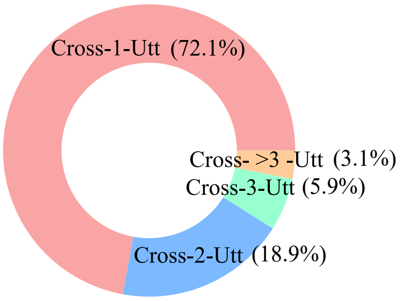


Figure 3: The ratio of cross-utterance quadruples. We define the max utterance-level distance between every two items in one quadruple as the number of cross-utterance. For example, the first quadruple in Fig. 1 crosses two utterances.

tio of 8:1:1. The Chinese version of the dataset contains a total of 7,452 utterances, and 5,742 sentiment quadruples, while the English version contains 5,514 quadruples, which are far bigger numbers than the existing quadruple and quintuple ABSA datasets (Cai et al., 2021; Zhang et al., 2021a). Also, there is an average of one sentimental expression in each utterance. Such annotation density makes it quite convenient for task prediction. The data statistics are shown in Table 2. 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. In Fig. 3, we show the ratio of quadruples of the dataset under different cross-utterance levels. More data statistics are shown in Appendix § B.

4 Grid-tagging Task Modeling with Renewed Label Scheme

The input of the DiaASQ task includes a dialogue $D = \{u_1, \cdots, u_n\}$ with the corresponding replying record $l = \{l_1, \cdots, l_n\}$ of utterances, where l_i denotes that i-th utterance replies to l_i-th utterance. Each $u_i = \{w_1, \cdots, w_m\}$ denotes i-th utterance text and m is the length of utterance u_i. The replying record l reflects the hierarchical tree structure of D. Based on the input D and l, DiaASQ aims to extract all possible (*target*, *aspect*, *opinion*, *sentiment*) quadruples, denoted as $Q = \{t, a, o, p\}_{k=1}^{K}$. The *target*, *aspect* or *opinion* term (t_k, a_k, o_k) is a sub-string of an utterance text u_i. The sentiment p_k is a category label $\in \{pos, neg, other\}$.

DiaASQ naturally includes four subtasks. Different popular end-to-end ABSA systems can be utilized to solve our DiaASQ, such as the graph-based (Zhou et al., 2021; Chen et al., 2022a), seq-to-seq (Zhang et al., 2021c; Mukherjee et al., 2021) and grid-tagging models (Wu et al., 2020). Yet enumerating all possible terms with graph-based methods will cost computational efficiency, while seq-to-seq methods suffer from exposure bias. The grid-tagging method advances in higher efficiency, i.e., $\mathcal{O}(n^2)$ complexity, where n denotes the sequence length. However, the labeling scheme in (Cai et al., 2021; Zhang et al., 2021a) only supports term-pair extraction (i.e., *aspect* and *opinion* terms), which fails to directly solve our DiaASQ that requires term-triple extraction (i.e., *target*, *aspect* and *opinion* terms). Here we inherit the success of the grid-tagging method for an end-to-end solution and re-design the labeling scheme to fit our needs.

To reach the goal, we re-decompose the task into three joint jobs: detections of the entity boundary, entity pair, and sentiment polarity. We renew the labeling scheme of grid-tagging in support of these jobs, which is shown in Fig. 4.

- **Entity Boundary Labels**: We use *tgt, asp, opi* to denote the token-level relations between the

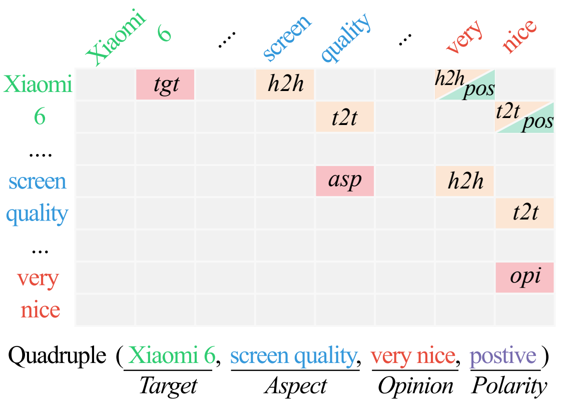


Figure 4: Tagging scheme for quadruple extraction.

head and tail of a *target*, *aspect*, and *opinion* term, respectively. For example, the *tgt* between '*Xiaomi*' and '*6*' denotes a *target* term of '*Xiaomi 6*'.

• **Entity Pair Labels**: We then need to link different types of terms together as a combination. To represent the relation between entities, we devise two labels: *h2h* and *t2t*, both of which align the head and tail tokens between a pair of entities in two types. For example, the head words of '*Xiaomi*' (*target*) and '*screen*' (*aspect*) is connected with *h2h*, while the tail words of '*6*' (*target*) and '*quality*' (*aspect*) is connected with *t2t*. By labeling a chain of term pairs in different types, we form a triplet of (t_k, a_k, o_k).

• **Sentiment Polarity Labels**: By adding a sentiment category label p_k, we then form a quad q_k=(t_k, a_k, o_k, p_k). Since the *target* and *opinion* terms together determine a unique sentiment, we assign the category label between the heads and tails of these two terms, as shown in Fig. 4.

5 DiaASQ Model

We present a DiaASQ model to accomplish the task based on the above grid-tagging label scheme. Fig. 5 shows the overall architecture.

5.1 Base Encoding

We adopt a pre-trained language model (PLM), e.g., BERT (Devlin et al., 2019), to encode the dialogue utterances. However, the length of a whole dialogue may far exceed the max length that BERT can accept. We thus encode each utterance with a separate PLM one by one. We use the [CLS] and [SEP] tokens to separate each utterance u_i.

$$u_i^{'} = < \text{[CLS]}, w_1, \cdots, w_m, \text{[SEP]} > , \tag{1}$$

$$\boldsymbol{H_i} = \boldsymbol{h}_{cls}, \boldsymbol{h}_1, \cdots, \boldsymbol{h}_m, \boldsymbol{h}_{sep} = \text{PLM}(u_i^{'}) , \tag{2}$$

where $\boldsymbol{h}_m$ is the contextual representation of w_m.

5.2 Dialogue-specific Multi-view Interaction

To strengthen the awareness of the dialogue discourse, we then introduce a multi-view interaction layer to learn the dialogue-specific features. This layer is built upon the multi-head self-attention (Vaswani et al., 2017). Inspired by (Shen et al., 2021; Zhao et al., 2022), we use three types of features: dialogue threads, speakers, and replying. Specifically, we realize the idea by constructing attention masks $\boldsymbol{M}^c$ that carry the bias of such prior features, controlling the interactions between tokens. And $c \in \{Th, Sp, Rp\}$ represents different types of token interaction, i.e., thread, speaker, and replying, respectively.

$$\begin{aligned} \boldsymbol{H}^c &= \text{Masked-Att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{M}^c) \\ &= \text{Softmax}(\frac{(\boldsymbol{Q}^T \cdot \boldsymbol{K}) \odot \boldsymbol{M}^c}{\sqrt{d}}) \cdot \boldsymbol{V} , \end{aligned} \tag{3}$$

where $\boldsymbol{Q}$=$\boldsymbol{K}$=$\boldsymbol{V}$=$\boldsymbol{H} \in \mathbb{R}^{N \times d}$ is the representation of the whole dialogue sequence obtained by concatenating token representations of each utterance ($\boldsymbol{H}_i$ in Eq. (2)), N is the token-level length of D, and $\odot$ is element-wise production. The value of $\boldsymbol{M}^c \in \mathbb{R}^{N \times N}$ is defined as follows:

• **Thread Mask**: $\boldsymbol{M}_{ij}^{Th}$=1 if the i^{th} and j^th token belong to the same dialogue thread.

• **Speaker Mask**: $\boldsymbol{M}_{ij}^{Sp}$=1 if the i^{th} and j^{th} token are derived from the same speaker.

• **Reply Mask**: $\boldsymbol{M}_{ij}^{Rp}$=1 if the two utterances containing the i^{th} and j^{th} token respectively have a replying relation.

We then conduct Max-Pooling over the masked representations, followed by a tag-wise MLP layer to yield the final feature representation $\boldsymbol{v}_i^c$:

$$\boldsymbol{H}^f = \text{Max-Pooling}(\boldsymbol{H}^{Th}, \boldsymbol{H}^{Sp}, \boldsymbol{H}^{Rp}) , \tag{4}$$

$$\boldsymbol{v}_i^r = \text{MLP}^r(\boldsymbol{h}_i^f) , \tag{5}$$

where $r \in \{tgt, \cdots, h2h, \cdots, pos, \cdots, \epsilon_{ent}, \cdots\}$ indicates a specific label, and ϵ_{ent} denotes the non-relation label in the entity boundary matrix.

5.3 Integrating Dialogue Relative Distance

Limited by the PLM, we can only encode each utterance separately, potentially hurting the conversational discourse. To compensate for it, we consider fusing the Rotary Position Embedding (RoPE) (Su et al., 2021) into token representations. RoPE dynamically encodes the relative distance globally between utterances at the dialogue level. Introducing such distance information can help guide better

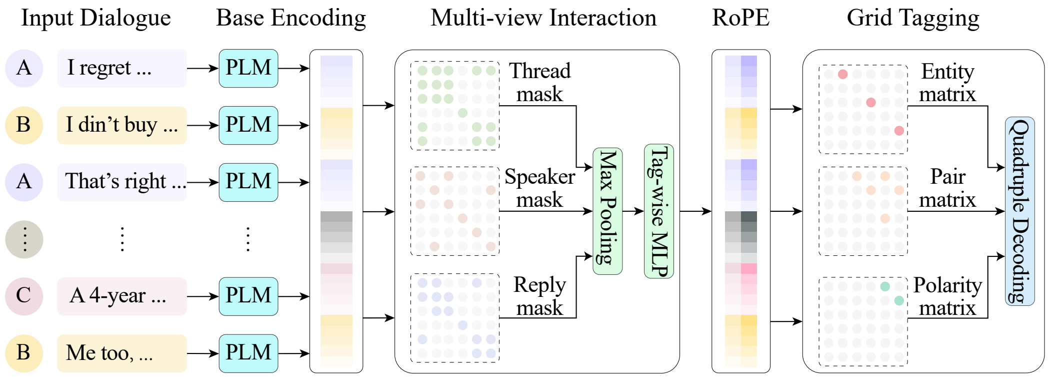


Figure 5: The overall framework of our DiaASQ model. First, the base encoder learns base contextual representations for the input dialogue texts. The multi-view interaction layer then aggregates dialogue-specific feature representations, such as the threads, speakers, and replying information. We further fuse the Rotary Position Embedding (RoPE), where the relative dialogue distance information helps guide better discourse understanding. Finally, the system decodes all the quadruples based on the grid-tagging labels.

discourse understanding.

$$\boldsymbol{u}_i^r = \boldsymbol{\mathcal{R}}(\theta, i)\boldsymbol{v}_i^r\,, \tag{6}$$

where $\boldsymbol{\mathcal{R}}(\theta, i)$ is a positioning matrix parameterized by θ and the absolute index i of $\boldsymbol{v}_i^r$.

5.4 Quadruple Decoding

Based on each tag-wise representation $\boldsymbol{u}_i^r$, we finally calculate the unary score between any token pair in terms of label r:

$$s_{ij}^r = (\boldsymbol{u}_i^r)^T \boldsymbol{u}_j^r\,, \tag{7}$$

where s_{ij}^r is the probability that the relation label between w_i and w_j is r. Then we put a softmax layer over all elements in each matrix to determine the relation label r. For example, the probability of entity boundary matrix can be obtained via:

$$p_{ij}^{ent} = \text{Softmax}([s_{ij}^{\epsilon ent}; s_{ij}^{tgt}; s_{ij}^{asp}; s_{ij}^{opi}])\,. \tag{8}$$

Obtaining all the labels in the grid, we decode all the quadruples based on the rules stated in § 4.

5.5 Learning

The training target is to minimize the cross-entropy loss of each subtask:

$$\mathcal{L}_k = -\frac{1}{G \cdot N^2}\sum_{g=1}^{G}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\alpha}^k\, y_{ij}^k \log(p_{ij}^k)\,, \tag{9}$$

where $k \in \{ent, pair, pol\}$ indicates a subtask, N is the total token length in a dialogue, and G is the total training data instances. y_{ij}^k is ground-truth label, p_{ij}^k is the prediction. The label types (stated in Section 4) are imbalanced. Thus we apply a tag-wise weighting vector $\boldsymbol{\alpha}^k$ to counteract this. We then add up all three loss items as the final one:

$$\mathcal{L} = \mathcal{L}_{ent} + \beta\mathcal{L}_{pair} + \eta\mathcal{L}_{pol}\,. \tag{10}$$

6 Experiment

6.1 Settings

We conduct experiments on our DiaASQ dataset to evaluate the efficacy of our proposed model. We mainly measure the performances in terms of three angles: 1) *span match*: the boundary of three types of term spans; 2) *pair extraction*: the detection of span pair, i.e., *Target-Aspect*, *Aspect-Opinion* and *Target-Opinion*; 3) *quadruple extraction*: recognizing the full quad of DiaSAQ task. We use the *exact F1* as the metric: for span, a correct prediction should match both the left and right boundaries; for pair, match both two spans and the relation; for quad, match all four elements exactly. The performance of quadruple extraction is our main focus. We thus take the *micro F1* and *identification F1* respectively for measurements, where the micro F1 measures the whole quad, including the sentiment polarity. In contrast, *identification F1* (Barnes et al., 2021) does not distinguish the polarity.

We take the Chinese-Roberta-wwm-base (Cui et al., 2021) and Roberta-Large (Liu et al., 2019) as our base encoders for the Chinese and English datasets, respectively. We put a 0.2 dropout rate on the BERT output representations. MLP in Eq. (5) has a 64-d hidden size. The testing results are given by the models tuned on the developing set. All experiments take five different random seeds, and the final scores are averaged over five runs.

As no prior method is deliberately designed for DiaASQ, we consider re-implementing several strong-performing systems closely related to the task as our baselines, including **CRF-Extract-Classify** (Cai et al., 2021), **SpERT** (Eberts and Ulges, 2020) **Span-ASTE** (Xu et al., 2021) and

		Span Match (F1)			Pair Extraction (F1)			Quadruple (F1)	
		T	A	O	T-A	T-O	A-O	Micro	Iden.
ZH	CRF-Extract-Classify	**91.11**	75.24	50.06	32.47	26.78	18.90	8.81	9.25
	SpERT	90.69	76.81	54.06	38.05	31.28	21.89	13.00	14.19
	ParaPhrase	/	/	/	37.81	34.32	27.76	23.27	27.98
	Span-ASTE	/	/	/	44.13	34.46	32.21	27.42	30.85
	w/o PLM	/	/	/	28.36	24.81	22.50	8.96	11.21
	Ours	90.23	**76.94**	**59.35**	**48.61**	**43.31**	**45.44**	**34.94**	**37.51**
	w/o PLM	85.52	75.21	47.15	34.72	26.16	30.73	14.21	17.55
EN	CRF-Extract-Classify	88.31	71.71	47.90	34.31	20.94	19.21	11.59	12.80
	SpERT	87.82	74.65	54.17	28.33	21.39	23.64	13.07	13.38
	ParaPhrase	/	/	/	37.22	32.19	30.78	24.54	26.76
	Span-ASTE	/	/	/	42.19	30.44	**45.90**	26.99	28.34
	w/o PLM	/	/	/	27.26	20.63	44.62	13.84	14.17
	Ours	**88.62**	**74.71**	**60.22**	**47.91**	**45.58**	44.27	**33.31**	**36.80**
	w/o PLM	83.02	68.89	53.87	32.53	31.09	35.59	15.68	19.57

Table 3: Main results of the DiaASQ task. 'T/A/O' represent Target/Aspect/Opinion, respectively. All the scores are averaged values over five runs under different random seeds. Since ParaPhrase and Span-ASTE do not distinguish the term types, we here do not measure the performances of span match. Note that 'w/o PLM' indicates that we use randomly initialized word2vec to encode the text.

ParaPhrase (Zhang et al., 2021a). All baselines take the same PLM as used in our model except that **ParaPhrase** uses mT5-base (Xue et al., 2021).

6.2 Main Comparisons

Table 3 compares the performances of different models on the DiaASQ task. We see that our proposed method achieves the overall best results under almost all measurements. Besides, we have the following observations.

First, the performance divergences of different models on span detection are not significant, and all the methods perform well on the subtask. We think this is mainly because, without considering the inter-relation between each type of term (T/A/O), recognizing the mentions is a pretty simple task.

Second, it is clear that our model starts surpassing the baselines on pair-wise detection. Our system outperforms the second-best models over average 9% of F1 score on almost all cases, i.e., T-A, T-O, and A-O. This result verifies that our model is more effective than baselines on sentiment information extraction under the conversational scenario. One exception is that the Span-ASTE slightly exceeds our model on A-O pair extraction in the English version dataset. The possible reason is that aspect and opinion pair usually co-occur closely, and it has been a classical task for which span-aste can achieve competitive results.

Finally and most importantly, our system shows huge wins on the quadruple extraction, with 7.52% micro F1(=34.94-27.42) and 6.66% identification F1(=37.51-30.85) improvements on the Chinese dataset, with 6.32% micro F1(=33.31-26.99) and 8.46% identification F1(=36.80-28.34) improvements on the English dataset, respectively. This result evidently shows our model's efficacy on the task. We also find that stripping off the PLMs hurts the task performances very prominently, even for the strong models.

6.3 Ablation Study

We now take a further step, examining the efficacy of several key designs in our method, including the dialogue-specific multi-view interaction, the relative distance embedding (RoPE), and the label-wise weighting mechanism. The ablating results are shown in Table 4.

First, we see that the different type of dialogue-specific interaction shows the varying influence. For example, thread features show the overall most negligible impacts, which improve the F1 score of Inter-Utt by no more than 1% in the two datasets. In contrast, the speaker-aware and reply-aware interactions are more important that improve the score Inter-Utt by more than 1%. Interestingly, some ablations increase the performances in the intra-utterance case but decrease rapidly in the cross-utterance case.

Then, we witness the most significant performance drops when removing the RoPE feature. Significantly, the F1 score of cross-utterance drops 2.99% and 3.54% in the Chinese and English datasets, respectively. This result demonstrates the importance of modeling dialogue-level discourse information. Finally, we see that the label-wise

	ZH			EN		
	Overall	Intra-Utt.	Inter-Utt.	Overall	Intra-Utt.	Inter-Utt.
Ours	34.94	37.95	23.21	33.31	37.65	15.76
w/o All-Interaction	34.04(↓0.90)	37.40(↓0.55)	20.95(↓2.26)	32.51 (↓0.80)	37.23 (↓0.32)	12.98 (↓2.78)
w/o Speaker	34.43(↓0.51)	37.82(↓0.13)	21.90(↓1.31)	33.06 (↓0.25)	37.68 (↑0.03)	14.20 (↓1.56)
w/o Thread	34.52(↓0.42)	37.61(↓0.34)	22.62(↓0.59)	33.09 (↓0.22)	37.33 (↓0.32)	15.09 (↓0.67)
w/o Reply	34.26(↓0.68)	37.06(↓0.89)	22.91(↓0.30)	32.82 (↓0.49)	37.46 (↓0.21)	13.50 (↓2.26)
w/o RoPE	33.10(↓1.84)	36.42(↓1.53)	20.22(↓2.99)	31.59 (↓1.72)	36.44 (↓1.21)	12.22 (↓3.54)
w/o Lab.Wei. (α^k)	33.52(↓1.42)	36.63(↓1.32)	20.93(↓2.28)	32.54 (↓0.77)	37.06 (↓0.59)	13.50 (↓2.26)

Table 4: Ablation results (Micro F1). ‘w/o All-Interaction’: removing all three multi-view interaction items.

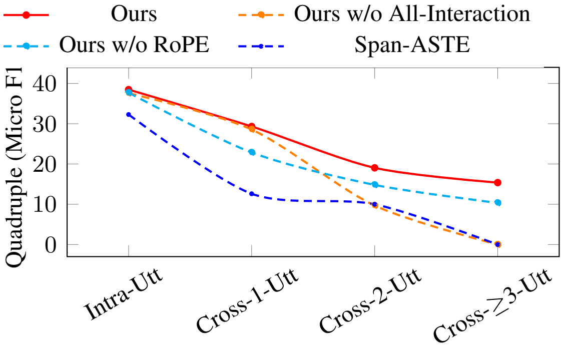


Figure 6: results on different cross-utterance levels.

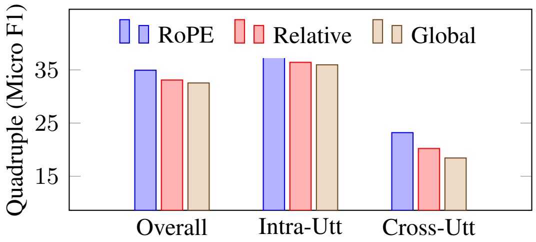


Figure 7: Influences of using difference distance-encoding methods.

weighting mechanism used for task learning is also much crucial. This finding is reasonable because the labels of different types in the grid among the whole dialogue are imbalanced and sparse, e.g., the positive tags are far less than the negative ones (i.e., ϵ_{ent}). Label-wise weighting helps effectively solve the label imbalance issue.

6.4 Further Analysis

In this section, we consider diving into the model performances and carry on an in-depth analysis to better understand the strengths of our method.

Cross-utterance Quadruple Extraction. Earlier in Table 3, we verify the superiority of our model. We mainly credit its capability to effectively model the cross-utterance features. Here we directly examine this attribute by observing the performances under different levels of the cross-utterance quad extraction. As plotted in Fig. 6, we observe the patterns that the more utterances quadruple across, the lower the performances all models can achieve. Especially when the cross-utterance level $\geq$3, the baseline systems fail to recognize any single quad. Nevertheless, our system can still well resolve the challenge, even in case of cross-$\geq$3-utterance. Also, by comparing two of our ablated models, we learn that the dialogue-specific interaction features are more beneficial for handling the super-long-distance cross-utterance. But the RoPE that carries discourse information contributes more to the short-range case (i.e., cross-1-utterance).

Impact of Dialogue-level Distance Encoding. We equip our framework with dialogue-level relative distance embeddings (i.e., RoPE, a dynamic positioning feature), so as to enhance conversational discourse understanding. Here we study the influence of using different dialogue-level distance embeddings. We consider two other alternative solutions: 1) *Relative position encoding*, which is a type of dense embedding of relative distances of utterance; We directly add the embedding to the token relation probability vector in Eq. (8) to introduce this information. 2) *Global position encoding*, which is an absolute position embedding of the token. We utilize the global position by adding to the token representation $\boldsymbol{v}_i^r$ in Eq. (5).

We study the performance changes on quadruple extraction by using the alternatives, as shown in Fig. 7. We see that the Global position strategy shows the lowest helpfulness consistently, compared to the relative position methods. This finding suggests that relative distance may be more helpful in modeling the conversation discourse. Moreover, the RoPE gives the best usefulness, especially under inter-utterance cases. Intuitively, such dynamic position information offers more flexible bridging knowledge for easing the long-range dependence issue of term pairing where the entities are separated in different utterances in distance.

7 What To Do Next?

In this work, we propose an initial method to solve the DiaASQ task. Although achieving stronger performances than baselines, it could be further benefited from many angles. To facilitate the follow-up research in this direction, we try to shed light on several potential future works.

▶ **Making Better Use of The Dialogue Discourse Structure Information.** The core challenge of the DiaASQ task lies in handling conversation contexts. Compared to the typical case of single sentences, the dialogue utterances are syntactically disjoint. Thus, it is critical to carefully model the dialogue discourse structure information (Fei et al., 2022b), so as to better capture the dialogue semantics, for better recognition of the cross-utterance quadruples. Although we leverage the dialogue relative distance information (RoPE) in this work, without treating the dialogue utterances as a whole, our method may still lose some important discourse information. As seen in Fig. 6, our model's performance on the super cross-utterance quads is still far from satisfaction, i.e., zero F1 score on the cross->3-utterance case. Intuitively, constructing an explicit conversational discourse structure (i.e., the tree or graph structure) for the task is promising.

▶ **Enhancing Coreference Resolution.** In the conversation scenario, the speaker and target coreference is one of the biggest issues. In the DiaASQ task, the bundled sentiment elements (e.g., target, aspect, and opinion term) of one quad may be yielded by different users or maybe one individual. Besides, the sentiment terms may be coreferred by pronouns, for example, '*the screen quality of it*' where '*it*' refers to the target term '*Xiaomi 6*' mentioned in the previous context. Without correctly understanding the coreference, it is problematic for a system to precisely capture the context semantics, and thus leads to a wrong pairing between sentiment elements and unexplainable predictions.

▶ **Extracting Overlapped Quadruple.** It is common in our DiaASQ dataset that one sentiment term of one quad overlaps with other terms of another quad. For example, different electronic devices (targets) may have the same aspects, e.g., battery life, screen or size, etc. A sound DiaASQ system should also well solve the quadruple overlap issue. We note that the overlapped quads can essentially share certain structural information, and thus it is favorable to use such shared knowledge effectively.

▶ **Transferring Well-learned Sentiment Knowledge from Existing System.** The sentiment analysis community has developed a great amount of powerful ABSA systems well-trained on the large-scale free texts or existing sentiment corpora (Xu et al., 2020; Tian et al., 2020; Li et al., 2021). Since this work still inherits the basic spirit of ABSA, it is naturally a promising idea to transfer the existing well-trained sentiment-enriched ABSA model for enhancing the understanding of the DiaASQ task.

▶ **Multi-/Cross-lingual Dialogue ABSA.** One of the key challenges for more accurate multi-/cross-lingual ABSA is the missing of parallel annotations in different languages, i.e., causing troubles for label alignments (Feng and Wan, 2019; Fei and Li, 2020; Zhang et al., 2021b). As we annotate the DiaASQ dataset in two languages (i.e., Chinese and English) with parallel sentences, this paves the way for the research of more effective multi-lingual or cross-lingual dialogue-level ABSA.

8 Conclusion

This work introduces a new task of conversational aspect-based sentiment quadruple analysis, namely DiaASQ, which aims to detect the sentiment quadruple of *target-aspect-opinion-sentiment* structure in the conversation texts. DiaASQ bridges the gap between conversational opinion mining and fine-grained sentiment analysis. We manually construct a large-scale, high-quality dataset with Chinese and English versions for the task, with 1,000 dialogue snippets, including 7,452 utterances. We then benchmark the DiaASQ task with an end-to-end neural model, which effectively models the dialogue utterance interactions. Experiments demonstrate the advantages of our method in effectively learning the dialogue-specific features for better cross-utterance sentiment quadruple extraction.

Acknowledgment

This work is supported by the National Key Research and Development Program of China (No. 2022YFB3103602, No. 2017YFC1200500), the National Natural Science Foundation of China (No. 62176187), the Research Foundation of Ministry of Education of China (No. 18JZD015), China Scholarship Council (CSC), and Sea-NExT Joint Lab.

Limitations

Our paper has the following potential limitations. First, our current DiaASQ dataset is limited to only the domain of digital devices. We plan to further extend the DiaASQ texts to other domains, e.g., foods/restaurants, hotel/trips, etc. Secondly, our proposed model may be limited to insufficient modeling of the dialogue-level discourse structure information, which would somehow prevent us from obtaining further task improvements. Third, in DiaASQ task, it is more difficult to recognize the opinion terms, compared to the extraction of target and aspect terms. This may largely deteriorate the overall performance due to the fact that opinion expressions are much more flexible and sometimes are subject to satirical expression.

Ethical Considerations

Here we discuss the primary ethical considerations of the DiaASQ dataset.

Intellectual Property Protection. Our dataset is collected from the open Chinese social media platform via the officially open API.[4] Permissions are granted to copy, distribute and modify the contents under the terms of Weibo API distribution.

Privacy Claim. The user-specific information in the data is anonymized during preprocessing, and no personal information of the user or customer is included. The data collection procedure is designed for factual knowledge acquisition and does not involve privacy issues.

Annotator Information and Compensation. The crowd-sourcing annotators are the senior postgraduate students who are trained before annotating. We estimated that a skillful annotator needs 3 to 5 minutes to finish an annotation for each dialogue utterance. Therefore, we paid annotators 1 yuan ($0.15) for each utterance. The salaries for linguistic and computer science experts are determined by the average time they devote.

References

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. In *Proc. of ACL*, pages 3387–3402.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proc. of ACL*, pages 340–350.

Erik Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intell. Syst.*, 31(2):102–107.

Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022a. Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In *Proc. of ACL*, pages 2974–2985.

Shaowei Chen, Jie Liu, Yu Wang, Wenzheng Zhang, and Ziming Chi. 2020. Synchronous double-channel recurrent network for aspect-opinion pair extraction. In *Proc. of ACL*, pages 6515–6524.

Yuqi Chen, Chen Keming, Xian Sun, and Zequn Zhang. 2022b. A span-level bidirectional network for aspect sentiment triplet extraction. In *Proc. of EMNLP*, pages 4300–4309.

Zhexue Chen, Hong Huang, Bang Liu, Xuanhua Shi, and Hai Jin. 2021. Semantic and syntactic enhanced aspect sentiment triplet extraction. In *Proc. of ACL Findings*, pages 1474–1483.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese BERT. *IEEE ACM Trans. Audio Speech Lang. Process.*, pages 3504–3514.

Zehui Dai, Cheng Peng, Huajie Chen, and Yadong Ding. 2020. A multi-task incremental learning framework with category name embedding for aspect-category sentiment analysis. In *Proc. of EMNLP*, pages 6955–6965.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proc. of EMNLP*, pages 2112–2128.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *Proc. of ECAI*, pages 2006–2013.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proc. of EMNLP*, pages 3433–3442.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proc. of NAACL*, pages 2509–2518.

Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. 2022a. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proc. of IJCAI*, pages 4096–4103.

[4] https://open.weibo.com/wiki/API

Hao Fei, Jingye Li, Shengqiong Wu, Chenliang Li, Donghong Ji, and Fei Li. 2022b. Global inference with explicit syntactic and discourse structures for dialogue-level relation extraction. In *Proc. of IJCAI*, pages 4082–4088.

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proc. of ACL*, pages 7014–7026.

Hongliang Fei and Ping Li. 2020. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proc. of ACL*, pages 5759–5771.

Yanlin Feng and Xiaojun Wan. 2019. Learning bilingual sentiment-specific word embeddings without cross-lingual supervision. In *Proc. of NAACL*, pages 420–429.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. In *Proc. of ACL*, pages 7042–7052.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proc. of EMNLP*, pages 6279–6284.

Jingye Li, Hao Fei, and Donghong Ji. 2020. Modeling local contexts for joint dialogue act recognition and sentiment classification with bi-channel dynamic convolutions. In *Proc. of COLING*, pages 616–626.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *Proc. of AAAI*, pages 6714–6721.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. In *Proc. of IJCAI*, pages 4194–4200.

Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. Emocaps: Emotion capsule based model for conversational emotion recognition. In *Proc. of ACL Findings*, pages 1610–1618.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In *Proc. of EMNLP*, pages 246–256.

Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. Mmconv: An environment for multimodal conversational search across multiple domains. In *Proc. of SIGIR*, pages 675–684.

Lizi Liao, Ryuichi Takanobu, Yunshan Ma, Xun Yang, Minlie Huang, and Tat-Seng Chua. 2022. Topic-guided conversational recommender in multiple domains. *IEEE Trans. Knowl. Data Eng.*, 34(5):2485–2496.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021a. Towards emotional support dialog systems. In *Proc. of ACL*, pages 3469–3483.

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. *CoRR*.

Ziheng Liu, Rui Xia, and Jianfei Yu. 2021b. Comparative opinion quintuple extraction from product reviews. In *Proc. of EMNLP*, pages 3955–3965.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Proc. of ACL*, pages 432–439.

Rajdeep Mukherjee, Tapas Nayak, Yash Butala, Sourangshu Bhattacharya, and Pawan Goyal. 2021. PASTE: A tagging-free decoding framework using pointer networks for aspect sentiment triplet extraction. In *Proc. of EMNLP*, pages 9279–9291.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2022. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, pages 1–101.

Bo Pang and Lillian Lee. 2007. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

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 *Proc. of AAAI*, pages 8600–8607.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia V. Loukachevitch, Evgeniy V. Kotelnikov, Núria Bel, Salud María Jiménez Zafra, and Gülsen Eryigit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proc. of SemEval*, pages 19–30.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proc. of SemEval*, pages 486–495.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proc. of SemEval*, pages 27–35.

Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016. Context-sensitive twitter sentiment classification using neural network. In *Proc. of AAAI*, pages 215–221.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021. Dialogxl: All-in-one xlnet for multiparty conversation emotion recognition. In *Proc. of AAAI*, pages 13789–13797.

Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. 2022. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proc. of ACL*, pages 4232–4241.

Linfeng Song, Chunlei Xin, Shaopeng Lai, Ante Wang, Jinsong Su, and Kun Xu. 2022. CASA: conversational aspect sentiment analysis for dialogue understanding. *J. Artif. Intell. Res.*, 73:511–533.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *CoRR*.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective lstms for target-dependent sentiment classification. In *Proc. of COLING*, pages 3298–3307.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: sentiment knowledge enhanced pre-training for sentiment analysis. In *Proc. of ACL*, pages 4067–4076.

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 *Proc. of NeurIPS*, pages 5998–6008.

Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2019. Aspect sentiment classification towards question-answering with reinforced bidirectional attention network. In *Proc. of ACL*, pages 3548–3557.

Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. 2022. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proc. of AAAI*, pages 11513–11521.

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 *Proc. of IJCAI*, 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 *Proc. of EMNLP Findings*, pages 2576–2585.

Hu Xu, Lei Shu, Philip S. Yu, and Bing Liu. 2020. Understanding pre-trained BERT for aspect-based sentiment analysis. In *Proc. of COLING*, pages 244–250.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. In *Proc. of ACL*, pages 4755–4766.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proc. of NAACL*, pages 483–498.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Proc. of EMNLP*, pages 9209–9219.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021b. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proc. of EMNLP*, pages 9220–9230.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021c. Towards generative aspect-based sentiment analysis. In *Proc. of ACL*, pages 504–510.

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. Spanmlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In *Proc. of ACL*, pages 3239–3248.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ED: multi-modal multi-scene multi-label emotional dialogue database. In *Proc. of ACL*, pages 5699–5710.

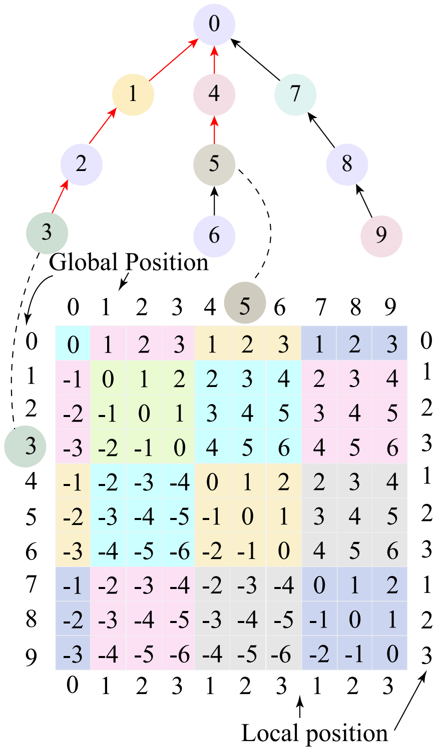
Ranran Zhen, Rui Wang, Guohong Fu, Chengguo Lv, and Meishan Zhang. 2021. Chinese opinion role labeling with corpus translation: A pivot study. In *Proc. of EMNLP*, pages 10139–10149.

Yuxiang Zhou, Lejian Liao, Yang Gao, Zhanming Jie, and Wei Lu. 2021. To be closer: Learning to link up aspects with opinions. In *Proc. of EMNLP*, pages 3899–3909.

A Model and Setup Specification

Algorithm 1 Calculating global indices of tokens in two threads

```
Require: $P_t$; Two thread $T_i, T_j$, where $i, j$ are thread id.
  if $i * j$ == 0 or $i$==$j$ then
    $P_t^{ij} = P_t(t \in T_i, T_j)$
  else if i < j then
    $P_t^{ij} = -P_t(t \in T_i)$
    $P_t^{ij} = P_t(t \in T_j)$
  else
    $P_t^{ij} = P_t(t \in T_i)$
    $P_t^{ij} = -P_t(t \in T_j)$
  end if
```



	0	1	2	3	4	5	6	7	8	9	
0	0	1	2	3	1	2	3	1	2	3	0
1	-1	0	1	2	2	3	4	2	3	4	1
2	-2	-1	0	1	3	4	5	3	4	5	2
3	-3	-2	-1	0	4	5	6	4	5	6	3
4	-1	-2	-3	-4	0	1	2	2	3	4	1
5	-2	-3	-4	-5	-1	0	1	3	4	5	2
6	-3	-4	-5	-6	-2	-1	0	4	5	6	3
7	-1	-2	-3	-4	-2	-3	-4	0	1	2	1
8	-2	-3	-4	-5	-3	-4	-5	-1	0	1	2
9	-3	-4	-5	-6	-4	-5	-6	-2	-1	0	3
	0	1	2	3	1	2	3	1	2	3	

Figure 8: Relative token distance calculation over the dialogue tree structure. To simplify the problem, we assume that one utterance has only one token.

Distance Encoding Details. In our data, tokens may distribute in different dialogue threads. Therefore, the relative distance between tokens cannot be calculated by subtracting their absolute position ids. However, the RoPE uses the global index to represent relative distance:

$$(\boldsymbol{\mathcal{R}}_m \boldsymbol{q})^\top (\boldsymbol{\mathcal{R}}_n \boldsymbol{k}) = \boldsymbol{q}^\top \boldsymbol{\mathcal{R}}_m^\top \boldsymbol{\mathcal{R}}_n \boldsymbol{k} = \boldsymbol{q}^\top \boldsymbol{\mathcal{R}}_{n-m} \boldsymbol{k},$$

where m and n are the absolute positions of two tokens. And m-n is not the relative distance of two tokens, which means the RoPE can not work typically. Therefore, we develop a method to calculate the relative token distance for different thread pairs. In detail, for each token t, we define the distance between t and the root node as its local position id P_t. For each two threads T_i and T_j, the absolute positions to represent their relative distance can be calculated by the Algorithm 1.

For example, as the block with different color shows in Fig. 8, we can see that $(P_t^{ij} - P_{t'}^{ij})(t \in T_i, t' \in T_j)$ is the relative distance of t and t'. Then we use the calculated P_t^{ij} as the absolute position to perform the RoPE operation.

Specification of Baselines. As no prior method is deliberately designed for DiaASQ, we consider reimplementing several strong-performing systems closely related to the task as our baselines. Here we give a complete description on these baseline systems.

- **CRF-Extract-Classify** is a three-stage system (extract, filter, and combine) proposed for the sentence-level quadruple ABSA by Cai et al. (2021). Here we retrofit the model to further support *target* term extraction.
- **SpERT** is proposed by Eberts and Ulges (2020) for joint extraction of entity and relation based on a span-based transformer. Here we slightly modify the model to support triple-term extraction and polarity classification.
- **Span-ASTE** is a span-based approach for triplet ABSA extraction (Xu et al., 2021). Similarly, we change it to be compatible with the DiaASQ task by editing the last stage of Span-ASTE to enumerate triplets.
- **ParaPhrase** is a generative seq-to-seq model for the quadruple ABSA extraction (Zhang et al., 2021a). We modify the model outputs to adapt to our DiaASQ task.

In particular, ParaPhrase (Zhang et al., 2021a) is a generative model proposed for the quadruple ABSA task. We re-implement the model and modify the output to fit it with our task. In short, given the source dialogue, we expect the model to output a sentiment-aware string:

"`Target` is *great/bad/ok*, because the `Aspect` of it is `Opinion` ...",

where `Target/Aspect/Opinion` is a term palaceholder, and *greate/bad/ok* is an opinionated expression indicating the specific sentiment polarity, i.e., positive/negative/other. For the dialogue in Fig. 1, a promising output is:

Xiaomi 6 is *great* because the *screen quality* of it is *very nice*.

Param.	Value
Learning rate(BERT)	1e-5
Learning rate(Other)	1e-3
Batch size	4 (dialogues)
Max grad norm	1.0
Weight decay	0.01
Epoch size	20
θ	10,000
$\boldsymbol{\alpha}$	[1, 5, 5, 5]
β	0.5
η	0.5
Parameter scale	210M
Training time / epoch	3min20s
CPU	Intel i9
GPU	NVIDIA RTX 3090

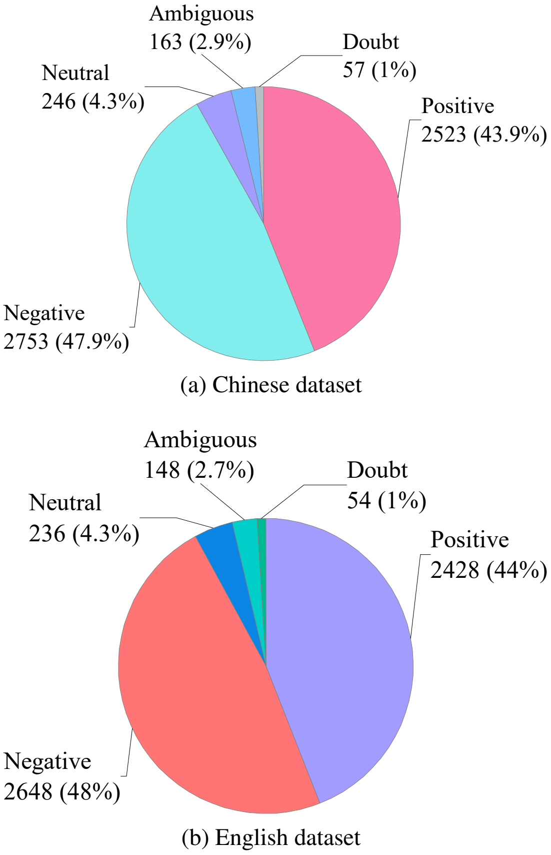
Table 5: Detail of the hyper-parameter setting.

Hyper-Parameters. Here we detail the experimental setups. The testing results are shown by our model tuned on the developing set to achieve the best developing performances. Hyper-parameters are listed in Table 5. We adopt AdamW as BERT optimizer. Our model is implemented with PyTorch and trained on the Ubuntu-20.04 OS with the Intel i9 CPU and NVIDIA RTX 3090 GPU.

B Extended Data Specification

Polarity Distribution. We statistics the polarity of quadruples in both Chinese and English datasets. As illustrated in Fig. 9, most of the quadruple express the clear sentiment tendency, which is constituent with the users' speaking habits on social media. Then, the positive and negative sentiment rates are near, indicating that our data sampling is balanced. Furthermore, due to the left three polarity being quite a few in our dataset, we merge them as a new category, others, for the convenience of extraction.

Cross-utterance Quadruples. We also analyzed the categories and numbers of cross-utterance quadruples. As shown in Fig. 10, most of the cross-utterance quadruples existed at two utterances next to each other, which reminds us that replying relationship can provide critical clues for cross-utterance quadruple extraction. Besides, the speaker and thread information needs to be further explored as they also indicate quite a few cross-utterance quadruples.



(a) Chinese dataset

(b) English dataset

Figure 9: Distribution of quadruple polarities.

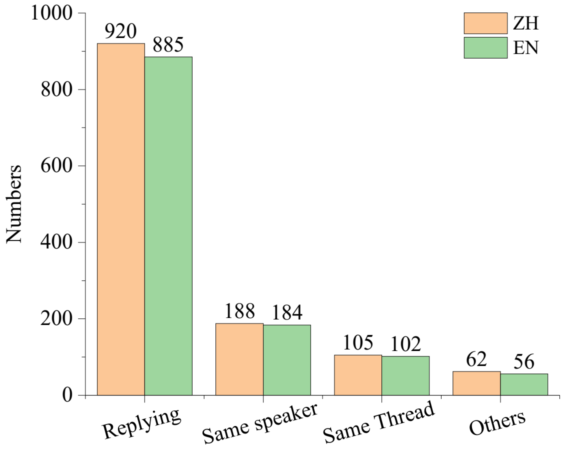


Figure 10: The number of different types of cross-utterance quadruples, whose elements at least come from two different utterances. 'Replying' denotes that the two utterances have replying relationship and 'Same speaker' indicates the two utterances spoken by the same person. 'Same thread' denotes the two utterances belonging to the same dialogue thread. 'Others' mainly contains very rare cases, e.g., one quadruple contains elements from different threads.

Quadruple Overlapping. As we cast earlier, there are a good number of quadruples overlapping between each others in our DiaASQ dataset, which is not specially described in our main article due to space limitations. As the first case shown in

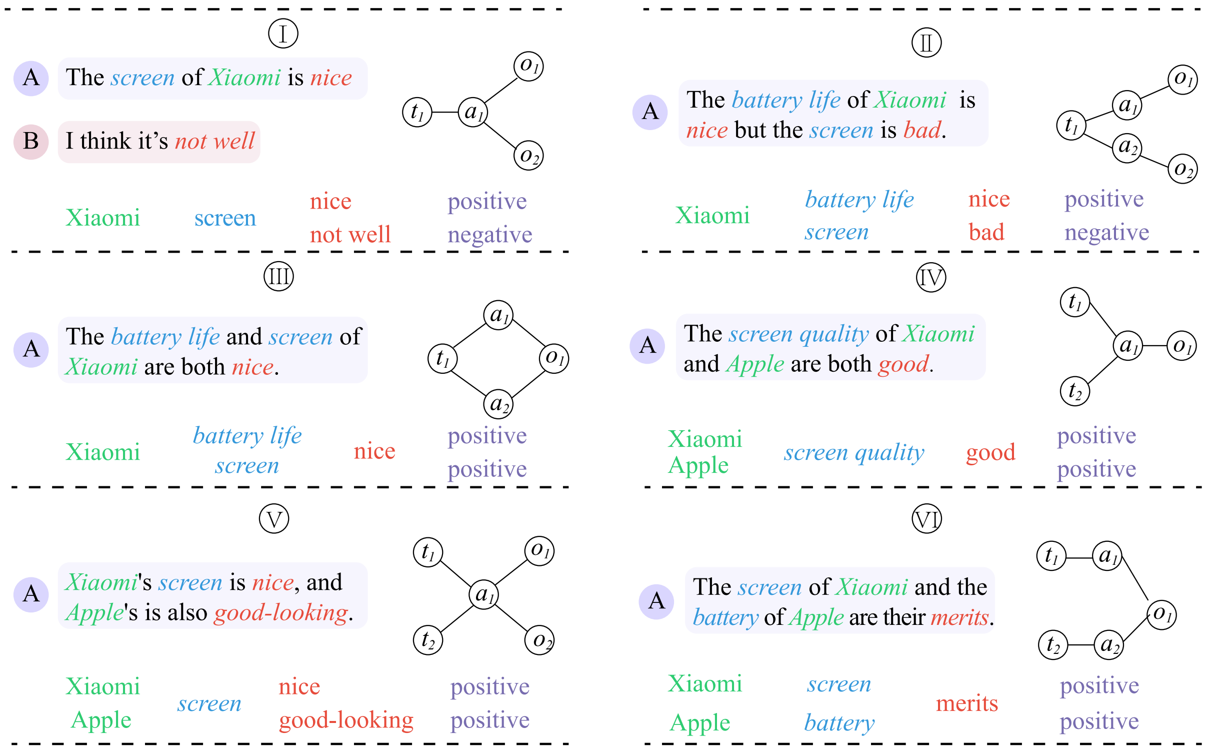

Ⅰ
A The screen of Xiaomi is nice
B I think it's not well
t_1 a_1 o_1 o_2
Xiaomi screen nice not well positive negative
Ⅱ
A The battery life of Xiaomi is nice but the screen is bad.
t_1 a_1 a_2 o_1 o_2
Xiaomi battery life screen nice bad positive negative
Ⅲ
A The battery life and screen of Xiaomi are both nice.
t_1 a_1 a_2 o_1
Xiaomi battery life screen nice positive positive
Ⅳ
A The screen quality of Xiaomi and Apple are both good.
t_1 t_2 a_1 o_1
Xiaomi Apple screen quality good positive positive
Ⅴ
A Xiaomi's screen is nice, and Apple's is also good-looking.
t_1 t_2 a_1 o_1 o_2
Xiaomi Apple screen nice good-looking positive positive
Ⅵ
A The screen of Xiaomi and the battery of Apple are their merits.
t_1 a_1 t_2 a_2 o_1
Xiaomi Apple screen battery merits positive positive

Figure 11: Quadruple overlap in our DiaASQ dataset, including a total of six cases.

Lang	Index	Same Polarity	Different Polarity
ZH	I	90	71
	II	1,684	1,366
	III	548	4
	IV	178	399
	V	70	106
	VI	59	84
EN	I	102	67
	II	1,578	1,260
	III	549	4
	IV	170	382
	V	67	99
	VI	62	77

Table 6: Statistics of overlapped quadruples. The second column of each row is the index of the subplot in Fig. 11.

Fig. 11, two quadruples may contain the same target and opinion term. The overlap information can actually provide valuable clues for better extraction. Here we show in Fig. 11 all types of overlap cases of the Chinese version dataset and their statistics information in Table 6.

C Specification on Data Construction

This part describes the details that we constructed the DiaASQ dataset, including the data acquisition and annotation projection.

C.1 Data Acquisition

Fig. 12 illustrates the overall workflow that we obtain a high-quality original corpus from social media.

First, based on the official leaderboard, we collected the top 100 influential digital-domain bloggers on Weibo and crawl their history tweets as many as possible. Meanwhile, we also built a mobile-phone related keywords library and crawled tweets and their comment searched by these keywords. After this step, we obtained nearly 9 million tweets and comments, and the replying relation was also recorded. Then, we conduct a preliminary screening to exclude posts with less than ten replies or no father node. About 1.2 million posts were retained after this procedure. Next, according to the replies relation, we combine these posts into dialogue trees, whose root nodes are level-1 comments below each primary tweet, and the maximum depth is no more than 4. Based on the phone-related keywords and a collection of abusive words, a more strict filtering rule is performed at the tree level. In detail, a thread will be kept if it contains any two of the phone-related keywords and does not contain any abusive words. Once a dialogue tree has three valid threads and the total number of nodes is between 6 and 10, it will be selected as the candidate dialogue. Around 6,000 dialogue

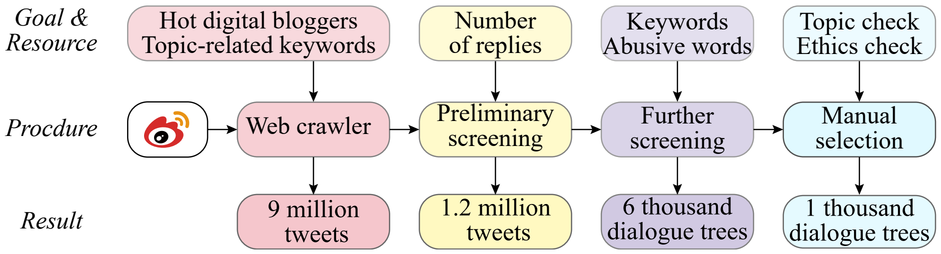


Figure 12: The workflow of data acquisition and preprocesssing.

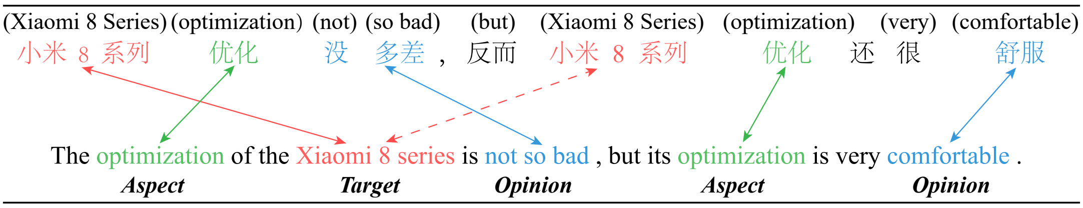

(Xiaomi 8 Series) (optimization) (not) (so bad) (but) (Xiaomi 8 Series) (optimization) (very) (comfortable)
小米 8 系列 优化 没 多差 , 反而 小米 8 系列 优化 还 很 舒服
The optimization of the Xiaomi 8 series is not so bad , but its optimization is very comfortable .
Aspect
Target
Opinion
Aspect
Opinion

Figure 13: A example for projection correction. The red dotted line denote manually added alignment relation.

trees are left after these steps. Finally, we manually checked the candidate dialogue, and dialogues that are indeed phone-related and do have no ethical issue are selected as the final corpus. After a very rigorous processing, we obtained 1,000 pieces of high-quality tree-like dialogues.

Item	Text
Source	所以我还是买了*12x*，虽然性价比不高
Translated	*So I still bought 12x,* *although the price is not high*
Revision	*So I still bought 12x,* *although the cost-effective is not high.*
Source	9带*dc*肯定香，不过夏天你再试试
Translated	*9 with DC is definitely* *fragrant, but you try again in summer*
Revision	*9 with DC is definitely* *nice, but you could try it again in summer*

Table 7: Two typical translation revision examples. The first one is token-level translation error correction. And the second one shows a more proper statement.

C.2 Parallel-Language Data Construction

We also constructed an English version dataset based on the Chinese corpus via the annotation projection method. Following Fei et al. (2020) and Zhen et al. (2021), the entire process contains two steps: text translation and annotation projection. We manually revise the process result after each step to ensure the corpus quality.

Step1: Text Translation We first utilize Google Translate API to translate the Chinese text into English.[5] Despite the stunning performance of NMT(neural machine translation), it still makes some mistakes during translation. The main reason is that our corpus is collected from social media and full of non-grammatical sentences, which has brought challenges for the NMT system to generate correct and elegant translations. Therefore, we carefully revise the translation to eliminate errors and meanwhile improve readability. Table 7 lists one of the errors and revision results.

Step2: Annotation Projection Then we conduct projection to obtain English versions corpus based on original Chinese annotation. Specifically, we achieve corpus projection with the help of awesome-align (Dou and Neubig, 2021), an excellent alignment tool based on large-scale multilingual language models. We found that the alignment tool is not good at aligning named entities, and a representative error and correction are shown in Fig. 13. After manually correcting all of the projection results, we obtained the final annotated corpus.

C.3 Data Instances

In Table 8 we illustrate a full piece of data instance (a conversation) with our annotation (English version is shown).

[5] https://cloud.google.com/translate

Key	value	
Dialogue-ID	0002	
Dialogue	0 *This phone is not very good , but compared to the iPhone , I think it is better than the iPhone except for the processor [laughs cry]*	
	1 *The iPhone is excellent as the processor and iOS , and others have been beaten by Android for many years .*	
	2 *Really . Sales also beat Android . Android manufacturers claim to be high - end and high - end every day , but they are just children in front of Apple .*	
	3 *Samsung , Xiaomi does not all exceed Apple ? Because there are too many Android systems , there is only one iOS . If there is only one Android , what do you think of the result ?*	
	4 *As you say , I have n't used Xiaomi , so I can 't comment . But traveling , my friend 's Xiaomi phone never took good photos . Especially when went to Malinghe Waterfall this week, we had to take pictures . Every photo taken by my brother 's Mi 11 was blurry . This experience is also speechless .*	
	5 *Xiaomi 11 is really not good [black line] [black line] [black line] .*	
	6 *The parameters overwhelm every year , and the experience is general every year ... that 's all . The phone is yours , who uses it , who knows .*	
Replies	(-1, 0, 1, 2, 0, 4, 0, 6)	
Speakers	(0, 1, 2, 1, 3, 0, 3, 0)	
Targets	(20, 21, *iPhone*)	(30, 31, *iPhone*)
	(82, 83, *Samsung*)	(84, 85, *Xiaomi*)
	(169, 171, *Mi 11*)	(180, 182, *Xiaomi 11*)
	(236, 237, *iPhone*)	(248, 249, *iPhone*)
Aspects	(52, 53, *Sales*)	(175, 176, *experience*)
	(207, 208, *experience*)	(199, 201, *The parameters*)
	(35, 36, *processor*)	(24, 25, *processor*)
	(37, 38, *iOS*)	(231, 232, *experience*)
	(244, 246, *image system*)	(145, 146, *photos*)
Opinions	(17, 18, *better*, pos)	(201, 202, *overwhelm*, pos)
	(54, 55, *beat*, pos)	(184, 186, *not good*, neg)
	(172, 173, *blurry*, neg)	(32, 33, *excellent*, pos)
	(178, 179, *speechless*, neg)	(88, 89, *exceed*, pos)
	(209, 210, *general*, neg)	(243, 244, *backward*, neg)
	(233, 234, *better*, neg)	
Quadruples	(20, 21, 24, 25, 17, 18, pos, *iPhone, processor, better*)	
	(30, 31, 35, 36, 32, 33, pos, *iPhone, processor, excellent*)	
	(30, 31, 37, 38, 32, 33, pos, *iPhone, iOS, excellent*)	
	(30, 31, 52, 53, 54, 55, pos, *iPhone, Sales, beat*)	
	(82, 83, 52, 53, 88, 89, pos, *Samsung, Sales, exceed*)	
	(84, 85, 52, 53, 88, 89, pos, *Xiaomi, Sales, exceed*)	
	(180, 182, 145, 146, 184, 186, neg, *Xiaomi 11, photos, not good*)	
	(248, 249, 244, 246, 243, 244, neg, *iPhone, image system, backward*)	
	(236, 237, 231, 232, 233, 234, neg, *iPhone, experience, better*)	
	(169, 171, 145, 146, 172, 173, neg, *Mi 11, photos, blurry*)	
	(169, 171, 175, 176, 178, 179, neg, *Mi 11, experience, speechless*)	

Table 8: An instance of our annotated corpus. The start and end positions of each entity are their global positions in the tokenized dialogue. '-1' in "Replies" row indicate the corresponding utterance is the root of dialogue tree.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8 Limitations

☒ A2. Did you discuss any potential risks of your work?
My research poses no potential risk to participants or the general public as the dataset used is publicly available and the annotation is solely for scholarly study.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Appendix E.2

☑ B1. Did you cite the creators of artifacts you used?
Appendix E.2.1

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
9 Ethical Considerations

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix E.2.1

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
9 Ethical Considerations

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix E.2.3 Detailed Guidance for Annotation

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3 Data Construction (Data Insights).

C ☑ Did you run computational experiments?

6 Experiment

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
6.1 Settings

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B Model and Setup Specification

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
6.1 Settings

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We have mentioned that we use BERT to encode the text, and BERT contains the tokenizer. Thus, we didn't claim the preprocessing tool repeatedly.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

3 Data Construction

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
E.2 Data Annotation Manual

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
9 Ethical Considerations

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
9 Ethical Considerations

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.