SGCD: Subtask-Guided Causal-Debiasing Framework for Robust Cross-Utterance Sentiment Quadruple Extraction in Dialogues

Xiang Li, Keyu Yao, Gang Shen*
School of Software Engineering
Huazhong University of Science and Technology, Wuhan, China
{lixiang511, keyuyao, gang_shen}@hust.edu.cn

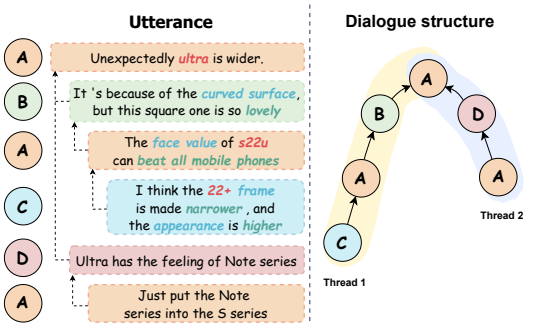
Abstract

The rise of digital social media has generated a vast amount of conversational data on platforms like Twitter and Reddit, allowing users to express sentiments through multi-turn dialogues. Dialogue-level aspect-based sentiment quadruple analysis (DiaASQ) seeks to extract structured information in the form of quadruples from these dialogues. However, it encounters challenges related to cross-utterance elements and focus bias. To address these issues, we introduce the Subtask-Guided and Causal-Debiasing (SGCD) framework. This framework leverages subtask-specific features to guide the learning of token-level features, which are then adaptively combined at the utterance level to meet specific semantic requirements. The SGCD framework employs multi-granularity attention paths to enhance cross-utterance matching and dialogue structure modeling. It also incorporates structural causal graphs and inverse probability weighting to mitigate biases from speakers and thread structures. Experimental results demonstrate that SGCD outperforms state-of-the-art methods, improving semantic modeling and bias robustness. This approach provides an effective solution for structured sentiment analysis in complex dialogues.

1 Introduction

The advent of digital socialization attracts users to frequent interactive online platforms such as Twitter, Reddit, Facebook groups, and various forum communities. They engage in sustained, multi-turn dialogues on product experiences, social events, or personal emotions. As of 2025, global social media users have reached 5 billion, accounting for more than 60% of the world population (Singh, 2025). The extensive conversational content produced on these platforms exhibits a variety of linguistic styles and rich expressions, displaying emotional transitions among speakers and utterances. This emerging trend presents challenges for computers in understanding sentiments.

Dialogue-level aspect-based sentiment quadruple analysis (DiaASQ) (Li et al., 2023), derived from aspect-based sentiment analysis (ABSA), has garnered significant attention in sentiment understanding. DiaASQ targets extracting structured sentiment quadruples (Target, Aspect, Opinion, Sentiment) from complete dialogues. It faces challenges from complex dialogue structures, cross-turn utterances, fragmented information distribution, and the possibility that quadruple components may span multiple utterances or speakers (see Figure 1).



Type	Traget	Aspect	Opinion	Sentiment
Cross-utt	ultra	curved surface	lovely	Positive
Intra-utt	s22u	face value	beat all mobile phones	Positive
Intra-utt	22+	frame	narrower	Negative
Intra-utt	22+	appearance	higher	Positive

Figure 1: An example of cross-utterance scenario.

Specifically, DiaASQ introduces two more complex settings than sentence-level ABSA. First, emotional elements exhibit significant cross-utterance distribution, requiring the model to perform cross-layer matching in context. As shown in Figure 1, the opinion word "lovely" mentioned by User B in the reply, and the aspect word "curved surface" need to be cross-matched to the target word "Ultra"

*This work was supported by the National Key Research and Development Project of China under Grant 2023YFC3304504. Xiang Li and Keyu Yao contributed to this work equally (*Corresponding author: Gang Shen*).

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11430–11440

mentioned by User A. These elements distribute unevenly across utterances, with some containing no elements. Second, quadruples in dialogue often concentrate on specific targets and aspects, leading to significant focus bias in prediction perspectives, which affects the model's generalization and robustness in low-resource settings. For example, in Thread 1 of Figure 1, all aspects focus on the appearance of mobile phones.

Recently, researchers have attempted to improve performance using multi-task learning, multi-granularity modeling, or context-sharing mechanisms. Models like DMIN (Huang et al., 2024), DMCA (Li et al., 2024b), and H2DT (Li et al., 2024a) execute semantic sharing by parallelly training multiple subtasks (e.g., target extraction and opinion detection). However, these methods typically control learning priorities by adjusting loss function weights, lacking explicit instruments to characterize feature discrepancies between subtasks, and struggling to balance modeling effects across all subtasks. Besides, they often use fixed-proportion multi-view feature aggregation strategies, ignoring the actual contribution of different utterances to predictions, hindering models from capturing critical information in cross-utterance structures and rendering them susceptible to dialogue structure bias.

To address these problems, this study proposes a Subtask-Guided and Causal-Debiasing (SGCD) framework for collaborative learning of three subtasks: entity extraction (ent), sentiment relation classification (rel), and sentiment polarity classification (pol), while enhancing structural awareness and bias robustness. The framework relies on two core hypotheses: First, the deep semantic features required by all subtasks can form a latent semantic space, with subtasks differing in their priority on distinct subsets of the features. Second, the semantic contributions of cross-utterance features to final predictions vary significantly, requiring dynamic selection and modeling at the structural level. Based on these hypotheses, the SGCD framework has three key designs to improve model performance:

Subtask-Guided Feature Selection: Leveraging dialogue thread modeling, this module guides the learning of deep token-level features through subtask-specific guidance signals. It adaptively selects and fuses these features at the utterance level to extract diverse deep feature representations relevant to individual subtasks.

Subtask-Specific Attention Fusion: Aiming at the heterogeneity of semantic requirements across subtasks, SGCD constructs multi-granularity attention paths to dynamically fuse semantic representations at the token and utterance levels, enhancing cross-utterance matching and dialogue structure modeling capabilities.

Bias Reduction with Backdoor Adjustment: This module establishes a structural causal graph, including variables of speakers and thread structures, and adjusts biases at both the sample and token levels via inverse probability weighting (IPW) to approximate unbiased training distributions, effectively mitigating interference from structural and pragmatic biases in dialogues.

Experimental results on the DiaASQ datasets demonstrate that the proposed SGCD method outperforms state-of-the-art approaches across multiple standard and bias-sensitivity metrics, particularly in cross-utterance scenarios. The source code of SGCD can be accessed at https://github.com/hustselab511/SGCD.

2 Related Work

Aspect-based sentiment analysis (ABSA) has evolved from target-aspect identification (Pontiki et al., 2014) to more structured extractions such as triplets (ASTE) (Hua et al., 2024) and quadruplets (ASQP) (Hua et al., 2024; Zhang et al., 2022a). To improve semantic modeling, researchers have incorporated graph structures and attention mechanisms, including GCNs (Zhang et al., 2019), Transformers (Sun et al., 2019), and multi-granularity denoising approaches (Luo et al., 2024; Zhang et al., 2024), which also inspire dialogue-level sentiment extraction.

To address the problem caused by the emotional elements spread across multiple utterances, the DiaASQ task targets the joint extraction of targets, aspects, opinions, and sentiments in multi-turn dialogues. MVQPN (Wu et al., 2020) captures emotional flow via query propagation, DMIN (Huang et al., 2024) integrates multi-level discourse features, while DMCA (Li et al., 2024b) and H2DT (Li et al., 2024a) enhance cross-utterance modeling through hierarchical fusion and token-level graph reasoning. However, most approaches adopt a "shared encoder + weighted loss" framework (Cai et al., 2023; Zhang et al., 2023), lacking explicit semantic disentanglement across tasks. We address this by introducing task-guided represen-

tation learning and dynamic feature selection to improve interpretability and generalization.

The causal inference has become increasingly relevant for debiasing in NLP, offering a principled way to model confounders and causal paths (Feder et al., 2021). DINER (Wu et al., 2024) applies SCM, IPW, and counterfactual reasoning to reduce ABSA bias. MCIS (Yang et al., 2024) mitigates multimodal bias via causal graph purification, and generative counterfactual augmentation improves polarity robustness. While effective in static and multimodal settings, causal modeling in multi-turn dialogues is underexplored. Prior work such as cfVQA (Niu et al., 2020) shows promise but lacks structural bias modeling for DiaASQ-specific variables like thread and speaker. Our work bridges this gap through integrated causal and task-structured learning.

3 Proposed Method

We represent a dialogue as $D = \{u_1, u_2, \ldots, u_n\}$, where u_i denotes the i-th utterance and n is the total number of utterances. Each utterance $u_i = \{t_1, \ldots, t_{m_i}\}$ consists of a sequence of tokens, with m_i indicating its length. Each utterance also includes two structural metadata: a speaker label sequence $s = \{s_1, \ldots, s_n\}$ and a reply relationship record $r = \{l_1, \ldots, l_n\}$, where l_i specifies the index of the preceding utterance it replies to. Our objective is to predict labels for each token across subtasks $\gamma \in \{\text{ent}, \text{rel}, \text{pol}\}$, following the sentiment quadruple prediction setup of DiaASQ (Li et al., 2023). The overall framework of our proposed method is illustrated in Figure 2, comprising three consecutive steps.

Subtask-Guided Feature Selection: Guiding the learning of deep token-level features related to subtasks, and utterance-level features are obtained from the token-level through dynamic selective fusion.

Subtask-Specific Attention Fusion: Constructs task-specific attention aggregation paths for token-level and utterance-level features, flexibly combining information from both levels to produce final representations and classification results.

Bias Reduction via Backdoor Adjustment: Performs backdoor adjustment on two bias paths ($S \rightarrow L$ and $T \rightarrow L$) within a Structural Causal Model (SCM). Inverse Probability Weighting (IPW) adjusts training losses at both the dataset-wide and sample-specific levels to approximate bias-free distributions.

3.1 Subtask-guided Feature Selection

Tokens are the basic units for final label prediction in dialogue, while utterances carry speakers' implicit prior knowledge and fundamental units for cross-utterance interactions. Extracting information from these units is critical for solving the three subtasks. Balancing multi-task losses through weight adjustment is fragile. Therefore, we design subtask-guided features to explicitly direct the shared network across subtasks to focus on learning more generic and deeper semantic features.

Subtask-guided encoder (SGEncoder): SGEncoder builds on DMIN's thread modeling approach, treating dialogue threads as utterance-level directed chains input into Pre-trained Language Models (PLMs) (Devlin et al., 2019) to enhance cross-utterance interaction modeling. Each utterance is represented as $u'_i = \{[\text{CLS}], u_i, s_i\}$, where [CLS] is a special start token in PLMs and s_i is the corresponding speaker label. The first utterance u'_1 of each thread is designated as the root node and repeated at the beginning of each thread. For thread $t_k = \{u'_1, u'_i, u'_{i+1}, \ldots, u'_j\}$, we can derive the PLM-encoded representation, $H_{u'_i} = \text{PLMs}(u'_i)$.

Subtask-guided features, matching the semantic needs of specific subtasks, serve as non-overlapping portions of deep semantic features required by subtasks, explicitly reducing subtasks' dependence on overlapping feature modeling. To achieve this, different PLM layer features are selected to match each subtask's semantic requirements (Jawahar et al., 2019), and Average Pooling (AvgPool) (Lecun et al., 1998) with a linear transformation is used to obtain the unified representation H_g^γ for subtask γ.

For the token-level representations, interactions between nodes are crucial for capturing semantic flow and information transmission. Inspired by prior work by Zhang et al. (2022b), we use graph convolutional networks (GCNs) (Chen et al., 2022) to systematically model interaction dependencies between nodes in dialogue threads, with features fused via residual connections (He et al., 2016) and layer normalization (LN) (Xu et al., 2019a). The fused token-level representation for subtask γ is:

$$\begin{aligned} H_{\text{tok}}^\gamma = &\text{LN}(H_t + \text{GCNs}(A_{\text{sem}}, H_t) + \\ &\text{GCNs}(A_{\text{int}}, H_t) + H_g^\gamma) \end{aligned} \quad (1)$$

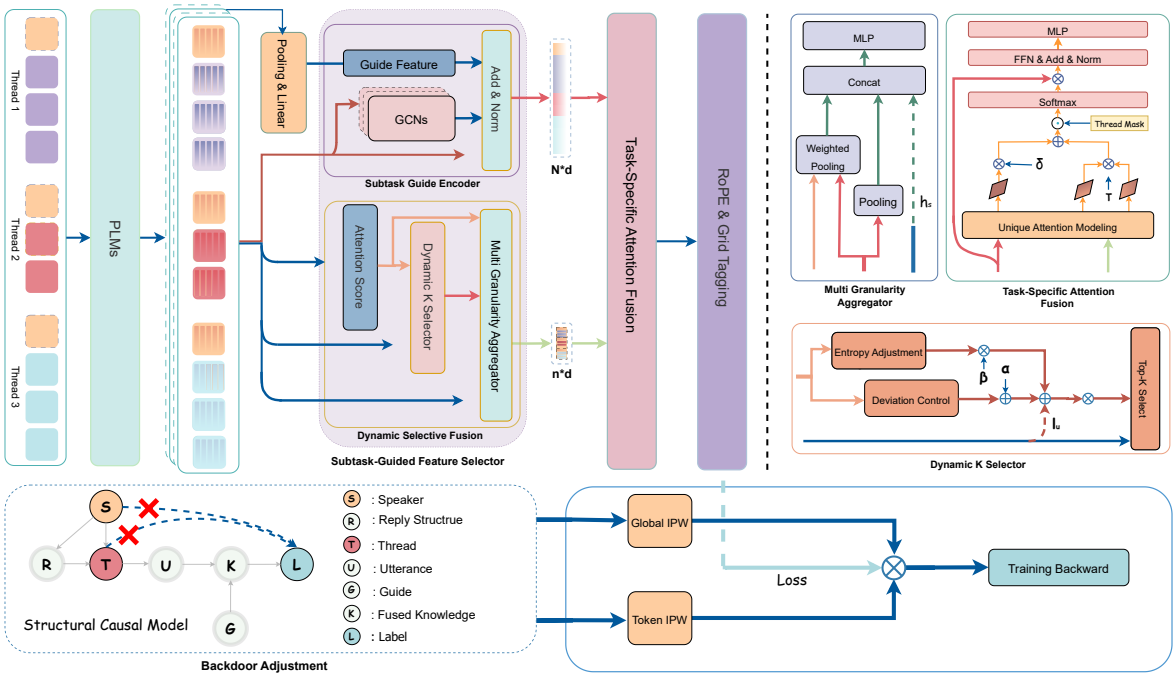


Figure 2: The architecture of the proposed SGCD model.

Here, A_{sem} is a semantic adjacency matrix built via self-attention (Vaswani et al., 2017) to capture implicit semantic connections within and across utterances, A_{int} is an interaction adjacency matrix based on reply structures, and H_t are features output by PLMs.

Dynamic Selective Fusion (DSFusion). The proportion of tokens containing valid information varies across utterances, inspiring dynamic selective fusion. Specifically, using an attention mechanism on $H_{u_i'}$, query $Q_{u_i'}$ and key $K_{u_i'}$ are derived. The global representation of $K_{u_i'}$ is obtained via average pooling and dotted with $Q_{u_i'}$ to calculate the attention score of each token for the utterance:

$$S_{u_i'} = W_{\text{score}} \left(\sigma \left(\text{LN} \left(\text{avg}(K_{u_i'}) \cdot Q_{u_i'} \right) \right) \right) \quad (2)$$

Dynamic K selector (DKSelector). DKSelector is used to dynamically select features. For an utterance i with length m_i, the entropy of $S_{u_i'}$ measures distribution uniformity (Caticha and Preuss, 2004), adjusting the K value in the Top-K method. Entropy adjustment (normalized by $\log(m_i)$) yields K_i^{ad}.

We thereby introduce a deviation control: if the highest token score exceeds twice the average score, $K_i^{\text{red}} = \zeta$ (where ζ is a hyperparameter), otherwise 0. With hyperparameters α and β for base ratio control, the final dynamic K value is:

$$K_i^{\text{dyn}} = \max \left(1, \text{int} \left((\alpha + \beta K_i^{\text{adj}} - K_i^{\text{red}}) m_i \right) \right) \quad (3)$$

After Top-K selection using K_i^{dyn}, the representation H_i^{du} of $H_{u_i'}$ is obtained. A Multi Granularity Aggregator (Wang et al., 2020) concatenates speaker features, weighted and unweighted pooling features, which are then aggregated via a multilayer perceptron (MLP) to generate the final utterance representation h_{oi}. We then model the global inter-utterance relationships via GCNs:

$$H_{\text{utt}} = \text{LN}(\text{GCNs}(A_{\text{utt}}, H_o) + H_o) \quad (4)$$

where $H_o = \{h_{o1}, h_{o2}, \ldots, h_{on}\}$ and A_{utt} is an adjacency matrix where $A_{\text{utt}} = 1$ if two utterances have a reply relationship, otherwise 0.

3.2 Subtask-Specific Attention Fusion

Due to the semantic disparities across subtasks, the model must flexibly aggregate contextual information for specific subtasks. To address this, we propose subtask-specific attention fusion (SAF) to construct dedicated attention aggregation paths for each subtask to explicitly model multi-granular interactions between tokens and utterances (Vaswani et al., 2017).

Following Huang et al. (2024), we use token-level features for subtask γ $H_{\text{tok}}^{\gamma} \in \mathbb{R}^{N \times d}$ and

utterance-level features $H_{\text{utt}} \in \mathbb{R}^{n\times d}$ as queries (Q) and keys (K) to construct three attention score matrices: $S_{\text{tok-tok}} \in \mathbb{R}^{N\times N}$, $S_{\text{tok-utt}} \in \mathbb{R}^{N\times n}$, and $S_{\text{utt-tok}} \in \mathbb{R}^{n\times N}$. Learnable coefficients $\tau, \delta \in [0, 1]$ adapt to each task's preference for attention paths:

$$A_{\text{itg}} = \tau \cdot (S_{\text{tok-utt}} \cdot S_{\text{utt-tok}}) + \delta \cdot S_{\text{tok-tok}} \quad (5)$$

To ensure effective fusion, a thread mask $M^{\text{th}} \in \{0, 1\}^{N\times N}$ is applied: $M^{\text{th}}_{ij} = 1$ if tokens i and j belong to the same thread, otherwise 0. The fused representation is the result of linear transformation:

$$H_{\text{itg}} = W_{\text{itg}} \left(\text{softmax}(A_{\text{itg}} \odot M^{\text{th}}) \cdot H^{\gamma}_{\text{tok}}\right) \quad (6)$$

The fused representation H_{itg} goes through a feedforward network with residual connections. Task-specific MLP layers combined with RoPE (Su et al., 2024), generate the feature output v^{γ}_i of the i-th token for subtask γ:

$$\begin{cases} H_f = \text{LN}(\text{FFN}(H_{\text{itg}}) + H_{\text{itg}}) \\ v^{\gamma}_i = R(\theta, i) \cdot \text{MLP}(H_{f,i})\} \end{cases} \quad (7)$$

Here, $R(\theta, i)$ is the position encoding matrix based on index i and hyperparameter θ.

Following Li et al. (2023), we adopt a grid tagging method to derive final relation predictions. During training, we optimize using weighted cross-entropy loss across three subtasks. For task γ, let G = total samples, N = tokens per sample, y^{γ}_{ij} and p^{γ}_{ij} = ground truth and predicted probabilities, and α_{γ} = task weight. The loss is defined as:

$$\mathcal{L}_{\gamma} = -\frac{1}{G \cdot N^2} \sum_{g=1}^{G} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{\gamma} y^{\gamma}_{ij} \log(p^{\gamma}_{ij}) \quad (8)$$

3.3 Bias Reduction with Backdoor Adjustment

We construct the structural causal model (SCM) for DiaASQ by extending Wu et al. (2024)'s SCM for ABSA with dialogue-specific causal variables S (speaker habits/patterns), R (reply relations), and T (thread structures), as illustrated in Figure 1. Speaker-related implicit dialogue habits and knowledge in S form a causal path S→L, while reply-driven opinion convergence/divergence patterns in T create a T→L path. Leveraging backdoor adjustment (Fang et al., 2024) and Inverse Probability Weighting (IPW) for these paths, the model mitigates biases from S and T in finite samples, achieving "pseudo-randomness" akin to randomized trials (Chesnaye et al., 2022) for robust causal inference in dialogue sentiment analysis.

For a dialogue sample D_i with n_i utterances, larger n_i increases the risk of bias from speaker-message quantity disparities and implicit prior differences. Thus, IPW is used to downweight the loss contribution of such samples during training. The global inverse probability weight for sample u_i is:

$$w^{(i)}_{\text{global}} = \frac{\exp(1/n_i)}{\sum_{j=1}^{N} \exp(1/n_j)} \cdot N \quad (9)$$

For each token t in utterance u_i, with speaker $s_i \in \{0, 1, \ldots, S_{\max}\}$ and belonging to thread T_j (containing $|T_j|$ utterances), the joint inverse probability weight for token t is:

$$w_t = \left[\frac{S_{\max} \cdot \exp(1/r_{s_i})}{\sum_{j=0}^{S_{\max}} \exp(1/r_j)}\right] \cdot \left[\frac{|T_h| \cdot \exp(1/|T_j|)}{\sum_{k=1}^{|T_h|} \exp(1/|T_k|)}\right] \quad (10)$$

The IPW-weighted loss function for sample D_i is:

$$\mathcal{L}_{\text{ipw}} = \sum_{i=1}^{N} w^{(i)}_{\text{global}} \cdot \left(\frac{1}{n_i} \sum_{t=1}^{n} w_t \cdot \mathcal{L}^{(i)}_t\right) \quad (11)$$

Here, $w^{(i)}_{\text{global}}$ suppresses the overall influence of long-dialogue samples, while w_t mitigates biases from speakers (S) and thread structures (T). The combined weighting approximates training under a re-sampled debiased distribution.

4 Experiment and Results

4.1 Datasets

To evaluate the effectiveness of the proposed SGCD method, we conducted experiments using DiaASQ (Li et al., 2023), a multilingual benchmark dataset specifically developed for extracting dialogue-based aspect sentiment quadruples. DiaASQ covers user-generated conversations related to mobile devices on Chinese social media platforms. It includes Chinese (ZH) and English (EN) versions, featuring real-world dialogue threads annotated with information about aspects, opinions, sentiments, and targets. The Chinese version contains a total of 1,000 dialogues, 7,452 utterances, and 5,742 sentiment quadruples, while the English version contains 5,514 quadruples. On average, each dialogue involves approximately five speakers. The datasets contain a wide range of sentiment

expressions and context-dependent relationships, including 1,275 (22.2%) cross-utterance quadruples in the Chinese version and 1,227 (22.3%) in the English version. As for pair labels, the training set of dataset ZH includes 4,699 T-A, 5,931 T-O, and 3,989 A-O instances, while the training set of dataset EN contains 4,823 T-A, 6,062 T-O, and 4,297 A-O instances. We adhered to the standard partitioning protocol, splitting the data into training, validation, and testing sets in an 8:1:1 ratio. The evaluation was performed across multiple subtasks, including span extraction, pair identification, and quadruple prediction.

4.2 Model Implementation

The hyperparameter α in the DSFusion controls the proportion of top-k components. Through statistical analysis of the sample distribution in the dataset and referencing the proportion of task-relevant key tokens within utterances, α is set to 0.7 for ZH and 0.5 for EN to cover primary information. The values of β and ζ are set to 0.1 and 0.2, respectively.

To avoid the over-smoothing problem in graph neural networks (Li et al., 2018) and integrate the experimental conclusions (Xu et al., 2019b), the MGTEncoder uses 3 GCN layers, and the Top-K Selector module employs 2 GCN layers, achieving optimal semantic modeling performance.

We set the dropout rate to 0.2 to mitigate overfitting risks, and the training duration is 60 epochs to ensure the model fully converges and maintains stable performance under mini-batch training. In line with existing studies on the DiaASQ task, we use RoBERTa-Large (Liu et al., 2019) for English (EN) and Chinese-RoBERTa-wwmext-base (Cui et al., 2021) for Chinese (ZH) as the PLMs, with all other settings consistent with DMIN to ensure fair comparisons. The model has 1,664M parameters. It took 5 hours to train the model for 60 epochs on an NVIDIA RTX 3090 GPU, and the inference of all test sets took 2 minutes.

4.3 Evaluation Metrics

We use exact-match F_1 as our primary metric, following Li et al. (2023). Specifically, we report Micro-F_1 and Identification-F_1 (Barnes et al., 2021) to evaluate performance on the DiaASQ task. Micro-F_1 assesses the overall correctness of quadruples, including sentiment polarity, while Identification-F_1 focuses on element boundaries and structural accuracy by excluding polarity.

Furthermore, we also analyze the model's performance on sub-tasks using span and pair F_1 scores, which include Target-Aspect (T-A), Aspect-Opinion (A-O), and Target-Opinion (T-O) pairings. Performance assessment covers scenarios defined by Intra-Utt. (within a single utterance), Inter-Utt. (multi-utterance dependencies in the same thread), and Cross-Utt. (typically across threads or loosely associated utterances). The result allows us to investigate how effectively the model captures complex dependencies and dialogue structure.

Finally, we explicitly assess the model's robustness in extracting quadruples across utterances, a particularly challenging aspect of the DiaASQ task. We categorize test samples based on their cross-utterance distance (the number of utterances spanned by the quadruple components) and report the model's performance at each distance level. This approach allows us to examine how effectively the model captures long-range dependencies and discourse-level structures.

4.4 Baselines

We compared our proposal against several state-of-the-art methods, including the generic methods and approaches specifically tailored for the DiaASQ task.

ChatGPT: ChatGPT-3.5-turbo is a large language model built on GPT-3 (Brown et al., 2020). The experimental findings used for comparison were reported by (Zhou et al., 2024) and (Huang et al., 2024).

DiaASQ (Li et al., 2023): designed to perform end-to-end extraction of sentiment quadruples in dialogue settings.

H2DT (Li et al., 2024a): enhancing discourse feature modeling through a heterogeneous token-level graph and a triadic scorer that strengthens the cohesion of related tokens, improving quadruple extraction performance.

DMCA (Li et al., 2024b): introducing a multi-scale context aggregation strategy using dialogue windows and dynamic hierarchical fusion to better capture inter-utterance dependencies.

DMIN (Huang et al., 2024): integrating token-level and utterance-level features using a thread-based structure and a multi-granular integration module, aiming for a more holistic understanding of dialogue context. These baselines represent a range of paradigms and design philosophies for DiaASQ, offering a comprehensive benchmark for comparison

Speaker	Uttrance	ID	Type	Gold Label [P, T, A, O]	DMIN	SGCD
0	Taking pictures sucks	1	Cross	["neg","Xiaomi", "Taking pictures", "sucks"]	✗	["neg", "Xiaomi", "pictures", "sucks"]
1	100 million pixel super wide - angle may have, about others you can guess [laugh cry]	2	Cross	["pos","neo5", "camera", "reputation"]	["pos", "neo5", "reputation", "good"]	["pos", "neo5", "reputation", "good"]
0	But it's useless, software optimization sucks	3	Cross	["pos","Xiaomi 11", "camera", "beats"]	✗	["pos","Xiaomi 11","reputation","beats"
1	Optimization is Xiaomi's weakness [allow sadness]	4	Cross	["neg","Xiaomi", "software optimization", "sucks"]	✗	✓
3	The price of neo5 is 500 more expensive than k40, and it is normal to take better pictures.	5	Intra	["pos","neo5", "pictures", "better"]	✓	✓
4	With all due respect, is there any phone with a good camera that costs about 2,000 yuan? [doge]	6	Intra	["neg", "Xiaomi","Optimization", "weakness"]	✗	✓
0	Friends business neo5 has a good reputation					
5	Xiaomi 11 beats it [doge]					

Figure 3: A case for the quadruples extracted from a complex dialogue. Colors represent target, aspect, and opinion.

with our method.

4.5 Main Results

In the case described in Figure 3, we compared the proposed method and DMIN on a complex 3-thread dialogue with cross/intra-utterance quadruples. DMIN only predicted one quadruple, struggling with complex structures. Our method extracted two intra- and one cross-quadruples, despite minor deviations like "Taking Pictures" → "pictures". Its capability of relationship modeling handles diverse structures (short/long-distance cross, multi-thread interference), validating its superiority in complex dialogues.

Tables 1 and 2 present the results of our model on the DiaASQ datasets, demonstrating its superior performance compared to the state-of-the-art models across key metrics and dialogue structure dimensions.

(1) For the core quadruple extraction, our model outperforms the second-best approach in both Micro and Identification F_1 scores. Specifically, it achieves improvements of approximately 2.03% and 1.61% on the ZH dataset and 1.72% and 0.4% on the EN dataset. This increase indicates powerful representation and modeling capabilities for boundary detection and element matching in complete sentiment quadruples.

(2) Our model outperforms the second-best approach in Intra-Utterance and Inter-Utterance setups, improving 1.8% to 5.4% on the ZH and EN datasets, validating its robustness in predicting quadruple relationships in conventional dialogue structures.

(3) In the Cross-Utterance (Cross-1) scenario, our model gains 8.23% for ZH and 5.43% for EN over DMIN, demonstrating effective dialogue context modeling in complex scenarios. While some models perform slightly better in A-O (P) pairwise relationships, their weaker performance in Inter-Utterance and Cross-Utterance cases confirms our model's strength in handling complex multi-utterance dependencies.

While some models slightly exceed SGCD in A-O (P) relationships, they lag significantly in Inter- and Cross-Utterances. This suggests our model prioritizes learning complex, multi-utterance relationships over simple pairwise associations.

Our method delivers superior performance compared to LLM-based baselines. It surpasses DMIN in overall quadruple extraction accuracy, structural adaptability, and complex context perception, showcasing broad applicability and strong generalization in dialogue sentiment analysis tasks.

The significance analysis between SGCD and the baseline demonstrated that SGCD improved the performance with statistical significance. Please refer to Appendix B for further details.

4.6 Ablation Study

To determine the role of each module in model structure and performance improvement, we conducted ablation experiments on overall performance and Cross-Utterance metrics. Tables 3 and 4 list the results.

Impact of Modules on Overall Performance: We first evaluated the contribution of each module to the three primary DiaASQ subtasks. Removing the Bias Reduction module leads to significant declines in Cross-Utterance performance (5.71% in ZH and 2.3% in EN), validating the critical role of the causal debiasing mechanism in enhancing the model's ability to model complex cross-context relationships.

Further ablation of SGEncoder and DSFusion causes all three primary metrics to drop by over 1.4%, indicating their substantial contributions to deep semantic feature learning. The neg-

Model	ZH Dataset					EN Dataset				
	T-A (P)	T-O (P)	A-O (P)	Micro (Q)	Ident. (Q)	T-A (P)	T-O (P)	A-O (P)	Micro (Q)	Ident. (Q)
ChatGPT zero-shot (Huang et al., 2024)	23.86	10.55	15.81	13.77	18.15	23.26	16.07	14.34	10.98	12.99
ChatGPT one-shot (Huang et al., 2024)	29.90	17.48	25.59	18.26	20.56	26.18	20.33	21.20	13.20	14.67
ChatGPT 5-shot ICL (Zhou et al., 2024)	34.98	42.48	27.43	20.59	18.41	28.76	37.24	25.36	17.17	15.26
DiaASQ (Li et al., 2023)	48.61	43.31	45.44	34.94	37.51	47.91	45.58	44.27	33.31	36.8
H2DT (Li et al., 2024a)	50.48	48.8	52.4	40.34	42.44	48.69	48.84	**52.47**	39.01	43.92
DMCA (Li et al., 2024b)	56.88	51.7	52.8	42.68	43.56	53.08	50.99	52.4	37.96	41
DMIN (Huang et al., 2024)	57.62	51.65	56.16	44.49	47.50	53.49	52.66	52.09	39.22	42.31
SGCD	**59.44**	**52.55**	**57.65**	**46.52**	**49.11**	**53.82**	**52.85**	52.12	**40.94**	**44.37**

Table 1: Comparison of the F_1 scores (%) of our method against baseline modelst. The symbols T, A, and O represent Target, Aspect, and Opinion, respectively.

Model	ZH Dataset			EN Dataset		
	Intra	Inter	Cross	Intra	Inter	Cross
DiaASQ	37.95	23.21	*29.9*	37.65	15.76	*23.47*
H2DT	*43.2*	*25.55*	N.A.	*42.52*	*19.15*	N.A.
DMCA	*44.22*	*30.88*	*29.56*	*37.97*	*20.97*	*21.28*
DMIN	47.4	31.69	31.23	41.62	22.09	25.56
SGCD	**49.26**	**34.74**	**39.46**	**44.12**	**27.43**	**30.99**

Table 2: Comparison of F_1 (%) for different methods on subtasks. Italicized numbers indicate results obtained by reproducing the models, as these results were unavailable in the original publications.

	ZH Dataset			EN Dataset		
	Micro (Q)	Ident.(Q)	Cross	Micro (Q)	Ident.(Q)	Cross
complete model	46.52	49.11	39.46	41.18	43.93	27.97
w/o SGEncoder	44.71	47.33	38.03	38.83	42.28	21.58
w/o GCNs	45.49	47.06	40.99	40.36	41.88	26.25
w/o Guided Features	45.9	48.24	36.6	40.67	43.12	27.13
w/o DSFusion	44.51	46.47	36.36	39.43	42.07	26.57
w/o DKSelector	45.62	47.79	36.24	40.57	42.81	24.97
w/o SAF	43.85	46.38	33.42	39.04	41.63	24.86
w/o Bias Reduction	44.74	48.02	33.75	40.04	43.83	25.67
w/o Token IPW	45.94	49.34	37.91	40.76	43.62	26.74
w/o Global IPW	45.01	47.16	35.9	40.23	43.39	25.72

Table 3: Comparison of the F_1 scores (%) for different model settings.

ative impact on Quadruple F_1 shows their importance in modeling long-range cross-utterance relationships.

Evaluating the SAF module, we observe drastic drops in Micro-F_1, Identification-F_1, and Cross-Utt metrics (with maximum drops exceeding 6%), demonstrating the module's indispensability in fusing diverse features and focusing on task-relevant semantics.

Specific Contributions of Modules Across Tasks: To analyze the fine-grained contributions, we took three key tasks, i.e., entity recognition, entity pairing, and complete quadruple prediction, and reported the changes in average F_1 in Table 4.

Experimental results showed that for the relatively simple entity recognition task (ent), structural modeling could slightly interfere with performance. Removing individual modules led to minor improvements, indicating that single-entity extraction has a low dependency on complex architectures.

Guided features significantly enhanced performance in the entity pairing task (pair), demonstrating their effectiveness in directing the model to identify semantically related entity pairs within structural contexts.

The DSFusion and the dynamic top-k selection proved critical to the quadruple extraction (quad). Their removal caused significant performance declines (with the largest drop exceeding 2%), indicating their role in modeling complex, long-range, and cross-utterance relationships.

Analysis of Causal Debiasing: We further investigated the specific impacts of the two-level IPW causal debiasing strategy: Global-IPW and Token-IPW. As shown in Figure 4, we recorded the performance of using single debiasing strategies on complex tasks and plotted their F1 score ratios relative to the complete model.

Results indicate that Global-IPW primarily enhances Cross-Utterance extraction capabilities, particularly in the EN dataset, reflecting its advantages in global structural distribution adjustment. Token-IPW, meanwhile, positively impacts all tasks (Intra-Utt, Inter-Utt, Cross-Utt), demonstrating that fine-grained adjustment of token-level training contributions improves modeling stability.

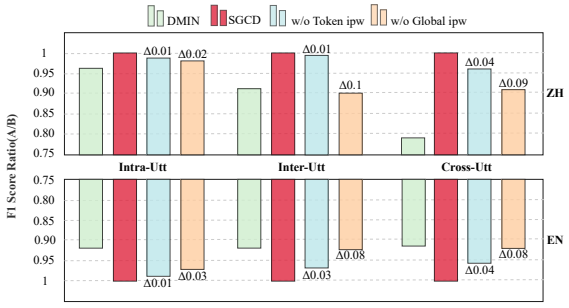


Figure 4: Quadruple extraction scores on Intra-Utt., Inter-Utt., and Cross-Utt., reported as F_1 ratios relative to the SGCD model (B model). Δx represents the gap compared to the our SGCD method.

	Ent	Pair	Quad
Complete model	77.15	56.55	43.82
w/o guided features	77.05	55.62	42.9
w/o DSFusion	77.12	55.94	41.23
w/o DKSelector	77.34	56.31	42.05

Table 4: Comparison of the average F_1 scores (%) for different model settings.

4.7 Statistical Significance Analysis

To determine if the performance improvement of the SGCD model compared to existing methods is statistically significant, we conducted a systematic significance test. We established the null hypothesis (H_0) as follows: the performance of the SGCD method is not superior to that of the comparison models. Conversely, the alternative hypothesis (H_1) states that the performance of the SGCD method is superior to that of the comparison models. This test was carried out by analyzing the prediction results from end-to-end models.

In the evaluation process, we observed that while the models produced consistent rankings at the dialogue level, the order of their internal prediction results (such as triplets, t-a, etc.) appeared random. To address this issue, we used the ground truth as a ranking benchmark to align the items in the prediction results that could be accurately or partially matched. Specifically, for each predicted triplet, we extracted seven-dimensional matching values, which included six position indices corresponding to the target, aspect, and opinion, as well as the sentiment polarity. These values formed the basis for determining the accuracy of the predictions.

Using this alignment strategy, we developed evaluation metrics for significance analysis. Specifically, we applied two non-parametric methods, the one-sided paired t-test and the Wilcoxon signed-rank test, to assess the performance differences between SGCD and each of the baseline models. The results of the test are presented in Table 5, indicating that H_0 should be rejected.

Model	DMIN	DMCA	H2DT	DiaASQ
paired t-test	0.047	0.036	0.018	0.013
Wilcoxon	6.00×10^{-4}	4.28×10^{-6}	5.48×10^{-7}	1.36×10^{-7}

Table 5: Results of p-values for statistical significance tests of the SGCD model compared to baselines.

5 Conclusion

This study presents the SGCD framework to address the challenges of cross-utterance quadruple extraction, varying effectiveness of tokens in utterances, and non-negligible focal biases in DiaASQ. Previous research typically relies on subtask loss weights to regulate shared deep feature learning across subtasks. In contrast, we introduce a subtask-guided feature selector to direct the modeling of deep features at the token and utterance levels shared among subtasks. We then dynamically select the features according to each utterance's need to generate utterance-level representations. For complex cross-utterance interactions, subtask-specific attention fusion flexibly aggregates token-level and utterance-level features based on subtask-specific needs, producing prediction features tailored to each subtask. Finally, causal debiasing reduces interference from confounding variables (speaker and thread) at both the dataset and sample levels, enhancing the model's adaptability to complex scenarios. Experimental results on Chinese and English DiaASQ benchmark datasets fully validate the effectiveness and robustness of the proposed SGCD method.

Limitations

Despite the proposed SGCD framework achieving superior performance on the DiaASQ dataset, several limitations deserve further investigation. First, the causal debiasing modeling remains incomplete. Beyond the explicit structural biases of speaker and thread already considered, biases arise from uneven distributions of intra- and cross-utterance quadruples in the dataset and unobservable latent factors in dialogues. The latent factors include temporal semantic shifts like evolving language trends and popular discourse patterns, which are challenging to model and incorporate into the causal graph. Second, for constructing subtask-guided features, we need to explore more focused and explicit task-specific feature representations tailored to individual subtasks beyond mere PLM layer-wise features.

Our future work will address these gaps through improvements such as more comprehensive causal structure modeling for dialogues, more precise task-specific feature design, and dataset structure cleaning/enhancement. We expect these will further promote the model's generalization and cross-domain adaptability.

References

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. 33:1877–1901.

Chenran Cai, Qin Zhao, Ruifeng Xu, and Bing Qin. 2023. Improving conversational aspect-based sentiment quadruple analysis with overall modeling. *Lecture Notes in Computer Science*, pages 149–161.

Ariel Caticha and Roland Preuss. 2004. Maximum entropy and bayesian data analysis: Entropic prior distributions. *Physical Review E*, 70.

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

Nicholas C. Chesnaye, Vianda S. Stel, Giovanni Tripepi, Friedo W. Dekker, Edouard L. Fu, Carmine Zoccali, and Kitty J. Jager. 2022. An introduction to inverse probability of treatment weighting in observational research. *Clinical Kidney Journal*, 15:14–20.

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*, pages 1–1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North*, 1.

Junpeng Fang, Gongduo Zhang, Qing Cui, Caizhi Tang, Lihong Gu, Longfei Li, Jinjie Gu, and Jun Zhou. 2024. Backdoor adjustment via group adaptation for debiased coupon recommendations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:11944–11952.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, Brandon M Stewart, Victor Veitch, and Diyi Yang. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition.

Yan Cathy Hua, Paul Denny, Jörg Wicker, and Katerina Taskova. 2024. A systematic review of aspect-based sentiment analysis: domains, methods, and trends. *Artificial Intelligence Review*, 57.

Peijie Huang, Xisheng Xiao, Yuhong Xu, and Jiawei Chen. 2024. Dmin: A discourse-specific multi-granularity integration network for conversational aspect-based sentiment quadruple analysis. *Findings of the Association for Computational Linguistics ACL 2024*, pages 16326–16338.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language?

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324.

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. pages 13449–13467.

Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024a. Harnessing holistic discourse features and triadic interaction for sentiment quadruple extraction in dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:18462–18470.

Qimai Li, Zhichao Han, and Xiao-ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.

Yuqing Li, Wenyuan Zhang, Binbin Li, Siyu Jia, Zisen Qi, and Xingbang Tan. 2024b. Dynamic multi-scale context aggregation for conversational aspect-based sentiment quadruple analysis. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11241–11245.

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.

Xianlong Luo, Meng Yang, and Yihao Wang. 2024. Overcome noise and bias: Segmentation-aided multi-granularity denoising and debiasing for enhanced quarduples extraction in dialogue.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2020. Counterfactual vqa: A cause-effect look at language bias.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis.

Shubham Singh. 2025. Social media users — how many people use social media in 2022.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. *Proceedings of the 2019 Conference of the North*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. 2020. Disentangled graph collaborative filtering. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jialong Wu, Linhai Zhang, Deyu Zhou, and Guoqiang Xu. 2024. Diner: Debiasing aspect-based sentiment analysis with multi-variable causal inference.

Zhen Wu, Chengcan Ying, Fang Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. *arXiv (Cornell University)*.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. 2019a. Understanding and improving layer normalization.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019b. How powerful are graph neural networks? *arXiv:1810.00826 [cs, stat]*.

Dingkang Yang, Mingcheng Li, Dongling Xiao, Yang Liu, Kun Yang, Zhaoyu Chen, Yuzheng Wang, Peng Zhai, Ke Li, and Lihua Zhang. 2024. Towards multimodal sentiment analysis debiasing via bias purification.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Heng Zhang, Lingpeng Zhong, Hua Li, Yunling He, and Qin Hang. 2024. Multi-granularity short circuit fault diagnosis for tokamak power converters under small sample conditions and noise interference. *Fusion Engineering and Design*, 203:114475.

Hua Zhang, Zeqi Chen, Bi Chen, Biao Hu, Mian Li, Cheng Yang, and Bo Jiang. 2022a. Complete quadruple extraction using a two-stage neural model for aspect-based sentiment analysis. *Neurocomputing*, 492:452–463.

Zheng Zhang, Zili Zhou, and Yanna Wang. 2022b. Ssegcn: Syntactic and semantic enhanced graph convolutional network for aspect-based sentiment analysis.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2023. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods.

Changzhi Zhou, Zhijing Wu, Dandan Song, Linmei Hu, Yuhang Tian, and Jing Xu. 2024. Span-pair interaction and tagging for dialogue-level aspect-based sentiment quadruple analysis. page 3995–4005.