TAG: Dialogue Summarization Based on Topic Segmentation and Graph Structures

Yatian Shen[1*] **Qichao Hao**[1] **Guosong Deng**[1] **Songyang Wang**[1] **Eryan Zhang**[2]
[1] School of Computer and Information Engineering, Henan University, Zhengzhou 450046, China
[2] School of Information Engineering and Big Data, Zhengzhou Technical College, Zhengzhou 450044, China
{ytshen, haoqichao, guosongdeng, sywang}@henu.edu.cn
zhangeryan@zzy.edu.cn

Abstract

In recent years, dialogue summarization has emerged as a rapidly growing area of research in natural language processing. Dialogue summarization is challenging due to dispersed key information, redundant expressions, ambiguous topic identification, and difficult content selection. To address these challenges, we propose an innovative approach to dialogue summarization that integrates topic segmentation and graph-structured modeling. Specifically, we first perform topic segmentation of the dialogue through clustering and quantify the key information in each utterance, thereby capturing the dialogue topics more effectively. Then, a redundancy graph and a keyword graph are constructed to suppress redundant information and extract key content, thereby enhancing the conciseness and coherence of the summary. Evaluations were conducted on the DialogSum, SAMSum, CSDS, and NaturalConv datasets. The experimental results demonstrate that the proposed method significantly outperforms existing benchmark models in terms of summary accuracy and information coverage. The Rouge-1 scores achieved were 48.03%, 53.75%, 60.78%, and 81.48%, respectively, validating its effectiveness in the dialogue summarization task. Our code is available at https://anonymous.4open.science/r/TAG-E64A.

Keywords: Dialogue Summarization , Topic Segmentation , Redundancy Graph , Keyword Graph.

1 Introduction

With the rapid development of the internet and social media, the volume of textual information has been growing exponentially. As an important research direction in natural language processing, dialogue summarization has gained widespread attention in recent years. However, dialogue summarization is more challenging than traditional document summarization due to its dynamic and highly interactive nature(Adilazuarda et al., 2024; Purwarianti et al., 2025). Dialogue data is typically long and often contains redundant content from multiple speakers, making redundancy elimination a key issue(Zhong et al., 2021). Moreover, the structure and linguistic style of dialogue are highly variable, and generating coherent summaries that capture the core topics remains an open problem.

To address these challenges, existing dialogue summarization methods often leverage external linguistic tools such as keyword extraction or discourse parsing to build pre-computed graphs representing inter-utterance relationships(Tang et al., 2023; Zhao et al., 2020). These graphs go beyond the linear sequence of dialogues by connecting distant semantically related utterances, allowing the model to capture non-sequential information and enabling cross-turn reasoning(Hua et al., 2023). Despite their effectiveness, current graph-based methods suffer from two key limitations. First, they rely heavily on external tools trained on limited domains or fixed rules, which lack robustness in handling open-domain dialogues with informal expressions and complex pragmatics, resulting in semantic deviations and error propagation(Huang et al., 2023; Chen et al., 2022). Second, the graph construction process is typically decoupled

* Corresponding Author

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1108-1121, Jinan, China, August 11-14, 2025.

from the downstream summarization task, lacking adaptability to context and failing to model dynamic interactions effectively(Rennard et al., 2024; Park and Lee, 2022).

Recent studies have attempted to alleviate these issues by integrating static and dynamic graphs (Tang et al., 2023), however, they still fall short in redundancy modeling and topic guidance capabilities. In this paper, we enhance graph-based dialogue summarization by introducing a topic segmentation module and redundancy-aware static graph structures. The topic segmentation module captures semantic transitions and topic shifts across utterances, guiding the model to focus on essential topical clues. Meanwhile, the redundancy graph explicitly models semantic overlap between utterances to help filter out redundant or irrelevant information, improving the accuracy, conciseness, and coherence of the generated summaries. As shown in Figure 1, we take a dialogue-summary paired sample from the DialogSum dataset as an example to intuitively illustrate the key processes of the dialogue preprocessing module in this study, including topic segmentation, redundant utterance detection, and keyword extraction.

Our main contributions are as follows:

- Topic segmentation is performed using clustering methods to identify different themes within the dialogue and extract the core utterances from each theme.
- A graph structure is constructed based on the semantic similarity between utterances to explicitly model redundant information, assisting the decoder in identifying repetitive and irrelevant content during summary generation.
- A keyword graph structure is introduced to connect semantically related keyword nodes, guiding the model to focus on the core semantic regions of the dialogue.

dialogue	Person1: You only have an hour for lunch? Person2: No, now I only have 45 minutes. Person1: That's not enough. Where are we going? Person2: We can go to a place near the mall. Person1: Oh, alright, let's go across the street. We can eat at Tony's Italian restaurant. I love their pizza. Person2: I love their food, too. But they are really slow. Last week I waited 30 minutes for my food. Person1: OK. Let's have sushi at Dave's. We can be in and out in 20 minutes. Person2: Today is Thursday, Dave's isn't open. Person1: Oh, right. Then, let's go to the Jungle Cafe. We can be there in 60 seconds. Person2: Great idea.	Topic 1 : brown Topic 2 : blue Redundancy : green keywords : red
summary	Person1 and Person2 talk about where to have lunch. Person2 only has 45 minutes and they decide to the Jungle Cafe.	

Figure 1: A sample from the DialogSum dataset. Topic segments are shown in different colors, redundant utterances are marked in green, and keywords are highlighted in red.

2 Related Work

2.1 Dialogue Summarization

Dialogue summarization aims to extract accurate and coherent key information from multi-turn conversations involving multiple speakers. Unlike traditional document summarization, dialogue text is typically unstructured, highly redundant, and prone to rapid topic shifts, creating greater challenges for summarization. Significant efforts have been made across domains to construct high-quality datasets covering diverse scenarios. (Feng et al., 2020) and (Chen and Yang, 2020) explore integrating various types of semantic information into summarization models to enhance performance. While these methods have shown promising results, real-world dialogues still present numerous challenges, especially in accurately identifying core topics(Wang et al., 2025; Belwal et al., 2023), suppressing redundancy(Rahman and Borah, 2021), and maintaining coherence across multi-turn, multi-speaker conversations. (Liang et

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1108–1121, Jinan, China, August 11–14, 2025.

al., 2023) introduce topic segmentation to capture cross-turn semantic connections and improve summarization performance. (Feng et al., 2021) leverage DialoGPT to extract keywords, topics, and redundant utterances, incorporating various information types in summary generation. While such methods improve the reasoning capabilities of models to some extent, challenges such as redundant content and topic transitions in dialogue remain insufficiently addressed.

2.2 Graph Neural Networks

In recent years, Graph Neural Networks (GNNs) have gained significant attention in Natural Language Processing due to their ability to represent graph-structured data across tasks such as social network modeling(Mitra and Paul, 2025), sequence labeling(Ezquerro et al., 2024), relation classification(Lei and Huang, 2024), and text generation(Yang et al., 2024). In the domain of dialogue summarization, modeling dialogue structures using graph-based approaches has become increasingly popular. Early traditional methods constructed sentence-level graphs via cosine similarity and selected representative utterances using graph-based ranking algorithms such as LexRank(Erkan and Radev, 2004) and TextRank(Mihalcea and Tarau, 2004). Others utilized discourse-based relations to form approximated Abstract Discourse Graphs (ADG)(Yasunaga et al., 2017) or Rhetorical Structure Theory (RST) graphs(Xu et al., 2019). However, these approaches often rely on external tools, leading to error propagation due to fixed rule-based processing. SDDS attempts to capture semantic relationships dynamically with dynamic graphs, thereby mitigating the limitations of static structures. However, it does not explicitly consider topic shifts or redundancy within the dialogue.

3 Task Definition

Given a dialogue $D = \{x_1, x_2, \ldots, x_{L_d}\}$ consisting of L_d utterances, where the i-th utterance is denoted as $x_i = \{w_{i,1}, w_{i,2}, \ldots, w_{i,L_i^u}\}$ and contains L_i^u words. Each utterance x_i is spoken by a speaker s_i, and there are $|S|$ unique speakers in total. The goal is to generate a summary $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{L_y}\}$ of length L_y, which covers the key information of the dialogue while maintaining semantic coherence and concise content. During training, the objective is to minimize the loss between the generated summary $\hat{Y}$ and the reference summary Y, thereby improving the quality of the generated summary. The task can be formally defined as learning a mapping function $f : D \rightarrow \hat{Y}$, which generates an informative and coherent summary based on the input dialogue D.

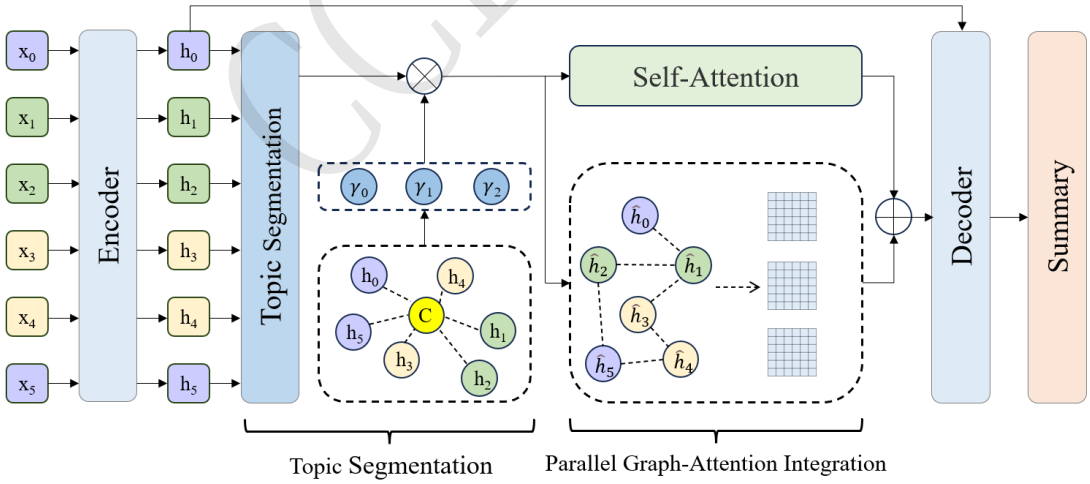

Figure 2: The main structure of our proposed model. Here, x_i denotes the utterances in a sample; the purple, green, and brown colors represent different topics to which the utterances belong. h_i is the encoded representation of each utterance. The yellow C indicates the utterance-level centrality vector, γ denotes the topic salience score, and $\hat{h}_i$ is the final weighted utterance representation.

4 Method

In this section, we introduce our dialogue summarization model based on Topic Segmentation and Graph structures (TAG). The model architecture is illustrated in Figure 2, which consists of four key components: (1) encoder; (2) topic segmentation; (3) parallel graph-attention integration; (4) summary generation.

4.1 Encoder

To acquire semantic representations of individual utterances, we utilize the pre-trained language model BART (Lewis et al., 2019a) to encode each utterance separately. Specifically, let the i-th utterance be composed of L_i tokens, denoted as:

$$\mathrm{x_i} = [\mathrm{w_i^{(1)}}, \mathrm{w_i^{(2)}}, \ldots, \mathrm{w_i^{(L_i)}}] \tag{1}$$

A special start-of-sequence token `[CLS]` is prepended to each utterance and the sequence is then fed into the BART encoder to obtain contextualized token embeddings h_i:

$$\mathrm{h_i} = \mathrm{BART_{Encoder}}([\mathrm{CLS}] \oplus \mathrm{x_i}) \tag{2}$$

Here, h_i refers to the hidden state of the `[CLS]` token, which is used as the semantic representation of the corresponding utterance. As a result, the entire dialogue is represented as a set of embeddings at the utterance level: $\mathrm{H} = \mathrm{h_1}, \mathrm{h_2}, \ldots, \mathrm{h_{L_d}}$. During the pre-processing stage, the dialogue is segmented into utterances in their original order and encoded sequentially.

4.2 Topic Segmentation

The goal of topic segmentation is to identify latent topics in a conversation and assign utterance weights based on Degree Centrality Theory(Freeman and others, 2002) to guide downstream summarization tasks. The overall process consists of three stages: semantic clustering, topic strength and utterance weighting, and regulation and enhancement.

4.2.1 Semantic Clustering

For all utterances H in a given conversation, the K-Means algorithm is applied for clustering, with the number of clusters set as $K = \min(3, \lfloor n/5 \rfloor)$ based on the conversation length. The cluster centers C are computed as:

$$\mathrm{C} = [\mathrm{c_1}, \ldots, \mathrm{c_K}] \tag{3}$$

where $\mathrm{c_k} = \frac{1}{|\mathrm{C_k}|} \sum_{i \in \mathrm{C_k}} \mathrm{h_i}$, representing the center of the k-th cluster. The cluster assignment for utterance $\mathrm{h_i}$ is obtained by minimizing the Euclidean distance:

$$\mathrm{z_i} = \arg\min_k \|\mathrm{h_i} - \mathrm{c_k}\|_2^2 \tag{4}$$

4.2.2 Topic Strength and Utterance Weighting

We treat the cluster center set $\mathcal{C}$ as a collection of topic representations in the semantic space, rather than constructing an explicit topic graph. To estimate the importance of each topic, we borrow the idea of degree centrality and compute the overall similarity between each center and all others:

$$\alpha_k = \frac{\sum_j \mathbf{c}_k^\top \mathbf{c}_j}{\left\| \sum_j \mathbf{c}_k^\top \mathbf{c}_j \right\|_2} \tag{5}$$

This "centrality" is used purely as a scalar importance score for each topic cluster. We do not construct a learnable graph or apply any message-passing mechanism in this step, thus avoiding additional graph modeling complexity.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1108–1121, Jinan, China, August 11–14, 2025.

This global strength measures the influence of the k-th topic in the semantic space. To compute the importance of each utterance within its cluster, an utterance weight β_i is introduced and normalized:

$$\beta_i = \frac{\sum_{j \in \mathrm{C}_{z_i}} \mathrm{h_i}^\top \mathrm{h_j}}{\| \sum_{j \in \mathrm{C}_{z_i}} \mathrm{h_i}^\top \mathrm{h_j} \|_2} \tag{6}$$

This is combined with the corresponding cluster's topic strength to form a joint centrality score: $\gamma_i = \alpha_{z_i} \cdot \beta_i$. This score comprehensively reflects the utterance's importance in both the topic distribution and within its cluster, ultimately used as a semantic enhancement weight.

4.2.3 Regulation and Enhancement

The joint weights are used to regulate low-level utterance representations, generating new utterance representations $\hat{\mathrm{h}}_i$ via linear interpolation:

$$\hat{\mathrm{h}}_i = \lambda \cdot \gamma_i \cdot \mathrm{h_i} + (1 - \lambda) \cdot \mathrm{h_i} = [1 + \lambda(\gamma_i - 1)] \cdot \mathrm{h_i} \tag{7}$$

where $\lambda = 0.5$ if $n > 15$, otherwise $\lambda = 0.3$. When the number of utterances n in the conversation is less than 10, the clustering step is skipped, and $\gamma_i = \beta_i$ is used directly. The enhanced conversation representation $\hat{\mathrm{H}} = [\hat{\mathrm{h}}_1, \ldots, \hat{\mathrm{h}}_n]$ is obtained, serving as input to the subsequent summarization module with stronger topic expressiveness and structural guidance.

4.3 Parallel Graph-Attention Integration

4.3.1 Redundancy Graph

Dialogues often contain redundant utterances, which negatively impact the compression quality and information density of summaries. To enhance the model's redundancy detection ability, TAG employs an external tool, DialoGPT(Zhang et al., 2019b), to identify redundant utterances while capturing non-critical information in the dialogue using a redundancy graph.

The DialoGPT model is used to encode each utterance in the dialogue, obtaining vector representations for each utterance. Suppose a sample dialogue consists of n utterances, denoted as $\mathrm{Z} = \{\mathrm{z_1}, \mathrm{z_2}, \ldots, \mathrm{z_n}\}$. These n utterances are concatenated as a continuous input sequence, with a special boundary token appended at the end of each utterance. The hidden layer states are extracted as the representation of each utterance, resulting in a set of utterance vectors: $\mathrm{V} = \{\mathrm{v_1}, \mathrm{v_2}, \ldots, \mathrm{v_n}\}$.

A semantic similarity graph is constructed within the dialogue. A recursive backtracking mechanism is adopted, traversing from the last utterance $\mathrm{z_n}$ backward to the second utterance $\mathrm{z_2}$. For the current utterance $\mathrm{z_i}$, its cosine similarity with all preceding utterances $\{\mathrm{z_1}, \ldots, \mathrm{z_{i-1}}\}$ is computed:

$$\mathrm{sim}(\mathrm{z_i}, \mathrm{z_j}) = \frac{\mathrm{v_i}^\top \mathrm{v_j}}{\|\mathrm{v_i}\| \cdot \|\mathrm{v_j}\|}, \quad \text{for all } j < i \tag{8}$$

This generates an asymmetric similarity matrix $\mathrm{S} \in \mathbb{R}^{n \times n}$, retaining valid similarity values in the upper triangular region, as each utterance is compared only with prior utterances. A redundancy threshold $\theta \in (0, 1)$ is set. If the maximum similarity between the i-th utterance $\mathrm{z_i}$ and its historical utterances is no less than θ, then there exists redundancy between $\mathrm{z_j}$ and $\mathrm{z_i}$, where $\mathrm{z_j}$ is the redundancy source and $\mathrm{z_i}$ is the redundant utterance. To ensure the uniqueness of redundancy edges, the index j^* that maximizes the similarity is recorded, and redundancy is attributed to utterance $\mathrm{z_i}$. Ultimately, a redundancy index set is constructed:

$$\mathrm{R} = \left\{ i \mid \exists i > j^*, \max_{j<i} \mathrm{sim}(\mathrm{z_i}, \mathrm{z_j}) \geq \theta \right\} \tag{9}$$

This set represents the collection of utterances that are repetitive with the dialogue history. In this paper, the threshold is set to $\theta = 0.99$, filtering out utterances with low semantic overlap to ensure that redundancy edges exhibit strong semantic repetition.

The original dialogue D, its corresponding summary Y, and the redundancy index R are combined to form a redundancy graph annotation triplet: $\mathrm{G_{RD}} = [\mathrm{D}, \mathrm{Y}, \{\text{“RD”} : \mathrm{R}\}]$. The redundancy graph serves as an auxiliary encoding input for the graph structure module, explicitly marking ignorable content in the dialogue to enhance the model's ability to model redundant utterances.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1108-1121, Jinan, China, August 11-14, 2025.

4.3.2 Keyword Graph

The keyword graph captures semantic consistency and key information across utterances in a dialogue, improving the model's ability to detect semantic cues across utterances. In contrast to the redundancy graph, the keyword graph focuses on identifying semantically similar utterances with consistent topics, constructing a graph structure using keyword co-occurrence information.

Similar to the keyword extraction method in GPT-Anno(Feng et al., 2021), we use DialoGPT for part-of-speech tagging of utterances. The approach mirrors that of the redundancy graph: unpredictable words are selected as keywords, with the correspondence between keyword groups and original utterances preserved in their sequential order.

The original dialogue D, its corresponding summary Y, and the keyword labels K are combined to form a keyword graph annotation triplet: $\mathrm{G_{kw}} = [\mathrm{D}, \mathrm{Y}, \{\text{“keywords”} : \mathrm{K}\}]$, where K represents the set of keyword groups extracted for each utterance. The keyword graph serves as an input to the graph attention mechanism, working in parallel with the redundancy graph to model both redundant information and global semantic consistency in the dialogue.

4.3.3 Attention Mechanism

To capture the semantic relationships between utterances, an attention mechanism is employed to compute the relationship representations between each pair of utterances. This module draws inspiration from the graph attention structure used for utterance modeling in SDDS, calculating the attention weight matrix A through the following formula:

$$\mathrm{Q} = \hat{\mathrm{H}}\mathrm{W_Q}, \quad \mathrm{K} = \hat{\mathrm{H}}\mathrm{W_K}, \quad \mathrm{A} = \frac{\mathrm{QK}^\top}{\sqrt{d_k}} \tag{10}$$

$\hat{\mathrm{H}}$ represents the utterance representations for each turn, $\mathrm{W_Q}$ and $\mathrm{W_K}$ are trainable parameters, and A denotes the attention weight matrix derived based on semantic relationships.

4.3.4 Graph-Attention Integration

This model extends the existing graph fusion framework SDDS(Gao et al., 2023), incorporating multi-source dialogue information into the attention mechanism. The original design features a semantic relation graph, a keyword graph, a positional relation graph, and a speaker graph. However, dialogues often have redundant information. Thus, we introduce a redundancy graph to better capture structural clues in the dialogue. For the adjacency matrix constructed based on redundancy, the construction process is

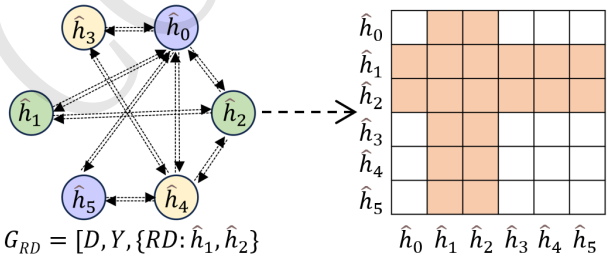

Figure 3: Construction of the redundancy graph, where $\hat{h}_1$ and $\hat{h}_2$ are identified as redundant sentences by DialoGPT. The orange regions in the matrix are set to "$-\infty$" to sever the connections between redundant and other sentences, while the white regions are set to "1" to strengthen the relations among non-redundant sentences.

illustrated in Figure 3. These matrices are normalized separately. The normalized matrices are then concatenated along the channel dimension and fused with the relation matrix A, yielding a fused structural relation representation:

$$\mathrm{G} = \mathrm{A} \oplus \mathrm{Conv}\left(\mathrm{softmax}\left(\mathrm{A_{RD}}\right) \oplus \mathrm{softmax}\left(\mathrm{A_{KW}}\right)\right) \tag{11}$$

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1108–1121, Jinan, China, August 11–14, 2025.

A_{RD} and A_{KW} represent the relation matrices constructed based on redundancy and keyword heuristics, respectively, $\oplus$ denotes concatenation along the channel dimension, and the Conv operation integrates dialogue information from different sources.

The model normalizes the fused matrix G and uses it as attention scores to weight the utterance representations, producing a graph-guided semantic representation H^{graph}. To further enhance the model's representation capability, the standard attention output is concatenated with the graph-guided semantic output, resulting in the final fused representation:

$$H^{graph} = \text{softmax}(G)\hat{H}W_V, H^{fused} = [H^{attn} \parallel H^{graph}] \tag{12}$$

This fusion approach enables the model to leverage key information in the dialogue while modeling semantic information, providing richer input features for summary generation.

4.4 Summary Generation

The decoder generates the target summary sequence $\{y_1, y_2, \ldots, y_T\}$ in an autoregressive manner, predicting one word at a time. During training, we use the standard cross-entropy loss(Sutskever et al., 2014) to maximize the consistency between the generated summary and the ground-truth reference:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}) \tag{13}$$

y_t denotes the t-th target word, and T is the length of the target summary.

5 Experimental Setup

5.1 Datasets

We evaluate our model on four widely used dialogue summarization datasets: DialogSum, SAMSum, NaturalConv, and CSDS. Details of the four datasets are summarized in Table 1. **DialogSum**(Chen et al., 2021) is a large-scale English dataset of face-to-face spoken dialogues on various everyday topics, including school, work, medication, shopping, leisure, and travel. Most conversations occur between friends, colleagues, or service providers and customers. **SAMSum**(Gliwa et al., 2019) is an English dialogue summarization dataset annotated by linguists. It features informal conversations, including chit-chat between friends, gossip, meeting arrangements, political discussions, and academic consultations among colleagues. **NaturalConv**(Duan and Lu, 2025) is a multi-turn, topic-driven Chinese dialogue dataset spanning domains such as sports, entertainment, technology, gaming, education, and health. The dialogues are grounded in specific scenarios, which enhances their contextual realism. **CSDS**(Lin et al., 2021) is a Chinese dataset specifically for customer service dialogue summarization. It was created by the NLP&CC team at the Institute of Automation, Chinese Academy of Sciences.

Ethical and Privacy Statement: All four dialogue summarization datasets used in this study were either anonymized by the original authors prior to release or synthetically constructed, and thus do not involve any real users' personal or sensitive information. Therefore, this research does not pose additional ethical or privacy risks. Furthermore, we strictly adhere to ethical standards in academic research on data usage, and no new human data collection or human subject experiments were conducted in this work.

	Number of samples	Avg_Dia	Avg_Tur	Avg_sum
DialogSum	13460	121.56	9.50	22.64
SAMSum	16369	120.24	11.11	22.79
NaturalConv	19917	244.8	20.1	110.7
CSDS	10701	399.46	25.93	83.21

Table 1: Avg_Dia is the average number of words per dialogue, Avg_Tur is the average number of turns per dialogue, and Avg_sum is the average number of words in the summary.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1108–1121, Jinan, China, August 11–14, 2025.

5.2 Baseline Methods

We compare our proposed **TAG** model with several representative baselines to evaluate its performance. **Longest-3**(Gliwa et al., 2019) is a commonly used extractive method in both news and dialogue summarization tasks. **BART**(Lewis et al., 2019a) and **UniLM**(Bao et al., 2020) are powerful pretrained language models designed for abstractive summarization. **Transformer**(Vaswani et al., 2017) adopts a purely attention-based architecture for sequence-to-sequence tasks. **GPT-Anno**(Feng et al., 2021) uses DialoGPT to annotate topics, keywords, and redundancy to guide the summary generation process. **SDDS**(Gao et al., 2023) fuses static and dynamic graph information to enhance dialogue summarization performance. **GLC**(Liang et al., 2023) segments topics via clustering. **PGN**(See et al., 2017) enables the copying of important tokens and the generation of new words. **OmniVec2**(Srivastava and Sharma, 2024) proposes a modality-switching pretraining strategy to unify heterogeneous input spaces. **CriSPO 3-shot**(He et al., 2025) introduces a critique-suggestion-based prompt optimization method for few-shot settings. **SICK**(Kim et al., 2022) leverages external commonsense reasoning to select plausible inferences based on similarity. **ChatGPT**(Qin et al., 2023) is well-regarded for its general-purpose capabilities across diverse tasks; however, in the absence of prompt constraints, it is prone to generating summaries that contain hallucinations and redundant content.

	DialogSum				SAMSum			
Model	Rouge-1	Rouge-2	Rouge-L	BS	Rouge-1	Rouge-2	Rouge-L	BS
TextRank	21.19	6.49	23.91	–	15.70	2.87	10.63	–
Longest-3	24.15	6.25	22.73	–	32.46	10.27	29.92	–
UniLM	47.04	21.13	45.04	69.40	–	–	–	–
PGN	33.77	9.24	32.18	70.21	32.27	14.42	34.36	80.67
GPT-Anno	47.12	20.88	44.56	–	53.70	28.79	50.81	90.04
SDDS	47.96*	21.68*	47.87*	91.25*	53.34*	28.64*	54.66*	92.08*
SICK	46.26	20.95	41.05	71.32	53.73	28.81	49.50	71.92
OmniVec2	47.60	22.10	41.40	72.80	–	–	–	–
CriSPO 3-shot	–	–	–	–	47.20	20.80	38.20	91.30
ChatGPT	38.40	12.90	29.80	86.69*	32.70	12.30	24.70	32.50
TAG (ours)	**48.03**	**22.68**	**49.19**	**91.33**	**53.75**	**29.01**	**55.25**	**92.13**

Table 2: Experimental results of various models on the DialogSum and SAMSum datasets. Results marked with * are re-evaluated by us.

	NaturalConv			CSDS		
Model	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
PGN	52.28	43.96	48.62	59.00	58.68	58.23
BART	79.66*	65.31*	79.74*	60.11	59.86	58.75
GLC	80.84*	66.89*	80.97*	60.32	61.03	59.02
TAG (ours)	**81.48**	**68.31**	**82.08**	**60.78**	**61.58**	**60.72**

Table 3: Performance of various models on the NaturalConv and CSDS datasets. Scores marked with * indicate results re-evaluated by us.

6 Experimental Results

6.1 Main Results

We evaluate the generated summaries using standard ROUGE metrics—Rouge-1, Rouge-2, and Rouge-L F1 scores(Lin, 2004)—which assess the overlap of unigrams, bigrams, and the longest common subsequence between the generated and reference summaries,and compute the semantic similarity score BS (BERTScore)(Zhang et al., 2019a) between the generated summary and the reference summary. Table 2

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1108-1121, Jinan, China, August 11-14, 2025.

presents the results from the DialogSum dataset. Our TAG model generates summaries with ROUGE scores of 48.03%, 22.69%, and 47.87%, demonstrating significant improvements over baseline models. This indicates that the topic segmentation and redundancy graph-based approach effectively integrates human prior knowledge, leading to better identification of key content in dialogues. It strikes a good balance between conciseness and semantic coherence. The TAG model also demonstrates strong generalization on the SAMSum dataset. Compared to the state-of-the-art dialogue summarization models GPT-Anno and SDDS, TAG improves performance, suggesting that modeling topic segmentation and redundancy structures enhances the model's semantic understanding.

The TAG model achieves state-of-the-art results on the Chinese datasets NaturalConv and CSDS, as shown in Table 3. On the NaturalConv dataset, which involves frequent topic shifts and long multi-turn dialogues, the TAG model effectively reduces redundant content interference, identifies more information-dense utterances, and generates higher-quality summaries. Compared to the GPT-Anno model, which relies on heuristic graph construction, and the pretrained language model UniLM, the TAG model better integrates structural and contextual information, improving the accuracy and coverage of generated summaries, while enhancing the model's adaptability across various scenarios and languages. Figure 4 illustrates the dimensionality reduction and visualization of high-dimensional utterance representations using spherical 3D t-SNE(Van der Maaten and Hinton, 2008),The resulting visualization intuitively reveals the clustering distribution of utterances in the semantic space, highlighting clear separability between different topics and strong cohesion within individual topics.

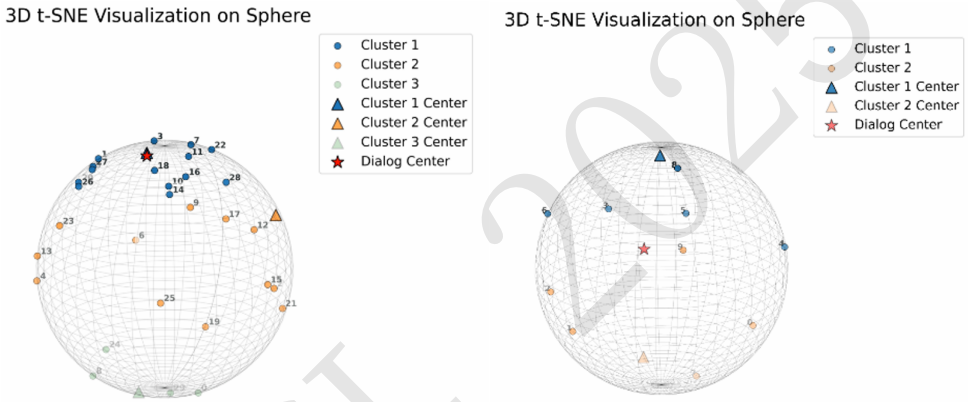

Figure 4: Different colors represent distinct topics identified within the dialogue, and each point corresponds to the embedding of a single utterance. Triangular markers denote the centroids of individual topic clusters, while the red pentagram indicates the global semantic center of the entire dialogue, reflecting the overall semantic orientation of the conversation.

6.2 Ablation Study

Model	Rouge-1	Rouge-2	Rouge-L
TAG w/o TS	46.60	21.85	48.76
TAG w/o RD	47.21	21.40	47.35
TAG w/o G	45.95	20.13	45.07
TAG	**48.03**	**22.68**	**49.19**

Table 4: Ablation study results on the DialogSum dataset show that removing any key component of the TAG model results in a noticeable performance drop.

We conduct an ablation study on the DialogSum dataset to evaluate the contribution of each TAG component. Specifically, we compare the full model with three variants: (1) **TAG w/o TS** (no topic segmentation), (2) **TAG w/o RD** (no redundancy graph), and (3) **TAG w/o G** (no graph module). As shown in Table 4, removing topic segmentation significantly degrades performance, underscoring its role in

capturing topic shifts. Excluding the redundancy graph also harms results, as it helps eliminate repetitive content. Removing all graph modules causes the largest drop, showing their importance in modeling long-range dependencies. Overall, both topic segmentation and redundancy modeling are essential for high-quality summaries.

6.3 Case Study

Figure 5 presents a comparative analysis of the summaries from three models on the same dialogue instance, highlighting differences in information coverage and semantic consistency. Although the SDDS model identifies the main events, it misses specific interpersonal details, such as congratulations, appreciation, and celebration. The ChatGPT summary shows significant improvement in content coverage, capturing the entire event trajectory—including expressions of congratulation, surprise, gratitude, admiration, and celebration. However, its summary is verbose, lacking conciseness and failing to highlight key points precisely. In contrast, our TAG model performs better in dialogue summarization. By incorporating topic segmentation and graph-based representations for enhanced semantic understanding, the TAG model generates concise and informative summaries, balancing brevity and content richness.

Case Study	
Dialogue	*Person1: Tom, I've got good news for you.* *Person2: What is it?* *Person1: Haven't you heard that your novel has won The Nobel Prize?* *Person2: Really? I can't believe it. It's like a dream come true. I never expected that I would win The Nobel Prize!* *Person1: You did a good job. I'm extremely proud of you.* *Person2: Thanks for the compliment.* *Person1: You certainly deserve it. Let's celebrate!*
Reference	*Person1 congratulates Tom for achieving the Nobel Prize.*
SDDS	*Person1 tells Tom that his novel has won The Nobel Prize. Tom is excited.*
ChatGPT	*Person1 congratulates Person2 on winning the Nobel Prize for their novel. Person2 is shocked and overjoyed, expressing disbelief and gratitude. Person1 praises Person2's achievement and suggests celebrating together.*
TAG(ours)	*Person1 congratulates Tom that his novel has won The Nobel Prize. Person1 is proud of him.*

Figure 5: A comparative example of summaries generated by different models. The red text denotes the reference summary, while the green text indicates the summary generated by our model.

7 Conclusion

In this paper, We propose a model that combines topic segmentation and graph structure, leveraging the autoregressive BART model for automatic dialogue summarization. Clustering methods are applied to segment the dialogue into topics and analyze subtopic structures. We also introduce redundancy and keyword graphs to capture redundancy relations and key information between utterances, enhancing the conciseness and coverage of the summaries. Experimental results demonstrate the superiority of our approach in conciseness, informational completeness, and factual consistency, validating the effectiveness of topic structure and graph modeling. In the future, we aim to incorporate dynamic structure modeling or more powerful pre-trained models to enhance the model's generalization ability and semantic expression.

Acknowledgements

We appreciate the anonymous reviewers for their insightful comments. This research was partially supported by Henan Province Science and Technology Development Plan Project under the following grants: Research on disease prediction based on heterogeneous graph neural network models (Project No. 242102210065), Construction of dynamic medical knowledge graph based on electronic medical records (Project No. 242102210093).

References

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, Genta Indra Winata, Ayu Purwarianti, and Alham Fikri Aji. 2024. LinguAlchemy: Fusing typological and geographical elements for unseen language generalization. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3912–3928, Miami, Florida, USA, November. Association for Computational Linguistics.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International conference on machine learning*, pages 642–652. PMLR.

Ramesh Chandra Belwal, Sawan Rai, and Atul Gupta. 2023. Extractive text summarization using clustering-based topic modeling. *Soft Computing*, 27(7):3965–3982.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. *arXiv preprint arXiv:2010.01672*.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

Jiaao Chen, Mohan Dodda, and Diyi Yang. 2022. Human-in-the-loop abstractive dialogue summarization. *arXiv preprint arXiv:2212.09750*.

Jiaxin Duan and Fengyu Lu. 2025. DialogES: A Large Dataset for Generating Dialogue Events and Summaries. https://github.com/Lafitte1573/NLCorpora/tree/main/DialogES.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Ana Ezquerro, David Vilares, and Carlos Gómez-Rodríguez. 2024. Dependency graph parsing as sequence labeling. *arXiv preprint arXiv:2410.17972*.

Xiachong Feng, Xiaocheng Feng, Bing Qin, Xinwei Geng, and Ting Liu. 2020. Dialogue discourse-aware graph convolutional networks for abstractive meeting summarization. *arXiv preprint arXiv:2012.03502*, 17.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring dialogpt for dialogue summarization. *arXiv preprint arXiv:2105.12544*.

Linton C Freeman et al. 2002. Centrality in social networks: Conceptual clarification. *Social network: critical concepts in sociology. Londres: Routledge*, 1(3):238–263.

Shen Gao, Xin Cheng, Mingzhe Li, Xiuying Chen, Jinpeng Li, Dongyan Zhao, and Rui Yan. 2023. Dialogue summarization with static-dynamic structure fusion graph. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13858–13873.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Han He, Qianchu Liu, Lei Xu, Chaitanya Shivade, Yi Zhang, Sundararajan Srinivasan, and Katrin Kirchhoff. 2025. Crispo: Multi-aspect critique-suggestion-guided automatic prompt optimization for text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24014–24022.

Yilun Hua, Zhaoyuan Deng, and Kathleen McKeown. 2023. Improving long dialogue summarization with semantic graph representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13851–13883.

Kung-Hsiang Huang, Siffi Singh, Xiaofei Ma, Wei Xiao, Feng Nan, Nicholas Dingwall, William Yang Wang, and Kathleen McKeown. 2023. Swing: Balancing coverage and faithfulness for dialogue summarization. *arXiv preprint arXiv:2301.10483*.

Seungone Kim, Se June Joo, Hyungjoo Chae, Chaehyeong Kim, Seung-won Hwang, and Jinyoung Yeo. 2022. Mind the gap! injecting commonsense knowledge for abstractive dialogue summarization. *arXiv preprint arXiv:2209.00930*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yuanyuan Lei and Ruihong Huang. 2024. Sentence-level media bias analysis with event relation graph. *arXiv preprint arXiv:2404.01722*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019b. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Xinnian Liang, Shuangzhi Wu, Chenhao Cui, Jiaqi Bai, Chao Bian, and Zhoujun Li. 2023. Enhancing dialogue summarization with topic-aware global-and local-level centrality. *arXiv preprint arXiv:2301.12376*.

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. Csds: A fine-grained chinese dataset for customer service dialogue summarization. *arXiv preprint arXiv:2108.13139*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Anirban Mitra and Subrata Paul. 2025. Analyzing social networks with dynamic graphs: Unravelling the ever-evolving connections. In *Applied Graph Data Science*, pages 195–214. Elsevier.

Seongmin Park and Jihwa Lee. 2022. Unsupervised abstractive dialogue summarization with word graphs and pov conversion. *arXiv preprint arXiv:2205.13108*.

Ayu Purwarianti, Dea Adhista, Agung Baptiso, Miftahul Mahfuzh, Yusrina Sabila, Aulia Adila, Samuel Cahyawijaya, and Alham Fikri Aji. 2025. Nusadialogue: Dialogue summarization and generation for underrepresented and extremely low-resource languages. In *Proceedings of the Second Workshop in South East Asian Language Processing*, pages 82–100.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Nazreena Rahman and Bhogeswar Borah. 2021. Redundancy removal method for multi-document query-based text summarization. In *2021 International Symposium on Electrical, Electronics and Information Engineering*, pages 568–574.

Virgile Rennard, Guokan Shang, Michalis Vazirgiannis, and Julie Hunter. 2024. Leveraging discourse structure for extractive meeting summarization. *arXiv preprint arXiv:2405.11055*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Siddharth Srivastava and Gaurav Sharma. 2024. Omnivec2-a novel transformer based network for large scale multimodal and multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 27412–27424.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin. 2023. Enhancing dialogue generation via dynamic graph knowledge aggregation. *arXiv preprint arXiv:2306.16195*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

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.

Huiyao Wang, Peifeng Li, Yaxin Fan, and Qiaoming Zhu. 2025. Simulating dual-process thinking in dialogue topic shift detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2592–2602.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Discourse-aware neural extractive text summarization. *arXiv preprint arXiv:1910.14142*.

Yizhe Yang, Heyan Huang, Yang Gao, and Jiawei Li. 2024. Building knowledge-grounded dialogue systems with graph-based semantic modelling. *Knowledge-Based Systems*, 298:111943.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. *arXiv preprint arXiv:1706.06681*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019b. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Lulu Zhao, Weiran Xu, and Jun Guo. 2020. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th international conference on computational linguistics*, pages 437–449.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. *arXiv preprint arXiv:2104.05938*.

A Implementation Details

For English datasets, we use **BART-large**(Lewis et al., 2019b) with a maximum generation length of 100, beam size of 5, and train for 5 epochs using the **Adam**(Kingma and Ba, 2014) optimizer (learning rate 2e-5, batch size 1). For Chinese datasets, we adopt **BART-base-Chinese**(Shao et al., 2021) with a generation length of 150 and beam size of 4, keeping other settings consistent.

Experiments on DialogSum with an NVIDIA A4000 GPU (16GB) show that baseline BART training (5 epochs) takes 9.6 hours (2.12 samples/s, 9.8,GB memory). Adding redundancy and keyword graphs slightly increases memory (10.1,GB) with negligible time overhead. Adding topic segmentation extends training to 11 hours (1.86 samples/s, 10.5,GB). The full TAG model takes 13.5 hours (1.52 samples/s, 10.8,GB). Overall, the computational overhead is moderate and acceptable.

B Sensitivity Analysis of Cluster Number

To investigate the impact of the number of clusters on model performance, we conducted a small-scale sensitivity analysis on the DialogSum dataset. Specifically, we varied the number of clusters $K = \min(\alpha, \lfloor n/5 \rfloor)$, where $\alpha \in \{1, 2, 3, 4, 5\}$ and n denotes the number of utterances in the dialogue. We evaluated the model under each cluster setting using four metrics: Rouge-1, Rouge-2, Rouge-L, and BERTScore. The results are illustrated in the line charts shown in Figure 6. The experimental results indicate that both excessively small and large values of K can lead to performance degradation. This suggests that an appropriate number of topic clusters is beneficial for capturing the underlying thematic structure of dialogues: too many clusters may introduce noise, while too few may overlook fine-grained topic shifts.

C Redundancy Threshold Hyper-parameter Tuning

Table 5 reports the hyper-parameter tuning results for the redundancy similarity threshold θ on the DialogSum dataset, with values ranging from 0.95 to 0.99. Based on the observed performance, we set $\theta = 0.99$ to achieve an optimal balance between precision and recall in redundancy detection.

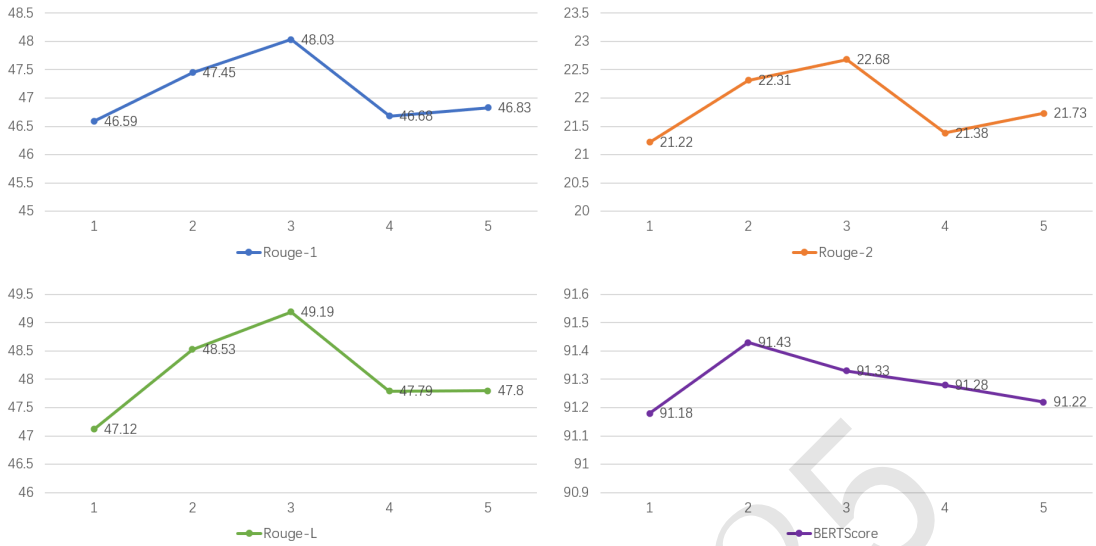

Figure 6: Model performance on the DialogSum dataset under different cluster settings, where $\alpha \in \{1, 2, 3, 4, 5\}$ in $K = \min(\alpha, \lfloor n/5 \rfloor)$. The x-axis denotes the value of α, and the y-axis represents the evaluation scores for Rouge-1, Rouge-2, Rouge-L, and BERTScore.

θ	**0.95**	**0.96**	**0.97**	**0.98**	**0.99**
Rouge-1	47.10	47.85	47.55	47.35	**48.03**
Rouge-2	21.14	22.31	21.75	21.48	**22.68**
Rouge-L	46.88	48.50	48.16	47.85	**49.19**

Table 5: Hyper-parameter tuning results for the redundancy similarity threshold θ on the DialogSum dataset.