A Unified Supervised and Unsupervised Dialogue Topic Segmentation Framework Based on Utterance Pair Modeling

Shihao Yang[1], Ziyi Zhang[1], Yue Jiang[1], Chunsheng Qin[2], Shuhua Liu[1,*],

[1] School of Information Science and Technology, Northeast Normal University
[2]Academy for Research in teacher Education, Northeast Normal University
{yangsh861, ziyizhang, jiangyue, qincs, liush129}@nenu.edu.cn
Correspondence: liush129@nenu.edu.cn

Abstract

The Dialogue Topic Segmentation task aims to divide a dialogue into different topic paragraphs in order to better understand the structure and content of the dialogue. Due to the short sentences, serious references and non-standard language in the dialogue, it is difficult to determine the boundaries of the topic. Although the unsupervised approaches based on LLMs performs well, it is still difficult to surpass the supervised methods based on classical models in specific domains. To this end, this paper proposes UPS (Utterance Pair Segment), a dialogue topic segmentation method based on utterance pair relationship modeling, unifying the supervised and unsupervised network architectures. For supervised pre-training, the model predicts the adjacency and topic affiliation of utterances in dialogues. For unsupervised pre-training, the dialogue-level and utterance-level relationship prediction tasks are used to train the model. The pre-training and fine-tuning strategies are carried out in different scenarios, such as supervised, few-shot, and unsupervised data. By adding a domain adapter and a task adapter to the Transformer, the model learns in the pre-training and fine-tuning stages, respectively, which significantly improves the segmentation effect. As the result, the proposed method has achieved the best results on multiple benchmark datasets across various scenarios.

1 Introduction

The Dialogue Topic Segmentation is an important task in Natural Language Processing (NLP), which helps to better understand the structure and content of a dialogue by dividing it into multiple different “topics". It can be applied to a variety of application scenarios such as dialogue generation (Li et al., 2016), dialogue summarization (Liu et al., 2019b), and knowledge selection (Yang et al., 2022).

Dialogue topic segmentation methods are divided into supervised and unsupervised types. Supervised methods typically model topic segmentation as an utterance sequence labeling problem (Barrow et al., 2020; Arnold et al., 2019). It predicts whether each sentence is beginning of a topic paragraph. This approach typically requires a large amount of labeled data to train the model and may perform better when dealing with specific domains or specific types of dialogue. Large Language Models(LLMs) have performed well on many general-purpose tasks in recent years, but there is still a gap with supervised learning methods using small language models. Recent research has demonstrated the greater potential and advantages of using small language models for supervised learning in dialogue topic segmentation tasks (Fan et al., 2024). Therefore supervised learning remains the optimal choice when high-quality labeled data is available. However, in specialized domains lacking such data, supervised methods are not applicable. Unsupervised methods (Xu et al., 2021) do not rely on labeled datasets and train coherent models to evaluate the similarity of continuous discourse, and then employ global segmentation algorithms to compute topic segmentation points. Unsupervised methods do not directly use segmentation annotations as supervisory signals but instead learn the similarity of continuous utterances to indirectly identify topic boundaries. Although unsupervised methods (Xing and Carenini, 2021) (Wang et al., 2017) can address the lack of training data in certain domains, but, due to the limitations of their network structures, existing unsupervised methods cannot utilize segmentation annotations as supervisory signals, making it difficult to extend them to broader data scenarios such as supervised and few-shot learning.

“Sentence pair” relationship prediction studies how to recognize and understand the relationship between two sentences. The goal of topic segmentation is to divide the text into smaller topics. Understanding the relationships between utterances

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4898–4908
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

can help to more accurately determine the boundaries of these segments. To this end, this paper proposes a method of dialogue topic segmentation based on “utterance pair” relation modeling, which unifies supervised and unsupervised network architectures.At the same time, we propose both supervised and unsupervised pre-training strategies. In the supervised pre-training strategy, utterance pairs are categorized into four relationships based on whether they are adjacent and whether they belong to the same topic. Unsupervised pre-training utilizes two utterance pair relation prediction tasks: dialogue-level (Same Dialogue Prediction, SDP) and utterance-level (Next Sentence Prediction, NSP). These two pre-training strategies serve two purposes: enabling the model to adapt to the dataset’s domain and acquiring the ability to predict utterance pair relations.

The main contributions of this paper are listed as follows:

(1) This study proposes a unified framework for supervised and unsupervised dialogue topic segmentation. This framework offers excellent scalability as it can utilize any pre-trained language model with an encoder-decoder structure as its backbone.

(2) We introduce two pre-training strategies, supervised and unsupervised, according to the amount of labeled data available. These strategies enable the model to acquire domain adaptation capabilities for the corresponding datasets and enhance its ability to predict utterance pair relationships.

(3) The proposed method demonstrates superior performance in supervised, few-shot, and unsupervised settings, achieving the state-of-the-art results. Experimental results validate the effectiveness of our proposed unified framework.

2 Related Work

Dialogue topic segmentation divides a dialogue into sections by identifying different topics within the dialogue, including both unsupervised (Park et al., 2023; Artemiev et al., 2024) and supervised methods. Previous studies have often adopted unsupervised methods due to the lack of dialogue topic segmentation datasets.Unsupervised methods generally included two steps: assessing topic similarity and determining segmentation boundaries. Popular unsupervised methods were TextTiling (Hearst, 1997) and its improvements (Song et al., 2016). Discourse similarity was often taken into account when assessing topic similarity. TextTiling (Hearst, 1997) measured discourse similarity based on patterns of lexical co-occurrence and distribution. (Xu et al., 2021) leveraged BERT model embeddings to enhance TextTiling, taking advantage of the rich semantic information contained in pre-trained models. (Xing and Carenini, 2021) proposed training a utterance pair coherence scoring model to measure topic relevance between utterances. (Gao et al., 2023) introduced DialSTART, which considered both utterance similarity and coherence.

Despite the advantages of unsupervised methods in terms of data cost, their performance was often difficult to match that of supervised methods (Jiang et al., 2023). (Xia et al., 2022) proposed the PEN-NS method, which utilized a parallel extraction network for segment extraction and optimized the bipartite matching cost to capture inter-segment dependencies. (Barrow et al., 2020) proposed Segment Pooling LSTM (S-LSTM) model for joint topic segmentation and labeling. (Arnold et al., 2019) proposed SECTOR, an end-to-end model that learned latent topic embeddings. Relevant experiments have shown that using the entire text as input can lead to a loss of coherence in dialogue. (Wang et al., 2017) demonstrated the importance of text pairs modeling for topic segmentation. They addressed the challenge of data scarcity by constructing a dataset with pairs labeled as internal paragraph or internal document. (Xing and Carenini, 2021) extended this to dialogue by building a corpus of consecutive and non-consecutive utterance pairs. (Zhou et al., 2022) proposed Dialogue Sentence Embedding (DSE), employing consecutive utterances within the same dialogue as positive pairs for contrastive learning.

In order to decrease the cost of fine-tuning pre-trained models to downstream tasks, prompt learning becomes a mainstream method. In topic segmentation tasks, when using LLMs for topic segmentation, prompt templates can be used to guide models such as GPT-3.5 (Fan et al., 2024) and GPT-4 (Hwang et al., 2024). Topic shift is a task that detects if the topic changes in a dialogue and has similarities with topic segmentation. (Lin et al., 2023) utilized label-level, topic-level, and turn-level prompts to enhance a model’s ability to understand and predict topic shift in a dialogue.

In summary, this paper proposes a topic segmentation method based on utterance pair relationship modeling, which unifies the supervised and

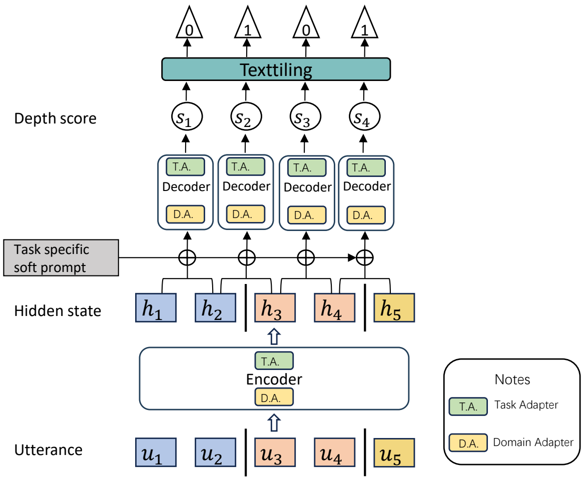


Figure 1: Framework of UPS. "u_i" denotes an utterance, and the same color indicates belonging to the same topic. For instance, the boundaries between u_2 and u_3, and between u_4 and u_5 should be topic segment points.

unsupervised learning frameworks. At the same time, prompt learning is used to fine-tuning the pre-trained model of the backbone network, making full use of the advantages of the above methods.

3 Methodology

3.1 Model Overview

This paper presents UPS (Utterance Pair Segment), a unified framework for both supervised and unsupervised topic segmentation. UPS adopts utterance pair modeling to predict whether two consecutive utterances constitute a topic boundary. By adopting different training strategies based on data annotation availability, UPS can be applied in supervised, few-shot, and unsupervised scenarios. The model architecture is shown in Figure 1. To retain the knowledge from the pre-trained model and facilitate knowledge transfer between pre-training and fine-tuning, domain and task adapters are added to the Transformer. Furthermore, to better adapt the pre-trained model to new tasks, different prompt templates are designed for different utterance relation prediction tasks. The model operates as follows: an encoder-decoder pre-trained model (e.g., BART, T5) is employed. The encoder first encodes the entire dialogue to obtain representations for each dialogue turn.

$$H = encoder\left(U\right) \tag{1}$$

Where $H = (h_1, h_2, \ldots, h_m)$ denotes the hidden states of the utterances. The decoder predicts the score s_i of the utterance pair based on the designed prompt template and the hidden states of the utterance pair. A higher score indicates higher coherence and a higher probability of belonging to the same topic.

$$s_i = decoder\left(prompt_{ts-task}, (h_i, h_{i+1})\right) \tag{2}$$

The prompt is learnable, and its parameters are initialized by the following template: $Prompt^{ts}$*=utterance1:(Text1)* `</s>` *utterance2:(Text2)*`</s>` *Should utterance1 and utterance2 be topic segmentation points?*[mask]

Replace "Text1" and "Text2" with the utterance pair to be predicted. "[mask]" represents the token to be predicted. The token-to-class verbalizer is defined as follows:*Verbalizer*={"0" : $[$"no", "$false$", "$negative$"$]$, "1" : $[$"yes", "$true$", "$positive$"$]$}. Class "0" and "1" include three tokens respectively. The definition of "Verbalizer" remains consistent throughout the following text. Finally, the TextTiling algorithm is used to identify topic boundaries. This algorithm analyzes the depth scores between adjacent utterances, and selects the score at valley points as segmentation points.

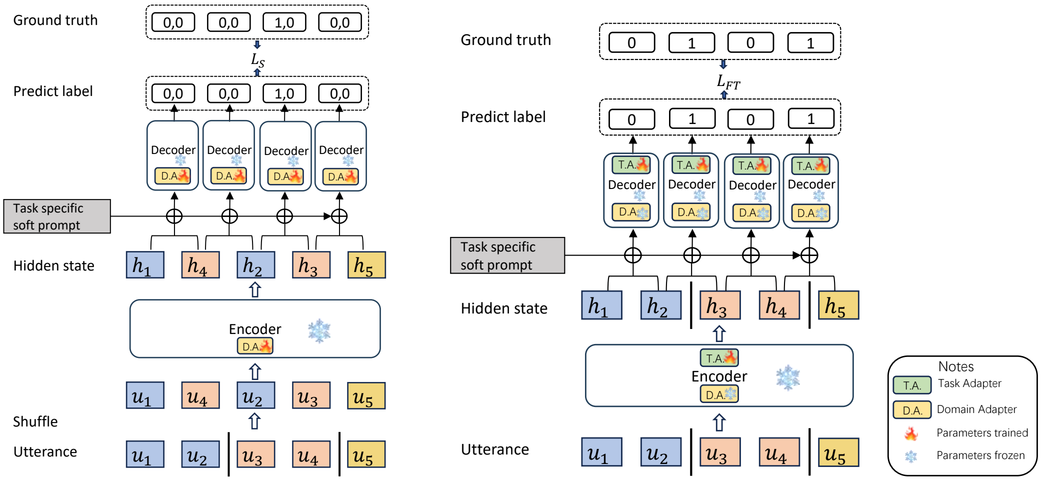
$$B = Texttiling\left(S\right) \tag{3}$$

Where $S = (s_1, s_2, \ldots, s_{m-1})$, and m is the number of utterances in a dialogue.

Within this unified framework, this paper proposes different pre-training and fine-tuning strategies to effectively and flexibly address varying amounts of labeled data, thereby enhancing the overall performance and practicality of the model. In supervised and few-shot scenarios, the model employs a two-stage training strategy: pre-training followed by fine-tuning. In unsupervised scenarios, the model is pre-trained solely on the task of predicting relationships between unlabeled utterance pairs. The training strategies for different data scenarios are elaborated below.

3.2 Supervised Setting

In a supervised setting with abundant labeled data, the model employs a two-stage training strategy: pre-training followed by fine-tuning. During pre-training, we propose a pre-training task called Supervised Utterance Pair Relation Prediction(SUPRP). First the input dialogue is shuffled to randomize the order of utterances. Utterance pairs are then categorized into four relationships based on adjacency and topic coherence: (1) adjacent and belonging to the same topic; (2) adjacent but



(a) Supervised Pre-training Stage

(b) Supervised Fine-tuning Stage

Figure 2: Utterance pair relation prediction model in supervised setting

belonging to different topics; (3) non-adjacent but belonging to the same topic; and (4) non-adjacent and belonging to different topics. Subsequently, utterances are encoded, and adjacent utterance as a pair, along with a prompt, are fed into a decoder to predict the relationship of utterance pair. This calculation process is detailed in equations (4) and (5).

$$H^s = encoder(shuffle(U)) \tag{4}$$

$$label^{nsp}_{i,i+1}, label^{stp}_{i,i+1} = Verbalizer(decoder (prompt_{nsp+stp}, (h^s_i, h^s_{i+1}))) \tag{5}$$

Where $H^s = (h^s_1, h^s_2, \ldots, h^s_m)$ denotes the hidden states of each utterance after shuffling. $label^{nsp}_{i,i+1}$ is the next sentence prediction(NSP) label, indicating whether h^s_{i+1} is the next utterance of h^s_i. $label^{stp}_{i,i+1}$ is the same topic prediction(STP) label, indicating whether h^s_i and h^s_{i+1} belong to the same topic. The prompt is learnable, and its parameters are initialized with the following template: $Prompt^{nsp+stp}$=*utterance1:(Text1)* </s> *utterance2:(Text2)*</s> *Should utterance2 be the next utterance of utterance1?*[mask] *Are these two utterances discussing the same topic?*[mask]

This approach combines the NSP and STP tasks using a single template, eliminating redundant encoding and accelerating training. [mask] denotes the token to be predicted.

Both tasks use Binary Cross Entropy Loss:

$$L_{nsp} = BCE(y^{nsp}_i, label^{nsp}_{i,i+1}) \tag{6}$$

$$L_{stp} = BCE(y^{stp}_i, label^{stp}_{i,i+1}) \tag{7}$$

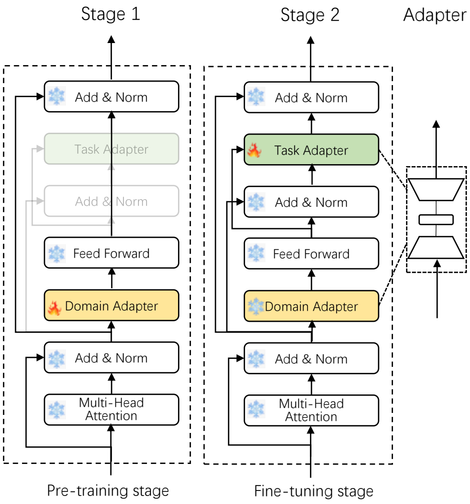


Figure 3: Two-stage training architecture. In the first stage, only the domain adapter is trainable, while other parameters are frozen. In the second stage, only the task adapter is trainable, while other parameters are frozen.

$$L_s = L_{nsp} + L_{stp} \tag{8}$$

The supervised pre-training loss is the sum of the losses from two subtasks.

The fine-tuning process for topic segmentation is illustrated in Figure 2(b). An encoder-decoder pre-trained model is employed. First, the encoder processes the entire dialogue to generate representation for each utterance. During decoding, consecutive utterance pairs are fed into the decoder. The decoder then predicts whether an utterance pair constitutes a topic boundary based on the prompt and the hidden states of the utterance pair.

$$H = encoder(U) \tag{9}$$

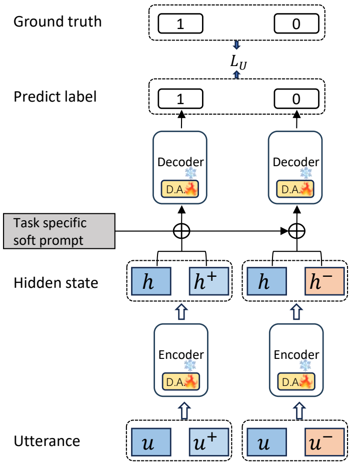


Figure 4: Unsupervised pre-training in few-shot and unsupervised setting

$$label^{ts}_{i,i+1} = Verbalizer(decoder(prompt_{ts-task}, (h_i, h_{i+1}))) \tag{10}$$

The fine-tuning loss is calculated using Binary Cross Entropy:

$$L_{ts} = BCE(y^{ts}_i, label^{ts}_{i,i+1}) \tag{11}$$

To facilitate knowledge transfer between pre-training and fine-tuning, this paper introduces two adapters to the model: a domain adapter and a task adapter, as illustrated in Figure 3. Adapters (Houlsby et al., 2019; Pfeiffer et al., 2020; Hu et al., 2021) are a parameter-efficient technique commonly used during model fine-tuning, primarily to enhance flexibility and reduce training costs. In this paper, in the first stage of training, only the domain adapter is trainable while other parameters are frozen. In the second stage of training, only the task adapter is trainable while other parameters are frozen. The adapter training in this paper is inspired by (Diao et al., 2023). However, unlike Diao et al., who used MLM and L2 loss to train the domain adapter, this paper pre-trains the domain adapter through different utterance pair relation prediction tasks. Compared to MLM and L2 loss, our method more directly enables the model to learn the semantic coherence of utterance pairs, which is more suitable for the topic segmentation task.

After the model training is complete, inference is performed according to the framework UPS shown in Figure 1.

3.3 Few-shot Setting

In this scenario, with abundant unlabeled data and limited labeled data, we propose a pre-training task called Unsupervised Utterance Pair Relation Prediction(UUPRP). UUPRP contains two self-supervised tasks: dialogue-level Same Dialogue Prediction (SDP) and utterance-level Next Sentence Prediction (NSP), as illustrated in Figure 4. These self-supervised tasks, not requiring segmentation point annotations, leverage the vast unlabeled data for pre-training, enabling the model to acquire domain adaptability. For sample selection: in SDP, positive sample u^+ and anchor sample u come from the same dialogue, while negative sample u^- and anchor sample u come from different dialogues. In NSP, positive sample u^+ is the next utterance of anchor sample u, and negative sample u^- is not. A sample pair is denoted as $(u, \bar{u})$, u denotes anchor utterance, $\bar{u}$ denotes positive/negative sample. The model first encodes $(u, \bar{u})$ to obtain hidden states, which are then fed into the decoder along with the prompt.

$$(h, \bar{h}) = encoder(u, \bar{u}) \tag{12}$$

$$label^{sdp/nsp} = Verbalizer(decoder(prompt_{sdp/nsp}, (h, \bar{h}))) \tag{13}$$

The prompt template is designed as follows: *prompt^{nsp}=utterance1:(Text1)* `</s>` *utterance2: (Text2)* `</s>` *Should utterance2 be the next utterance of utterance1?* [mask]. *prompt^{sdp}=utterance1:(Text1)*`</s>` *utterance2: (Text2)* `</s>` *Are these two utterances belonging to the same dialogue?* [mask]. The choice of which prompt to use depends on the task.

Loss is calculated as follows:

$$L_u = BCE(y^{sdp/nsp}_i, label^{sdp/nsp}_{i,i+1}) \tag{14}$$

The fine-tuning stage is same as in the supervised setting. The key difference is that the supervised setting utilizes the full training datasets for fine-tuning, while the few-shot setting employs only a limited number of examples. Figure 2(b) illustrates this fine-tuning process. In this scenario, the placement and learning methodology of both the domain adapter and the task adapter remain consistent with the supervised scenario, as depicted in Figure 3. The inference process is shown in Figure 1.

3.4 Unsupervised Setting

In unsupervised training, only unlabeled data is available. Therefore, only pre-training for unsupervised utterance pair relation prediction is performed, without fine-tuning. The pre-training process is the same as in the few-shot setting, as shown

Scenario	Pre-training	Fine-tuning	Inference
Supervised	SUPRP	TS	UPS
Few-shot	UUPRP	TS	UPS
Unsupervised	UUPRP	×	UPS

Table 1: Usage of different modules in various scenarios. SUPRP denotes Supervised Utterance Pair Relation Prediction. UUPRP denotes Unsupervised Utterance Pair Relation Prediction. TS denotes topic segmentation. "×" denotes that it's not executed.

Dataset	TIAGE	SuperDialseg	Doc2Dial	ZYS
words	188.4	200.5	186.4	740.0
language	English	English	English	Chinese
#sent/seg	3.7	2.3	2.3	3.9
#seg/doc	4.3	5.7	5.7	6.5

Table 2: Statistics of four datasets. #sent/seg is the number of sentences per segment, and #seg/doc is the number of segments per document.

in Figure 4. Two pre-training tasks, dialogue-level same dialogue prediction and utterance-level next sentence prediction, enable the model to acquire utterance pair relation prediction abilities at both dialogue and utterance levels. During inference, a topic segment prompt is used to guide the model in topic segmentation, as shown in Figure 1.

In summary, the UPS framework proposed in this paper unifies the inference processes for supervised, few-shot, and unsupervised learning. This unified framework can reduce the complexity of designing and implementing different models in various scenarios, making the development and maintenance of models more efficient. Additionally, we propose different training strategies based on the scale of the labeled data. We list the modules used in different scenarios in Table 1.

4 Experiments and Analysis

4.1 Datasets

We conducted experiments on four publicly available dialogue topic segmentation datasets: TIAGE (Xie et al., 2021), SuperDialseg (Jiang et al., 2023), Doc2Dial (Feng et al., 2020) and ZYS (Xu et al., 2021). Statistics of datasets are in Table 2. More details of datasets are in Appendix A.

4.2 Metrics

To evaluate model performance, we utilize the following metrics: (1) P_k error (Beeferman et al., 1999) and (2) WindowDiff (WD) (Pevzner and Hearst, 2002). These are the two most commonly used evaluation metrics, with lower scores indicating better performance. Detailed introduction in Appendix B. (3) F1 and macro F1 scores. Previous work has employed one of the two metrics. For consistency, we adopt F1 for supervised and few-shot settings, and macro F1 for the unsupervised setting. (4) $Score$. By considering the soft errors and the F1 score simultaneously, we use the following $Score$ metric for convenient comparison:

Datasets	TIAGE				SuperDialseg			
Metrics	$P_k\downarrow$	WD $\downarrow$	F1$\uparrow$	Score$\uparrow$	$P_k\downarrow$	WD $\downarrow$	F1$\uparrow$	Score$\uparrow$
BERT	41.8	43.5	12.4	34.9	21.4	22.5	72.5	75.3
RetroT5	28.0	31.7	57.6	63.9	22.7	23.7	73.3	75.1
RoBERTa	26.5	28.7	57.2	64.8	18.5	19.2	78.4	79.8
BART	27.3	28.9	57.3	64.6	18.8	19.4	78.3	79.6
T5	27.1	28.9	57.5	64.7	18.4	19.2	78.5	79.8
GPT-3.5	49.6	56.0	36.2	41.7	31.8	34.7	65.8	66.3
Ours$_{\text{BART}}$	25.8	28.5	59.0	65.9	17.8	18.5	79.8	80.8
Ours$_{\text{T5}}$	**25.6**	**28.1**	**59.1**	**66.1**	**17.7**	**18.5**	**80.2**	**81.1**

Table 3: Experimental results in the supervised setting.

$$Score = \frac{2 * F1 + (1 - P_k) + (1 - WD)}{4} \quad (15)$$

suggested by the ICASSP2023 General Meeting Understanding and Generation Challenge (MUG).

4.3 Experimental Setup

The experiment is conducted on a single NVIDIA RTX 4090 GPU, powered by an Intel i9-12900K CPU with 64GB of memory. We use T5-base for English dataset and mT5-base for Chinese dataset. During adapter-only optimization, the learning rate is set to 5e-4 with a batch size of 32. In ablation studies, where all parameters are optimized, a learning rate of 3e-5 and a batch size of 16 are used. Both pre-training and fine-tuning are performed for 50 epochs, with early stopping employed during fine-tuning.

4.4 Experimental Results

The paper experiments in three different data scenarios: supervised, few-shot, and unsupervised. The following is a detailed explanation of the experimental results.

Supervised Setting. We design several strong baselines based on pre-trained language models (PLMs), including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019a), BART (Lewis et al., 2020), T5 (Raffel et al., 2020), and Retrot-T5. BERT, BART, T5 and RoBERTa model topic segmentation as a sequence labeling task. Retrot-T5 is

Datasets	**TIAGE**				**SuperDialseg**			
Metrics	P_k ↓	WD ↓	F1↑	Score↑	P_k ↓	WD ↓	F1↑	Score↑
BERT	43.5(1.2)	45.0(1.3)	33.0(4.2)	44.4	39.2(2.1)	42.3(1.8)	35.6(2.7)	47.4
RetroT5	44.4(0.8)	45.6(1.2)	33.1(2.8)	44.1	41.2(1.0)	43.6(1.2)	33.0(2.5)	45.3
RoBERTa	42.4(1.6)	44.6(1.4)	35.2(7.3)	45.9	42.0(0.8)	43.1(1.0)	30.8(4.3)	44.1
BART	44.5(0.8)	47.3(1.3)	33.2(5.5)	43.7	42.3(1.3)	47.9(1.6)	31.0(3.1)	43.0
T5	42.3(0.6)	44.3(1.2)	35.5(1.9)	46.1	39.6(0.8)	42.0(0.7)	37.0(1.5)	48.1
GPT-3.5	49.6	56.0	36.2	41.7	31.8	34.7	65.8	66.3
Ours(T5)	**34.5(3.1)**	**36.2(3.8)**	**47.6(5.2)**	**56.1**	**27.2(3.2)**	**28.3(2.8)**	**73.4(5.5)**	**72.8**

Table 4: Experimental results in the few-shot setting. The quality of samples chosen in few-shot scenarios significantly impacts model performance, so we experiment with three different random seeds. The table shows the mean and standard deviation.

a generative model for dialogue segmentation proposed by (Xie et al., 2021). We also compare the performance of GPT-3.5 reported by (Jiang et al., 2023), who used a defined template as a prompt. This prompt consists of task instructions, dialogue input, and an output example specifying the output format.

We conduct experiments on the TIAGE and SuperDialseg datasets. The results are shown in Table 3. Our models are first pre-trained on a supervised relation prediction task and then fine-tuned on the topic segmentation task. As can be seen, our proposed method achieves the state-of-the-art (SOTA) results and significantly outperforms GPT-3.5 across all metrics. Compared with baseline models, our method comprehensively outperforms the best baseline model RoBERTa. There is also a large gap between GPT-3.5 and baseline models such as RoBERTa and Retro T5, indicating that the pre-training and then fine-tuning paradigm on the topic segmentation task is still superior to general-purpose large language models.

Few-shot setting. The model is first pre-trained on an unsupervised utterance relation prediction task and then fine-tuned with a few-shot approach. The baseline models used for comparison follow the same supervised setting. We randomly select 16 labeled samples for fine-tuning. Our method achieves the state-of-the-art results. The proposed method achieves better performance than GPT-3.5 (Fan et al., 2024), outperforming it by 14.4% and 6.5% on the *Score* metric for the two datasets, respectively. Although GPT-3.5's performance is based on zero-shot learning, the significant difference in parameter size (175B vs. 0.22B) cannot be ignored.

Datasets	**Doc2Dial**			**ZYS**		
Metrics	P_k ↓	WD ↓	macro F1 ↑	P_k ↓	WD ↓	macro F1 ↑
Random	55.6	65.3	42.0	52.8	67.7	39.8
GreedySeg^{2021}	50.7	51.6	40.6	44.1	48.3	50.2
TextTiling (TeT)^{1997}	52.0	57.4	53.9	45.9	49.3	48.5
TeT + Embedding^{2016}	53.7	55.7	60.2	43.9	45.1	51.0
TeT + CLS^{2021}	54.3	57.9	51.8	43.0	43.6	50.2
TeT + NSP^{2021}	50.8	54.9	55.0	42.6	44.0	50.0
Mutual Learning^{2024}	48.3	52.7	/	/	/	/
CohereSeg^{2021}	45.2	47.3	66.0	41.0	41.3	52.1
DynamicCOCO^{2023}	42.0	45.1	70.1	38.1	40.1	54.9
DialSTART^{2023}	38.1	40.7	/	/	/	/
Llama3.2-3B^{2024}	51.5	53.4	47.9	43.3	58.0	54.8
GPT-3.5^{2024}	47.4	49.3	53.6	56.2	58.2	49.1
Mistral-8B^{2024}	43.5	47.0	58.6	41.3	56.7	53.2
Ours	**35.1**	**36.5**	**78.4**	**36.0**	**36.8**	**55.9**

Table 5: Experimental results of unsupervised methods.

Datasets	**TIAGE**				**SuperDialseg**			
Metrics	P_k ↓	WD ↓	F1↑	Score↑	P_k ↓	WD ↓	F1↑	Score↑
T5(SL)	27.1	28.9	57.5	64.7	18.4	19.2	78.5	79.8
T5(UP)	27.8	29.3	56.3	63.9	19.5	20.2	77.8	79.0
+NSP	26.8	28.8	58.1	65.2	18.2	18.9	78.4	79.9
+STP	26.0	28.3	58.7	65.8	18.2	19.0	78.4	79.9
+NSP+STP	25.8	28.1	58.8	65.9	17.9	18.5	80.1	81.0
Ours	**25.6**	**28.1**	**59.1**	**66.1**	**17.7**	**18.5**	**80.2**	**81.1**

Table 6: Ablation Study under Supervised Setting. SL refers to Sequence Labeling, and UP refers to Utterance Pair. Ours denotes T5(UP)+NSP+STP+Adapter.

Unsupervised setting. In this setting, the model is only pre-trained on the unsupervised utterance relation prediction task, without a fine-tuning process. The pre-training stage is shown in Figure 4, and the inference process is illustrated in Figure 1. The final output of the model serves as the depth score for the utterance pair, and topic boundaries are calculated using Texttiling algorithm. We compare our approach with the following baseline methods under the unsupervised setting:

GreedySeg (Xu et al., 2021) minimized intra-segment similarity using a greedy approach and a threshold to avoid over-segmentation. CohereSeg (Xing and Carenini, 2021) segmented dialogue by capturing coherence. TextTiling (Hearst,

1997) segmented text by identifying subtopic shifts via lexical co-occurrence patterns. TeT + Embedding (Song et al., 2016), TeT + CLS (Xu et al., 2021) and TeT + NSP (Xing and Carenini, 2021) were extended from TextTiling. We compare our method with unsupervised methods from the past two years: DynamicCOCO (Pu and Wang, 2023), Mutual Learning (Xu et al., 2024), DialSTART (Gao et al., 2023). In addition to above unsupervised algorithms, this paper introduces a topic segmentation baseline based on GPT-3.5 (Fan et al., 2024) and a random segmentation baseline. (Fan et al., 2024) only provided the P$_k$ and macroF1 scores on the ZYS dataset. We run their publicly available code to obtain the remaining GPT-3.5's results. Our method achieves the state-of-the-art (SOTA) performance on both the English dataset Doc2Dial and the Chinese dataset ZYS. GPT-3.5 still does not outperform carefully designed unsupervised methods, although the gap is not as significant as in the supervised setting. Llama3.2-3B and Mistral-8B are the latest LLMs. UPS still outperforms them. It is worth noting that the latest Mistral-8B has 8 billion parameters, surpassing GPT-3.5, which has 175 billion parameters, demonstrating the advancements in large language models.

4.5 Ablation Study

We explore the impact of different modules in three scenarios. The experimental results and analysis are as follows.

In the supervised setting, as shown in Table 6, both NSP and STP pre-training tasks improve the model's performance, with the greatest improvement observed when using both tasks simultaneously. The $Score$ metric on the two datasets increased by 2% and 2.1%, respectively. Adding Adapter to +NSP+STP only yields a small improvement of 0.2% and 0.1%, suggesting that the knowledge retained by Adapter plays a limited role when abundant labeled data is available.

In the few-shot setting, as shown in Table 7, the proposed utterance pair modeling method, pre-training tasks, and adapter all contribute to improved model performance. Utterance pair modeling demonstrates a significant advantage over sequence labeling, boosting the $Score$ metric by 3.0% and 12.7% on the two datasets, respectively. This difference stems from the increased training samples: in utterance pair modeling, an utterance pair is a sample. But in sequence labeling, an entire dialogue is a sample. Consequently, with the same training set size, utterance pair modeling benefits from exponentially more samples, leading to significant performance improvements, especially in few-shot setting. However, SDP negatively impacts performance on the TIAGE dataset. This may be because there is no clear thematic distinction between the different dialogues in TIAGE.

Datasets	TIAGE				SuperDialseg			
Metrics	P$_k$ ↓	WD ↓	F1↑	Score↑	P$_k$ ↓	WD ↓	F1↑	Score↑
T5(SL)	42.3	44.3	35.5	46.1	39.6	42.0	37.0	48.1
T5(UP)	40.6	42.2	39.5	49.1	35.5	38.4	58.6	60.8
+NSP	38.3	40.0	42.3	51.6	34.0	36.6	63.2	64.0
+SDP	41.2	42.8	39.4	48.7	34.5	36.8	61.2	62.8
+NSP+SDP	43.8	46.2	33.4	44.2	32.0	34.2	65.2	66.1
Ours	**34.5**	**36.2**	**47.6**	**56.1**	**27.2**	**28.3**	**73.4**	**72.8**

Table 7: Ablation Experiments under Few-shot Setting. Ours denotes T5(UP)+NSP+SDP+Adapter.

Datasets	Doc2Dial				ZYS			
Metrics	P$_k$ ↓	WD ↓	macro F1 ↑	Score↑	P$_k$ ↓	WD ↓	macro F1 ↑	Score↑
T5(SL)	53.8	62.1	45.3	43.7	44.7	48.3	51.2	52.4
T5(UP)	46.6	56.4	44.4	46.5	40.9	42.0	51.6	55.1
+NSP	38.9	40.7	57.8	59.0	37.0	37.9	54.9	58.7
+SDP	47.6	51.2	57.3	54.0	37.7	38.5	55.8	58.9
+NSP+SDP	40.0	43.6	60.1	59.2	37.5	38.5	56.0	59.0
Ours	**35.1**	**36.5**	**78.4**	**71.3**	**36.0**	**36.8**	**55.9**	**59.8**

Table 8: Ablation Study in an unsupervised setting. Ours denotes T5(UP)+NSP+SDP+Adapter.

In the unsupervised setting, as shown in Table 8, pre-training tasks significantly improve performance on both datasets. The adapter shows greater improvement on Doc2Dial than on ZYS. ZYS includes specific text related to finance and banking. The original T5 contains less relevant knowledge, so the adapter's retention of the original model's knowledge has a minimal impact on performance.

5 Conclusion

This paper proposes a novel approach for topic segmentation in dialogue texts based on "utterance pair" relationship modeling, unifying supervised and unsupervised network architectures. To address the availability of annotated data, both supervised and unsupervised pre-training strategies are employed. To facilitate knowledge transfer during the two-stage training process, the model incorporates domain and task adapters. Experiments conducted across three data scenario settings validate the effectiveness of the proposed method. Future work will explore more complex multi-modal scenarios, and investigate the applicability across different languages and cultural contexts to enhance the method's generalizability and robustness.

Limitations

This work has two limitations. First, considering the diversity of real-world text datasets, the proposed utterance pair modeling method may only be applicable to dialogue datasets. Second, although this study introduces an adapter for two-stage training, there is still scope to explore different adapter mechanisms.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under the Grant 62277009, the project of Jilin Provincial Science and Technology Department under the Grant 20220201140GX, the project of Jilin province development and reform commission under the Grant 2024C008-8.

References

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Aleksei Artemiev, Daniil Parinov, Alexey Grishanov, Ivan Borisov, Alexey Vasilev, Daniil Muravetskii, Aleksey Rezvykh, Aleksei Goncharov, and Andrey Savchenko. 2024. Leveraging summarization for unsupervised dialogue topic segmentation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4697–4704, Mexico City, Mexico. Association for Computational Linguistics.

Joe Barrow, Rajiv Jain, Vlad Morariu, Varun Manjunatha, Douglas Oard, and Philip Resnik. 2020. A joint model for document segmentation and segment labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 313–322, Online. Association for Computational Linguistics.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34:177–210.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shizhe Diao, Tianyang Xu, Ruijia Xu, Jiawei Wang, and Tong Zhang. 2023. Mixture-of-domain-adapters: Decoupling and injecting domain knowledge to pre-trained language models' memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5113–5129, Toronto, Canada. Association for Computational Linguistics.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2024. Uncovering the potential of ChatGPT for discourse analysis in dialogue: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16998–17010, Torino, Italia. ELRA and ICCL.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Haoyu Gao, Rui Wang, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. 2023. Unsupervised dialogue topic segmentation with topic-aware contrastive learning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2481–2485, New York, NY, USA. Association for Computing Machinery.

Marti A Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

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

Yerin Hwang, Yongil Kim, Yunah Jang, Jeesoo Bang, Hyunkyung Bae, and Kyomin Jung. 2024. Mp2d:

An automated topic shift dialogue generation framework leveraging knowledge graphs. *Preprint*, arXiv:2403.05814.

Junfeng Jiang, Chengzhang Dong, Sadao Kurohashi, and Akiko Aizawa. 2023. SuperDialseg: A large-scale dataset for supervised dialogue segmentation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4086–4101, Singapore. Association for Computational Linguistics.

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

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Jiangyi Lin, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2023. Multi-granularity prompts for topic shift detection in dialogue. In *International Conference on Intelligent Computing*, pages 511–522. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F Chen. 2019b. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821. IEEE.

Seongmin Park, Jinkyu Seo, and Jihwa Lee. 2023. Unsupervised dialogue topic segmentation in hyperdimensional space. In *Interspeech 2023*.

Lev Pevzner and Marti A Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Hengfeng Pu and Liqing Wang. 2023. Dialogue segmentation based on dynamic context coherence. In *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval*, pages 190–195.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Yiping Song, Lili Mou, Rui Yan, Li Yi, Zinan Zhu, Xiaohua Hu, and Ming Zhang. 2016. Dialogue Session Segmentation by Embedding-Enhanced TextTiling. In *Proc. Interspeech 2016*, pages 2706–2710.

Liang Wang, Sujian Li, Yajuan Lü, and Houfeng Wang. 2017. Learning to rank semantic coherence for topic segmentation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1340–1344.

Jinxiong Xia, Cao Liu, Jiansong Chen, Yuchen Li, Fan Yang, Xunliang Cai, Guanglu Wan, and Houfeng Wang. 2022. Dialogue topic segmentation via parallel extraction network with neighbor smoothing. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2126–2131.

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

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

Jiahui Xu, Feng Jiang, Anningzhe Gao, and Haizhou Li. 2024. Unsupervised mutual learning of dialogue discourse parsing and topic segmentation. *Preprint*, arXiv:2405.19799.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021. Topic-aware multi-turn dialogue modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14176–14184.

Chenxu Yang, Zheng Lin, Jiangnan Li, Fandong Meng, Weiping Wang, Lanrui Wang, and Jie Zhou. 2022. TAKE: Topic-shift aware knowledge sElection for dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 253–265, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhihan Zhou, Dejiao Zhang, Wei Xiao, Nicholas Dingwall, Xiaofei Ma, Andrew Arnold, and Bing Xiang. 2022. Learning dialogue representations from consecutive utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 754–768, Seattle, United States. Association for Computational Linguistics.

A Dataset Specifications and Split Details

Supervised datasets:**TIAGE** (Xie et al., 2021) is a manually annotated dataset based on personal chat conversations, consisting of 500 samples: 300 for training, 100 for testing, and 100 for validation. **SuperDialseg** (Jiang et al., 2023), a document-grounded dialogue system dataset, leverages data from Doc2Dial (Feng et al., 2020) and Multi-Doc2Dial (Feng et al., 2021), further annotated to comprise 6,863 training samples, 1,310 test samples, and 1,305 validation samples.

Unsupervised Datasets:Doc2Dial (Feng et al., 2021) comprises 4,130 human-machine dialogues sourced from the document-grounded, goal-oriented dialogue corpus Doc2Dial. This dataset is generated by automatically constructing dialogue flows based on document content elements, followed by crowd-sourced annotation to compose utterance sequences aligned with human-like conversation flow. Topic segments are extracted based on document text spans providing the corresponding utterance information. **ZYS** (Xu et al., 2021) is a real-world Chinese dataset containing 505 dialogues recorded from bank customer service calls, with topic segments manually annotated.

B Metrics Details

P_k error (Beeferman et al., 1999) which assesses the agreement of segmentation points within a sliding window between the prediction and reference. WindowDiff (WD) (Pevzner and Hearst, 2002), similar to P_k but examining the agreement in the number of segmentation points within the sliding window. Both P_k and WD calculate soft errors of segmentation points within a sliding window, with lower scores indicating better performance.

we use scikit-learn 1.0.2 to compute F1 and macro F1. We use segeval 2.0.11 to compute Pk and WindowDiff and set the window size to half the length of the segment, consistent with past work.