Dialogue Topic Shift Based on Prompt Learning

Anonymous ACL submission

Abstract

The task of dialogue topic shift detection aims to identify whether a topic shift occurs in the current sentence relative to the preceding context during a conversation. Current research often treats n-gram features as equally important; however, the significance of these features actually depends on the specific context, which influences the model's semantic understanding of the entire text. To address this issue, we propose a model based on prompt learning with multi-scale feature attention. Under the guidance of the prompt learning module, the multi-scale feature attention layer is better able to capture textual semantic features, thereby improving the accuracy of topic shift detection in dialogues. The proposed model was evaluated on the Chinese CNTD dataset and the English TIAGE dataset. Experimental results demonstrate that our model achieves significant performance improvements compared to existing approaches. Furthermore, we compared multi-scale and single-scale feature attention models and found that the optimal performance was achieved when k was set to 4. Finally, we conducted ablation studies and analyses to validate the effectiveness and robustness of the model, resulting in performance enhancements to varying degrees.

1 Introduction

Daily conversations typically revolve around specific topics. A single dialogue may also encompass multiple topics, with discussions within each topic being relatively coherent. Topic shift, as a common phenomenon in conversations, plays a crucial role in maintaining the fluency and engagement of dialogues. Research on topic shift has gradually emerged as a focal point in fields such as linguistics, psychology, and artificial intelligence. For example, in Figure 1, a dialogue between an AI customer service agent and a customer is illustrated. In the second round of the conversation, the customer shifts the topic from the color of the clothing to its price. However, the AI fails to detect this topic shift, leading to an incorrect response.

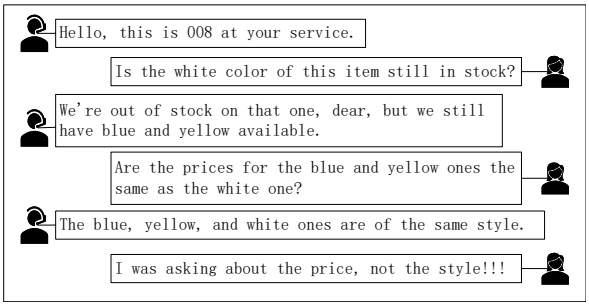

Figure 1: Example of a Conversation Between a Customer Service AI and a Customer.

Dialogue topic shift detection refers to the process of identifying whether a topic shift occurs during a conversation, which is somewhat analogous to topic classification but differs in key aspects. Topic classification involves segmenting and categorizing a piece of text or dialogue content into predefined topic categories, representing a static classification process. In contrast, topic shift refers to the dynamic behavior of participants changing the current topic of discussion and transitioning to another topic during a conversation. However, topic shift can also be viewed as a binary classification problem. Many mainstream text classification models are typically based on Convolutional Neural Networks (CNN) (Kim, 2014), Recurrent Neural Networks (RNN) (Bhowmik et al., 2018), and Transformer (Vaswani et al., 2017). For instance, some scholars (Wang et al., 2017) proposed using a symmetric CNN to perform topic segmentation by evaluating semantic coherence, thereby enabling topic classification. As some scholars (Sabour et al., 2017) enhanced CNN by introducing capsule networks, which replace neurons in neural networks with tensors, granting capsule networks more powerful feature learning capabilities. However, dia-

logue texts differ from general texts (news, novels) in that they exhibit temporal characteristics and variable lengths. Although CNNs can capture local features and contextual information through convolutional layers, they often struggle with long-range dependencies and fail to fully understand contextual features. RNNs are naturally suited for sequential data, but traditional RNNs may encounter difficulties when modeling long sequences. CNNs typically require fixed-length inputs, which may necessitate additional padding or truncation when processing variable-length dialogue texts. RNNs can handle variable-length inputs, but in practice, long sequences may increase computational burden. As a result, a number of studies have adopted BiLSTM models for topic classification and topic shift detection. Some scholars (Xing et al., 2020) proposed an improved BiLSTM model for context modeling, which better captures semantic consistency between sentences and restricts the attention scope, enabling the model to focus more on the local contextual information of the current sentence and better utilize information from adjacent sentences.

The "prompt" is a kind of cue provided to the pre-trained language model to allow the pre-trained language model to better understand human problems. The goal is to better utilize the knowledge in the pre-trained model means to add additional text to the input.

To better detect topic shifts in dialogue texts in real-time, this paper proposes a BERT-BiLSTM model based on prompt learning with multi-scale attention convolution. The multi-scale feature attention mechanism, initially proposed by other scholars (Wang et al., 2018), involves attention convolution across different window sizes to capture diverse n-gram features of the text. Building on prompt learning, the multi-scale feature attention mechanism introduced in this paper operates at the sentence level, utilizing varying window sizes to extract syntactic features of sentences at different scales. This model designs prompt templates tailored to the dialogue context, enabling a more effective understanding of topic shift expressions in sentences from large-scale pre-trained language models, thereby determining whether a sentence constitutes a topic shift. To facilitate real-time detection of topic shifts during conversations, the model eschews the sentence-by-sentence input approach in favor of a method where each input consists of the preceding discourse plus the current sentence. Finally, we conducted topic shift experiments on the Chinese CNTD and English TIAGE datasets, with the results demonstrating the superior performance of our model in topic shift detection.

The main contributions of this paper are as follows: (1) Firstly, this paper, for the first time, approaches the identification of topic shifts in dialogues from a novel perspective by proposing a prompt learning-based method for topic shift detection tasks. (2) Secondly, the multi-scale feature attention mechanism is integrated and enhanced with BERT-BiLSTM to accurately extract sentence-level multi-grammatical features, endowing the model with robust feature learning capabilities. (3) Lastly, topic shift experiments were conducted on the Chinese CNTD and English TIAGE datasets, respectively, with the experimental results demonstrating that our model outperforms the baseline methods.

2 Related Work

2.1 Topic Shift

Current research on dialogue topic shift is still in its preliminary stages (Erlin et al., 2013; Wang et al., 2020; Khatri et al., 2018; Huang et al., 2013; Zeng et al., 2018). Some scholars (Xie et al., 2021) were the first to construct the TIAGE dataset specifically for open-domain dialogues. Their study proposed three tasks to investigate topic shift behaviors in dialogue scenarios: topic shift detection, topic shift-triggered response generation, and topic-aware dialogue generation. Experimental results demonstrated that the topic shift labels in TIAGE are beneficial for generating topic-shifted utterances. However, most existing topic shift detection models rely on predefined topic sets, which are dynamically changing in open-domain dialogue systems, making these models difficult to apply. Some scholars (Konigari et al., 2021) proposed an XLNet-based model to detect and correct topic shifts. Their study utilized the Switchboard dataset, manually annotating 74 dialogues into three categories: main topics, subtopics, and off-topic discussions. Additionally, they introduced a System Initiative (SI) module that guides users back to the main topic when the XLNet model detects a topic shift. Experimental results showed that the XLNet model achieved the best performance in terms of precision, recall, and F1-score.

Due to the issue of monotonous response generation in general open-domain dialogue generation techniques, an increasing number of researchers

have focused on knowledge-based dialogue generation. By integrating external knowledge bases with generative models as supplementary information, the generated responses become more engaging and informative. However, previous studies on knowledge selection in dialogues overly relied on dialogue context, neglecting the intrinsic connections and transitions between knowledge pieces. This often led to models ignoring the selected knowledge (Lian et al., 2019), resulting in responses unrelated to the knowledge (Zhou et al., 2018). To address these issues, some scholars (Zhan et al., 2021) proposed a method combining "unsupervised learning + supervised learning + neural learning". Their paper introduced the SKT-KG model, which utilizes a BiLSTM-CRF model to simulate the transition probabilities between knowledge labels and employs unsupervised learning to pre-train a Transformer model, enabling it to better learn the language model. The study used two public datasets, DuConv and Wizard of Wikipedia, and the experimental results demonstrated superior performance compared to baseline models. The approach generated more informative and diverse responses and achieved higher accuracy and F1 scores in knowledge selection, proving the effectiveness of the sequential knowledge transition module. Some scholars (Yang et al., 2022) integrated a topic shift module with a knowledge selection module, creating the TAKE model, which leverages topic transition information to guide knowledge selection, thereby more accurately selecting knowledge relevant to the dialogue content and improving the accuracy of topic shift detection. Experimental results on the Wizard of Wikipedia (WoW) dataset showed that, compared to strong baseline models, the TAKE model not only selected knowledge more accurately, especially on unseen test sets, but also generated more informative and engaging responses.

But Lin et al. (2023b) addressed the challenge of unknown responses in dialogue topic shift tasks by proposing a hierarchical contrastive learning framework based on a teacher-student relationship to predict topic shifts without responses. They annotated the Chinese Natural Topic Dialogue (CNTD) corpu, which contains 1,308 dialogues, thereby filling a gap in Chinese natural dialogue topic corpora. Experimental results demonstrated that the proposed model outperformed baseline models in precision, recall, and F1-score on the CNTD dataset. In response to prior work that primarily focused on encoding utterances using pre-trained models for topic shift detection without delving into the granularity of topics at various levels or understanding the dialogue content, Lin et al. (2023a) further proposed a multi-granularity prompt-based dialogue topic shift detection model. This model leverages prompt learning techniques to extract dialogue information at three granularity levels—label, turn, and topic—constructs target sentences, and guides the model to learn deeper topic information within dialogues. Additionally, it combines the strengths of classification and generative models, using the classification model for topic shift prediction while employing the generative model to better understand dialogue topics and generate more natural language expressions. The proposed model achieved superior performance over baseline models on both the Chinese CNTD and English TIAGE datasets.

2.2 Prompt Learning

With the advancement of pre-trained models, we are currently undergoing a second major transformation, where the "pre-train, fine-tune" paradigm is being replaced by what we refer to as the "pre-train, prompt, and predict" paradigm Liu et al. (2023). In this new approach, downstream tasks are no longer adapted to pre-trained language models (LMs) through objective engineering. Instead, downstream tasks are reformulated to resemble tasks solved during the original LM training with the assistance of textual prompts, a method known as prompt learning. In recent years, research based on prompt learning has gained significant momentum. For instance, Brown et al. (2020) proposed the "pre-train, prompt, predict" paradigm and demonstrated the substantial potential of GPT-3 in few-shot learning. The advantage of this method is that, given an appropriate set of prompts, a single LM trained in a fully unsupervised manner can be used to solve a wide range of tasks. And Shin et al. (2020) introduced the AutoPrompt method, which utilizes automatically generated prompts to guide language models in knowledge probing. AutoPrompt's prompts enable more accurate factual knowledge extraction from masked language models (MLMs). So Schick and Schütze (2021) proposed a semi-supervised training method called PET (Pattern-Exploiting Training) to address natural language processing tasks in low-resource settings. The core idea of PET is to transform task descriptions into cloze-style phrases and use pre-trained language models (PLMs) to predict the

most suitable words to fill in the blanks.

3 Model

We propose a topic shift model based on prompt learning with multi-scale feature attention. This model capitalizes on the strengths of large-scale pre-trained models to design contextual prompt modules that incorporate both preceding context and the current sentence. The aim is to enable the model to better learn topic shift features through the utilization of pre-trained models. As shown in Figure 2, the proposed prompt learning-based multi-scale feature attention topic shift model consists of the following modules: (1) the prompt learning module, (2) the pre-trained model, and (3) the model training module.

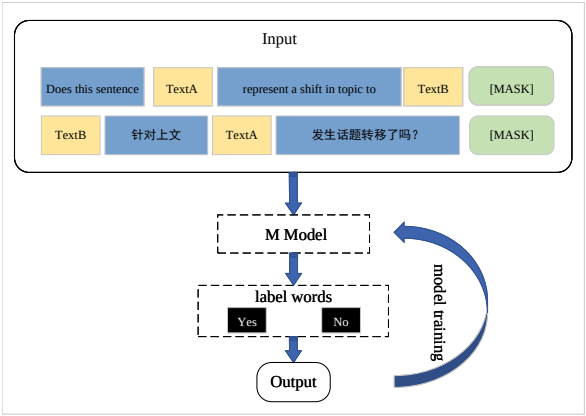

Figure 2: Flow chart of the Topic Shift Model Based on Prompt Learning.

3.1 Prompt Learning Module

To adapt the prompt learning module to the topic shift detection task, this paper builds upon the prompt-based template proposed by Liang et al. (2022), constructing topic-oriented ironic expression templates in the form of prefix prompt templates for input samples. Consequently, this paper utilizes the masked language model of pre-trained language models to fill in the [MASK] token positions with appropriate words. The advantage of using the masked language model lies in its ability to leverage large-scale pre-trained corpora, utilizing features from the non-masked regions to predict suitable words for the [MASK] position, thereby predicting the appropriate category labels. The constructed prompt learning template for the topic shift detection task is defined as follows:

$$T = \text{Does this sentence } \{a\} \text{ represent a shift in topic to } \{\text{b}\}\text{? [MASK]} \quad (1)$$

In the T, "ch" and "en" represent the definitions of the prompt learning templates for Chinese and English, respectively; "a" denotes all the preceding context before "b"; and "b" represents the current sentence. Here, the words "Yes" and "No" are used as the category label words for the model. That is, the label word set t = "Yes; No", corresponding to the topic shift category and the no-topic-shift category, respectively.

Based on this, the prompt learning template for the topic shift detection task can be derived for the input sample. Subsequently, it is necessary to utilize the pre-trained language model to predict the category label word at the [MASK] position, thereby determining whether the current sentence constitutes a topic shift relative to the preceding context.

3.2 Pre-trained Model

This module uses an improved BERT-BiLSTM model based on multi-scale feature attention for feature learning. This module is illustrated in Figure 3.

This paper initially employs a Chinese pre-trained language model (BERT-base-Chinese) to model the input samples. The input representation of the model is as follows:

$$\text{text} = (x_1, x_2, x_3, \ldots, x_m)_{m \times t} \quad (2)$$

$$r = [\text{CLS}]T[\text{SEP}] \quad (3)$$

In the equation, T includes both ch and en, representing two languages.

Subsequently, the model inputs the data into the pre-trained model M (The BERT-BiLSTM model based on multi-scale feature attention), and predicts the distribution of category label words at the [MASK] position in a masked language modeling manner:

$$P_M = M(r) \quad (4)$$

For the model M, the first component is the BERT model, which excels at extracting features of words within sentences. The BERT model adopts an encoder-only architecture, which allows it to focus more on understanding the input sequence rather than generating an output sequence. This is

one of the reasons why the BERT model is widely used in the field of NLP.

For the output of BERT model, after that it will be accessed to BiLSTM neural network, which is intended to have a better learning and representation of the Token that already has a certain feature representation in the temporal relationship, BiLSTM is combined by Forward LSTM and Reverse LSTM, for the output of BERT, BiLSTM can learn the dependency between the front and back of the sentence in a deeper level. After two models, the output size at this time is [B, S, 2*H] (B is Batch size, S is sequence length, H is the Hidden dim of LSTM, which is 2*H by using BiLSTM). The output is represented as:

$$c = \text{BiLSTM}(r) \tag{5}$$

Multi-scale feature attention aims to enable the model to adaptively select multi-gram features for each word. This paper employs this method to precisely capture the multi-gram features present in the text. Multi-scale feature attention consists of two steps: convolutional feature aggregation and scale feature weighting. Convolutional feature aggregation aims to represent the l-gram feature vector c of x with a scalar z_l; scale feature weighting uses z_l as input, and outputs a softmax distribution of attention weights to re-weight the multi-gram features at different scales.

For a convolutional kernel of size k, the convolution operation can be expressed as:

$$z_l^i = F(c_l^i) \tag{6}$$

In the equation, F(•) denotes the summation of each component of the input vector.

Weighted Summation: Based on the computed attention weights, a weighted summation is performed for each scale of n-gram features to obtain the comprehensive feature representation at that position. The process of weighted summation can be expressed as:

$$c_{\text{att}}^i = \sum_{l=1}^{L} c_l^i \times w_l^i \tag{7}$$

$$w_l^i = \text{softmax}(\text{MLP}(z_l^i)) \tag{8}$$

In the equation, c_l^i represents the comprehensive feature representation at position i, where l denotes the size and quantity of the convolutional kernel, and w_l^i represents the corresponding attention weight. MLP stands for Multi-Layer Perceptron.

At this stage, the output incorporates both temporal information and textual feature information. This output is then fed into the multi-scale feature attention module. Through multi-scale attention, the calculation can be performed using Equation 6. By applying convolutional kernels of different sizes to the input, the system logically integrates data for words of varying lengths. For instance, when the text contains the phrase "not cute", it is necessary to place more emphasis on the convolutional kernel of size 3 rather than size 2, which would only capture the word "cute".

After the convolutional operations are completed, Equations 7 and 8 are used to compute the attention scores for all convolutional results. These scores evaluate the contributions of different words across various convolutional kernels. Subsequently, a weighted average of the results from each convolutional kernel is computed to obtain the final sequence information. This sequence information is then passed through a linear layer to calculate the final output. To facilitate seamless connections between different models, a linear layer is added at the end of each model to transition feature information by converting between different hidden dimension sizes.

3.3 Model Training Module

By utilizing the pre-trained language model M, we can predict the probability distribution of each label word t in the label word set within the text X. To convert word probabilities into label probabilities, this paper introduces a mapping function f (Liang et al., 2022), which maps words from the label word set t to the category distribution space Y : $\{f : t \rightarrow Y\}$. Consequently, for the input text X, the calculation of the category distribution P(y|X) corresponding to the predicted label word t is as follows:

$$P(y|X) = Q(P_M(t|X)) \tag{9}$$

In the equation, Q(•) represents the function that transforms the probability of label words into the probability of category labels.

The proposed model is trained and optimized by minimizing the cross-entropy loss:

$$\tau = -\sum_{i=1}^{N}\sum_{j=1}^{L} y_i^j \log(\hat{y}_i^j) + \lambda \, ||\theta||^2 \tag{10}$$

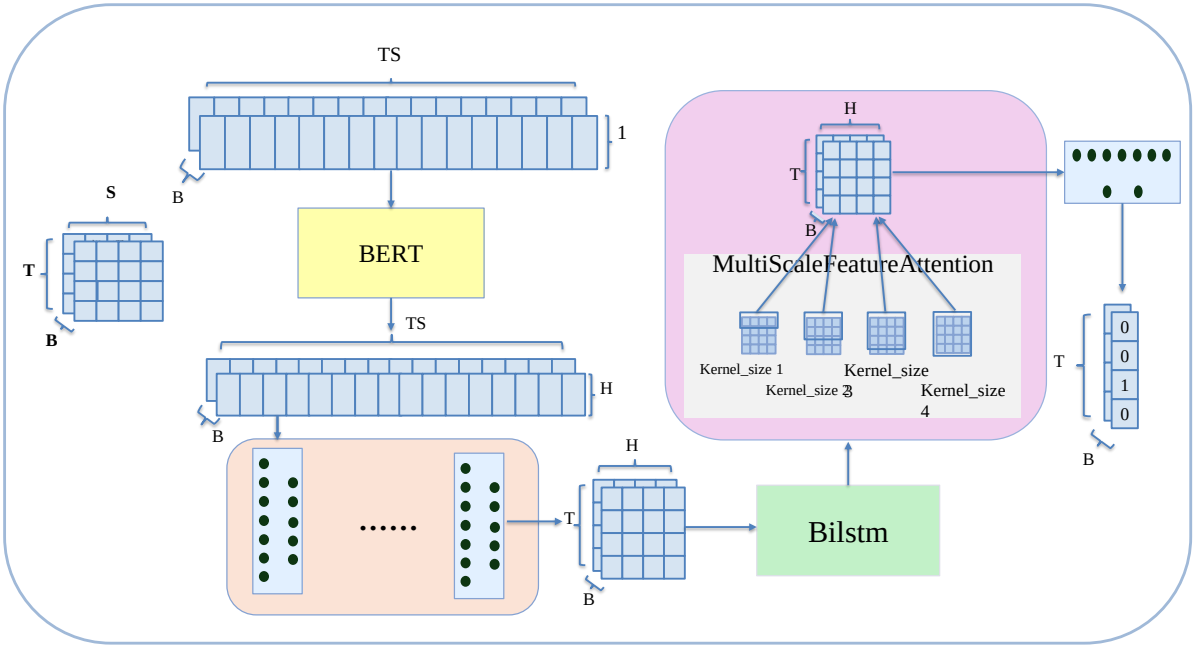

Figure 3: Architecture of our BERT-BiLSTM Model Based on Multi-Scale Feature Attention.

In the equation, N denotes the size of the training set, and L represents the number of categories. y_i^j and $\hat{y}_i^j$ correspond to the true category distribution and the predicted category distribution, respectively, for the training sample i. θ encompasses all trainable parameters in the model, and λ is the coefficient for L2 regularization.

4 Experiments

4.1 Experiment Settings

We evaluated our model on two datasets: the Chinese dataset CNTD and the English dataset TIAGE. The partitioning of the Chinese CNTD dataset and the English TIAGE dataset follows the same approach as Lin et al. (2023a). For both the Chinese CNTD dataset and the English TIAGE dataset, we selected the context (previous sentences and the current sentence) of each dialogue as input, and fed the context (previous sentences and the current sentence) along with the labels into the prompt module for enhanced learning. In our experiments, the label for the first sentence of each dialogue was set to 0 (indicating no topic shift). For the evaluation of all experiments in this study, we utilized Precision (P), Recall (R), and Macro-F1 scores. All experiments were conducted on a 4090 GPU, with a batch size of 8 and a training epoch of 20 for each experiment. The Adam optimizer was employed, and different dropout rates were set for different experiments.

4.2 Experiment Results

The most of representative models for the topic shift detection task are still BERT and T5. Therefore, we selected the following models (Lin et al., 2023a) for comparative experiments: (1) RoBERTa (Liu et al., 2019), an improved version of BERT; (2) BERT (Devlin et al., 2019), a Transformer-based bidirectional encoder used for text encoding; (3) Hier-BERT (Zhang et al., 2019), a hierarchical structure based on the Transformer model; (4) T5 (Xie et al., 2021), a modified architecture based on the Transformer; (5) BERT+BiLSTM (Lukasik et al., 2020), a bidirectional long short-term memory network integrated with BERT; (6) Ours, BERT+BiLSTM+Multi-Scale+Prompt (scale=4; size=1-4; D=0.50).

Model	P	R	F1
RoBERTa	84.4	75.4	78.6
BERT	82.9	79.2	80.8
Hier-BERT	85.6	79.0	81.7
T5	83.0	79.7	81.1
BERT+BiLSTM	82.8	82.0	82.4
Ours	**86.1**	**83.8**	**84.9**

Table 1: Experimental Results on the Chinese CNTD Dataset ($p < 0.01$).

Model	P	R	F1
RoBERTa	84.4	75.4	78.6
BERT	82.9	79.2	80.8
Hier-BERT	85.6	79.0	81.7
T5	83.0	79.7	81.1
BERT+BiLSTM	82.8	82.0	82.4
Ours	**86.1**	**83.8**	**84.9**

Table 2: Experimental Results on the Chinese CNTD Dataset ($p < 0.01$).

The experimental results are presented in Table 1. Comparative experiments on the Chinese CNTD dataset indicate that the RoBERTa model among the pre-trained models performed the worst, with a P-value of only 84.4, an R-value of 75.4, and an F1-score of 78.6. The BERT+BiLSTM and Hier-BERT models exhibited the best performance, with Hier-BERT achieving a P-value as high as 85.6, and BERT+BiLSTM attaining R-value and F1-score of 82.0 and 82.4, respectively. Clearly, compared to single pre-trained models, the combination of multiple models is more suitable for tasks such as text classification and topic shift. Furthermore, when comparing our experimental model, it is evident that our model outperformed the others, achieving a P-value of 86.1, an R-value of 83.8, and an F1-score of 84.9.

The comparative experimental results on the English TIAGE dataset are shown in Table 2. The results demonstrate that among the pre-trained models, the BERT model performed the worst on the topic shift task in the English dataset, with a P-value of only 68.5, an R-value of 65.4, and an F1-score of 66.6. The T5 model performed the best, even achieving a P-value of 76.5. Additionally, compared to all the aforementioned models, our model surpassed them in terms of the average of the three experimental metrics, thereby proving the effectiveness of our experimental model.

Moreover, we observed that our model performed better on the Chinese dataset, while its experimental results on the English dataset were less satisfactory.

4.3 Comparative Experiments Between Multi-Scale and Single-Scale Approaches

To evaluate the impact of multi-scale feature attention convolution sizes on model performance, we conducted comparative experiments on the CNTD dataset. For the multi-scale feature attention convolution, we selected different values of k=4, 5, 6 and various dropout rates of 0.4, 0.45, 0.5, and 0.55 for comparative experiments. The results are presented in Table 3.

Scale	SIZE	D	P	R	F1
4	1-4	0.40	84.9	83.4	84.1
4	1-4	0.45	82.3	83.0	82.6
4	**1-4**	**0.50**	**86.1**	**83.8**	**84.9**
4	1-4	0.55	85.4	81.9	83.4
5	1-5	0.50	85.3	82.7	83.9
6	1-6	0.50	85.9	81.7	83.5

Table 3: Comparative Experiments on Attentional Convolution of Multiscale Features.

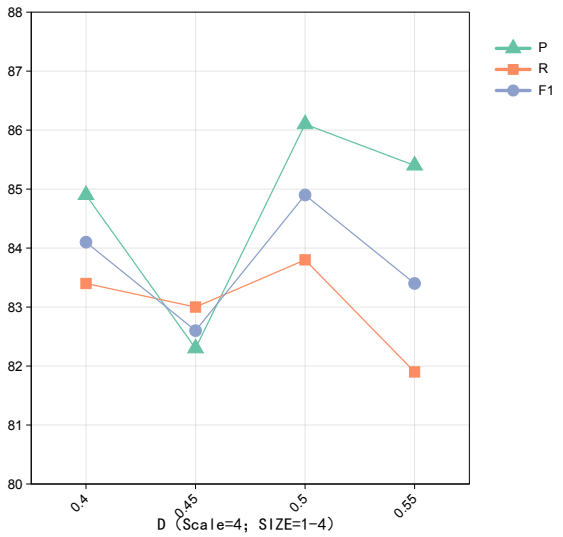

Figure 4: Multi-scale Feature Attention Convolution Comparative Experiment Diagram (Scale=4; SIZE=1-4; D=0.40, 0.45, 0,50, 0.55).

The experimental results presented in Table 3 and Figure 4 indicate that for feature attention convolutions with four scales (1-4), the best performance is achieved when the dropout rate is set to 0.5, while a dropout rate of 0.45 yields relatively poorer results. Furthermore, when comparing feature attention convolutions with four (1-4), five (1-5), and six (1-6) scales, the four-scale (1-4) feature attention convolution demonstrates the best performance. As the number of scales increases, the performance gradually declines, suggesting that for each dialogue in this dataset, a four-scale sentence-level feature attention convolution is the most effective one.

To assess the impact of different single-scale

convolution sizes on the model, we conducted experiments on the CNTD dataset. For the single-scale feature attention convolution experiments, we selected five different single-scale features, namely 1, 2, 3, 4, and 5, for comparative analysis. The dropout rate was consistently set to 0.5 to ensure optimal performance (demonstrated by the results in Table 3), which confirmed that a dropout rate of 0.5 yields the best outcomes. The experimental results are summarized in Table 4.

SIZE	D	P	R	F1
1	0.5	83.5	83.2	83.3
2	0.5	83.0	82.8	82.9
3	0.5	84.3	82.9	83.6
4	**0.5**	**86.1**	**81.5**	**83.5**
5	0.5	84.4	81.3	83.2

Table 4: Comparative Experiments on Single-Scale Feature Attention Convolution (Scale=1).

The experimental results of the single-scale feature attention convolution are presented in Table 4. The findings reveal that the feature attention convolution with a scale of 4 achieves the best performance in the single-scale experiments. Furthermore, when the single-scale value is less than 4, the performance improves progressively as the scale increases. However, when the single-scale value exceeds 4, the performance gradually declines. These results highlight the importance of selecting an appropriate scale for single-scale feature attention convolution to optimize model performance.

4.4 Ablation Study

To evaluate the effectiveness of various components of the model, this paper conducted experiments on the CNTD dataset to investigate the contributions of the model's key components. Theoretically, these components can be categorized into the following types: (1) multi-scale feature attention, and (2) prompt learning module.

Model	P	R	F1
BERT+BiLSTM	82.8	82.0	82.4
+mul	85.8	82.7	84.8
+mul+prompt	**86.1**	**83.8**	**84.9**

Table 5: Ablation Experiments on the Chinese Dataset CNTD.

The ablation experiments were conducted to examine the impact of the multi-scale feature attention layer and the prompt learning module on the model's performance. As shown in Table 5, both the multi-scale feature attention convolution and the prompt learning module significantly influence the model's performance. Specifically, the P-value increased from 82.8 to 85.8 after incorporating the multi-scale feature attention convolution and further rose to 86.1 upon adding the prompt learning module. These results demonstrate the effectiveness of our model.

5 Conclusion

In this paper, we address the issue that the significance of multi-scale feature attention convolution size should be determined by specific contextual information by proposing a topic shift model based on prompt learning. This model is composed of two key components: (1) prompt learning (2) multi-scale feature attention convolution. The multi-scale feature attention convolution in this paper is based on a multi-scale feature attention-enhanced BERT-BiLSTM model (MLM version), which captures diverse syntactic features according to specific contextual information. We conducted comparative experiments on both the Chinese CNTD dataset and the English TIAGE dataset. The results demonstrate that our model achieves significant performance improvements compared to current state-of-the-art approaches. Additionally, we performed comparative experiments with different sizes of multi-scale feature attention convolution and various types of single-scale feature attention convolution. Our findings reveal that multi-scale feature attention convolution outperforms single-scale feature attention convolution, proving that the importance of multi-scale feature attention convolution size is context-dependent, with the optimal performance achieved when k is set to 4. Finally, we conducted ablation experiments to validate the effectiveness and robustness of our model.

Limitations

However, upon analyzing and observing the information extracted at different scales, it is evident that these key pieces of information are subject to errors and noise. Our future work will focus on enhancing the reliability of conversational information mining and exploring finer granularity in scenarios of topic shifts within dialogues.

References

Showmik Bhowmik, Ram Sarkar, Mita Nasipuri, and David Doermann. 2018. Text and non-text separation in offline document images: a survey. *Int. J. Doc. Anal. Recognit.*, 21(1–2):1–20.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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.

Erlin, Unang Rio, and Rahmiati. 2013. Text message categorization of collaborative learning skills in online discussion using support vector machine. In *2013 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*, pages 295–300.

Shu Huang, Wei Peng, Jingxuan Li, and Dongwon Lee. 2013. Sentiment and topic analysis on social media: a multi-task multi-label classification approach. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, page 172–181, New York, NY, USA. Association for Computing Machinery.

Chandra Khatri, Rahul Goel, Behnam Hedayatnia, Angeliki Metanillou, Anushree Venkatesh, Raefer Gabriel, and Arindam Mandal. 2018. Contextual topic modeling for dialog systems. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 892–899.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Rachna Konigari, Saurabh Ramola, Vijay Vardhan Alluri, and Manish Shrivastava. 2021. Topic shift detection for mixed initiative response. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 161–166, Singapore and Online. Association for Computational Linguistics.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5081–5087. International Joint Conferences on Artificial Intelligence Organization.

Bin Liang, Zijie Lin, Bing Qin, and Ruifeng Xu. 2022. Topic-oriented sarcasm detection: New task, new dataset and new method. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 557–568, Nanchang, China. Chinese Information Processing Society of China.

Jiangyi Lin, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2023a. Multi-granularity prompts for topic shift detection in dialogue. In *Advanced Intelligent Computing Technology and Applications*, pages 511–522, Singapore. Springer Nature Singapore.

Jiangyi Lin, Yaxin Fan, Feng Jiang, Xiaomin Chu, and Peifeng Li. 2023b. Topic shift detection in chinese dialogues: Corpus and benchmark. In *Document Analysis and Recognition - ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part III*, page 166–183, Berlin, Heidelberg. Springer-Verlag.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

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

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text segmentation by cross segment attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4707–4716, Online. Association for Computational Linguistics.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic routing between capsules. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3859–3869, Red Hook, NY, USA. Curran Associates Inc.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4222–4235, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Jiancheng Wang, Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2020. Sentiment classification in customer service dialogue with topic-aware multi-task learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9177–9184. AAAI Press.

Liang Wang, Sujian Li, Yajuan Lv, 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, Copenhagen, Denmark. Association for Computational Linguistics.

Shiyao Wang, Minlie Huang, and Zhidong Deng. 2018. Densely connected cnn with multi-scale feature attention for text classification. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4468–4474. International Joint Conferences on Artificial Intelligence Organization.

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, Brad Hackinen, Giuseppe Carenini, and Francesco Trebbi. 2020. Improving context modeling in neural topic segmentation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 626–636, Suzhou, China. Association for Computational Linguistics.

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.

Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R. Lyu, and Irwin King. 2018. Topic memory networks for short text classification. In *Conference on Empirical Methods in Natural Language Processing*.

Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021. Augmenting knowledge-grounded conversations with sequential knowledge transition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5621–5630, Online. Association for Computational Linguistics.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4623–4629. AAAI Press.

A Datasets

Category	Train	Val	Test	Sum
Health	85	11	11	107
Education	167	22	21	210
Technology	176	22	22	220
Sports	347	45	46	438
Games	86	11	11	108
Entertain-ment	180	23	22	225
Total	1041	134	133	1308

Table 6: Category and proportion of the CNTD corpus.

	Min	Max	Avg
Dialogue Turns	20	26	20.1
Utterance Words	1	141	21.0
Dialogue Words	194	888	421.7
Dialogue Topics	2	9	5.2
Topic Turns	1	17	4.2

Table 7: Details of CNTD.

Tables 6 and 7 provide a detailed explanation of the CNTD dataset used in our study, along with its division into training, testing, and validation sets.

	Train	Dev	Test
Dialogs	300	100	100
Instances	4,767	1,546	1,548
AvgTurns	15.6	15.5	15.6

Table 8: Details of TIAGE.

The details of the TIAGE dataset are presented in Table 8, which includes information on Dialogs, Instances, and AvgTurns. The dataset is divided into Train, Dev, and Test sets.

B The template for prompt learning

$T1 =$ Does this sentence $\{a\}$ represent a shift in topic to $\{\text{b}\}$? [MASK]

$T2 =$ Is there a change in topic from $\{a\}$ to $\{\text{b}\}$? [MASK]

$T3 =$ Does $\{a\}$ indicate a transition in subject matter to $\{\text{b}\}$? [MASK]

We conducted tests on the templates used in prompt learning by employing different phrasings while maintaining the same meaning, as illustrated by T1, T2, and T3. Our findings revealed no significant difference in their effectiveness. Consequently, we concluded that the choice of prompt learning templates does not influence the detection of topic shifts in dialogues.