Simulating Dual-Process Thinking in Dialogue Topic Shift Detection

Huiyao Wang, Peifeng Li*, Yaxin Fan, Qiaoming Zhu
School of Computer Science and Technology, Soochow University, Suzhou, China
20235227091@stu.suda.edu.cn, pfli@suda.edu.cn
yxfansuda@stu.suda.edu.cn, qmzhu@suda.edu.cn

Abstract

Previous work on dialogue topic shift detection has primarily focused on shallow local reasoning, overlooking the importance of considering the global historical structure and local details to elucidate the underlying causes of topic shift. To address the above two issues, we introduce the dual-process theory to this task and design a novel Dual-Module Framework DMF (i.e., intuition module and reasoning module) for the task of dialogue topic shift detection to emulate this cognitive process. Specifically, the intuition module employs Large Language Models (LLMs) to extract and store the global topic structure of historical dialogue, while the reasoning module introduces an LLM to generate reasoning samples between the response and the most recent topic of historical dialogue, thereby providing local detail explanations for topic shift. Moreover, we distill the dual-module framework into a small generative model to facilitate more precise reasoning. The experimental results on three public datasets show that our DMF outperforms the state-of-the-art baselines.

1 Introduction

Topic shift detection in dialogue is a real-time classification task that determines whether a topic shift has occurred between the context (i.e., the historical dialogue) and the current response (i.e., the current utterance). As an essential subtask of target-guided proactive dialogue systems (Deng et al., 2023), topic shift detection generates shift signals that can assist in many downstream tasks, such as topic planning and topic-aware response generation (Yang et al., 2022; Gupta et al., 2022).

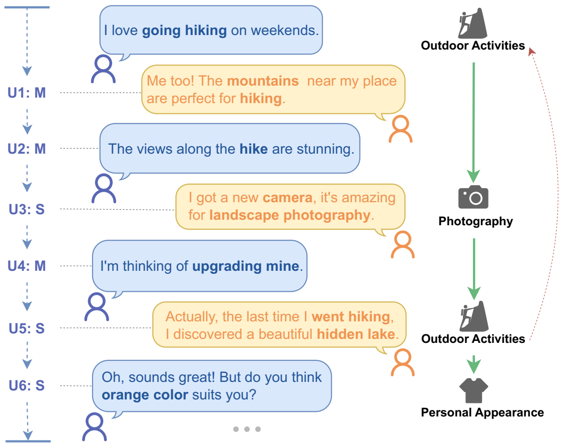

I love **going hiking** on weekends.
U1: M
Me too! The **mountains** near my place are perfect for **hiking**.
U2: M
The views along the **hike** are stunning.
U3: S
I got a new **camera**, it's amazing for **landscape photography**.
U4: M
I'm thinking of **upgrading mine**.
U5: S
Actually, the last time I **went hiking**, I discovered a beautiful **hidden lake**.
U6: S
Oh, sounds great! But do you think **orange color** suits you?
Outdoor Activities
Photography
Outdoor Activities
Personal Appearance

Figure 1: An example of the topic shift detection task, where real-time determination is made on whether each new response introduces a new topic relative to the preceding context. 'U_i' represents the i-th utterance in a dialogue, 'M' denotes a maintained topic (i.e., no topic shift), and 'S' denotes a topic shift.

As a new task, only a few studies focused on dialogue topic shift detection. Existing methods typically model shallow semantics between the context and the response for local reasoning (Xie et al., 2021), or utilize surface-level auxiliary information such as keywords (Lin et al., 2023a) or knowledge entities (Hwang et al., 2024). They overlook the capture of long-term topic evolution trajectories and the global historical structure. Moreover, such shallow local reasoning methods still face limitations in capturing the latest topic dynamics and addressing local details.

Topic changes occur in real-time and gradually accumulate in speaker's memory, which means that the speaker can make quick matches in relevant memory when determining whether a new topic has arisen. Hence, the global historical structure serves as the speaker's memory mechanism, organizing the thematic context and allowing for quick matching and judgment when faced with new responses. As the conversation progresses, the topics under discussion continue to evolve. Moreover, to capture the latest dynamics of the dialogue and the specific content and details, it is also essential to

[*] Corresponding author

Proceedings of the 31st International Conference on Computational Linguistics, pages 2592–2602
January 19–24, 2025. ©2025 Association for Computational Linguistics

carefully analyze the local reasons for "shift" or "maintain".

The global historical structure of dialogue is of paramount importance for the detection of topics. Figure 1 shows an example of the topic shift detection task. During the 5-th turn of shift detection (i.e., identifying whether U_5 has changed the topic), the global historical structure reveals that U_0-U_2 discussed outdoor activities, U_3-U_4 transitioned to photography equipment, and U_5 returned to the topic of outdoor activities. By aligning the response semantics with the historical theme of outdoor activities, it can be swiftly determined that the present response is inconsistent with the most recent topic photography equipment. This suggests that a topic shift has occurred. Furthermore, the latest topic dynamics and local details can also improve topic shift detection. In Figure 1, by extracting and analysing the relationships between the latest topic "Outdoor Activities" with its specific details "hiking" and the topic of the response (i.e., u_6) "Personal Appearance" with its details "orange color", we can gain a more detailed understanding of the reasons behind topic shifts.

Inspired by the dual-process theory (Kahneman, 2003), which proposes two modes of human cognition: System 1 (intuitive judgments based on existing memory) and System 2 (slower, analytical reasoning), we designed a novel Dual-Module Framework (DMF) to emulate this cognitive process for topic shift detection, which consists of an intuition module that continuously stores the global historical structure and a detailed reasoning module containing the latest topic dynamics and local details.

Specifically, the intuition module (corresponding to System 1) initially employs LLMs to extract and store the global historical structure. Subsequently, the response is matched with the historical memory in a small model, thereby enabling rapid judgments based on the global historical structure and the principal topics of the conversation.

Moreover, the reasoning module (corresponding to System 2) employs an LLM to generate reasoning samples between the response and the most recent topic of historical dialogue, thereby providing local detail explanations for topic shift in the current turn. This more profound examination and rational reasoning facilitate the precise identification of alterations and nuances within utterances. The combination of global historical structure storage with local detail reasoning analysis has the potential to significantly enhance the accuracy of topic shift detection.

Finally, we adopt Fine-tune-CoT (Chain-of-Thought) (Ho et al., 2023), which generates reasoning processes using zero-shot CoT prompts and fine-tunes small models to further distill the dual-system thinking learned from LLMs into a smaller generative model. This enables the small model to employ the dual-system thinking process to store the global historical structure and the comprehend reasoning details, thereby facilitating more precise reasoning and identification.

2 Related Work

The task of topic segmentation is similar to topic shift detection, which aims to divide a complete dialogue into topic blocks. and it can be divided into unsupervised and supervised methods. The former usually achieves segmentation by leveraging semantic similarity or dialogue coherence (Hearst, 1997; Xing and Carenini, 2021; Gao et al., 2023), while the latter also treated this task as a sequence labeling problem (Jiang et al., 2023).

In contrast, only a limited number of studies have addressed the issue of dialogue topic shift detection. This new task is more challenging, as it requires the ability to make real-time judgments regarding the shift in relation between each reply and its preceding historical context.

Currently, there are two public datasets specifically designed for the dialogue topic shift detection tasks. The first is TIAGE (Xie et al., 2021), an English dataset based on PersonaChat (Zhang et al., 2018). The second is CNTD (Lin et al., 2023b), a Chinese natural dialogue dataset annotated on NaturalConv (Wang et al., 2021). Moreover, some corpora originally used for topic segmentation have also been adapted for topic shift, such as SuperDialseg (Jiang et al., 2023). Earlier studies on topic shift detection mainly relied on textual relevance analysis and failed to fully leverage semantic similarity (Sun and Loparo, 2019). Current studies can be broadly divided into two types: those that directly classify by extracting sentence-level semantic information, and those that leverage auxiliary information such as keywords and knowledge entities to complete the task. For example, Xie et al. (2021) employed sentence-level semantic analysis to detect topic shift. Lin et al. (2023a) introduced a prompt-based approach to extract multi-granularity information (e.g., keywords and semantic roles) to

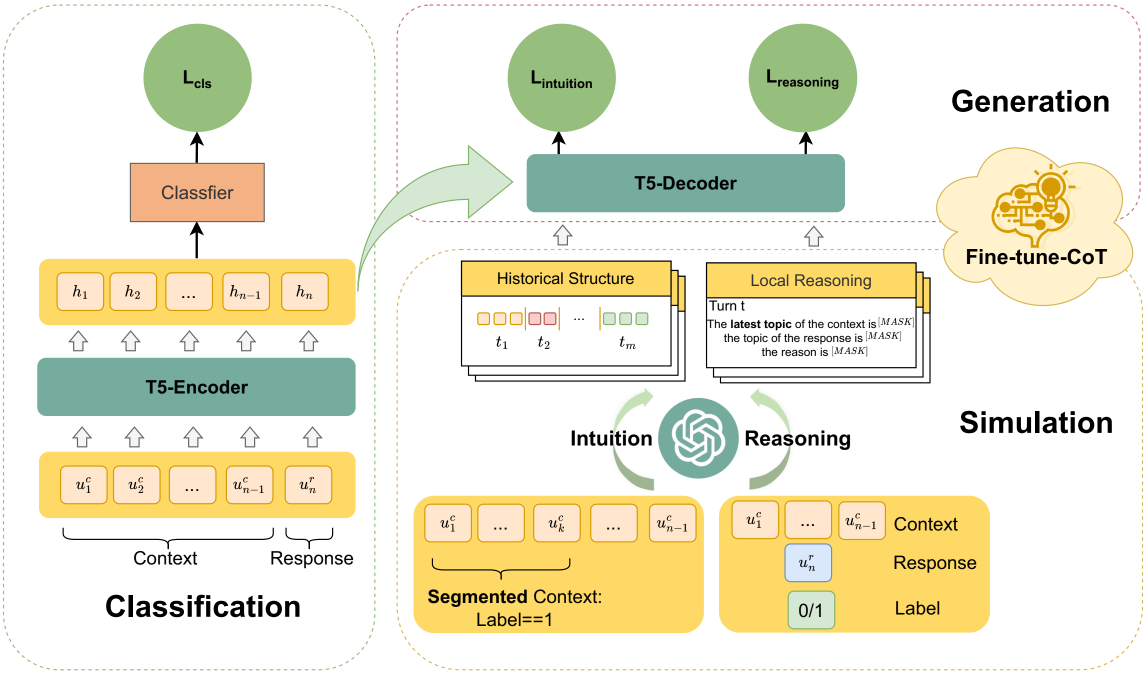


Figure 2: The overview of our framework DMF.

assist the model in comprehending dialogue topics. Hwang et al. (2024) used entity relations from the knowledge graphs to map the flow of topics, and annotated the TS-WikiDialog dataset. However, this dataset has not been publicly released and they did not evaluate their model on the public datasets.

In conclusion, the existing methods are largely dependent on surface features or rely on auxiliary information for local judgments, without adequately capturing the historical structure from a global perspective. Concurrently, it is important to capture the most recent dynamics and details as the dialogue progresses. Our approach simulates the dual processes of human cognition: intuition and analysis. By extracting the global historical structure, we capture the overall framework and thematic evolution, while in-depth reasoning allows us to focus on local transitions and detail extraction.

3 Methodology

As illustrated in Figure 2, our framework DMF consists of three main components, i.e., simulation, generation, and classification.

Simulation To simulate System 1, we leverage GPT-4 to extract the global historical structure, which serves as the knowledge base for intuitive judgments in the small model. Additionally, we infer local reasons for the current topic shift to simulate System 2.

Generation We apply Fine-tune-CoT (Ho et al., 2023) to distill GPT-4's capabilities in extracting global historical structures and reasoning over local details into the small model.

Classification We regard dialogue topic shift detection as a supervised classification task and train the model for accurately identifying and classifying topic shifts in dialogues.

3.1 Task Formulation

In this study, we treat the topic shift detection task as a classification task. Formally, given a context sequence of the historical dialogue $C = \{u_1^c, \ldots, u_i^c, \ldots, u_{n-1}^c\}$, which consists of n-1 historical utterances, and the response utterance $R = u_n^r$ at turn n, our goal is to train a model to learn the probability distribution $P(Y|C, R)$, where Y belongs to the label set $\{0, 1\}$ (0 indicates no topic shift, and 1 indicates a topic shift).

3.2 Dual-Process Thinking Simulation

Dual-process theory (Kahneman, 2003) proposes two cognitive modes: System 1 (rapid intuitive judgments based on existing memory) and System 2 (slower analytical reasoning). Inspired by previous work (He et al., 2024) on dialogue planning, we apply the dual-process theory to topic shift detection as follows: **System 1 is intuitive inference** which is simulated by quickly matching the historical memory with the topic of the current response using stored global historical structures. **System 2 is detailed analysis** which is modeled through in-depth reasoning of topic shift on local details.

Intuitive Inference: Extracting Historical Topic Structure at Global Level The global historical structure functions is similar to a speaker's memory system, enabling the model to rapidly access and match pertinent memories by elucidating the overarching context. In this paper, we use topic structure to represent the global historical structure. By segmenting the historical dialogue into finer substructures, the model is able to focus on shorter and relatively independent dialogue segments, which facilitates the identification of key elements within the larger context. This enables the model to capture essential memories and primary themes while extracting the global historical structure. This process enables small models to rapidly align topics during training, facilitating the system's capacity to make intuitive judgments in the initial stages.

Specifically, the context C is segmented using historical ground truth labels, resulting in $\{u_1^c, \ldots, u_k^c, \ldots, u_{n-1}^c\}$, where k indicates the point at which the current utterance is labeled 1, meaning a topic shift has occurred. It is worth noting that the process of segmenting the context using ground truth labels is only applied during training.

Using the prompt shown in Appendix A, the segmented historical dialogue is fed to GPT-4 to extract the global historical structure, generating a sequence of global historical (topic) structure: $\{t_1, \ldots, t_i, \ldots, t_m\}$, where t represents a topic which contains several tokens, and m corresponds to the number of historical topics in the context C. Table 1 shows a specific example from the TIAGE training set.

Detail Analysis: Inferring Reasons for Shift at Local Level in the Current Turn As the dialogue progresses, it is important to capture the latest dynamics and details in a timely manner in order to infer the specific reasons for topic shifts and their subtle nuances. During this process, the model must not only identify obvious topic changes but also demonstrate an ability to comprehend the underlying logical connections in complex dialogue scenarios. By capitalizing on the reasoning capabilities of LLMs for analysis and distilling these abilities into small models, we construct the second analysis and reasoning module.

A detailed analysis at local level elucidates the reasons for topic shifts or continuations within the current turn. Specifically, considering a standard sample including the context C, the response R, and the corresponding ground truth label L in the train dataset, which is only used for training. By prompting GPT-4 to generate a corresponding COT explanation, the local reasoning generates a text sequence in the following form:

E: " *The **latest topic** of the context is* $[MASK]$*, the topic of the response is* $[MASK]$*, and the reason for the shift or continuation is* $[MASK]$."

This prompt can utilize the comprehension and summarization abilities of LLMs to extract the topic of the response and the latest topic in the context, thereby provide critical local details for shift detection. An example of the local reasoning E is shown in Table 1 and the prompt guides GPT-4's reasoning process is detailed in Appendix B.

3.3 Fine-tune-CoT Generation

This module draws upon three distinct types of information. Two categories of the inference samples extracted by LLMs are the global historical structure and local reasoning details in Subsection 3.2. The remaining data comprises the encoding results generated by the T5 encoder. Specifically, the global historical structure is represented as follows:

G: " ***Historical topic structure** is* $[MASK]$*, the current response topic is* $[MASK]$*, the topic [is maintained | has shifted].*"

The local reasoning is represented as E in Subsection 3.2. These pieces of information are fed into the decoder for supervised fine-tuning.

As a result of this process, the small generative model can not only segment and extract global historical structures during the validation and testing phases, but also match the most recent responses with historical memory, thereby enabling intuitive judgments during the training phase. Furthermore, the model develops the capacity to reason thoroughly about the underlying causes of local transitions, thereby enabling it to capture the nuances of real-time conversational dynamics. This approach also facilitates the distillation of Fine-tune-COT capabilities from a large model to a small one.

We employ a generation model using encoder-decoder architecture to generate the representation of the combination of the context C and the response R. Specifically, using the T5 model as an example, the encoder calculates the representation vector for each utterance by inputting the context C and the response R into the encoder, forming a

Dialogue	**Historical Structure**
A: “hey! do you love cats?” B: “hey... I am a dog person, I have two” A: “ah that is cool, I have two cats and got a collection of 1000 hats for them!” B: “wow!!! that is a lot lol”	Pets and Pet Accessories
A: “yeah, I have a weakness for cats and vanilla ice cream, they are the best!” B: “my weakness is eating when I am bored”	Personal Weaknesses
A: “what is your favorite season? mine is winter!” B: “I am a summer girl”	Preferences for Seasons
A: “have you ever watched the Olympics? I won a gold medal in 1992!” B: “cool what did you win for?”	Sports and Olympics
Historical topic structure: Pets and Pet Accessories,Personal Weaknesses, Preferences for Seasons,Sports and Olympics. **Historical structure** G: Historical topic structure is Pets and Pet Accessories,Personal Weaknesses, Preferences for Seasons, Sports and Olympics, the current response topic is Awarded Projects Inquiry, the topic is maintained. **Local reasoning** E: The latest topic of the context is: Olympics; The topic of response is: win; The reason is: The response is directly related to the context, asking for more details about the previously mentioned Olympic gold medal, hence no topic shift.	

Table 1: An example of historical structures and local reasoning, taken from the TIAGE train dataset. In the dialogue column, the black text represents the dialogue context, and the red text indicates the current turn’s response. The lable of current turn is 0.

complete turn of dialogue $\{u_1, \ldots, u_i, \ldots, u_n\}$.

$$HE = \text{T5-Enc}(S) \tag{1}$$

where $S = [CLS] \oplus u_i \oplus \cdots \oplus u_n$ is the concatenated sequence using predefined special token $< \backslash s >$, $HE \in \mathbb{R}^d$ is the hidden state of the utterances, and d denotes the hidden size of encoder.

Specifically, taking T5 for example, the global historical structure G and the local detail reasoning E are separately injected for autoregressively decoding during training stage as follows.

$$L_{\text{intuition}} = -\sum_{j=1}^{|G|} \log P(G_j | HE, G_{<j}) \tag{2}$$

$$L_{\text{reasoning}} = -\sum_{i=1}^{|E|} \log P(E_i | HE, E_{<i}) \tag{3}$$

The generation component is optimized through the application of two cross-entropy loss functions, enabling T5 to develop both intuitive and analytical reasoning abilities through monitoring global historical and local reasoning data.

3.4 Classification

HE is fed into the classifier to obtain the classification probability distribution P. The classification process is trained using cross-entropy loss, where N is the number of categories, y represents the gold label, and $y \in \{0, 1\}$.

$$L_{\text{cls}} = -\sum_{i=1}^{N} y_i \log P_i \tag{4}$$

3.5 Training Objective

The total loss for the final training is the sum of the losses from the classification and the generation, which integrate the shift signal into two thinking modes during teacher-student training to more accurately determine the final topic shift as follows.

$$L_{\text{total}} = L_{\text{cls}} + \lambda_1 L_{\text{intuition}} + \lambda_2 L_{\text{reasoning}} \tag{5}$$

where λ_1 and λ_2 are the hyperparameters that decide the importance of each component.

4 Experimentation

4.1 Experimental Settings

We conducted experiments on two English datasets **TIAGE** (Xie et al., 2021) and **SuperDialseg** (Jiang et al., 2023) and one Chinese dataset **CNTD** (Lin et al., 2023b). The details of the three datasets can be found in Table 2.

We evaluated the experiments using Precision (P), Recall (R), and Macro-F1 score for classification consistency, following previous work (Lin et al., 2023a). The experiments were conducted by using PyTorch and Transformers[1] library. Specifically, we used the T5 model for the English datasets

[1] https://github.com/huggingface/transformers

Dataset	TIAGE	CNTD	SuperDiaseg
Source	PersonaChat	NaturalConv	Doc2dial, MultiDoc2Dial
Language	English	Chinese	English
Dialogs	500	1308	9592
Splits	300/100/100	1041/134/133	6948/1322/1322
Turns	4092/1346/1364	19902/2558/2538	85203/16412/16006

Table 2: Dataset statistics for TIAGE, CNTD, and SuperDiaseg, respectively, where "Splits" represents the number of dialogues in the training, validation, and test sets, and "Turns" represents the number of turns requiring topic shift detection.

and the MT5 model for the Chinese CNTD dataset due to the absense of a Chinese version T5. Moreover, we trained the LLaMA-3.1-8B with LoRA fine-tuning using LLaMA-Factory[2], setting the rank r to 8, and the scaling parameter α to 16. For closed-source models, we evaluated them in a zero-shot manner. The full prompts used can be found in Appendix C. We employed AdamW optimizer with an initial warm-up stage. The training batch size is set to 4, and the learning rate is adjusted based on the dataset language: 1e-5 for English and 1e-4 for Chinese. The LLaMA model is trained for 5 epochs, while other models are trained for 20 epochs. The weights λ_1 and λ_2 for the generation components are both set to 1.

4.2 Experimental Results

Lin (Lin et al., 2023a) achieved the current state-of-the-art (SOTA) performance using a T5 model for this task. We use it as our SOTA baseline. Moreover, we introduce the following models as baselines:

1) **BERT** (Devlin et al., 2019): a bidirectional Transformer model pre-trained to generate deep semantic representations;

2) **Hier-BERT** (Lukasik et al., 2020): a hierarchical BERT, better suited for handling long documents;

3) **RoBERTa** (Liu et al., 2019): an optimized version of BERT, trained for a longer duration on more data;

4) **BART** (Lewis et al., 2020): a denoising autoencoder for pretraining sequence-to-sequence models;

5) **T5** (Raffel et al., 2020): a model that unifies all NLP tasks into a text-to-text format;

6) **LLaMA (LLaMA-3.1-8B)** (Touvron et al., 2023): a collection of foundation language models;

7) **GPT-3.5 (GPT-3.5-turbo-0125)** (Ouyang et al., 2022): a generative Transformer model trained on a larger dataset, with stronger general world knowledge, and improved context learning and reasoning capabilities;

8) **GPT-4.0 (GPT-4o)** (OpenAI, 2023): an upgraded version of GPT-3.5, with enhanced performance and capabilities, extending text input to multimodal signals.

The experimental results are shown in Table 3. Our proposed DMF achieves the best F1-scores with T5 as backbone and all improvements are significant (significance test: P <0.05). These results can be attributed to our proposed dual-process reasoning framework, which addresses issues related to the global historical structure, local detail reasoning and Fine-tune-CoT, which can distill the dual-process thinking learned by LLMs into the small generative models T5 or BART.

In comparison to the LLM LLaMA, our DMF demonstrates a notable enhancement in F1 scores, with an increase of 7.3 and 1.3 on TIAGE and CNTD, respectively. These findings suggest that our DMF, which employs dual-process reasoning, is capable of enhancing the detection of topic shifts. Despite the fact that it employs a considerably smaller number of parameters than LLMs, our small model, which utilises Fine-tune-CoT, is still capable of outperforming LLMs. Therefore, our DMF exhibits superior performance and scale advantages in comparison with LLMs.

Our DMF based on T5 demonstrates excellent performance, achieving improvements of 6.5, 4.0 and 0.9 on the TIAGE, CNTD, and SuperDialseg datasets compared to the T5 baseline, respectively. These results suggest that our proposed dual-process reasoning methodology can assist T5 in identifying topic shifts in dialogue. This is due to the fact that it introduces a global historical structure and local details to capture the difference between the response and the latest topic in a historical context.

The degree of improvement observed in the SuperDiaseg dataset is less significant in comparison to the other datasets. The reasons for this discrepancy are as follows: 1) The size of its training set is larger, which allows for the development of a high-performance base model. This, in turn, constrains the potential for performance improvement of our model. 2) This dataset was not designed with the explicit objective of detecting topic shifts. Consequently, even a relatively simple model can achieve

[2] https://github.com/hiyouga/LLaMA-Factory

Model	TIAGE			CNTD			SuperDialseg		
	P	R	F1	P	R	F1	P	R	F1
BERT	71.5	71.0	71.2	82.9	79.2	80.8	85.8	86.5	86.2
Hier-BERT	73.8	69.6	71.2	85.6	79.0	81.7	85.9	89.3	87.4
RoBERTa	77.3	77.7	77.5	84.4	75.4	78.6	86.5	89.1	87.7
BART	76.7	**80.2**	78.2	84.9	82.4	83.6	85.8	89.5	87.3
T5	76.5	72.2	73.9	83.0	79.7	81.1	86.0	89.3	87.4
LLaMA	75.1	71.7	73.1	83.9	83.6	83.8	85.9	88.9	87.2
GPT-3.5	79.5	28.9	22.7	58.6	54.2	27.8	63.5	67.6	62.9
GPT-4	64.6	63.7	64.1	85.3	70.9	74.9	63.5	67.5	63.0
Lin	73.8	77.2	76.2	85.7	**83.8**	84.7	-	-	-
DMF(BART)	80.9	77.9	79.3	**87.6**	82.9	84.9	86.1	**90.1**	87.8
DMF(T5)	**82.1**	79.0	**80.4**	86.9	83.4	**85.1**	**87.1**	89.7	**88.3**

Table 3: Results of the baselines and our DMF on TIAGE, CNTD, and SuperDialseg.

commendable results. Evaluating our model on this dataset is to demonstrate its robustness on large-scale datasets.

4.3 Ablation Study

To further elucidate the effectiveness of the global historical structure and the local detail reasoning, we conducted ablation experiments on the T5 and BART models in Table 4. The results show that both the global historical structure and the local detail reasoning contribute to improving the performance of topic shift detection.

In comparison to the T5 model, the incorporation of the global historical structure (+Historical Structure) enhances the F1 scores by 4.7, 3.3, and 0.5, respectively, on the three datasets TIAGE, CNTD, and SuperDialseg. Similarly, the global historical structure has the potential to enhance the F1 scores on the BART backbone to a limited extent. These results, particularly the notable enhancement on T5, substantiate the efficacy and generalizability of the global historical structure in topic shift detection due to the fact this mechanism can capture the connections between the response and the long-distance utterances through topic. Further analysis indicates that the improvement is attributable to the correction of samples whose responses are associated with one or more of the preceding topics.

In comparison to T5, the incorporation of local reasoning (+Local Reasoning) has been observed to enhance the F1 scores by 5.2, 3.7, and 0.7 on the three datasets, respectively. Similarly, local reasoning has also demonstrated the capacity to moderately improve the F1 scores on the BART backbone. These results, particularly the notable enhancement on T5, verify the efficacy and generalizability of local reasoning in topic shift detection.

When both the global historical structure and local reasoning are applied to T5 or BART, we can find that the performance can be further improved. This result shows that they can complement each other. Besides, it is worth mentioning that the above improvements are also due to Fine-tune-CoT, which cannot be removed from our model.

4.4 Analysis on “Shift” and “Maintain”

In the topic shift detection task, recognizing the “Shift” (1) category is more challenging in comparison to the “Maintain” (0) category. In the TIAGE, CNTD, and SuperDiaseg datasets, the “Shift” category accounts for only 21.8%, 21.8%, and 25.1%, respectively. We conducted a comparative analysis of the LLM LLaMA, the best-performing small model BART and our DMF, and the results are shown in Table 5.

The results indicate that all models consistently perform well on the “Maintain” category, with all metrics exceeding 90%. However, there is considerable variation in the performance of different models in recognizing the “Shift” category, with overall lower scores. To illustrate, the discrepancies of our DMF in F1 scores for TIAGE, CNTD, and SuperDiaseg are 23.2, 17.9, and 11.0, respectively. This primarily attributable to the fact that the majority of the “maintain” category and the relatively straightforward recognition of utterance relations within the same topic.

In regard to the “Shift” category, our DMF demonstrates superior performance compared to BART, with the improvements of 2.0, 2.0, and 1.2, respectively. Similarly, it exhibits a notable lead over LLaMA, with the improvements of 11.8,

Model	TIAGE			CNTD			SuperDialseg		
	P	R	F1	P	R	F1	P	R	F1
BART	76.7	**80.2**	78.2	84.9	82.4	83.6	85.8	89.5	87.3
+Historical Structure	79.1	77.7	78.3	85.3	82.8	84.0	**86.4**	88.8	87.5
+Local Reasoning	79.0	78.7	78.9	85.0	**83.2**	84.0	85.7	90.1	87.5
DMF(BART)	**80.9**	77.9	**79.3**	**87.6**	82.9	**84.9**	86.1	**90.1**	**87.8**
T5	76.5	72.2	73.9	83.0	79.7	81.1	86.0	89.3	87.4
+Historical Structure	79.1	78.2	78.6	86.8	82.6	84.4	86.6	89.6	87.9
+Local Reasoning	80.5	77.9	79.1	**88.7**	82.2	84.8	86.8	**89.7**	88.1
DMF(T5)	**82.1**	**79.0**	**80.4**	86.9	**83.4**	**85.1**	**87.1**	89.7	**88.3**

Table 4: Ablation study on TIAGE, CNTD, and SuperDialseg using BART and T5.

Model	TIAGE						CNTD						SuperDiaseg					
	Maintain (1066)			Shift (298)			Maintain (1985)			Shift (553)			Maintain (11986)			Shift (4020)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BART	**92.1**	87.2	89.5	61.4	**73.2**	66.8	92.2	93.9	93.1	76.7	71.6	74.1	**95.9**	90.4	93.1	75.6	**88.6**	81.6
LLaMA	87.2	91.5	89.3	63.0	52.0	57.0	92.8	93.1	92.9	75.0	**74.1**	74.5	95.2	91.4	93.1	76.6	86.3	81.2
DMF (T5)	90.4	**93.6**	**92.0**	**73.9**	64.4	**68.8**	**92.9**	**95.2**	**94.0**	**81.0**	71.7	**76.1**	95.6	**92.1**	**93.8**	**78.6**	87.4	**82.8**

Table 5: Performance comparison on the topic "Shift" and "Maintain" categories.

1.6, and 1.6 on TIAGE, CNTD, and SuperDiaseg, respectively. These figures are greater than the overall improvements presented in Table 3. This suggests that the model has a preference for detecting shifts and that enhancing the model's capacity to recognize the "Shift" category in data with a low proportion of instances.

A comparison of the performance of our DMF in the "Shift" category reveals that the F1 score on TIAGE is the lowest, at 68.8, while that on SuperDiaseg is the highest, reaching 82.8. This discrepancy is primarily attributable to the fact that SuperDiaseg has access to a considerably larger training dataset and engages in longer dialogues.

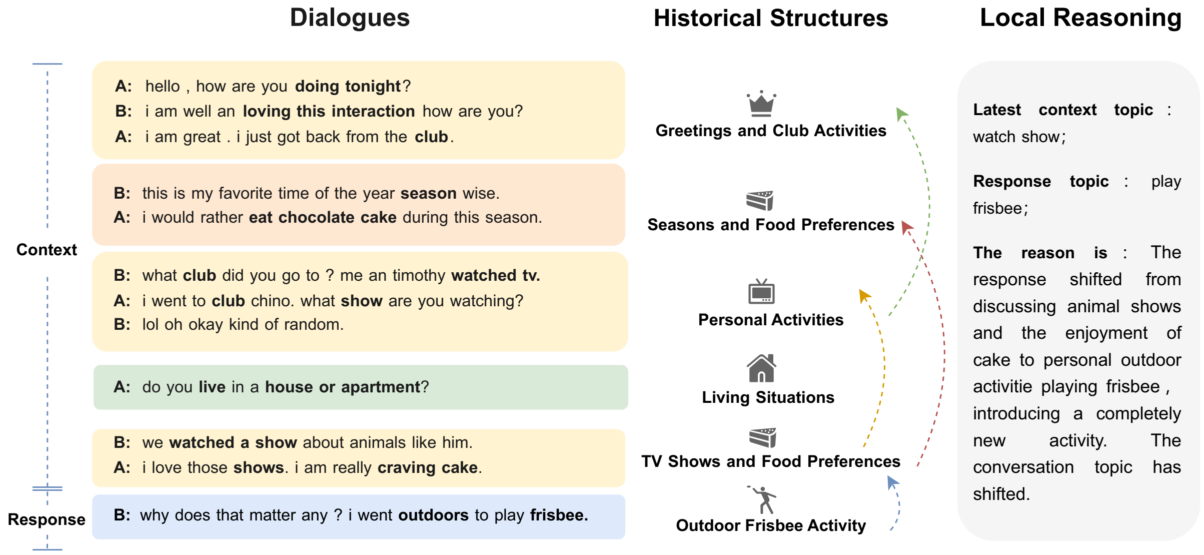
4.5 Case Study

Figure 3 presents a case of dialogue topic shift detection. In this case, the global historical structure captures the various subtopics within the conversation, tracing the overall trajectory of topic evolution. Each topic shift can be identified by referring to the historical structure, which recalls previous discussion themes, while the local reasoning focuses on the current context, detecting changes in conversational details.

For instance, in the final turn, local reasoning identifies a topic shift from discussing animal shows and expressing a preference for cake to outdoor activities, specifically playing frisbee, marking a clear change in topic. The historical structure, meanwhile, draws on earlier mentions of TV shows and seasonal preferences to provide global background. By integrating an understanding of the conversation's overarching thematic progression with real-time analysis of local topic shifts, our DMF achieves greater accuracy in detecting changes in conversation topics. This dual-process approach maintains global coherence while enhancing the capture of local relevance, ultimately improving topic shift detection in dialogue systems.

4.6 Error Analysis

In the task of dialogue topic shift detection, the model tends to misclassify casual conversations following opening greetings, such as "hello, how are you tonight?". This issue is evident across two datasets (e.g., TIAGE and CNTD). After a few greetings, new nouns or slight topic changes are often introduced, which the model frequently misidentifies as a topic shift. The core issue lies in the model's reliance on local lexical changes, failing to adequately capture semantic coherence, particularly during the natural transition between greetings and formal discussions. As a result, the model incorrectly labels subtle changes within the same topic as a topic shift. To address this, it is necessary to improve the model's ability to understand context, especially in modeling the transition



A: hello , how are you **doing tonight**?
B: i am well an **loving this interaction** how are you?
A: i am great . i just got back from the **club**.

B: this is my favorite time of the year **season** wise.
A: i would rather **eat chocolate cake** during this season.

B: what **club** did you go to ? me an timothy **watched tv.**
A: i went to **club** chino. what **show** are you watching?
B: lol oh okay kind of random.

A: do you **live** in a **house or apartment**?

B: we **watched a show** about animals like him.
A: i love those **shows**. i am really **craving cake**.

B: why does that matter any ? i went **outdoors** to play **frisbee.**

Latest context topic : watch show;

Response topic : play frisbee;

The reason is : The response shifted from discussing animal shows and the enjoyment of cake to personal outdoor activitie playing frisbee , introducing a completely new activity. The conversation topic has shifted.

Figure 3: A specific example from the TIAGE test set. The local reasoning corresponds to the final turn of dialogue. The arrows in the historical structure indicate the backtracking of topics, while the same background color in the dialogue denotes the discussion of the same topic.

between greetings and formal topics, in order to reduce such misclassifications.

5 Conclusion and Future Work

In this study, we simulate the dual-thinking process to explore the topic shift task. We utilize LLMs to extract global historical structure as a repository for intuitive judgments, while inferring the reasons for local topic shifts as detailed analysis. Subsequently, we employ a small model as the student model, using Fine-tune-CoT to learn the dual-thinking process and perform supervised training to enhance the final determination of transition relations. Our experiments on three datasets demonstrate the effectiveness of our proposed method. Future work will continue to explore fine-grained relations in the topic shift detection task.

Limitations

Our approach still has limitations in handling natural transitions, such as those between greetings and minor topics. In particular, the model's ability to detect subtle topic shifts and accurately identify the "Shift" category requires further optimization. From the perspective of topic shift granularity, the relationships between topic blocks need to be more finely delineated. Moreover, the quality of historical structure summarization and reasoning performed by LLMs remains suboptimal, leaving significant room for improvement. In the future, we aim to address these limitations through targeted enhancements.

Acknowledgements

The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62276177 and 62376181), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A survey on proactive dialogue systems: Problems, methods, and prospects. In *IJCAI*, pages 6583–6591.

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

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 *SIGIR*, pages 2481–2485.

Prakhar Gupta, Harsh Jhamtani, and Jeffrey P. Bigham. 2022. Target-guided dialogue response generation using commonsense and data augmentation. In *NAACL*, pages 1301–1317.

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen, and Bing Qin. 2024. Planning like human: A dual-process framework for dialogue planning. In *ACL*, pages 4768–4791.

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

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *ACL*, pages 14852–14882.

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. In *EMNLP*, pages 17682–17702.

Junfeng Jiang, Chengzhang Dong, Sadao Kurohashi, and Akiko Aizawa. 2023. Superdialseg: A large-scale dataset for supervised dialogue segmentation. In *EMNLP*, pages 4086–4101.

Daniel Kahneman. 2003. Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93:1449–1475.

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 *ACL*, pages 7871–7880.

Jiangyi Lin, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2023a. Multi-granularity prompts for topic shift detection in dialogue. In *ICIC*, pages 511–522.

Jiangyi Lin, Yaxin Fan, Feng Jiang, Xiaomin Chu, and Peifeng Li. 2023b. Topic shift detection in chinese dialogues: Corpus and benchmark. In *ICDAR*, pages 166–183.

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 *EMNLP*, pages 4707–4716.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*, pages 27730–27744.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Yingcheng Sun and Kenneth A. Loparo. 2019. Topic shift detection in online discussions using structural context. *COMPSAC*, 1:948–949.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. In *AAAI*, pages 14006–14014.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann A. Copestake. 2021. TIAGE: A benchmark for topic-shift aware dialog modeling. In *EMNLP*, pages 1684–1690.

Linzi Xing and Giuseppe Carenini. 2021. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In *SIGdial*, pages 167–177.

Zhitong Yang, Bo Wang, Jinfeng Zhou, Yue Tan, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2022. Topkg: Target-oriented dialog via global planning on knowledge graph. In *COLING*, pages 745–755.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*, pages 2204–2213.

A Prompts of Extracting Global Historical Structures

We use GPT-4 to extract the global historical structure from the segmented context of each training sample. Giving the segmented history, we design the specific prompts as follows.

---

*You are engaging in the task of automatic assistant Dialogue History Topic Detection based on dialogue content. Please detect the theme of each segmented part of the following dialogue history and provide a brief output. Just output the specific theme content without explanation. Dialog content START: **{SegmentedHistory:}** Dialog content END.*

SegmentedHistory:

1
A: (Label==0)
B: (Label==0)
A: (Label==0)

2
B: (Label==1)
A: (Label==0)
B: (Label==0)
......

B Prompts for Local Reasoning

Before training, we use GPT-4 to perform local reasoning on the context, replies, and label samples to identify the reasons for topic shifts. The prompts given to GPT-4 are as follows.

Given a dialogue context:
Dialog Context START
A:
B:
A:
B:
...
Dialog Context END
A current response:
***Response START** A/B:*
***Response** END*
and a label:
***Label** 0/1*
*You need to extract the topic of context and response, and briefly provide the reason for the topic shift or no shift. **{Instruction:}** The brief output format is as follows: The latest context topic: ; response topic: ; The reason is: .*
If label == 1:
***Instruction** = "The topic has shifted. Please explain the shift."*
else:
***Instruction** = "The topic has not shifted. Please explain why the topic remains the same."*

C Prompts for GPT as a Baseline

When we use GPT-3.5 and GPT-4 as baselines for the topic shift classification task under a zero-shot setting, the prompts used is as follows.

You are engaging in the task of automatic assistant topic shift detection. Dialogue topic shift detection refers to the task of detecting a shift in the topic when given a dialogue context and a new response. In the provided dialogue below, determine whether a topic shift has occurred. If a topic shift has occurred, output 1; if there has been no topic shift, output 0.
Dialog Context START
A:
B:
......
Dialog Context END
Response START
A/B:
Response END