MP2D: An Automated Topic Shift Dialogue Generation Framework Leveraging Knowledge Graphs

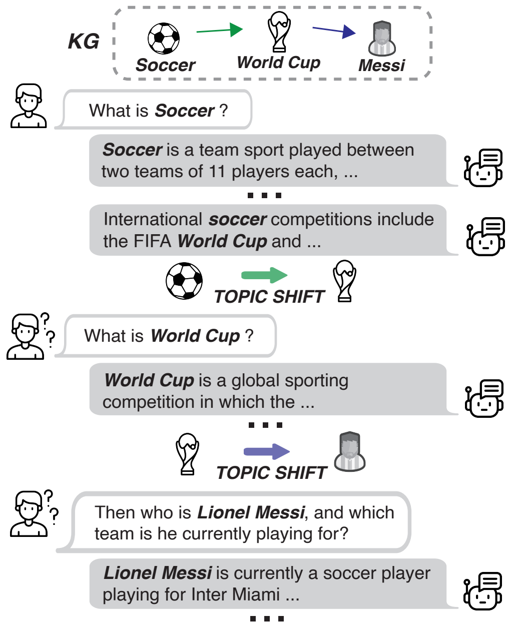
Yerin Hwang[1] **Yongil Kim**[2] **Yunah Jang**[3]
Jeesoo Bang[2] **Hyunkyung Bae**[2] **Kyomin Jung**[1,3,4†]
[1]IPAI, Seoul National University [2]LG AI Research
[3]Dept. of ECE, Seoul National University [4]SNU-LG AI Research Center
{dpfls589, vn2209, kjung}@snu.ac.kr
{yong-il.kim, jeesoo.bang, hkbae}@lgresearch.ai

Abstract

Despite advancements in on-topic dialogue systems, effectively managing topic shifts within dialogues remains a persistent challenge, largely attributed to the limited availability of training datasets. To address this issue, we propose Multi-Passage to Dialogue (MP2D), a data generation framework that automatically creates conversational question-answering datasets with natural topic transitions. By leveraging the relationships between entities in a knowledge graph, MP2D maps the flow of topics within a dialogue, effectively mirroring the dynamics of human conversation. It retrieves relevant passages corresponding to the topics and transforms them into dialogues through the passage-to-dialogue method. Through quantitative and qualitative experiments, we demonstrate MP2D's efficacy in generating dialogue with natural topic shifts. Furthermore, this study introduces a novel benchmark for topic shift dialogues, TS-WikiDialog. Utilizing the dataset, we demonstrate that even Large Language Models (LLMs) struggle to handle topic shifts in dialogue effectively, and we showcase the performance improvements of models trained on datasets generated by MP2D across diverse topic shift dialogue tasks.

1 Introduction

Dialogue systems (Chen et al., 2017; Ni et al., 2023), designed to respond to inquiries or provide relevant information, have gained considerable attention in academia and industry. These systems show promise in various applications, including virtual assistants, customer service, and chatbots (King, 2023; Zeng et al., 2024; Liu et al., 2024). Nonetheless, while numerous studies have advanced the field of conversational systems within on-topic dialogue scenarios, significant challenges persist concerning dialogues encompassing topic shifts (Holtzman et al., 2019; Zhang et al., 2019; Xu et al., 2021).



What is *Soccer* ?

Soccer is a team sport played between two teams of 11 players each, ...

...

International *soccer* competitions include the FIFA *World Cup* and ...

What is *World Cup* ?

World Cup is a global sporting competition in which the ...

...

Then who is *Lionel Messi*, and which team is he currently playing for?

Lionel Messi is currently a soccer player playing for Inter Miami ...

...

Figure 1: An example of a topic shift dialogue. The MP2D framework utilizes paths in a Knowledge Graph (KG) to extract entities and facilitates natural topic transitions based on the relations between these entities.

Unlike human conversation, which can naturally flow between topics and continue the discussion seamlessly, current dialogue systems often struggle to determine the appropriate timing for topic shifts or to execute these shifts fluently (Xie et al., 2021). A major challenge in these tasks lies in the scarcity of data. This issue is compounded by the fact that creating all existing topic shift dialogue datasets (Xu et al., 2021; Sevegnani et al., 2021) involves a laborious and costly human annotation process.

To address this challenge, we propose a novel framework, Multi-Passage to Dialogue (MP2D), specifically crafted for the automatic generation of Conversational Question-Answering (ConvQA) data featuring topic shifts. The framework lever-

[†] Corresponding author

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17682–17702
November 12-16, 2024 ©2024 Association for Computational Linguistics

ages the Passage-to-Dialogue (P2D) method (Dai et al., 2022; Hwang et al., 2023) in ConvQA dataset generation, which entails generating relevant questions by using the sentences within a passage as answers. Unlike existing methods that construct dialogues from a single passage, our proposed framework integrates multiple passages to create dialogues with topic shifts. To emulate the dynamic nature of real-world conversations, we employ a knowledge graph to identify paths connecting various entities and their relations, as shown in Figure 1. By retrieving passages using entities in the path as queries, the passages and the relation sentences between entities form a multi-passage structure. This structure is then segmented into sentences, which directly serve as answers, and a question generator generates suitable questions to complete dialogues with natural topic transitions.

We conduct experiments using various question generators to transform multi-passage into dialogue, employing diverse reference-free dialogue metrics for automatic evaluation. The results indicate that the generated topic shift dialogue datasets demonstrate high quality when utilizing Large Language Models (LLMs) (Brown et al., 2020) as a question generator within MP2D. Moreover, through qualitative assessments conducted via human and GPT-4 evaluation (Liu et al., 2023; Wang et al., 2023), we found that approximately 91% of the generated dialogues are deemed excellent in their handling of topic shifts.

Furthermore, we introduce a novel topic shift dialogue benchmark, TS-WikiDialog, to evaluate LLMs in handling topic shift dialogue tasks and to broaden the application of datasets generated via MP2D. The benchmark is constructed by paraphrasing the MP2D-generated dialogue by LLMs and human annotators. In experiments utilizing the TS-WikiDialog, we initially verify that various LLMs struggle to effectively address problems related to topic segmentation (Purver, 2011), topic shift detection (Holz and Teresniak, 2010), and topic shift ConvQA (Xie et al., 2021). Furthermore, the T5-base model (Raffel et al., 2020), trained on the MP2D-generated dataset, exhibits significantly better performance in topic segmentation and topic shift detection tasks compared to baseline LLMs. This validates the MP2D framework as an effective dataset-creation approach for tackling topic shift challenges. Additionally, we demonstrate that utilizing finetuned models on MP2D-generated datasets for the topic shift detection task can enhance the performance of LLM responses in ConvQA, especially in topic shift turns.

2 Related Works

2.1 Topic Shift Dialogue Systems

Topic shift dialogue refers to instances within multi-turn dialogues where the focus of the conversation changes mid-discussion (Garcia and Joanette, 1997). Such transitions are a natural aspect of human interaction. According to Soni et al. (2021), a change in topic happens every 12 conversational turns. Particularly in ConvQA, which aims to provide information to users (Zaib et al., 2022), the phenomenon of users changing topics is more commonly observed (Spink et al., 2002) and presents a significant challenge (Zaib et al., 2023). Topic transitions in ConvQA often occur because users pose follow-up questions to explore new curiosities that stem from previous answers.

Dialogue systems proficient in handling topic shifts must effectively detect such transitions (Galley et al., 2003; Somasundaran et al., 2020) and maintain the quality of their responses thereafter (Wang et al., 2021). However, current conversational systems struggle with this task (Holtzman et al., 2019; Xie et al., 2021). Despite the data scarcity problem being a major challenge for topic shift dialogue or ConvQA tasks, all existing frameworks for generating topic shift dialogue datasets (Xie et al., 2021; Yang et al., 2022) involve a process of human annotation. These methods have limitations: 1. They are prone to instability due to the subjective criteria on the scope of topics (Galley et al., 2003), and 2. they are labor-intensive and time-consuming. To the best of our knowledge, this study is the first to propose a framework capable of automatically generating a ConvQA dataset with natural topic transitions.

2.2 Passage to Dialogue (P2D)

Recently, among various research efforts addressing the issue of data scarcity in ConvQA (Dalton et al., 2020; Anantha et al., 2020; Kacupaj et al., 2021; Mulla and Gharpure, 2023), the Passage to Dialogue (P2D) frameworks hold potential as they enable the generation of dialogues from textual sources without loss of information. These frameworks segment passages into sentence units to function as "answers" and employ question generators trained on task-specific objectives to generate corresponding "questions" for each "answer." Dai et al.

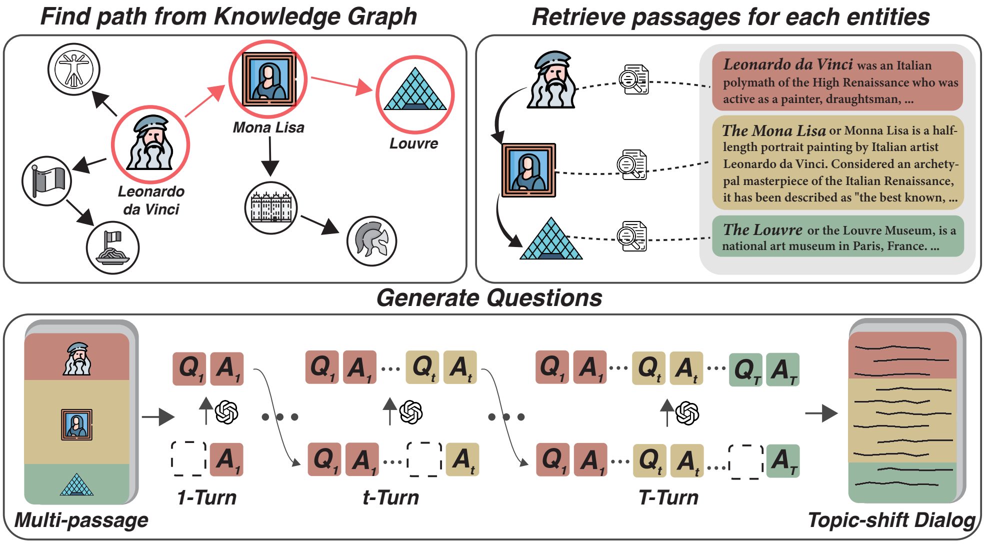

Find path from Knowledge Graph
Retrieve passages for each entities
Mona Lisa
Louvre
Leonardo da Vinci
Leonardo da Vinci was an Italian polymath of the High Renaissance who was active as a painter, draughtsman, ...
The Mona Lisa or Monna Lisa is a half-length portrait painting by Italian artist Leonardo da Vinci. Considered an archetypal masterpiece of the Italian Renaissance, it has been described as "the best known, ...
The Louvre or the Louvre Museum, is a national art museum in Paris, France. ...
Generate Questions
Multi-passage
1-Turn
t-Turn
T-Turn
Topic-shift Dialog

Figure 2: An overview of the MP2D framework. In the knowledge graph, paths are identified and passages are retrieved for entities within those paths. Then, the retrieved passages and their relations become the *"answers"*, and a LLM generates *"questions"* corresponding to each answer to create dialogues.

(2022) first proposed a P2D framework, employing a question generator trained through a dialogue reconstruction task. Additionally, Hwang et al. (2023) introduces supplementary tasks capable of learning sentence-level alignment, presenting a P2D framework that prioritizes contextual relevance.

These methods can transform specialized passages into dialogues without any loss of information, even without the participation of a domain expert. However, their application is limited to converting a single passage into a dialogue, lacking the capability to manage topic shifts or control the flow of the generated dialogues. This study introduces the first automatic framework designed to transform multi-passage content into dialogues with natural topic shifts.

3 Multi-Passage to Dialogue (MP2D)

The Multi-Passage to Dialogue (MP2D) framework represents a novel approach tailored to generating dialogues with topic shifts by incorporating multiple textual passages. Through the utilization of a knowledge graph, MP2D extracts paths to establish natural linkages among topics, facilitating smooth transitions within the dialogue. The framework retrieves passages by querying each entity within these paths, forming a multi-passage structure by pairing these passages with sentences that describe the relationships between entities (§3.1). Subsequently, it segments the multi-passage into sentence-sized units to serve as answers. Leveraging the capability of LLMs to generate contextually appropriate questions for each answer, MP2D employs LLMs as question generators in crafting dialogues (§3.2). An automatic post-processing phase follows, resulting in dialogues that seamlessly transition between topics. Figure 2 provides an illustrative overview of the proposed framework.

3.1 Find Path & Retrieve Passages

In our methodology's initial phase, we navigate a knowledge graph to establish connections among diverse entities, determining discourse subjects. We utilize the KELM dataset (Agarwal et al., 2020), structured around an entity-driven knowledge graph. Within this graph, knowledge is arranged into triplets (S, r, O): a subject entity S (e.g., "Leonardo da Vinci"), a relation r (e.g., "painted"), and an object entity O (e.g., "Mona Lisa"). The dataset additionally contains a relation sentence, denoted as R, which provides a detailed narrative of the relationship between two entities, offering more context than the simple relation r. The MP2D framework leverages this component to reconstruct the knowledge graph $K := (S_i, R_i, O_i)_{i=1}^{N}$ of N factual triplets.

The framework randomly selects a particular (S, R, O^*) triplet, facilitating the connection of more than two entities by identifying an (O^*, R, O) where O^* becomes the subject. This process is conducted auto-regressively to construct the finite walk ϕ, iterating until it reaches a point where no further triplets exist for the last object to transition into a subject. Consequently, the walk ϕ is represented as a path $\{S_1, R_1, O_1 = S_2, R_2, O_2 =$

$S_3, ..., O_n\}$, and by merging subjects and objects as entities denoted as e, the path is expressed as $\phi = \{e_1, R_1, e_2, R_2, ..., e_n\}$.

For each entity e_i identified along the path, using it as a query q_i results in the sequential retrieval of passages p_i, as described in Figure 2 (top-right). Each retrieved passage p_i consists of m_i sentences, denoted as $p_i = \{s_1, s_2, ..., s_{m_i}\}$. To prepare these passages for processing, we truncate each to a maximum length of k_i sentences, resulting in a truncated passage $p_i^{\dagger} = \{s_1, s_2, ..., s_{k_i}\}$, where $k_i = \min(m_i, \text{random}(3, 6))(k_i \ll m_i)$. Ultimately, alongside the existing relation sentence R for a natural connection between the entities, a multi-passage $\mathit{MP} = \{p_1^{\dagger}, R_1, p_2^{\dagger}, R_2, ..., p_n^{\dagger}\}$ is constructed.

The utilization of a knowledge graph in constructing the dialogue flow provides several advantages. Primarily, knowledge graphs inherently encapsulate relationships between various entities, offering a structured method for understanding the context surrounding each entity. In the MP2D framework, this facilitates the selection of topics that are not only pertinent to the ongoing dialogue but also interconnected in meaningful ways. This interconnectedness ensures that the dialogue flows logically from one topic to the next, mirroring natural human conversations where topics shift smoothly based on underlying relationships.

Moreover, employing a knowledge graph for the automatic generation of dialogues confers the distinct advantage of readily producing conversations that are current and up-to-date. As information evolves over time, manually updating dialogues can be challenging due to their inherently unstructured nature. However, knowledge graphs are dynamic entities that expand and adapt over time, with considerable research dedicated to enhancing them (Paulheim, 2017; Cohen et al., 2023). Since MP2D constructs dialogue flows using knowledge graphs, it becomes feasible to automatically generate dialogues using the latest version of the knowledge graph, ensuring that the dialogues remain current and relevant to time-variant information.

3.2 Generate Questions

Using the retrieved multi-passage *MP*, a Passage-to-Dialogue (P2D) model is employed to auto-regressively generate questions for each sentence within the passage $p_i^{\dagger} = \{s_1, s_2 ... s_{k_i}\}$ as an answer, thereby creating a per-passage dialogue, as illustrated at the bottom of Figure 2. In essence, for a given input passage $p_i^{\dagger}$, the output dialog from the P2D model is $D_i = \{(q_1, s_1), (q_2, s_2), \ldots, (q_{k_i}, s_{k_i})\}$, where q_j represents the generated question based on the answer s_j. Subsequently, it naturally transitions to the subsequent topic through R_i while seamlessly incorporating questions Q_{R_i} about R_i and repeats this process iteratively. Ultimately, the multi-passage $\mathit{MP} = \{p_1^{\dagger}, R_1, p_2^{\dagger}, ..., p_n^{\dagger}\}$ is transformed into a dialogue that encapsulates natural topic shifts in the form $\{D_1, Q_{R_1}, R_1, D_2, ..., D_n\}$.

We employ an LLM as the question generator for MP2D. This decision is based on observations from comparing datasets generated by various P2D models and LLMs, which indicate that LLMs may generate questions for topic shift turns more effectively. One possible approach to utilize LLMs as question generators is performing a dialogue reconstruction task by filling in [BLANK] without providing a specific prompt. However, this approach does not ensure the generation of contextually relevant questions for topic shift turns or subsequent topics. We often find that, despite the answer for the topic-shift turn including information about a new topic, the generated questions still pertain to the previous topic without recognizing the shift (§7). Therefore, during the question-generation process, an additional instruction indicating a change in topic is provided in topic shift turns; *"Note that the conversation topic has shifted to [next_topic] from [current_topic].* Detailed information, including the prompt for question generation and the post-processing steps, can be found in Appendix C.

4 Evaluating the MP2D Framework

In this section, we empirically demonstrate that the MP2D framework is capable of automatically generating high-quality dialogue with smooth topic shifts, both quantitatively (§4.1) and qualitatively (§4.2). First, we employ existing P2D methods and the LLM as question generators to compare these approaches using various reference-free dialogue metrics. Within the MP2D framework's passage retrieval component, we employ the Wikipedia dataset to generate 10,000 multi-turn dialogues, subsequently evaluating their quality. Furthermore, absolute evaluations are conducted using various criteria to illustrate that the generated topic shift dialogues exhibit a natural transition of topics.

	USR-DR (c)	USR-DR (f)	GPT2	QRelScore$_{\text{LRM}}$	QRelScore$_{\text{GRG}}$	RQUGE
Single Passage						
Dialog Inpainter (Dai et al., 2022)	0.9615	0.7227	0.5125	0.4887	0.4808	3.1255
Dialogizer (Hwang et al., 2023)	0.9641	0.7883	0.5386	0.5044	0.4852	3.2511
GPT-3.5	**0.9856**	**0.8960**	**0.5739**	**0.5369**	**0.5305**	**3.2923**
Multiple Passages						
Dialog Inpainter (Dai et al., 2022)	0.9389	0.7160	0.4972	0.4874	0.4732	2.9156
Dialogizer (Hwang et al., 2023)	0.9474	0.7738	0.5034	0.5098	0.4748	2.9363
GPT-3.5 (MP2D)	**0.9873**	**0.9199**	**0.5746**	**0.5366**	**0.5437**	**3.1034**

Table 1: Automatic evaluation results obtained by assessing the generated dialogues using reference-free metrics.

4.1 Automatic Evaluation

Passage to Dialogue Methods We conduct experiments to compare Dialog Inpainter (Dai et al., 2022), Dialogizer (Hwang et al., 2023), and the LLM as question generators. To ensure a fair comparison, we implement Dialog Inpainter and Dialogizer to align with the specific framework requirements. Both models utilize T5-base (Raffel et al., 2020) as their backbone and are trained with four datasets: Task Masker (Byrne et al., 2019), Daily Dialog (Li et al., 2017), OR-QUAC (Qu et al., 2020), and QReCC (Anantha et al., 2020). For LLM, we employ GPT-3.5 and integrate the instructions for topic shift turns to generate questions.

First, we conduct a basic comparison of the question generation performance for single passages by applying these three question generators to randomly retrieved passages from Wikipedia. Additionally, we compare the three models by employing them to transform the multi-passage content into dialogues.

Evaluation Metrics Given the one-to-many nature of dialogue systems (Zhao et al., 2017), reference-free metrics are acknowledged for their stronger correlation with human judgment compared to reference-based generation metrics (Gupta et al., 2019; Zhang et al., 2021) when assessing dialogue quality. Furthermore, as MP2D functions as a data generation framework, we utilize various reference-free metrics to evaluate multiple aspects of dialogue and ConvQA to assess the performance of MP2D quantitatively. Primarily, **USR-DR** (Mehri and Eskenazi, 2020) is a reference-free dialogue metric for evaluating dialogues on context maintenance, interest, and knowledge utilization. We utilize USR-DR(c), a metric that evaluates dialogues based on history and facts as inputs, along with USR-DR(f), which assesses dialogues using fact information or context as inputs. Additionally, **GPT-2** based metric (Pang et al., 2020) evaluates dialogs to assess the coherence between utterances. Furthermore, **RQUGE** (Mohammadshahi et al., 2022) assesses question answerability given the context, while **QRelScore** (Wang et al., 2022) evaluates context-aware question generation without extra training or human supervision. QRelScore is divided into QRelScore$_{\text{LRM}}$, which evaluates complex reasoning through word-level similarity analysis, and QRelScore$_{\text{GRG}}$, assessing factual accuracy by examining the confidence in generating contextually relevant content.

Results Table 1 compares three question generation methods in single- and multi-passage settings. In the single passage setting, we aim to evaluate the performance of question generators within a basic single passage context by randomly selecting an entity without the assistance of knowledge graph pathways for passage retrieval. The results indicate that using GPT-3.5 for question generation outperforms both Dialog Inpainter and Dialogizer across all metrics.

Moving to the multi-passage setting, which introduces additional complexity due to the need to manage topic transitions and generate responses that effectively bridge these shifts, GPT-3.5 consistently exhibits superior performance. Metrics assessing the dialogue's overall context or the relevance of question-answer (QA) pairs indicate a performance decline in finetuned models relative to the single-passage setting. Conversely, GPT-3.5-generated datasets maintain high performance even in multi-passage settings, suggesting its robustness in handling more complex question-generation tasks. However, in the case of the RQUGE, which measures the contextual relevance between the given passage and target QA pair, there is a decline in all methods when applied to multi-passage contexts. This reduction is attributed to the broader

	Human	GPT-4
Topic shift timing *Is the timing of topic-shifts is natural?*	95.67%	93%
Topic shift fluency *Does the topic shifts occur smoothly?*	87.67%	88%
Overall quality *Is the overall quality of the dialog is good?*	84.33%	89%
Toxicity *Does the whole text has any potential risk?*	0%	0%

Table 2: The human and GPT-4 evaluation results.

range of topics covered by multi-passage content as opposed to single-passage content, which leads to a lower relevancy score between the multi-passage content and an individual QA pair.

In summary, when creating dialogues from multi-passage content, LLM proves to be more effective as a question generator than task-specific finetuned models. Furthermore, MP2D demonstrates a robust ability to generate dialogues of comparable quality to those created from a single passage while incorporating topic shifts within the dialogue.

4.2 Human & GPT-4 Evaluation

To qualitatively assess the quality of topic shift dialogue data generated by MP2D, we employ human and GPT-4 (OpenAI, 2023) as evaluators. GPT-3.5 serves as the question generator for MP2D, and from the previously generated 10,000 dialogues, we randomly select 100 sample dialogues to form the evaluation dataset. The evaluation focuses on three criteria: the timing of the topic shift, the naturalness of the topic shift, and the overall quality of the dialogue.

Table 2 presents the human evaluation results, indicating that over 95.6% of the MP2D-generated topic shift dataset exhibits timely topic shifts, with more than 87.6% of the dialogues demonstrating smooth topic transitions, and over 84.3% rated as excellent in overall quality. Moreover, experimental results using GPT-4 as an evaluator (Liu et al., 2023; Gilardi et al., 2023) reflect a similar trend to that of the human evaluation. An analysis of dialogues rated with unnatural topic shifts reveals that such instances often arise from abrupt transitions between entities at different levels of specificity within the knowledge graph. A comprehensive case study on this observation is provided in Section 7. Beyond the primary evaluation criteria, we also screen the evaluation dataset for toxicity and confirm the absence of toxic content in any dialogues. More detailed information, including inter-annotator agreement, compensation details, instructions, and prompts, can be found in Appendices E and F.

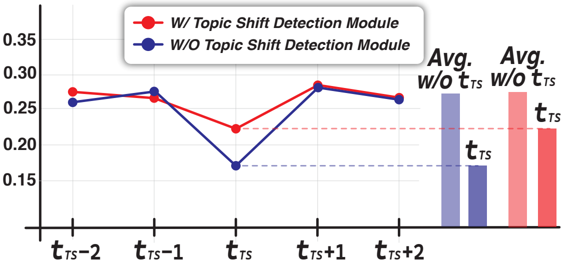


Figure 3: Results of the ConvQA response generation performance of GPT-3.5. Each score represents the BLEU-4 score, where t_{TS} denotes a topic shift turn.

5 Topic Shift Dialogue Benchmark

5.1 TS-WikiDialog

We introduce TS-WikiDialog, a benchmark designed to evaluate LLMs on topic shift tasks and to illustrate the usefulness of MP2D. The data generated by the P2D methods cannot be directly utilized for tasks such as ConvQA response generation because each answer sequentially spans across the passage. Therefore, to adapt the benchmark to the ConvQA response generation task, we paraphrase the answers in the MP2D-generated dialogues accordingly. The paraphrasing process involves providing GPT-4 with the dialogue history and target answer for paraphrasing, followed by manual adjustments by humans to ensure the context is naturally preserved. Furthermore, the few instances of unnatural topic flow present in the automatically generated data are filtered or revised to ensure a more natural progression. TS-WikiDialog encompasses 1,000 multi-turn dialogues, totaling 15,892 turns. Each dialogue covers approximately 2.136 topics, and more detailed statistical analysis can be observed in Appendix A.

5.2 The Struggle of LLM in Topic Shift turns

Using TS-WikiDialog, we assess how well the LLM maintains its response generation quality in the face of topic transitions. For the ConvQA response generation experiment, we set multi-passage content, dialogue history, and the target question as the input for the GPT-3.5 and evaluate the answers generated by the model. The results can be observed in Figure 3 (blue line), where each score represents the BLEU-4 score (Papineni et al.,

	Acc. *(per turns)*	***F1***	***P***	***R***	***EM***
Zero-shot (w/o In-Context Learning)					
GPT-3.5 (175B)	0.720	0.701	0.751	0.719	0.000
GPT-4 (>175B)	0.254	0.222	0.826	0.277	0.000
Few-shot (w/ In-Context Learning)					
GPT-3.5 (175B)	0.757	0.713	0.784	0.741	0.000
GPT-4 (>175B)	0.275	0.246	0.780	0.283	0.000
MP2D Random-shift					
T5-base (220M)	0.825	0.739	0.843	0.809	0.016
FLAN-T5-base (220M)	0.824	0.741	0.841	0.808	0.035
MP2D Knowledge-Graph (Ours)					
T5-base (220M)	**0.841**	**0.803**	**0.885**	**0.834**	**0.225**
FLAN-T5-base (220M)	**0.841**	0.792	0.873	0.833	0.205

Table 3: The results of Topic Shift Detection Task

	Acc. *(per turns)*	***F1***	***P***	***R***	***EM***
Zero-shot (w/o In-Context Learning)					
GPT-3.5 (175B)	0.720	0.701	0.751	0.719	0.000
GPT-4 (>175B)	0.254	0.222	0.826	0.277	0.000
Few-shot (w/ In-Context Learning)					
GPT-3.5 (175B)	0.757	0.713	0.784	0.741	0.000
GPT-4 (>175B)	0.275	0.246	0.780	0.283	0.000
MP2D Random-shift					
T5-base (220M)	0.825	0.739	0.843	0.809	0.016
FLAN-T5-base (220M)	0.824	0.741	0.841	0.808	0.035
MP2D Knowledge-Graph (Ours)					
T5-base (220M)	**0.841**	**0.803**	**0.885**	**0.834**	**0.225**
FLAN-T5-base (220M)	**0.841**	0.792	0.873	0.833	0.205

Table 4: The results of Topic Shift Detection Task

2002) between the candidate answer and the reference. We observe a decrease in the performance of GPT-3.5 in topic shift turns (t_{TS}), emphasizing that even LLMs face challenges in managing topic transitions.

6 Applications

To illustrate the efficacy of MP2D, we evaluate the performance of models trained on datasets produced by the MP2D framework in two topic shift tasks: topic segmentation and topic shift detection. Since MP2D is the first framework for automatically generating topic shift ConvQA data, we experiment with baseline models that include zero-shot and few-shot settings of LLMs, as well as models trained on dialogues constructed from randomly selected topics without the use of knowledge graphs (*MP2D Random-shift*). Our analysis aims to validate that the MP2D framework is an effective data generation tool, producing datasets that enable models to achieve high performance on diverse topic shift tasks. We generate 10,000 dialogues using the MP2D framework for training purposes and employ TS-WikiDialog as the test set, ensuring there is no overlap in topics or retrieved passages between the training and test sets. Further details, including the prompts for each task, can be found in Appendix G.

6.1 Topic Segmentation

The objective of topic segmentation is to partition the given dialogue into segments based on the topics being discussed (Arguello and Rosé, 2006). Given a dialogue D composed of utterances $\{u_1, u_2, ..., u_n\}$, where u_i represents an utterance, the model's output would be a sequence of segment labels, for example, $\{0, 0, 1, 1, ..., 2\}$, with each numeral indicating a distinct topic segment. The topic segmentation task has the potential to enhance the performance of tasks such as information retrieval and dialogue summarization by facilitating the identification of topic boundaries within a conversation (Lin et al., 2023a; Gao et al., 2023).

We measure the F1, Precision (P), Recall (R), Exact Match (EM). EM is required to delineate topics across all utterances within a dialogue precisely. GPT-3.5 (175B) and GPT-4 (>175B) exhibit poor performance in all metrics for both zero-shot and few-shot settings, as shown in Table **??**. Notably, they fail to achieve an Exact Match (EM) in all dialogues, demonstrating the considerable challenge of identifying topics within dialogues. In contrast, T5 and Flan-T5 models (220M), when finetuned on MP2D-generated datasets, exhibit superior performance across all metrics. Furthermore, models trained on *Knowledge-graph* outperform those trained on *Random-shift*, proving the effectiveness of our proposed method that leverages knowledge graphs in generating natural topic shifts.

6.2 Topic Shift Detection

Topic Shift Detection is the task of detecting topic transitions of a conversation in real-time (Lin et al., 2023b). This task is crucial for real-time dialogue systems, empowering them to adapt dynamically to the evolving conversation, thereby facilitating relevant responses based on the detected topic shifts. As the dialogue progresses with each sequential utterance, the objective is to identify whether a target utterance continues the current topic or initiates a new one. For instance, with the sequence

Case 1	Q:	What was *Lekain*'s education and how did it contribute to his early career as an actor?
	A:	He was educated at the Collège Mazarin, and joined an amateur company of players against which the Comédie-Française obtained an injunction.
	Q:	Was there no student of *Lekain*?
	A:	*Larive* was a student of *Lekain*.
	Q:	What can you tell me about *Larive*?
	A:	Jean Mauduit, stage name *Larive* or de La Rive was a French actor.
Case 2	Q:	What is the geological age of *Rhacheosaurus*?
	A:	The genus *Rhacheosaurus* is a fossil taxon, belonging to the *Metriorhynchidae* family.
	✗ Q:	What is the geographic distribution of *Rhacheosaurus?*
	✓ Q:	What is *Metriorhynchidae*, and during which geological periods and in which regions did it exist?
	A:	*Metriorhynchidae* is an extinct family of specialized, aquatic metriorhynchoid crocodyliforms from the Middle Jurassic to the Early Cretaceous period of Europe, North America and South America.
Case 3	Q:	When was the *Malcolm Group* founded?
	A:	*Malcolm Group*, a *logistics* business founded in 1960, is located in Linwood, Renfrewshire.
	Q:	What is the definition of *logistics*?
	A:	*Logistics* is the part of supply chain management that deals with the efficient forward and reverse flow of goods, services, and related information from the point of origin to ...

Table 5: Case Study. **Case 1**: A successful example. **Case 2**: An example of inaccurate question generation from lacking additional instruction in a topic shift turn. The question marked in red is generated without the instruction. **Case 3**: An example that might seem unnatural due to an abrupt change from specific to general topics.

$\{u_1, u_2, ..., u_i\}$, the system outputs '1' if the last utterance u_i introduces a shift in topic; otherwise, it produces a '0' if u_i continues the current topic.

Table 4 presents results consistent with those discussed in Section 6.1, demonstrating that models finetuned on MP2D-generated datasets uniformly surpass the baselines in all metrics. LLMs fail in all cases for EM and show a similar trend in accuracy per turn, which evaluates the correctness of detecting a topic shift at each utterance turn. Notably, GPT-4 exhibits significantly lower results compared to GPT-3.5 due to its sensitivity to topic changes and more detailed breakdown of topics.

6.3 Enhancing LLM in Topic Shift turns

We enhance the ConvQA response generation performance of LLMs, specifically at topic shift turns, by integrating the topic shift detection module. The finetuned Flan-T5 model is utilized as the detection module, which is trained on MP2D-generated data for topic shift detection tasks. Prior to feeding multi-passage content, dialogue history, and the target question into the LLM, they undergo processing through a topic shift detection model to ascertain if a topic shift has occurred in the current turn. This data is then integrated into the input for the LLM. Essentially, the LLM receives additional information regarding topic shifts to aid in generating the subsequent response. As depicted in Figure 3 (red line), the experimental findings validate that this information enhances the response generation performance at topic shift turns. Through this analysis, we illustrate the capability to enhance the performance of the LLM in handling topic shifts by utilizing data generated by MP2D without necessitating direct LLM training.

7 Case Study

We present examples drawn from the MP2D-generated topic shift dialogues in Table 5. Case 1 demonstrates a dialogue with a natural flow of the topic, and most dialogues generated by the MP2D framework include such topic transitions. Case 2, described in Section 3.2, demonstrates that converting multi-passage content to ConvQA with GPT-3.5 without clear instructions at topic shift turns leads to incorrect question generation. The question marked in red represents the question generated in a setting without instructions, illustrating that despite the topic shifting from *Rhacheosaurus* to *Metriorhynchidae* and the subsequent answer discussing the changed topic, it erroneously generates questions about *Rhacheosaurus*. The question presented below the red-marked question is generated using the MP2D framework, successfully creating a question relevant to the changed topic. Case 3 is an example of a dialogue flow that presents challenges, as cited in Section 4.2. The text transitions from discussing the *Malcolm Group* to *logistics*. Al-though there is a relationship between the two, the abrupt shift from a specific topic to a more general one might be considered unnatural. Controlling the relationship between such entities during the path construction process presents a promising avenue for future work.

8 Conclusion

In this work, we address the data scarcity issue in topic shift dialogues by proposing a framework that automatically generates dialogues with natural topic transitions. Our MP2D methodology utilizes the flow of relationships between entities in a knowledge graph to structure the dialogue flow and converts multi-passage content into a ConvQA format. Experimentally, we demonstrate that MP2D-generated topic shift dialogues are of high quality and prove their value as a training dataset for various topic shift dialogue tasks.

Limitations

This study employs an LLM as the question generator within the multi-passage to dialogue framework, opting for its superior performance over the finetuned T5 models. However, when employed as a question generator, a performance-efficiency trade-off exists between T5 and LLMs; while T5 may exhibit slightly lower performance, it incurs less cost (Patterson et al., 2021). Given that MP2D serves as a dataset generation framework, the inference cost can influence the size of the generated dataset. Therefore, when utilizing MP2D for dataset creation, it is crucial to make a context-appropriate judgment regarding the trade-off between quality and data size, allowing for selecting either a finetuned model or an LLM as the question generator.

The second limitation arises from the costs associated with LLMs (Musser, 2023), leading to the inability to employ GPT-4 in certain experiments due to financial constraints. For instance, when generating 10,000 dialogues in Section 4.1, we used GPT-3.5 instead of GPT-4. It remains an open question whether employing GPT-4 to generate the dialogues could have resulted in higher performance for topic shift tasks.

Additionally, this study did not tackle the challenge of disambiguating entities in a knowledge graph, such as distinguishing between "Python," the programming language, and "Python," the snake on Wikipedia (Chen et al., 2021). In addressing the existence of multiple links for a single entity, our approach simply disambiguates by utilizing the first page returned in search results. Even though we observed only a small fraction of entities (approximately 1.34%) exhibit such ambiguity, controlling the path to create more contextual topic shifts or to establish personalized topic flows represents a promising direction for future work.

Ethics Statement

In Section 4.2, we verify through crowd workers that the generated datasets are free of any potential ethical concerns. Such concerns include offensive, sexist, or racist remarks, toxic language, or any depictions of sexual behavior. The crowd workers received fair compensation for their evaluation of the dataset. A comprehensive description, the interface used for collecting human evaluations, and detailed information regarding compensation are provided in Appendix E.

Furthermore, we employ LLMs from the official website of OpenAI[*]. All models and datasets utilized in our research are sourced from publicly available websites or GitHub repositories.

Acknowledgements

This work was supported by LG AI Research. This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University) & RS-2021-II212068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)], and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2024. K. Jung is with ASRI, Seoul National University, Korea. The Institute of Engineering Research at Seoul National University provided research facilities for this work.

References

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2020. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. *arXiv preprint arXiv:2010.12688*.

[*] https://openai.com/

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2020. Open-domain question answering goes conversational via question rewriting. *arXiv preprint arXiv:2010.04898*.

Jaime Arguello and Carolyn Rosé. 2006. Topic-segmentation of dialogue. In *Proceedings of the analyzing conversations in text and speech*, pages 42–49.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. *arXiv preprint arXiv:1909.05358*.

Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating entity disambiguation and the role of popularity in retrieval-based nlp. *arXiv preprint arXiv:2106.06830*.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models. *arXiv preprint arXiv:2307.12976*.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International Conference on Machine Learning*, pages 4558–4586. PMLR.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. Trec cast 2019: The conversational assistance track overview. *arXiv preprint arXiv:2003.13624*.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569.

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*, pages 2481–2485.

Linda J Garcia and Yves Joanette. 1997. Analysis of conversational topic shifts: A multiple case study. *Brain and language*, 58(1):92–114.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. *arXiv preprint arXiv:1907.10568*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Florian Holz and Sven Teresniak. 2010. Towards automatic detection and tracking of topic change. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 327–339. Springer.

Yerin Hwang, Yongil Kim, Hyunkyung Bae, Jeesoo Bang, Hwanhee Lee, and Kyomin Jung. 2023. Dialogizer: Context-aware conversational-qa dataset generation from textual sources. *arXiv preprint arXiv:2311.07589*.

Endri Kacupaj, Barshana Banerjee, Kuldeep Singh, and Jens Lehmann. 2021. Paraqa: a question answering dataset with paraphrase responses for single-turn conversation. In *European semantic web conference*, pages 598–613. Springer.

Michael R King. 2023. The future of ai in medicine: a perspective from a chatbot. *Annals of Biomedical Engineering*, 51(2):291–295.

J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Haitao Lin, Junnan Zhu, Lu Xiang, Feifei Zhai, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2023a. Topic-oriented dialogue summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jiangyi Lin, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2023b. Multi-granularity prompts for topic shift detection in dialogue. *arXiv preprint arXiv:2305.14006*.

Shuijing Liu, Aamir Hasan, Kaiwen Hong, Runxuan Wang, Peixin Chang, Zachary Mizrachi, Justin Lin, D Livingston McPherson, Wendy A Rogers, and Katherine Driggs-Campbell. 2024. Dragon: A dialogue-based robot for assistive navigation with visual language grounding. *IEEE Robotics and Automation Letters*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. *arXiv preprint arXiv:2005.00456*.

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. 2022. Rquge: Reference-free metric for evaluating question generation by answering the question. *arXiv preprint arXiv:2211.01482*.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.

Micah Musser. 2023. A cost analysis of generative language models and influence operations. *arXiv preprint arXiv:2308.03740*.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4):3055–3155.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, Kewei Tu, et al. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508.

Matthew Purver. 2011. Topic segmentation. *Spoken language understanding: systems for extracting semantic information from speech*, pages 291–317.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 539–548.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Karin Sevegnani, David M Howcroft, Ioannis Konstas, and Verena Rieser. 2021. Otters: One-turn topic transitions for open-domain dialogue. *arXiv preprint arXiv:2105.13710*.

Swapna Somasundaran et al. 2020. Two-level transformer and auxiliary coherence modeling for improved text segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7797–7804.

Mayank Soni, Brendan Spillane, Emer Gilmartin, Christian Saam, Benjamin R Cowan, and Vincent Wade. 2021. An empirical study of topic transition in dialogue. *arXiv preprint arXiv:2111.14188*.

Amanda Spink, H Cenk Ozmutlu, and Seda Ozmutlu. 2002. Multitasking information seeking and searching processes. *Journal of the american society for information science and technology*, 53(8):639–652.

Hongru Wang, Mingyu Cui, Zimo Zhou, Gabriel Pui Cheong Fung, and Kam-Fai Wong. 2021. Topicrefine: Joint topic prediction and dialogue response generation for multi-turn end-to-end dialogue system. *arXiv preprint arXiv:2109.05187*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Xiaoqiang Wang, Bang Liu, Siliang Tang, and Lingfei Wu. 2022. Qrelscore: Better evaluating generated questions with deeper understanding of context-aware relevance. *arXiv preprint arXiv:2204.13921*.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. 2021. Tiage: A benchmark for topic-shift aware dialog modeling. *arXiv preprint arXiv:2109.04562*.

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.

Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. 2022. Conversational question answering: A survey. *Knowledge and Information Systems*, 64(12):3151–3195.

Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Subhash Sagar, Adnan Mahmood, and Yang Zhang. 2023. Learning to select the relevant history turns in conversational question answering. In *International Conference on Web Information Systems Engineering*, pages 334–348. Springer.

Yankai Zeng, Abhiramon Rajasekharan, Parth Padalkar, Kinjal Basu, Joaquín Arias, and Gopal Gupta. 2024. Automated interactive domain-specific conversational agents that understand human dialogs. In *International Symposium on Practical Aspects of Declarative Languages*, pages 204–222. Springer.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. Dynaeval: Unifying turn and dialogue level evaluation. *arXiv preprint arXiv:2106.01112*.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2019. Grounded conversation generation as guided traverses in commonsense knowledge graphs. *arXiv preprint arXiv:1911.02707*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.

A Generated Datasets Statistics

	Section 4	Section 5
# of dialogues	10,000	1,000
# of turns	156,856	15,892
Average # of topics	2.137	2.136
Average tokens per dialogue	22.50	23.26
# of unique tokens	25,546	15,938

Table 6: The statistics of the generated datasets in Sections 4 and 5.

The statistics for the multi-passage to dialogue datasets generated for automatic evaluation (Section 4) and TS-WikiDialog (Section 5) can be observed in Table 6.

B Reproducibility checklists

B.1 Dataset and Source code

Our experiment source code and configuration details are included as supplementary materials. The datasets produced and the complete codes, including weight parameters, will be made available to the public.

B.2 Computing Resources

For the experiments, Xeon 4210R (2.40 GHz) with RTX A6000 is employed. Four GPUs are utilized for the experimental setup. All codes are implemented on Python 3.7.13 and PyTorch 1.10.1.

B.3 Versions of the LLMs

The GPT-3.5 version utilized for MP2D framework, topic segmentation, and topic shift detection is *gpt-3.5-turbo-0613*. Moreover, the GPT-4 version employed for GPT-4 evaluation, ConvQA response generation, topic segmentation, and topic shift detection is *gpt-4-0613*.

C Details of MP2D Implementation

In the question generation process of the MP2D framework, the prompts used when employing GPT-3.5 as the question generator are provided in Table 7.

The topic shift dialog created through the MP2D framework undergoes a simple post-processing step. Since questions are generated to fill in the [BLANK] in "A: [BLANK]," a few samples are produced with "A: " prefixed to the question, which was then removed through a rule-based approach.

D Details of P2D models Implementation

In implementing the passage to dialogue methods(Dialog Inpainter and Dialogizer), we fundamentally adhere to the best-performing hyperparameters as proposed in their respective research for a fair comparison. The T5-base model[†] serves as the backbone for both models. To ensure the robustness of our findings, we conduct all experiments using three different seed numbers. Both models are trained with a batch size of 8 and a gradient accumulation step size of 8.

E Human Evaluation

The recruitment process for five crowd workers for Sections 4.2 and 5.1 was conducted through the university's online community, targeting individuals proficient in English. The crowd workers were provided with detailed task descriptions, evaluation guidelines, and illustrative examples, as depicted in Figures 4 and 5. Additionally, they were informed that the evaluation was intended for academic research purposes. After completing a sample evaluation and assessing the required time, the crowd workers were compensated fairly, ensuring a minimum hourly wage of $12 or more, as determined by the coworkers.

Inter-Annotator Agreement We evaluate the Inter-Annotator Agreement (IAA) among three crowd workers for human evaluation in Section 4.2, reporting the Cohen's kappa score (Cohen, 1960). The interpretation of these scores follows the guidelines (Landis and Koch, 1977), classifying them as substantial.

Cohen's kappa values are as follows:

A1-A2 Cohen's kappa score: 0.7039 (Substantial)
A1-A3 Cohen's kappa score: 0.6541 (Substantial)
A2-A3 Cohen's kappa score: 0.6592 (Substantial)
Average Cohen's kappa score: 0.6724
(A1,A2, and A3 stand for Annotator1, Annotator2, and Annotator3)

F GPT-4 Evaluation

The template and prompt for GPT-4 evaluation in Section 4.2 are based on (Liu et al., 2023), and the sample prompt can be found in Table 8.

[†] https://huggingface.co/t5-base

G Topic Shift Dialogue Tasks Details

We present the experimental details of finetuning the T5 and Flan-T5 models for topic segmentation and topic shift detection tasks in Section 6. Both models have approximately 220M parameters. They are trained with a batch size of 8, using a gradient accumulation step size of 8. We utilize AdamW (Loshchilov and Hutter, 2017) as optimizer with $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$. The max gradient norm for gradient clipping is set to 1.0. To determine the most effective model configuration, we conduct experiments with various combinations of hyper-parameters across three epochs: $per\_gpu\_batch\_size$: (1, 2), $initial\_learning\_rate$: ($1e-4$, $5e-5$, $2e-5$), $warmup\_step$: (0, 500). We repeat the experiment for three different seed numbers and report the mean values.

The prompts used for evaluating the capability of LLMs in handling topic shift tasks can be observed in Table 9 (Topic Segmentation Task) and Table 10 (Topic Shift Detection).

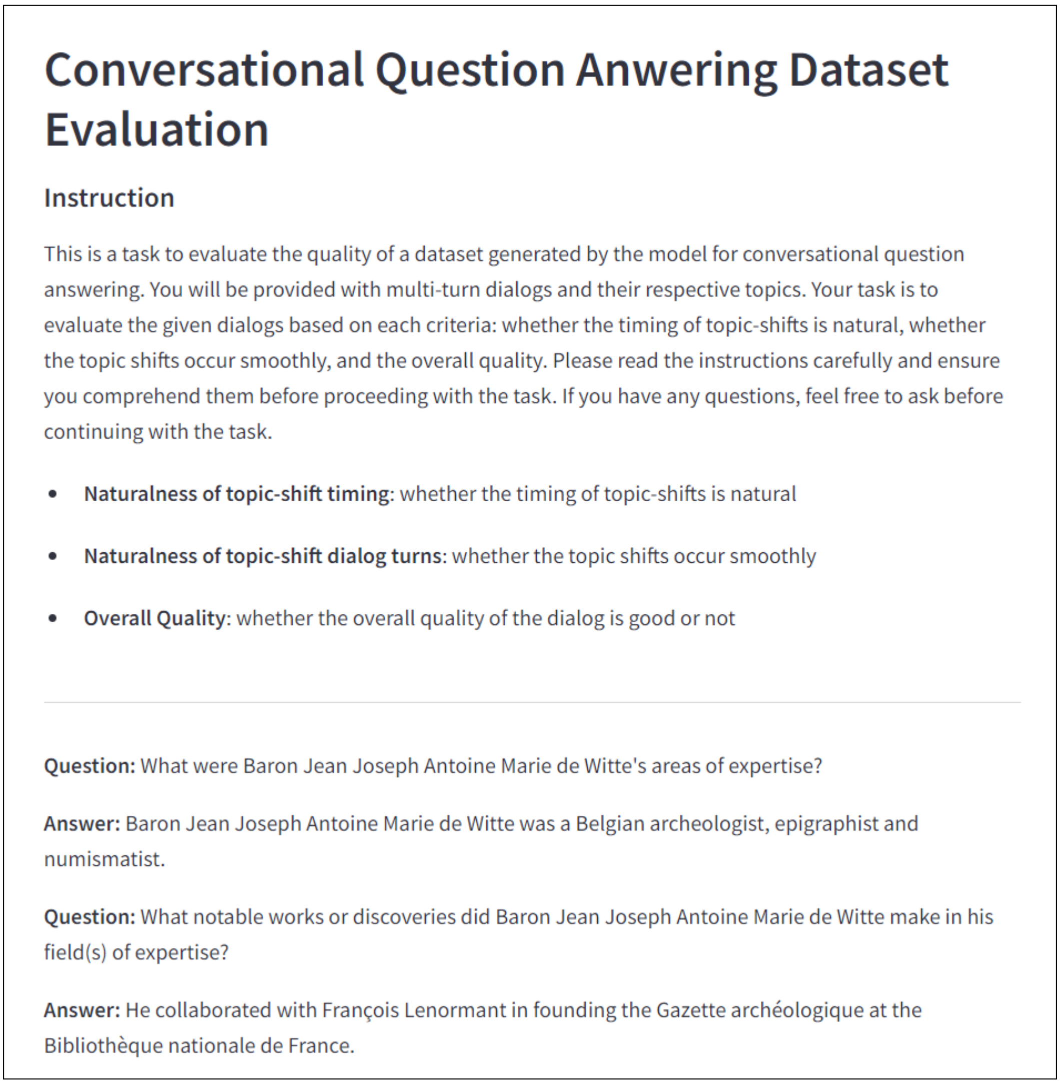
H Generated Topic shift Dialogue Examples

Examples of topic shift dialogues generated by the MP2D framework can be found in Tables 11 and 12.

You are an automatic assistant that generates appropriate question based on the predefined answer. Generate a single question that is most suitable for the given dialogue history and target answer. Please fill in only [BLANK] in the next dialogue.
Note that the conversation topic has changed into {next_topic} from {current_topic}.

START
A: {question 1}
B: {answer 1}
...
A: [BLANK]
B: {answer t}
END

Table 7: The template of the prompt used for question generation process in MP2D. The instructions marked in red are included exclusively in topic shift turns.



Conversational Question Anwering Dataset Evaluation

Instruction

This is a task to evaluate the quality of a dataset generated by the model for conversational question answering. You will be provided with multi-turn dialogs and their respective topics. Your task is to evaluate the given dialogs based on each criteria: whether the timing of topic-shifts is natural, whether the topic shifts occur smoothly, and the overall quality. Please read the instructions carefully and ensure you comprehend them before proceeding with the task. If you have any questions, feel free to ask before continuing with the task.

- **Naturalness of topic-shift timing**: whether the timing of topic-shifts is natural
- **Naturalness of topic-shift dialog turns**: whether the topic shifts occur smoothly
- **Overall Quality**: whether the overall quality of the dialog is good or not

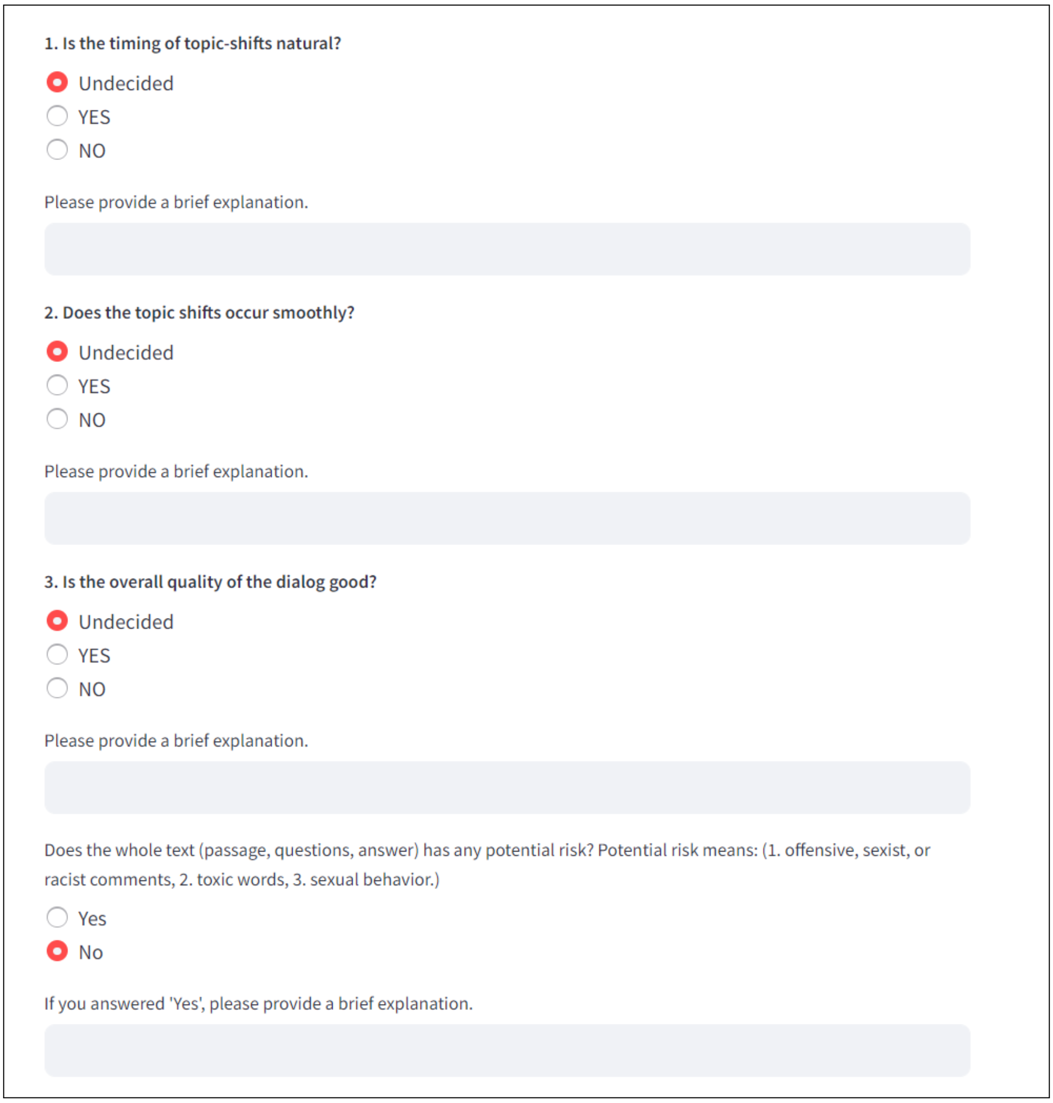
Question: What were Baron Jean Joseph Antoine Marie de Witte's areas of expertise?

Answer: Baron Jean Joseph Antoine Marie de Witte was a Belgian archeologist, epigraphist and numismatist.

Question: What notable works or discoveries did Baron Jean Joseph Antoine Marie de Witte make in his field(s) of expertise?

Answer: He collaborated with François Lenormant in founding the Gazette archéologique at the Bibliothèque nationale de France.

Figure 4: Interface of human evaluation. (1/2)



1. Is the timing of topic-shifts natural?

● Undecided

○ YES

○ NO

Please provide a brief explanation.

2. Does the topic shifts occur smoothly?

● Undecided

○ YES

○ NO

Please provide a brief explanation.

3. Is the overall quality of the dialog good?

● Undecided

○ YES

○ NO

Please provide a brief explanation.

Does the whole text (passage, questions, answer) has any potential risk? Potential risk means: (1. offensive, sexist, or racist comments, 2. toxic words, 3. sexual behavior.)

○ Yes

● No

If you answered 'Yes', please provide a brief explanation.

Figure 5: Interface of human evaluation. (2/2)

You will be given a multi-turn conversational question answering dialog, and
your task is to evaluate the quality of the given dialog based on four criteria.
The descriptions for each criterion are as follows.

1. **Naturalness of topic-shift timing:** whether the timing of topic-shifts is natural
2. **Naturalness of topic-shift dialog turns:** whether the topic shifts occur smoothly
3. **Overall Quality:** whether the overall quality of the dialog is good or not
4. **Toxicity:** whether the dialog contains any potential risk

Potential risk means: 1. offensive, sexist, or racist comments, 2. toxic words, 3. sexual behavior.)

- Question:
- Answer:
- Question:
- Answer:

...

Is the timing of topic-shifts natural?
options: [YES, NO]
Does the topic shifts occur smoothly?
options: [YES, NO]
Is the overall quality of the dialog good?
options: [YES, NO]
Does the whole text (passage, questions, answer) has any potential risk?
options: [YES, NO]

Table 8: The template of the prompt used for GPT-4 evaluation.

```
You are performing automatic assistant topic segmentation. In the context of a dialogue, the topic segmentation task involves segmenting the dialogue based on the points where topic shifts occur. In the provided dialogue below, identify the sections where topic shifts occur. Output the indices where the topics change, separated by spaces.

START
1 A:
  B:
2 A:
  B:
3 A:
  B:
...
END
```

Table 9: The template of the prompt used for topic segmentation task.

```
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 0; if there has been no topic shift, output 1.

Dialog Context START
A:
B:
A:
B:
...
Dialog Context END

New response START
A:
New response END
```

Table 10: The template of the prompt used for topic shift detection task.

Q:	Who is Benjamin Gibbard?
A:	Benjamin Gibbard is an American singer, songwriter and guitarist.
Q:	What are some of Benjamin Gibbard's notable accomplishments?
A:	He is best known as the lead vocalist and guitarist of the indie rock band Death Cab for Cutie, with whom he has recorded ten studio albums, and as a member of the supergroup The Postal Service.
Q:	Did Benjamin Gibbard release any solo albums?
A:	Gibbard released his debut solo album, Former Lives, in 2012, and a collaborative studio album, One Fast Move or I'm Gone with Jay Farrar.
Q:	Where was Benjamin Gibbard born?
A:	Gibbard was born to Allen and Margaret Gibbard in Bremerton, Washington.
Q:	Who is Benjamin Gibbard's wife?
A:	Zooey Deschanel, who was married to Ben Gibbard from September 19th 2009 until December 12th 2012, is his wife.
Q:	**What is Zooey Deschanel known for?**
A:	Zooey Claire Deschanel is an American actress and musician.
Q:	What are some notable roles in films that Zooey Deschanel has played?
A:	She made her film debut in Mumford and had a supporting role in Cameron Crowe's film Almost Famous.

Table 11: First example of a topic shift dialogue generated by MP2D. The topic changes from *Ben Gibbard* to *Zooey Deschanel*, as highlighted in the bold question.

Q:	What is VeraCrypt?
A:	VeraCrypt is a free and open-source utility for on-the-fly encryption .
Q:	What are the features of VeraCrypt?
A:	The software can create a virtual encrypted disk that works just like a regular disk but within a file.
Q:	What is the purpose of pre-boot authentication in VeraCrypt?
A:	It can also encrypt a partition or the entire storage device with pre-boot authentication.
Q:	What platforms is VeraCrypt available on?
A:	VeraCrypt is a free software on the fly encryption software that was created in 2012 for Microsoft Windows and macOS.
Q:	**What is macOS?**
A:	MacOS, originally Mac OS X, previously shortened as OS X, is an operating system developed and marketed by Apple Inc. since 2001.
Q:	Is it a primary operating system for Apple's Mac computers?
A:	It is the primary operating system for Apple's Mac computers.
Q:	Is it a most widely used operating system for desktop and laptop computers?
A:	Within the market of desktop and laptop computers, it is the second most widely used desktop OS, after Microsoft Windows and ahead of all Linux distributions, including ChromeOS.
Q:	What are the major influences on MacOS?
A:	MacOS is influenced by Linux.
Q:	**What is Linux based on?**
A:	Linux is a family of open-source Unix-like operating systems based on the Linux kernel, an operating system kernel first released on September 17, 1991, by Linus Torvalds.
Q:	What is Linux typically packaged as, and what does this package include?
A:	Linux is typically packaged as a Linux distribution , which includes the kernel and supporting system software and libraries, many of which are provided by the GNU Project.

Table 12: Second example of a topic shift dialogue generated by MP2D. The topic transitions from *VeraCrypt* to *MacOS*, and then to *Linux*, with each topic shift highlighted in bold.