SLDT: Sequential Latent Document Transformer for Multilingual Document-based Dialogue

Zhanyu Ma[1,2,4] **Zeming Liu**[3] **Jian Ye**[1,2,4*]
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China[1]
University of Chinese Academy of Sciences[2]
Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China[3]
Beijing Key Laboratory of Mobile Computing and Pervasive Device[4]
mazhanyu21s@ict.ac.cn zmliu@ir.hit.edu.cn jye@ict.ac.cn

Abstract

Multilingual document-grounded dialogue, where the system is required to generate responses based on both the conversation multilingual context and external knowledge sources. Traditional pipeline methods for knowledge identification and response generation, while effective in certain scenarios, suffer from error propagation issues and fail to capture the interdependence between these two sub-tasks. To overcome these challenges, we propose the application of the SLDT method, which treats passage-knowledge selection as a sequential decision process rather than a single-step decision process. We achieved the winner 3rd in dialdoc 2023 and we also validated the effectiveness of our method on other datasets. The ablation experiment also shows that our method significantly improves the basic model compared to other methods.

1 Introduction

The advancements in neural models and the development of large-scale dialogue datasets have significantly propelled dialog generation research (Huang et al., 2020; Liu et al., 2022a; Ma et al., 2022). Open-domain dialogue systems strive to produce more informative and fluent responses (Ke et al., 2018; Zhang et al., 2020; Liu et al., 2021; Meng et al., 2021), finding applications in a wide array of areas such as emotional companionship, mental health support, and social chatbots.

Despite demonstrating promising results, most existing dialogue generation systems (Liu et al., 2022b; Bao et al., 2020; Li et al., 2020) depend on substantial data resources. In real-world scenarios, dialogue corpora for many languages are not readily available, thereby restricting the applicability of dialogue systems for low-resource or even zero-resource languages. Consequently, it is crucial to develop methods capable of effectively transferring knowledge from a source language with ample resources to a target language.

One such task is multilingual document-grounded dialogue (Sannigrahi et al., 2023), where the system is required to generate responses based on both the conversation multilingual context and external knowledge sources, such as documents or databases (Glass et al., 2022; Qi et al., 2022). While various methods have been proposed to address the challenges of knowledge selection and response generation in this task (Kim et al., 2020; Lai et al., 2023), including sequential latent knowledge selection for document-grounded dialogue. There is a need for a novel approach that combines the advantages of these methods (Zhang et al., 2022b). In this paper, we propose a new method to address the problem of document dialogue by employing the Sequential Latent Document Transformer (SLDT) to select the most relevant knowledge for conversation from a multilingual document set.

The motivation behind focusing on multilingual document-grounded dialogue lies in its potential to provide more informative and engaging responses by leveraging external knowledge sources (Gao et al., 2022; Zhang et al., 2022a), thereby enabling the dialogue system to better assist users in satisfying their diverse information needs. Traditional pipeline methods for knowledge identification and response generation, while effective in certain scenarios, suffer from error propagation issues and fail to capture the interdependence between these two sub-tasks. To overcome these challenges, we propose the application of the SLDT method, which has shown promising results in knowledge-grounded dialogue, to the task of document dialogue. The use of SLDT in document conversations is expected to bring several advantages, such as better modeling the diversity in document-knowledge selection, more accurate leveraging of response information, and the ability to work even when

* Corresponding author.

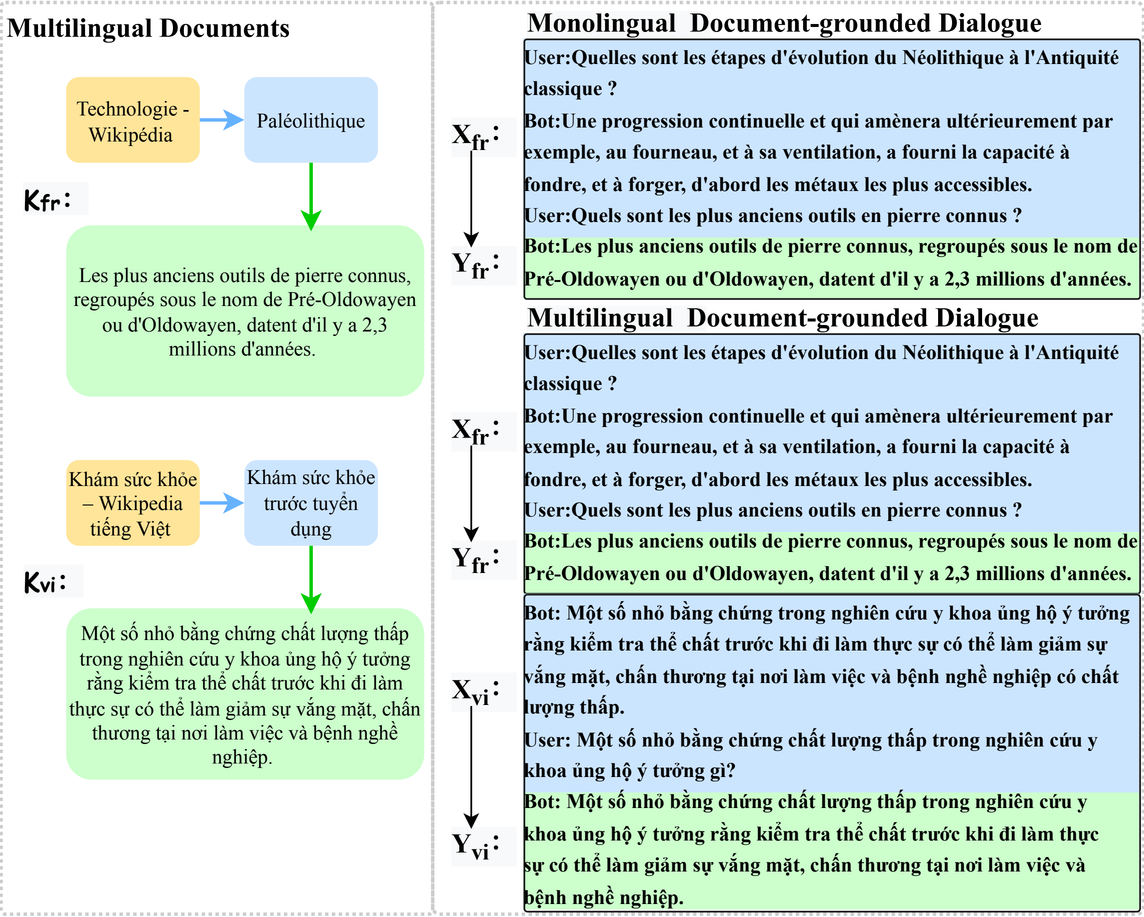

Figure 1: Introduction of Multilingual Document-grounded Dialogue.

knowledge selection labels for previous dialogues are not available . These properties of SLDT make it a suitable candidate for selecting relevant knowledge from documents to carry on the conversation. Our primary research goal is to develop an SLDT-based method for document dialogue that can effectively select the K most relevant documents from the document set based on the conversation history and input them into the generation module after concatenation.

Our method achieved excellent results on dialdoc 2023 Share Task. Obtained 208 points in online testing. We also validated the effectiveness of our method on other datasets. The ablation experiment also shows that our method significantly improves the basic model compared to other methods.

2 Related Works

Document-grounded dialogue systems (DGDS) categorize unstructured, semi-structured, and structured data in documents to facilitate the comprehension of human knowledge and interactions, thus fostering more natural human-computer interactions (HCI) (Zhou et al., 2018). The objective of DGDS is to generate conversational modes based on information (utterances, turns, context, clarification) supplied by a document or documents (Ma et al., 2020). DGDS are particularly advantageous in task-oriented and goal-oriented settings as they replicate the natural dialogue flow. A recent example of DGDS, closely related to our work, is Doc2Dial, a multi-domain DGDS dataset designed for goal-oriented dialogue that models hypothetical dialogue flows and scenes to simulate authentic interactions between a user and a machine agent in information-seeking contexts (Feng et al., 2020). In our proposed task, we adopt a similar approach, but we also permit users to pose clarification questions, the responses to which may not be directly grounded in the document. This aspect is crucial in the development of instruction-giving conversational agents, as the dialogue pipeline requires increased flexibility, as previously mentioned.

Multilingual dialogue tasks typically utilize a code-switching approach to achieve semantic alignment between various languages (Liu et al., 2020b;

Chapuis et al., 2021; Qin et al., 2021). This method of code-switching enables implicit semantic alignment without the need for parallel corpus pairs. Drawing inspiration from these studies, we apply the code-switching technique to transfer knowledge from English dialogue history to other target languages lacking training examples. In line with previous work (Chapuis et al., 2021) on multilingual representation, we implement code-switching at the utterance level, although code-switching at the word or span level is more prevalent (Banerjee et al., 2018; Bawa et al., 2020; Doğruöz et al., 2021).

3 Methodology

We utilize XLM-R (Conneau et al., 2020) as our retrieval model, employing a representation-based bi-encoder consisting of a dialogue query encoder, denoted as $q(\cdot)$, and a passage context encoder, represented by $p(\cdot)$.

For a given input query Q and a set of passages $\{P_i\}_{i=1}^{M}$, the encodings for the query and passage are computed as $q(Q)$ and $p\left(P_i\right)$, respectively. The similarity between these encodings is determined by the dot product $\langle q(Q), p\left(P_i\right)\rangle$, with the model being trained to minimize the negative log likelihood of the correct passage among L in-batch and challenging negatives.

Subsequently, we pre-calculate the representations for all passages and index them offline. During inference, the top-K passages are retrieved using Maximum Inner Product Search (MIPS) in conjunction with Faiss.

We introduce a Sequential Latent Document Transformer tailored for multilingual document-based dialogue, as illustrated in Figure 2. The objective of the model is to generate customized and informative responses by learning a probabilistic model $p(R|C, \mathcal{K}, \mathcal{P})$ that leverages passage-knowledge and context flowing (Kim et al., 2020).

We proceed by iterating through dialogue turns with $1 \leq t \leq T$, iterating over words in the utterances of the apprentice and wizard using $1 \leq m \leq M$ and $1 \leq n \leq N$, and denoting knowledge sentences in the pool with $1 \leq l \leq L$. Here, T represents the dialogue length, M and N correspond to the lengths of the apprentice and wizard's utterances, and L denotes the passage-knowledge pool size.

The input to the SLDT at turn t comprises previous conversation turns, which include user utterances $\mathbf{x}^1, ..., \mathbf{x}^t$, system responses $\mathbf{y}^1, ..., \mathbf{y}^{t-1}$, and the passage pool $\mathbf{k}^1, ..., \mathbf{k}^t$, where $\mathbf{k}^t = \{\mathbf{k}^{t,l}\} = \mathbf{k}^{t,1}, ..., \mathbf{k}^{t,L}$. The model's output consists of the chosen sample passage-knowledge $\mathbf{k}_s^t$ and the response $\mathbf{y}^t$. We provide an in-depth explanation of sentence embedding, passage-knowledge selection, and utterance decoding.

First, we consider passage-knowledge selection a sequential rather than a one-step decision-making process. Due to the diversity of passage-knowledge selection in dialogue, we model it with latent variables. Therefore, we can conduct joint inference for multiple turns of passage-knowledge selection and response generation, as opposed to distinct inference on a turn-by-turn basis.

Various studies have been conducted on sequential latent variable models. For instance, some have proposed a posterior attention model that represents the attention mechanism in seq2seq models as sequential latent variables. Drawing inspiration from these works, we factorize response generation with latent document passage-knowledge selection and derive the variational lower bound as follows. The conditional probability of generating response $\mathbf{y}^t$ given dialogue context $\mathbf{x}^{\leq t}$ and $\mathbf{y}^{<t}$:

$$p(\mathbf{y}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{<t}) \approx \prod_{i=1}^{t-1} \sum_{\mathbf{k}^i} q_\psi(\mathbf{k}^i) \Big(\sum_{\mathbf{k}^t} p_\gamma(\mathbf{y}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{<t}, \mathbf{k}^t) \pi_\gamma(\mathbf{k}^t) \Big) \tag{1}$$

Note that $p_\gamma(\mathbf{y}^t|\cdot)$ is a decoder network, $\pi_\gamma(\mathbf{k}^t)$ is a categorical conditional distribution of knowledge given dialogue context and previously selected knowledge, and $q_\psi(\mathbf{k}^t)$ is an inference network to approximate posterior distribution $p_\gamma(\mathbf{k}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{\leq t}, \mathbf{k}^{<t})$.

Eq.(1) means that we first infer from the knowledge posterior which knowledge would be used up to previous turn $t-1$, estimate the knowledge for current turn t from prior knowledge distribution and generate an utterance from the inferred knowledge. Figure 2 shows an example of this generation process at $t = 3$. We parameterize the decoder network p_γ, the prior distribution of knowledge π_γ, and the approximate posterior q_ψ with deep neural networks as will be discussed.

4 Experiments

For the retrieval training stage, we utilized a batch size of 128 and a learning rate of 1e-4 and 5e-5 for post-training and fine-tuning, respectively. And retrieval passage number top-k is 25. During the generation stage, we used a batch size of 32 with

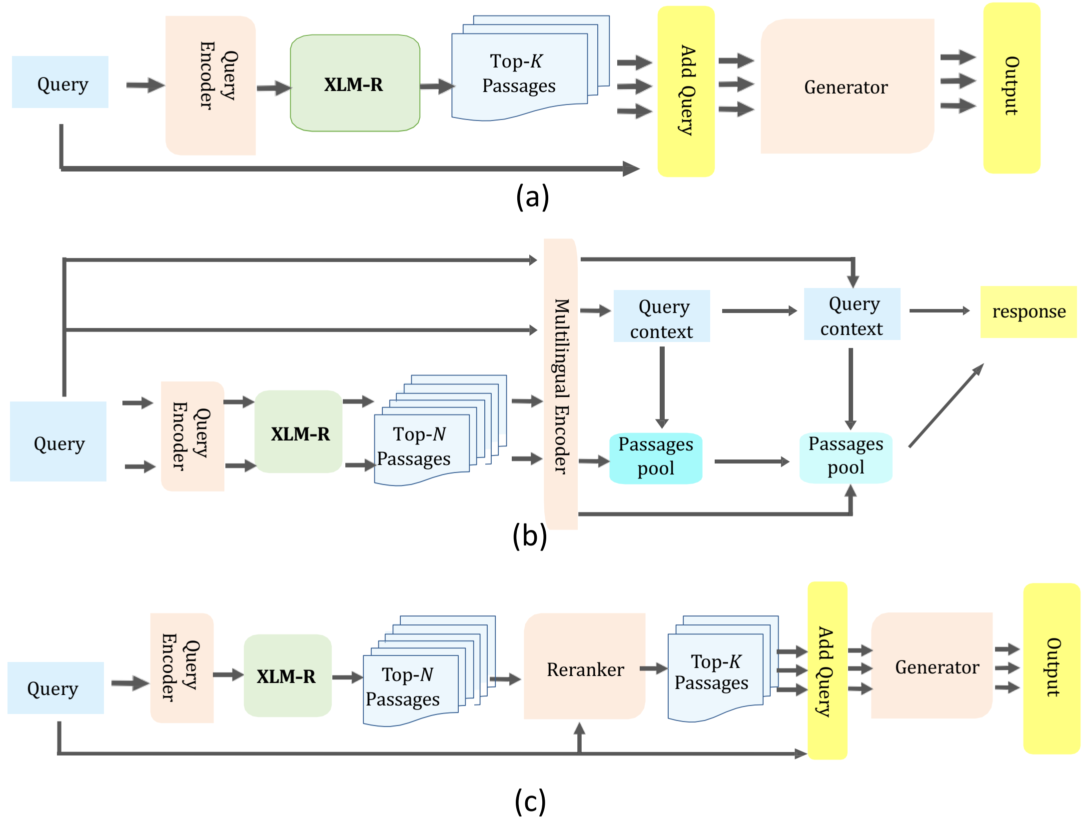

Figure 2: Subfigure (a) and (c) show the document-based dialogue in two-stage and three-stage (Glass et al., 2022), respectively, while subfigure (b) show our SLDT, a new paradigm between (a) and (c).

a learning rate of 1e-4 and 1e-5 for post-training and fine-tuning, respectively. For R-drop, we set the dropout rate to 0.1, and the KL-divergence loss weight 0.02.

4.1 Datasets

FrViDoc2Bot contains annotated Vietnamese and French document-grounded dialogue training data, the development data that the participants are required to provide the model predicts, as well as the passage corpus that the training and development data depend on (DAMO_ConvAI, 2023). Each piece of data in the training set contains three attributes: query, passages, and response. The query is a concatenation of the conversation history in reverse order, with the last turn marked as “<last_turn>" and the rest marked with “" for user input and “" for system output. The ‘passages‘ attribute contains the passage arranged according to reply dependencies, followed by a reverse-ordered chain of titles concatenated with “/" as the delimiter. The response attribute is the desired output, beginning with “". They have provided the ‘passage corpus‘ that all dialogues in the training, validation, and test sets rely on in passages.csv. We sampled 200 pieces of train data from it as a dev set during offline validation for Table 1 and 2.

Wizard of Wikipedia dataset is a large dataset with conversations directly grounded with knowledge retrieved from Wikipedia. It is used to train and evaluate dialogue systems for knowledgeable open dialogue with clear grounding. The dataset contains dialogues in which a bot needs to respond to user inputs in a knowledgeable way. Each response should be grounded on a sentence from Wikipedia that is relevant to the conversation topic. WoW encompasses a total of 18,430 dialogues for training, 1,948 dialogues for validation, and 1,933 dialogues for testing (Dinan et al., 2019a). The test set is divided into two subsets: *Test Seen*, containing 965 dialogues on topics overlapping with the training set, and *Test Unseen*, consisting of 968 dialogues on topics not previously encountered in the training and validation sets.

4.2 Automatic Evaluation

The F1 (Dinan et al., 2019b) value is used to evaluate the consistency between the predicted and golden responses when the golden response exists.

Model	Parameters	Response Generation									
		B-1	B-2	B-3	DIS-1	DIS-2	R-1	R-2	R-L	F1	S-BLEU
$MT5_S$ (golden kg)	300M	24.7	23.2	21.3	5.3	9.8	32.1	28.4	31.2	19.7	21.3
$MT5_S$ (no kg)		19.1	17.6	15.7	3.3	4.1	26.5	22.8	25.6	14.1	15.7
$MT5_B$ (golden kg)	580M	45.3	43.8	42.0	25.9	30.3	53.0	49.3	52.1	40.3	42.0
$MT5_B$ (no kg)		30.3	28.8	27.0	10.9	15.4	37.7	34.0	36.8	25.3	27.0
$MBART_B$ (golden kg)	170M	47.4	45.9	44.0	28.0	32.4	55.0	51.3	54.1	42.4	44.0
$MBART_B$ (no kg)		30.6	29.1	28.0	11.2	15.6	39.0	35.3	38.1	25.6	28.0
$MBART_L$ (golden kg)	680M	53.7	52.2	50.3	34.3	38.7	61.0	57.3	60.1	48.6	50.2
$MBART_L$ (no kg)		32.4	30.9	29.3	13.5	16.9	41.4	37.7	39.9	36.2	35.3

Table 1: Automatic evaluation results of different Pre-trained models on the FrViDoc2Bot dev set.

Model	Response Generation										
	B-1	B-2	B-3	DIS-1	DIS-2	R-1	R-2	R-L	F1	PPL	S-BLEU
MBART + No knowledge	31.4	27.9	19.3	10.5	12.9	21.4	20.7	19.9	16.2	-	35.3
MBART + Random knowledge	33.4	30.4	21.8	12.8	17.2	26.1	21.6	23.4	23.1	-	37.8
MBART + Repeat last utterance	35.5	32.9	23.3	15.1	18.5	31.4	22.5	26.9	26.7	103.7	40.3
MBART + Norm retrieval	47.6	45.3	39.8	27.4	25.8	46.2	38.4	38.4	37.0	88.3	42.8
MBART + XLM-R	49.6	47.6	43.3	29.8	30.0	51.0	44.3	45.8	45.4	83.5	45.3
MBART + XLM-R + SLDT	51.7	49.9	46.8	32.1	34.3	56.1	50.8	53.2	52.2	76.4	47.8
MBART + XLM-R + SLDT + Copy	53.7	52.2	50.3	34.3	38.7	61.0	57.3	60.1	58.6	64.4	50.3

Table 2: Automatic evaluation results of different models on the FrViDoc2Bot dev set.

Perplexity (PPL) (Meister and Cotterell, 2021) can determine the coherence of the predicted query to a certain extent. We additionally used BLEU (Papineni et al., 2002; Chen and Cherry, 2014; Post, 2018) to evaluate the consistency of predicted responses with standard responses, Distinct (Li et al., 2016a) to evaluate the diversity of responses in the test set (Li et al., 2016b).

4.3 Pre-training Models

XLM-R (Conneau et al., 2020) is an improved version of XLM based on the RoBERTa model (Liu et al., 2019). XLM-R is trained with a cross-lingual masked language modeling objective on data in 100 languages from Common Crawl. To improve the pre-training data quality, pages from Common Crawl were filtered by an n-gram language model trained on Wikipedia (Wenzek et al., 2020).

mBART (Liu et al., 2020a) is a multilingual encoder-decoder model that is based on BART (Lewis et al., 2020). mBART is trained with a combination of span masking and sentence shuffling objectives on a subset of 25 languages from the same data as XLM-R.

MT5 (Multilingual T5) is a massively multilingual pretrained text-to-text transformer model (Xue et al., 2021). It is trained following a similar recipe as T5. Current natural language processing (NLP) pipelines often make use of transfer learning, where a model is pre-trained on a data-rich task before being fine-tuned on a downstream task of interest.

4.4 Knowledge Access Methods

Weak correlation passage-knowledge in knowledge-based dialogue refers to the knowledge that is not directly related to the current dialogue context but is still useful for generating a response. It is called weak correlation because it is not directly related to the current dialogue context but is still useful for generating a response. For example, if you are talking about a movie and you mention that you like action movies,

Model	Response Generation (Test seen)										
	B-1	**B-2**	**B-3**	**DIS-1**	**DIS-2**	**R-1**	**R-2**	**R-L**	**F1**	**PPL**	**S-BLEU**
MBART + No knowledge	9.0	6.5	4.3	5.7	7.8	5.4	3.9	4.7	5.2	-	9.3
MBART + Random knowledge	8.9	6.1	4.0	6.8	8.5	6.9	4.2	5.8	6.6	-	9.6
MBART + Repeat last utterance	14.1	11.7	9.7	10.6	15.2	12.4	2.1	10.0	12.0	89.7	13.6
MBART + SLDT	17.3	15.3	13.4	15.4	20.9	16.9	5.4	14.4	16.4	64.4	18.6
MBART + SLDT + Copy	18.5	16.9	15.1	18.2	24.6	19.4	6.7	16.8	18.8	52.1	21.6

Table 3: Automatic evaluation results of different models on Wizard of Wikipedia test seen set.

Model	Response Generation (Test Unseen)										
	B-1	**B-2**	**B-3**	**DIS-1**	**DIS-2**	**R-1**	**R-2**	**R-L**	**F1**	**PPL**	**S-BLEU**
MBART + No knowledge	6.3	5.2	4.1	4.0	6.7	3.9	-	1.5	2.9	-	5.8
MBART + Random knowledge	6.4	5.8	5.6	4.4	7.5	7.3	-	4.6	6.3	-	6.6
MBART + Repeat last utterance	12.5	9.7	9.1	7.5	12.3	10.6	2.2	7.7	9.7	113.4	13.4
MBART + SLDT	15.6	12.6	12.6	11.6	17.1	14.9	3.7	11.8	11.0	90.5	15.2
MBART + SLDT + Copy	16.7	13.5	14.2	13.7	19.9	17.2	4.1	13.9	13.3	81.3	18.0

Table 4: Automatic evaluation results of different models on Wizard of Wikipedia test unseen set.

Model	Response Generation										
	B-1	**B-2**	**B-3**	**DIS-1**	**DIS-2**	**R-1**	**R-2**	**R-L**	**F1**	**PPL**	**S-BLEU**
Ours-FiD	43.6	51.5	58.4	32.3	36.0	58.4	53.9	56.3	46.2	73.2	46.3
Ours-R_drop	54.2	52.8	51.5	35.9	39.3	62.1	58.8	61.3	50.5	65.2	52.6
Ours-Prompt	56.6	54.0	52.2	36.4	39.6	63.2	56.0	61.8	53.2	63.8	53.1
Ours-Post_pretrain	58.9	56.2	53.9	40.5	44.3	64.6	57.2	62.1	61.8	47.3	58.6
Ours-Ensemble	60.7	58.5	55.6	43.4	48.8	67.0	61.4	66.3	66.7	38.9	60.5

Table 5: Automatic evaluation results of leaderboard submission which is based on the FrViDoc2Bot test set.

then the system can use this weak correlation passage-knowledge to recommend other action movies that you might like.

Norm retrieval means regaining the norm of the lost signal from its intensity measurements. It arises naturally from phase retrieval when one utilizes both a collection of subspaces and their orthogonal complements. Norm retrieval can be done using projections and can be used to extend certain results for frames.

4.5 Results and Analysis

4.5.1 Performance of Pre-training models

The experimental analysis presented in Table 1 aims to compare the performance of different pre-trained models on the FrViDoc2Bot dev set. The models investigated include MT5 and mBART, with small (S), base (B), and large (L) variants. The table further distinguishes between the models' performances when utilizing the golden knowledge (golden kg) and when relying solely on dialogue history information (no kg).

Upon analyzing the results, it is evident that the models' performance generally improves with the inclusion of the golden kg, as indicated by higher scores across most evaluation metrics. This implies that the utilization of external knowledge is beneficial for response generation tasks. For instance, the $MT5\_S$ model achieves a B-1 score of 24.7 with the golden kg, while the same model without the kg attains a B-1 score of 19.1. Similar improvements can be observed for other models and evaluation metrics.

Comparing the performance of MT5 and mBART models, it can be observed that mBART consistently outperforms MT5 for the same model size and knowledge condition. For example, mBART_B (golden kg) achieves a B-1 score of 47.4, while MT5_B (golden kg) scores 45.3. This trend is consistent across most of the evaluation metrics, indicating the superior performance of mBART models in this specific task.

Furthermore, it is noticeable that larger models generally yield better results than their smaller counterparts. For instance, mBART_L (golden kg) achieves a B-1 score of 53.7, outperforming both mBART_B (golden kg) and mBART_S (golden kg) with respective B-1 scores of 47.4 and 24.7. This suggests that larger model sizes can enhance the performance of response generation tasks.

4.5.2 Knowledge Access Methods

In this section, we analyze the performance of various knowledge acquisition methods on the FrVi-Doc2Bot dev set, as presented in Table 2. The models can be divided into several categories based on the knowledge acquisition strategy employed, and we will discuss the impact of these strategies on the performance of the knowledge dialogue system.

Performance of Basic Models The MBART + No knowledge model serves as the baseline, relying solely on the conversation history without incorporating any external knowledge. As expected, this model yields the lowest performance across all evaluation metrics. Introducing random knowledge (MBART + Random knowledge) provides some improvement, suggesting that even arbitrary knowledge can be useful in generating responses.

Incorporation of Targeted Knowledge When knowledge is specifically targeted to the conversation, such as with the MBART + Repeat last utterance model, we observe a significant improvement in performance. Repeating the last utterance as knowledge allows the model to generate more coherent responses by drawing on the context provided. However, this model's performance is still limited by its reliance on only one piece of knowledge.

Retrieval-Based Knowledge Acquisition The next category of models utilizes retrieval-based methods to acquire relevant knowledge from a knowledge base. The MBART + Norm retrieval model leverages a traditional retrieval model and exhibits a considerable performance boost compared to the previous models. This improvement underscores the importance of selecting appropriate knowledge to inform dialogue generation. The MBART + XLM-R model replaces the traditional retrieval model with XLM-R, a more advanced retrieval model. This change results in further performance gains across all metrics, highlighting the effectiveness of using powerful retrieval models to acquire relevant knowledge.

Sequential Latent Document Transformer The MBART + XLM-R + SLDT model incorporates the Sequential Latent Document Transformer (SLDT) into the knowledge selection process. This addition allows the model to perform a second stage of knowledge selection, leading to even better performance compared to the previous models. The SLDT mechanism effectively refines the retrieved knowledge, enabling the model to generate more accurate and coherent responses.

Incorporating Copy Mechanism Lastly, the MBART + XLM-R + SLDT + Copy model optimizes the decoding strategy by introducing a copy mechanism. This mechanism allows the model to copy or point to elements from the input sequence, leading to a more nuanced and accurate response generation. The introduction of the copy mechanism results in the best performance across all evaluation metrics, demonstrating the importance of a well-designed decoding strategy in knowledge dialogue systems.

Through the analysis of various knowledge acquisition methods and their impact on the knowledge dialogue system, we observe that incorporating targeted and relevant knowledge is crucial for generating coherent and accurate responses. Advanced retrieval models and techniques, such as XLM-R and SLDT, can significantly improve performance. Additionally, the incorporation of a copy mechanism in the decoding strategy leads to further enhancements. Overall, this analysis underscores the importance of effective knowledge acquisition and utilization in the development of high-performing knowledge dialogue systems.

4.5.3 Performance on the Wizard of Wikipedia

In this section, we examine the efficacy of various models on the Wizard of Wikipedia dataset, focusing on the impact of knowledge acquisition methods on knowledge dialogue systems. The performance of each model is evaluated on both seen and unseen test data.

Table 3 presents the results of the response generation for the test seen data. We observe that the MBART + SLDT + Copy model performs the best across most metrics. This demonstrates that the Sequential Latent Document Transformer model (SLDT), when combined with the copy mechanism (Li et al., 2019), significantly improves the efficacy of the knowledge dialogue system. The copy mechanism, which is inspired by the Pointer Network (Vinyals et al., 2015; Yang and Tu, 2022), allows the model to copy or point to input sequence elements, improving the generated output.

In contrast, the MBART + No knowledge and MBART + Random knowledge models exhibit lower performance in most metrics. This finding indicates that merely considering the conversation history or randomly selecting knowledge from the knowledge base is not sufficient for generating high-quality responses in a knowledge dialogue system.

Table 4 reports the results for the test unseen data. Similar to the test seen data, the MBART + SLDT + Copy model outperforms the other models across various metrics. This result confirms the robustness of the SLDT model combined with the copy mechanism, even when tested on unseen data.

The performance trends observed in this analysis are consistent with those reported in related research on the Wizard of Wikipedia dataset. For example, previous studies have shown that incorporating external knowledge and employing effective retrieval mechanisms enhance the response quality in knowledge dialogue systems.

4.5.4 Performance of Leaderboard Submission

In this section, we present a comprehensive analysis of various models' performance on the FrViDoc2Bot test set, focusing on response generation. Table 5 provides the automatic evaluation results for different models, showcasing their performance on metrics. The models in consideration are Ours-FiD, Ours-R_drop, Ours-Prompt, Ours-Post_pretrain, and Ours-Ensemble.

Ours-FiD is a model that leverages the Fusion-in-Decoder (FiD) (Izacard and Grave, 2021) mechanism, which has been demonstrated to improve knowledge integration and retrieval capabilities in large-scale language models. Despite the promise of the FiD mechanism, our implementation yields relatively modest performance in comparison to other models, suggesting that further optimization is required.

Ours-R_drop employs the R-drop (Wu et al., 2021) regularization technique, which encourages the model to generate diverse responses by minimizing the KL-divergence between two independently sampled outputs. This model exhibits improvements over Ours-FiD in various metrics, particularly DIS-1 and DIS-2, indicating that the R-drop technique contributes positively to response diversity.

Ours-Prompt focuses on utilizing prompt engineering to enhance the model's contextual understanding and control. The model's performance on most metrics surpasses that of Ours-FiD and Ours-R_drop, which highlights the effectiveness of prompt engineering in improving the model's ability to generate more contextually relevant and coherent responses.

Ours-Post_pretrain incorporates additional post-pretraining steps to fine-tune the model on the specific task of response generation in the Chinese and Englinsh of FrViDoc2Bot dataset. This model demonstrates superior performance across all metrics, especially in F1 and PPL scores, as compared to the previous models. The results support the notion that further task-specific pretraining can lead to significant performance gains.

Lastly, Ours-Ensemble combines the strengths of the aforementioned models by employing a voting-based ensemble method. This approach achieves the highest scores across all metrics, underlining the benefits of leveraging diverse model architectures and techniques in an ensemble setting.

5 Conclusion

In this paper, we present a novel SLDT method for multilingual document-grounded dialogue , with a focus on addressing the challenges of selecting the most relevant documents for conversation and generating informative responses based on the selected knowledge. We then present an extensive experimental evaluation of our method, demonstrating its effectiveness in comparison to existing approaches.

Acknowledgement

The research work is supported by National Key R&D Program of China (No.2022YFB3904700), Key Research and Development Program of in Shandong Province (2019JZZY020102), Key Research and Development Program of Jiangsu Province (No.BE2018084), Industrial Internet Innovation and Development Project in 2021 (TC210A02M, TC210804D), Opening Project of Beijing Key Laboratory of Mobile Computing and Pervasive Device.

Limitations

Our method relies on large-scale computing power and can only achieve the best results through NVIDIA-A100-80G training.

References

Suman Banerjee, Nikita Moghe, Siddhartha Arora, and Mitesh M Khapra. 2018. A dataset for building code-mixed goal oriented conversation systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3766–3780.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. Plato: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96.

Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23.

Emile Chapuis, Pierre Colombo, Matthieu Labeau, and Chloe Clavel. 2021. Code-switched inspired losses for generic spoken dialog representations. *arXiv preprint arXiv:2108.12465*.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 362–367.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

DAMO_ConvAI. 2023. French and vietnamese document-grounded dialogue data set. In *modelscope*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019a. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. *International Conference on Learning Representations*.

A Seza Doğruöz, Sunayana Sitaram, Barbara Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666.

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

Chang Gao, Wenxuan Zhang, and Wai Lam. 2022. UniGDD: A unified generative framework for goal-oriented document-grounded dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Dublin, Ireland. Association for Computational Linguistics.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Gautier Izacard and Édouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880.

Pei Ke, Jian Guan, Minlie Huang, and Xiaoyan Zhu. 2018. Generating informative responses with controlled sentence function. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1499–1508.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations*.

Tuan Lai, Giuseppe Castellucci, Saar Kuzi, Heng Ji, and Oleg Rokhlenko. 2023. External knowledge acquisition for end-to-end document-oriented dialog systems. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3615–3629, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 12–21.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

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. *arXiv preprint arXiv:1907.11692*.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. Durecdial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4335–4347.

Zeming Liu, Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, and Hua Wu. 2022a. Where to go for the holidays: Towards mixed-type dialogs for clarification of user goals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1024–1034.

Zeming Liu, Ding Zhou, Hao Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, Ting Liu, and Hui Xiong. 2022b. Graph-grounded goal planning for conversational recommendation. *IEEE Transactions on Knowledge and Data Engineering*.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020b. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8433–8440.

Longxuan Ma, Wei-Nan Zhang, Mingda Li, and Ting Liu. 2020. A survey of document grounded dialogue systems (DGDS). *CoRR*, abs/2004.13818.

Zhanyu Ma, Jian Ye, Xurui Yang, and Jianfeng Liu. 2022. Hcld: A hierarchical framework for zero-shot cross-lingual dialogue system. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4492–4498.

Clara Isabel Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1, pages 5328–5339. Association for Computational Linguistics.

Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and Maarten de Rijke. 2021. Initiative-aware self-supervised learning for knowledge-grounded conversations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 522–532.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Le Qi, Shangwen Lv, Hongyu Li, Jing Liu, Yu Zhang, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ting Liu. 2022. DuReader$_{\text{vis}}$: A Chinese dataset for open-domain document visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1338–1351, Dublin, Ireland. Association for Computational Linguistics.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2021. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*.

Sonal Sannigrahi, Josef Van Genabith, and Cristina España-bonet. 2023. Are the best multilingual document embeddings simply based on sentence embeddings? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2261–2271, Dubrovnik, Croatia. Association for Computational Linguistics.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 2692–2700.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Songlin Yang and Kewei Tu. 2022. Bottom-up constituency parsing and nested named entity recognition with pointer networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2403–2416.

Shiwei Zhang, Yiyang Du, Guanzhong Liu, Zhao Yan, and Yunbo Cao. 2022a. G4: Grounding-guided goal-oriented dialogues generation with multiple documents. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 108–114, Dublin, Ireland. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL (demo)*.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022b. Summ^{n}: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.