Enhancing Multilingual Document-Grounded Dialogue Using Cascaded Prompt-Based Post-Training Models

Anonymous ACL submission

Abstract

The Dialdoc23 shared task presents a Multilingual Document-Grounded Dialogue Systems (MDGDS) challenge, where system responses are generated in multiple languages using user's queries, historical dialogue records and relevant passages. A major challenge for this task is the limited training data available in low-resource languages such as French and Vietnamese. In this paper, we propose Cascaded Prompt-based Post-training Models, dividing the task into three subtasks: Retrieval, Reranking and Generation. We conduct post-training on high-resource language such as English and Chinese to enhance performance of low-resource languages by using the similarities of languages. Additionally, we utilize the prompt method to activate model's ability on diverse languages within the dialogue domain and explore which prompt is a good prompt. Our comprehensive experiments demonstrate the effectiveness of our proposed methods, which achieved the first place on the leaderboard with a total score of 215.40 in token-level F1, SacreBleu, and Rouge-L metrics[1].

1 Introduction

Document-Grounded Dialogue Systems (DGDS) have emerged as a research focus in the natural language processing field. They leverage documents to provide targeted information for specialized tasks such as question answering and recommendations (Chen et al., 2019; Rashkin et al., 2021). These systems ensure accuracy and reliability by leveraging comprehensive knowledge bases while enhancing real-time responsiveness and information retrieval efficiency (Gao et al., 2022). Additionally, they can accommodate the expanding scalability of new documents and knowledge sources (Rashkin et al., 2021). Nonetheless, these systems encounter challenges when operating with low-resource languages, including limited training data (Dabre et al., 2019; Gritta et al., 2022), and significant disparities in grammar, vocabulary, and semantics across languages (Artetxe et al., 2017). To address these challenges, researchers are developing multilingual approaches to improve the performance of low-resource languages in DGDS.

The DialDoc23 shared task introduces training and evaluation datasets for MDGDS in Vietnamese and French. The training dataset comprises three distinct components: query, passage, and response. Additionally, the dataset includes a set of documents for retrieval. The query combines the historical dialogue with the current inquiry. During inference, the intelligent agent retrieves the most relevant document from the document set based on the query and generates a response. Notably, this task focuses on low-resource languages, setting it apart from previous tasks.

In this paper, we propose cascaded prompt-based post-training models to solve MDGDS challenge. As inspried by Re2G (Glass et al., 2022) framework, our approach tackles the overall task by dividing it into three subtasks: Retrieval, Reranking, and Generation, with parallel training and sequential inference. As illustrated in Figure 1, the retrieval step identifies top k relevant passages, followed by reranking to select the most relevant passage, and in generation step the query and passage information are incorporated to generate the final response. To enhance the performance of retrieval and generation in low-resource languages such as French and Vietnamese, we conduct post-training on high-resource languages such as English and Chinese to learn language similarities. Additionally, we activate the models' capabilities in diverse languages within the dialogue domain by employing the prompt method. Besides, We employ domain loss function to align the domain of the query and passage during retrieval training. We conducted comprehensive experiments on the

[1]The entry information provided about the participating organization is inaccurate.

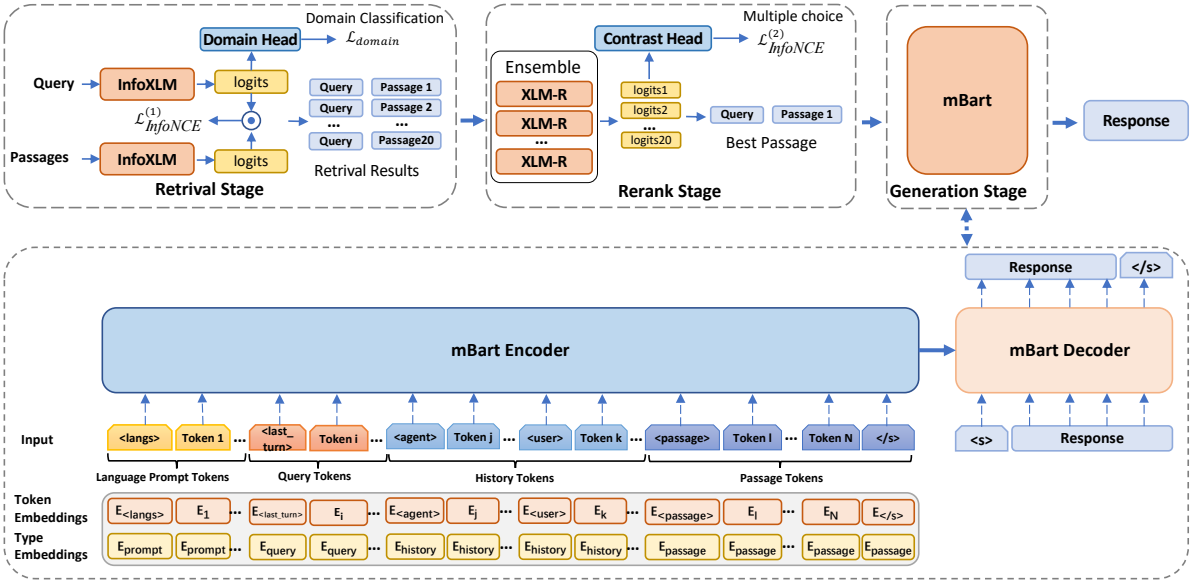

Figure 1: The framework comprises three main stages: (1) Retrieval Stage, which retrieves the top k relevant passages based on a dialogue context. (2) Reranking Stage, which reranks the top k retrieved passages to find the candidate passage. (3) Generation Stage, which generates a system response using user's query, historical dialogue, selected passage, and language prompts.

DialDoc23 shared task, which demonstrated the effectiveness of our proposed methods and resulted in the first place position in the competition.

2 Related Works

2.1 Document Grounded Dialogue Systems

In recent years, substantial advances have been made in DGDS, facilitated by high-quality annotated datasets like SQuAD 2.0 (Rajpurkar et al., 2018), CoQA (Reddy et al., 2019), and MultiDoc2Dial (Feng et al., 2021). Retrieval-and-Generation is a typical framework for implementing DGDS. The framework comprises two sequential stages: (i) retrieving relevant passages from knowledge bases, and (ii) generating responses based on the retrieved passages and users' input. To improve knowledge retrieval, scholars have proposed a variety of approaches such as learning sentence embeddings from dialogue (Liu et al., 2022, 2021a), adding a reranker after retriever retrieval (Re2G) (Glass et al., 2022), and using priori and posteriori knowledge selection (Chen et al., 2020). As for generation, recent studies have also introduced new techniques, such as improving dialogue generation via proactively querying grounded knowledge (Zhao et al., 2022) and leveraging fusion-in-decoder (FiD) (Izacard and Grave, 2021).

2.2 Multilingual Dialogue Generation

Multilingual dialogue is a new research topic in DGDS, aiming for high-quality and fluent communication across different languages. Pre-trained language models have the benefit of automatically learning similarities between languages and enabling unsupervised learning to improve performance in conversations across different languages. Models such as XLM-RoBERTa (Conneau et al., 2019), InfoXLM (Chi et al., 2020), mT5 (Xue et al., 2020), and mBART (Tang et al., 2020), can assist in implementing multilingual transfer learning to improve the performance and fluency of multilingual conversations. However, despite recent technological advancements, multilingual dialogues continue to face challenges. In particular, the lack of training data for many of the world's languages, especially those with limited resources and research, has significantly impeded the development of multilingual dialogue generation(Majewska et al., 2023). In order to combat the aforementioned challenges, our model leverages the similarities between language structures to augment post-training data, while incorporating prompt techniques to enhance language comprehension.

3 Method

The proposed method consists of three main stages: retrieval, reranking, and generation, as depicted

Figure 1. Using contrast learning techniques, the retrieval step efficiently identifies top k relevant passages, followed by reranking to select the most pertinent passage. In the generation step, the query and passage information are incorporated to generate the final response. Each step is described in detail in the following section.

3.1 Passage Retrieval

Passage retrieval constitutes a fundamental component of MDGDS. Given the historical dialogue records $\{u_1, u_2..., u_{T-1}\}$ and the user turn u_T, the passage retrieval identify the top-k relevant passages from a given document set $P = \{p_1, p_2, ..., p_M\}$.

Retriever For efficient passage retrieval, we implement a Bi-encoder architecture to encode the dialogue context and passages independently, as described in Zhang et al. (2023). Further, we leverage two InfoXLM cross-lingual models to derive semantic representations. During the inference phase, we regard the input dialogue context C as the search query and retrieve the top-k passages from the document set based on dot product similarity. In the training stage, each training sample consist of three attributes: the dialogue context, relevant passage, and non-relevant passage. In a training batch, the positive passage refers to the relevant passage, while the negative passage includes the non-relevant passage and the other passages in the batch. The objective function $\mathcal{L}_{InfoNCE}$ for the constrastive learning is formulated as:

$$\mathcal{L}^{(1)}_{InfoNCE} = -\log \frac{\exp\left(\boldsymbol{q} \cdot \boldsymbol{p}^{+}/\tau\right)}{\sum_{\boldsymbol{p}\in \boldsymbol{P}^{\pm}} \exp\left(\boldsymbol{q} \cdot \boldsymbol{p}/\tau\right)} \quad (1)$$

where $\boldsymbol{q}$ and $\boldsymbol{p}$ represent the semantic features of dialogue context and passage extracted by multilingual models, respectively.

Domain classification The Bi-encoder architecture has exhibited efficacy in text retrieval. Nonetheless, the absence of fine-grained supervision signals might hinder the alignment of semantic features between queries and passages. To surmount this constraint, we suggest incorporating domain classification information to guide the representation learning of queries and to align the encoding information of both queries and passages, without compromising the Bi-encoder architecture's efficiency. Technically, for a given dialogue context C, we derive its domain label y from the associated golden passage and employ a linear layer to classify the dialogue context's semantic feature accordingly. We subsequently train the model by minimizing the cross-entropy loss function $\mathcal{L}_{domain}$.

$$\mathcal{L}_{domain} = -\sum_{i=1}^{d} y_i \cdot log(p_i) \quad (2)$$

$$\mathcal{L} = \mathcal{L}_{InfoNCE} + \alpha \mathcal{L}_{domain} \quad (3)$$

where d represent the number of domain set $\mathcal{D}$, and p_i denote the probability of a given category i, α is a hyper-parameter weighting the domain classification loss.

Retrieval Post-training To further address the low-resource target language problem while leveraging the cross-lingual pretraining model's capabilities, we conducted a post-training on English and Chinese dialogue datasets for the same task. Techinically, we utilized the golden passage of the dialogue as the positive samples and retrieved the most relevant documents using the BM25 algorithm from the remaining document set as negative samples.

3.2 Passage Reranking

Passage reranking is the process of reordering the top-k highest-scoring passages C_p retrieved in the previous step, with the aim of improving the probability of the most relevant passages being retrieved correctly. To perform the reranking, We employ XLM-RoBERTa$_{large}$ as the encoder of the reranker, following the pipeline developed by (Zhang et al., 2023). The reranker concatenated the dialogue context C with the candidate passages $p \in P^{\pm}$, inserting a "<passage>" token between them as a separator. The reranker then utilized a contrastive loss function, known as the InfoNCE loss, to recalculate the scores of the passages. The highest-scoring passage is thereafter selected as an input for generation. The objective function $\mathcal{L}_{InfoNCE}$ is formulated as:

$$\mathrm{S}\left(C|p\right) = \mathrm{Sigmoid}\left\{\mathrm{linear}\left[\textit{XLM-R}\left([C, p]\right)\right]\right\} \quad (4)$$

$$\mathcal{L}^{(2)}_{InfoNCE} = -\log \frac{\exp\left(\mathrm{S}\left(C, p^{+}\right)/\tau\right)}{\sum_{p\in P^{\pm}} \exp\left(\mathrm{S}\left(C, p^{+}\right)/\tau\right)} \quad (5)$$

where $\mathrm{S}(C|p)$ represents the similarity between the dialogue context and passage, which is obtained by applying a Sigmoid activation function and a linear layer to the output of XLM-RoBERTa$_{large}$ model, τ is a temperature factor which is set to 1 in our experiment. To enhance the model's generalization ability, we apply an ensemble method in which multiple models receive the input, and the most relevant passages are voted on separately. Subsequently, the passage with the most votes is used as input for the generation stage.

Although both retrieval and reranking are methods used to evaluate the relevance of passage and dialogue context, they differ in the way they understand and score relevance. Retrieval method employs two encoders to encode the passage and dialogue and then calculates the similarity between them. In contrast, reranking methods prioritize sequence structure and semantic information, enabling a more profound comprehension of the content. Due to reranking requires greater computational resources, it is implemented after retrieval.

3.3 Response Generation

The main objective of response generation is to present the user with a system response u_{T+1} that is constructed using the historical dialogue records $\{u_1, u_2, ..., u_{T-1}\}$, a user turn u_T, and the selected passage p, while ensuring that it blends skillfully into the ongoing discourse.
We leverage the large pre-trained model mBART$_{large}$ (Liu et al., 2020) to deal with multilingual generation task. Our dataset contains a significantly greater amount of data in English and Chinese languages compared to French and Vietnamese. In order to improve the performance of low-resource languages utilizing data-rich languages, we employ prompt-based and post-training techniques.

3.3.1 Input Representation

Language Prompts The prompt method that aims to make better use of pre-trained knowledge has recently been successful in transferring pre-trained language models (PLMs) to downstream tasks (Liu et al., 2021b). Some researchers also find prompts can be effective in multilingual scenarios (Fu et al., 2022b; Huang et al., 2022). We leverage prompt techniques to activate model's capability of different languages. As inspired by Fu et al. (2022b), we design both in-lingual prompts(IP) and cross-lingual prompts(CP). In-lingual prompts refer to the prompts where the language used is identical to the target language. The prompts for the different languages are listed in Table 1. While cross-lingual prompts are the prompts templates which involve using the same language across various languages. We use Vietnamese prompts as the unified prompts for all languages.

Language	Prompts
En	Answer user questions based on document content and historical conversations.
Zh	根据文档内容和历史对话回答用户问题。
Fr	Répondre aux questions des utilisateurs sur la base du contenu des documents et de l'historique des conversations.
Vi	Trả lời câu hỏi của người dùng dựa trên nội dung tài liệu và các cuộc hội thoại lịch sử.

Table 1: In-lingual prompts in different languages

Input Setting For the input of generation model, we define our input to a concatenation:

$$x := [prompt; u_T; u_{T-1}...u_1; p] \quad (6)$$

where $prompt$, u_t, p is the prompt corresponding to the target language, the utterance of turn t, the chosen passage respectively.

Separator tokens We define several separator tokens to delimit different components of the input, as illustrated in Figure 1. We utilize the token $<Langs>\in S$ to correspond with the target language, where the set S is defined as $S=\{<En>,<Zh>,<Fr>,<Vi>\}$. We add $<last_turn>$ before u_T to identify the last query, we utilize $<agent>$ and $<user>$ tokens to specify historical system responses and user's utterance, respectively. $<passage>$ token is added to specify the selected passage. $<s>$ and $</s>$ tokens are used to specify the start and the end of generation tokens.

Type Embedding We use type embedding to distinguish prompt, query, history and passage as illustrated in Figure 1. This embedding comprises of four distinct values.

3.3.2 Training

Training objective Our approach use a sequence-to-sequence language model to achieve multilingual generation training. The objective function is to maximize the log-likelihood of the

Method	TestDev				Test
	F1	**S-BLEU**	**ROUGE-L**	**Score-all**	**Score-all**
Re2G(Baseline)	58.55	42.03	55.83	156.42	-
mBart$_{large}$	67.26	56.94	65.06	189.26	-
+FID	63.54	54.92	62.39	180.85	-
+CP	67.42	57.25	65.39	190.06	-
+CP+Post	69.27	58.39	66.39	194.05	-
+CP+Post+R-drop	69.19	59.13	66.85	195.17	214.46
+IP	68.09	57.56	66.06	191.71	-
+IP+Post	69.95	58.95	67.36	196.26	-
+IP+Post+R-drop	**70.25**	**59.73**	**68.48**	**198.46**	**215.40**

Table 2: Results of generation method on Leaderboard of MDGDS. The "+Fid" method denotes the application of the Fusion-in-Decoder model, while the "+Post" method refers to fine-tuning the model on the post-training model. "+IP" and "+CP" represent the usage of in-lingual prompts and cross-lingual prompts, respectively. Besides, the "+R-drop" method utilizes the R-drop technique.

output text and is defined as follows:

$$\mathcal{L} = -\sum_{i=1}^{|y|} \log P\left(y_i \mid y_{<i}, x; \theta\right) \tag{7}$$

Where $|y|$ is the number of tokens in the decoded text, y_i is the i_{th} token and $y_{<i}$ is the tokens before the time step i. Here, x denotes the input of the model specified by the Equation 6. The symbol θ represents the set of training parameters.

Generation Post-training The post-training method is used to transfer knowledge from high-resource languages to low-resource languages. To begin with, the model is post-trained on English, Chinese, French, and Vietnamese with a response generation task. Here, the French and Vietnamese data undergo translation from the English language. The model is then fine-tuned on our target languages, French and Vietnamese.

R-drop Regularization methods like the dropout technique are crucial in training a deep neural network as they prevent overfitting and enhance the generalization ability of deep models. However, dropout results in a unnegligible inconsistency between the training and inference stages (Ma et al., 2016). R-drop (Wu et al., 2021), which allows each data sample to go through the forward pass twice, is an effective measure to mitigate this inconsistency. R-Drop forces the two forward pass distributions for the same data sample outputted by the different dropout model to be consistent with each other, through minimizing the bidirectional Kullback-Leibler(KL) divergence between the two distributions.

4 Experiments

4.1 Dataset and Evaluation Metrics

We conduct our experiments on DialDoc23 shared task, which introduces multilingual document-grounded dialogue dataset in Vietnamese and French[2]. This dataset contains 797 dialogues in Vietnamese (3,446 turns), 816 dialogues in French (3,510 turns), and a corpus of 17272 paragraphs. Each turn utterance is annotated with a number of grounding passages and a corresponding response. And we incorporate additional English and Chinese datasets for post-training. Vietnamese language has a significant number of words derived from Chinese while English and French both belong to the Indo-European language family. We utilize the Doc2Bot dataset (Fu et al., 2022a), which comprises 5760 turns of dialogue in Chinese, and MultiDoc2Dial (Feng et al., 2021), containing 26,506 turns of dialogue in English.

The leaderboard evaluation method employs the token-level F1 score (F1), SacreBLEU (S-BLEU), and ROUGE-L metrics (Feng et al., 2021).

4.2 Experiment Detail

For the retrieval training stage, we utilized a batch size of 128 and a learning rate of 1e-4 and 2e-5 for post-training and fine-tuning, respectively. And retrieval passage number top-k is 20. In the reranking training stage, we set the batch size to 20 and the learning rate to 2e-5. During the generation stage, we used a batch size of 32 with a learning

[2]https://modelscope.cn/datasets/DAMO_ConvAI/FrViDoc2Bot

Model	$R@1$	$R@5$	$R@10$	$R@20$
XLM-R$_{base}$	48.75	68.25	76.25	81.25
XLM-R$_{large}$	55.75	73.25	80.25	88.00
InfoXLM$_{large}$	57.75	76.75	81.75	89.00
+Post	62.25	80.25	85.75	90.50
+Post+DomainCls	**64.50**	**82.50**	**87.00**	**91.25**

Table 3: Retrieval results on the development set. The "+Post" method refers to the use of the InfoXLM$_{large}$ multilingual pre-training model, followed by post-training with Chinese and English languages, and finally fine-tuning on the target language dataset. "DomainCls" represents the adoption of topic category optimization for sentence representations within dialogue records.

Model	$R@1$	$R@2$	$R@3$	$R@5$
XLM-R$_{base}$	80.50	86.25	86.25	94.50
XLM-R$_{large}$	92.50	97.00	98.25	99.00
+Ensemble	**93.75**	-	-	-

Table 4: Reranking results on the development set. The "+Ensemble" method involves the integration of 10 XLM-Roberta-large models created in the same training, and subsequently making the final selection through a voting process to identify the best passages.

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 α 0.02 (Wu et al., 2021). For post-training, we post-train the model on English, Chinese, French, and Vietnamese with a response generation task. Here, the French and Vietnamese data undergo translation from the English language. During each training session, AdamW is utilized as our optimizer with a 10% linear warmup technique. All experiments are conducted on an NVIDIA A100 GPU. To select the best model, we separated 200 French and 200 Vietnamese samples as our validation set. For testing, we utilize two test sets, referred to as DevTest and Test, obtained from the Leaderboard platform, each consisting of 194 dialogues. Since the Test dataset is not accessible to the public now and only the Score-all is visible on the leaderboard, we opted to present only the Score-all result. And due to the limit on the number of submissions for the Test dataset and the closure of the leaderboard, we only have the results of a relatively good performance.

4.3 Experimental Results and Analysis

Retrivel Results Table 3 presents the experimental results on the validation set for different multilingual models, with post-training and domain classification.

In the experimental setup, we evaluated the capabilities of XLM-R$_{base}$, XLM-R$_{large}$, and InfoXLM$_{large}$ multilingual models to identify the most suitable cross-lingual model. Furthermore, we conducted post-training on the InfoXLM$_{large}$ model using Chinese and English, and then fine-tuned it on the target language. Moreover, we assessed the effectiveness of optimizing dialogue content representation using topic category information based on the previous two steps.

Experimental results indicate that the performance of InfoXLM$_{large}$ surpasses that of XLM-R$_{large}$. Furthermore, post-training of the pre-trained model has improved the $\boldsymbol{R@20}$ score by 0.75. Additionally, introducing domain-specific supervision signals in the representation learning of dialogue content can enhance the semantic feature representation.

Reranking Results Table 4 presents the experimental results on the validation set for different multilingual models. The results illustrate that employing a model ensemble in the reranking stage yields an improvement of 1.25 in $\boldsymbol{R@1}$ score.

Generation Results Tabel 2 presents the results of different methods. Our generation methods employ the passage selected through the best retrieval and reranking models. Our method outperforms the baseline by a significant margin. This improvement can be attributed to both the generative model's design and the retrieval of the most relevant passages by the first two tasks.

To determine which type of prompt is more effective, we conducted experiments using both in-lingual and cross-lingual prompts. It is shown that in-lingual prompts outperform cross-lingual prompts in all settings. We think that the model's ability in various languages is triggered by distinct language prompts. This makes it easier to recall knowledge from the pre-training stage using in-lingual prompts.

To leverage the retrieval of multiple passages by the first two stages, we conducted an experiment using Fusion-in-Decoder (FiD) (Izacard and Grave, 2021). The FiD model employs the seq2seq framework to encode each passage independently with a query and subsequently decode all the encoded features to generate responses. Specifically, we configured the encoder to accept two passages

as input. The results indicate that the FiD model does not perform well in our generation task. We think this is due to the fact that the gold response is highly relevant to the retrieved passage, whereas FiD considers the top 2 passages, introducing noise to the model.

The results indicate that the method of post-training on datasets of English, Chinese, French, and Vietnamese followed by fine-tuning on the target languages, French and Vietnamese, enhances the performance a lot. The post-training method improves the performance of both cross-lingual prompts and in-lingual prompts considerably, yielding scores of 3.99 and 4.55 respectively. This suggests that using high-resource languages to enhance low-resource languages, by leveraging the similarities between the languages, can be an effective approach. Additionally, when combined with R-drop, it further enhances the performance of cross-lingual prompts and in-lingual prompts by 1.12 and 1.90, respectively, offering an effective solution to mitigate the inconsistency between training and inference.

5 Conclusion

In this paper, we propose a cascaded prompt-based post-training framework comprising Retrieval, Reranking, and Generation three-stage, to solve the MDGD challenge. To enhance the retrieval and generation performance in low-resource languages such as French and Vietnamese, we exploit the similarities between these and high-resource languages such as Chinese and English by applying post-training techniques. Prompt method are used to activate model's ability in a specific language and dialogue domain, and in-lingual prompts show superior results. Furthermore, we employ DomainCls loss function in retrieval, emsemble method in Rerank, and R-drop method to attain the best results in the Dialdoc23 shared task.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. *arXiv preprint arXiv:1908.05391*.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3426–3437.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1410–1416.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. Multidoc2dial: Modeling dialogues grounded in multiple documents. *arXiv preprint arXiv:2109.12595*.

Haomin Fu, Yeqin Zhang, Haiyang Yu, Jian Sun, Fei Huang, Luo Si, Yongbin Li, and Cam Tu Nguyen. 2022a. Doc2Bot: Accessing heterogeneous documents via conversational bots. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1820–1836, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022b. Polyglot prompt: Multilingual multitask promptraining. *arXiv preprint arXiv:2204.14264*.

Chang Gao, Wenxuan Zhang, and Wai Lam. 2022. Unigdd: A unified generative framework for goal-oriented document-grounded dialogue. *arXiv preprint arXiv:2204.07770*.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, rerank, generate. *arXiv preprint arXiv:2207.06300*.

Milan Gritta, Ruoyu Hu, and Ignacio Iacobacci. 2022. Crossaligner & co: Zero-shot transfer methods for task-oriented cross-lingual natural language understanding. *arXiv preprint arXiv:2203.09982*.

Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. Zero-shot cross-lingual transfer of prompt-based tuning with a unified multilingual prompt. *arXiv preprint arXiv:2202.11451*.

Gautier Izacard and Edouard 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, Online. Association for Computational Linguistics.

Che Liu, Rui Wang, Junfeng Jiang, Yongbin Li, and Fei Huang. 2022. Dial2vec: Self-guided contrastive learning of unsupervised dialogue embeddings. *arXiv preprint arXiv:2210.15332*.

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021a. Dialoguecse: Dialogue-based contrastive learning of sentence embeddings. *arXiv preprint arXiv:2109.12599*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Xuezhe Ma, Yingkai Gao, Zhiting Hu, Yaoliang Yu, Yuntian Deng, and Eduard Hovy. 2016. Dropout with expectation-linear regularization. *arXiv preprint arXiv:1609.08017*.

Olga Majewska, Evgeniia Razumovskaia, Edoardo M Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual dialogue dataset creation via outline-based generation. *Transactions of the Association for Computational Linguistics*, 11:139–156.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. *arXiv preprint arXiv:2107.06963*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

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. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yeqin Zhang, Haomin Fu, Cheng Fu, Haiyang Yu, Yongbin Li, and Cam-Tu Nguyen. 2023. Coarse-to-fine knowledge selection for document grounded dialogs.

Xiangyu Zhao, Longbiao Wang, and Jianwu Dang. 2022. Improving dialogue generation via proactively querying grounded knowledge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6577–6581. IEEE.