Cross-lingual Data Augmentation for Document-grounded Dialog Systems in Low Resource Languages

Qi Gou, Zehua Xia, Wenzhe Du
State Key Laboratory for Novel Software Technology, Nanjing University, China
{qi.gou, zehuaxia, gowott}@smail.nju.edu.cn

Abstract

This paper proposes a framework to address the issue of data scarcity in Document-Grounded Dialogue Systems(DGDS). Our model leverages high-resource languages to enhance the capability of dialogue generation in low-resource languages. Specifically, We present a novel pipeline CLEM (Cross-Lingual Enhanced Model) including adversarial training retrieval (Retriever and Reranker), and Fid (fusion-in-decoder) generator. To further leverage high-resource language, we also propose an innovative architecture to conduct alignment across different languages with translated training. Extensive experiment results demonstrate the effectiveness of our model and we achieved 4th place in the DialDoc 2023 Competition. Therefore, CLEM can serve as a solution to resource scarcity in DGDS and provide useful guidance for multilingual alignment tasks.

1 Introduction

Document-Grounded Dialogue System (DGDS) is a meaningful yet challenging task, which not only allows content accessible to end users via various conversational interfaces, but also requires generating faithful responses according to knowledge resources.

However, in real-world scenarios, we may not have abundant resources to construct an effective dialogue system due to the low resources of some minority languages such as Vietnamese and French. Previous works only consider building a DGDS in high-resource languages with rich document resources such as English and Chinese (Feng et al., 2021; Fu et al., 2022), which is contrary to real-world situations. Extensive minority languages struggle to build well-founded chatbots due to the low resource of documents.

Therefore, how to generate evidential responses under a scarce resources setting deserves our attention. To address this issue, we propose a novel architecture to leverage high-resource languages to supplement low-resource languages, in turn, build a fact-based dialogue system. Thus, our model can not only handle high-resource scenarios but also generate faithful responses under low-resource settings.Our key contributions can be split into three parts:

- We proposed a novel framework, dubbed as CLEM, including adversarial training Retriever, Re-ranker and Fid (fusion-in-decoder) generator.
- We presented the novel architecture of translated training and three-stage training.
- Extensive results demonstrated the effectiveness of CLEM. Our team won the 4th place in the Third DialDoc Shared-task competition.

2 Related Work

Document Grounded Dialogue System is an advanced dialogue system that requires the ability to search relevant external knowledge sources in order to generate coherent and informative responses. To evaluate and benchmark the performance of such systems, existing DGDS datasets can be broadly classified into three categories based on their objectives: 1) **Chitchat**, such as WoW (Dinan et al., 2019), Holl-E (Moghe et al., 2018), and CMU-DoG (Zhou et al., 2018). These datasets typically involve casual and open-ended conversations on various topics; 2) **Conversational Reading Comprehension (CRC)**, which requires the agent to answer questions based on understanding of a given text passage. Examples of CRC datasets include CoQA (Reddy et al., 2019), Abg-CoQA (Guo et al., 2021), and ShARC (Saeidi et al., 2018); and 3) **Information-seeking Scenarios**, such as Doc2dial (Feng et al., 2020), Multidoc2dial (Feng et al., 2021), and Doc2bot (Fu et al., 2022), where the agent needs to retrieve

relevant information from one or more documents to address a user's query.

Cross-lingual Data Augmentation has emerged as an effective approach to address the challenges of multilingual NLP tasks (Zhang et al., 2019; Singh et al., 2019; Riabi et al., 2021; Qin et al., 2020; Bari et al., 2021). Particularly in low-resource language settings, DA has demonstrated its usefulness (Liu et al., 2021; Zhou et al., 2022b,a). Explicit DA techniques mainly involve translation-based templates, such as word-level adversarial learning (Bari et al., 2020) and designed translation templates (Liu et al., 2021; Zhou et al., 2022b). Implicit data augmentation techniques, on the other hand, focus on modeling instead of expanding datasets like representation alignment (Mao et al., 2020), knowledge distillation (Chen et al., 2021) and transfer learning (Schuster et al., 2019).

3 Task Description

Formulation. We aim to improve the performance of DGDS in low-resource languages (Vietnamese and French). Formally, given labeled set $D = \{x_i, p_i, r_i\}, i \in [1, N_D]$, where N_D denotes the number of data and x_i, p_i, r_i denotes the input, grounding passage and response. Note that the input is obtained by concatenating the current turn and previous context. In addition, we have access to some high-resource language labeled datasets U with size N_U, where $N_U \gg N_D$. Our goal is to explore how to utilize high-resource datasets to enhance performance in low-resource languages (Vietnamese and French).

We have access to two large datasets, namely Multidoc2dial (Feng et al., 2021) for English and Doc2bot for Chinese(Fu et al., 2022). To fully take advantage of these high-resource datasets to enhance the performance in French and Vietnamese, we conducted translated training and generated pseudo-labeled training sets in Vietnamese and French. Specifically, we utilized the Baidu API[1] and Tencent API[2] to translate English and Chinese into French and Vietnamese, separately. Notably, English and French are Indo-European languages, indicating a common ancestral language, and Chinese and Vietnamese share historical and cultural connections and have influenced each other. Our methodology involved augmenting the training set by translating 5000 English examples into French and 5000 Chinese examples into Vietnamese. After filtering out instances of poor quality and excessive length, we ultimately derived 4980 En-Fr and 4908 Zh-Vi pseudo examples.

[1]https://fanyi-api.baidu.com/api/trans/product/index
[2]https://www.tencentcloud.com/products/tmt

data	number of turns
Chinese corpus	5760
English corpus	26506
Shared-Task/train	3446 (Vi) and 3510(Fr)
Zh-Vi	4908
En-Fr	4980
Shared-Task/dev	95(Vi) and 99(Fr)
Shared-Task/test	94(Vi) and 100(Fr)

Table 1: Statistics of provided datasets. Chinese and English corpus is provided by the third workshop committee of DialDoc. Zh-Vi and En-Fr means the number of translated data from Chinese to Vietnamese and from English to French respectively.

Now we have three training data, cross-lingual training data D, translated pseudo data $D^{'}$ and downstream fine-tuning data D^t. We will show how to use these data in Section 4.4. And the statistics are presented in Table 1.

4 Methodology

We adopt the Retrieve-Rerank-Generation architecture (Glass et al., 2022; Zhang et al., 2023) and incorporate adversarial training into both the Retriever and Re-ranker components. To address the low-resource DGDS scenario, we propose a novel three-stage training approach.

4.1 Passage-Retriever With FGM

Given an input x, the retriever aims to retrieve the most relevant top-k documents $\{z_i\}_i^k$ from a large candidate pool. We follow the schema of conventional Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) for passage retrieval:

$$s(q) = \text{XLM-R}_1(q)$$
$$s(z) = \text{XLM-R}_2(z)$$
$$p_\phi(z|q) \propto \text{dot}[s(q)^\top s(z)]$$

To improve multi-lingual performance further, where the encoder is initialized from XLM-RoBERTa (Conneau et al., 2019) denote as XLM-R which are used to convert question templates into dense embedding vectors for passage retrieval. Sub-linear time search can be achieved

with a Maximum Inner Product Search (MIPS) (Shrivastava and Li, 2014).

In addition, inspired by FGM (Miyato et al., 2017), we extend the adversarial training to document retrieval. We apply infinitesimal perturbations on word embeddings to increase the learning difficulty by constructing adversarial examples. Based on this, the passage retriever is regularized and has better generalization performance since it has to retrieve the correct relevant documents under the attack of adversarial examples.

4.2 Passage-Reranker with FGM

Given a shortlist of candidates, the goal of Reranker is to capture deeper interactions between a query x and a candidate passage p. Specifically, the query x and passage p are concatenated to form the input for XLM-RoBERTa (Conneau et al., 2019). And the pooler output of XLM-RoBERTa is considered as similarity score:

$$P(p|q) = \text{SoftMax}\left(\text{Linear}\left(\text{XLM-R}([p, q])\right)\right)$$

As in the previous stage, we still employed FGM (Miyato et al., 2017) to add perturbations to word embeddings.

4.3 Knowledge-Enhancement Generation

The generator aims to generate correct and factual responses according to the candidates of passages. The key problem is how to leverage the knowledge of passage candidates as much as possible. we adopt Fusion-in-Decoder(FiD) (Izacard and Grave, 2021) as our response generator. During generation, FiD will first encodes every input with multiple passages independently through encoder, and then decodes all encoded feature jointly to generate final response. Concisely, the decoder has extra Cross Attention on more passages feature. This is significant because it is equivalent to improve grounding passage accuracy from top-k to top-n. Note that $k \ll n$ due to the CUDA memory limitation.

Since prompt-learning is effective in generation proved by previous work (Wei et al., 2021), we also adopt this way by adding the prompt to the front of input query. We choose *"please generate the response:"* as our prompt, so the final input of generator is "`prompt` `<query>` `query` `<passage>` `passage`", where `<prompt>` and `<passage>` are special tokens.

Model	Total
Baseline	156.42
CLEM	**201.0913**

Table 2: Performance of CLEM on Test set

4.4 Training Process

Our training process consists of three stages. In the first stage, we use all available Chinese and English training corpora to pre-train the model, aiming to develop its primary cross-lingual perception capability. We incorporate downstream fine-tuning data in this stage as well. We denote this stage as $T(D + D^t)$, where T represents training.

In the second stage, we train the model using translated pseudo data, which includes both noisy data and downstream fine-tuning data. We denote this stage as $T(D^{'} + D^t)$.

Finally, we fine-tune the model from the second stage on downstream low-resource training data. We denote this stage as $F(D^t)$, where F represents fine-tuning.

Therefore, the complete training process can be represented as $T(D + D^t)T(D^{'} + D^t)F(D^t)$. In the Experiment section, we also explore other training processes, such as two-stage training and direct fine-tuning.

5 Experiments and Results

In this section, we will introduce our datasets and baseline system. Additionally, we will demonstrate the effectiveness of each component in our methodology, such as adversarial training and the novel training process.

5.1 Datasets

We train CLEM on the given shared task datasets, containing Vietnamese (3,446 turns), 816 dialogues in French (3,510 turns) and a corpus of 17272 paragraphs in ModelScope[3], where each dialogue turn is grounded in a paragraph from the corpus. Moreover, we also utilize Chinese (5760 turns) and English (26,506 turns) as additional training data.

5.2 Baseline System

The baseline follows the pipeline of Retrieval, Re-rank and Generation. It simply uses DPR

[3] https://modelscope.cn/

CLEM	F1	BLEU	ROUGE	Total
CLEM-Full	**66.51**	**57.45**	**64.38**	**188.34**
CLEM(two-stage)	65.52	55.23	63.15	183.9
CLEM(fine-tune)	63.76	53.41	61.47	178.64
CLEM(two-stage w/o Zh-Vi)	64.24	54.51	62.18	180.93
CLEM(two-stage w/o En-Fr)	61.99	51.21	60.28	173.48
CLEM(w/o prompt)	64.34	55.12	62.31	181.77

Table 3: Ablation results of Modelon Development set. Here, the best are marked with **Bold**. Two-stage means we do not use original Chinese and English data. fine-tune means we just use downstream training data.

CLEM	R@1	R@5	R@20	MRR@5
retrieval	0.57	0.78	0.87	0.65
retrieval†	0.62	0.77	0.87	0.68
re-rank	0.74	0.84	0.87	0.78
re-rank †	**0.76**	**0.85**	0.87	**0.79**

Table 4: Effect of FGM on Development set, where †means we use adversarial training

(Karpukhin et al., 2020) as retriever and Transformer Encoder (Vaswani et al., 2017) with a linear layer as re-ranker.

5.3 Result and Analysis

We evaluate the generation results based on token level F1, SacreBLEU and Rouge-L. The final result is the sum of them. As shown in Table 2, CLEM has a significant improvement by 28% on total result compared to strong baseline, which demonstrates the effectiveness of our method.

5.3.1 Ablation Study

We study the impact of different components of-CLEM, where the results are given in Table 3.

Training process we compare CLEM with two-stage training and fine-tuning directly. The former only contains translated corpus without original Chinese and English data, which can be denoted by $T(D^{'} + D^{t})F(D^{t})$. While the latter means we only use downstream fine-tuning data denoted by $F(D^{t})$. From the first three lines of Table 3, we can observe that CLEM has superior performance than two-stage training which means CLEM can leverage cross-lingual corpus to do a better language alignment for downstream training and get a better initialization. Not surprisingly, two-stage training outperforms fine-tuning directly which echos the Translated Training (Singh et al., 2019)

Different pseudo corpus As described in section 3, we leverage two translated pseudo corpus Zh-Vi and En-Fr. We also study the impact of each set with two-stage training. From 4th and 5th line of Table 3, the performance without Zh-Vi(Chinese to Vietnamese) and En-Fr(English to French) will decrease, which proved that the translated corpus is useful for shared task.

Without prompt We also run the experiments without prompt to explore the impact of prompt. From the last line of Table 3, the performance of CLEM will decrease sharply.

Without FGM We also explore the effectiveness of FGM (Miyato et al., 2017) at retriever and re-ranker. Results are listed in Table 4. We can observe significant improvements from retrieval to re-rank which prove the effectiveness of re-rank.

6 Conclusion

This paper introduces CLEM, a novel pipeline for document-grounded dialogue systems that uses a "retrieve, re-rank, and generate" approach. To address the issue of low performance due to limited training data, we extend the adversarial training to the document Retriever and Re-ranker components. Additionally, CLEM leverages high-resource languages to improve low-resource languages and develops a new training process under data-scarce settings.

Experimental results demonstrate that CLEM outperforms the strong, competitive baseline and achieved 4th place on the leaderboard of the third DialDoc competition. These findings provide a promising approach for generating grounded dialogues in multilingual settings with limited training data and further demonstrate the effectiveness of leveraging high-resource languages for low-resource language enhancement.

Acknowledgements

References

M. Saiful Bari, Shafiq R. Joty, and Prathyusha Jwalapuram. 2020. Zero-resource cross-lingual named entity recognition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7415–7423. AAAI Press.

M Saiful Bari, Tasnim Mohiuddin, and Shafiq Joty. 2021. UXLA: A robust unsupervised data augmentation framework for zero-resource cross-lingual NLP. 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 1978–1992, Online. Association for Computational Linguistics.

Weile Chen, Huiqiang Jiang, Qianhui Wu, Börje Karlsson, and Yi Guan. 2021. AdvPicker: Effectively Leveraging Unlabeled Data via Adversarial Discriminator for Cross-Lingual NER. 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 743–753, Online. Association for Computational Linguistics.

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*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Haomin Fu, Yeqin Zhang, Haiyang Yu, Jian Sun, Fei Huang, Luo Si, Yongbin Li, and Cam Tu Nguyen. 2022. 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.

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.

Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-coQA: Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction*.

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.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. 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 5834–5846, Online. Association for Computational Linguistics.

Xin Mao, Wenting Wang, Huimin Xu, Man Lan, and Yuanbin Wu. 2020. Mraea: An efficient and robust entity alignment approach for cross-lingual knowledge graph. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 420–428, New York, NY, USA. Association for Computing Machinery.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on*

Empirical Methods in Natural Language Processing, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. 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, IJCAI-20*, pages 3853–3860. International Joint Conferences on Artificial Intelligence Organization. Main track.

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.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2021. Synthetic data augmentation for zero-shot cross-lingual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Anshumali Shrivastava and Ping Li. 2014. Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips). *Advances in neural information processing systems*, 27.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. XLDA: cross-lingual data augmentation for natural language inference and question answering. *CoRR*, abs/1905.11471.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2019. Cross-lingual dependency parsing using code-mixed TreeBank. 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 997–1006, Hong Kong, China. Association for Computational Linguistics.

Yeqin Zhang, Haomin Fu, Cheng Fu, Haiyang Yu, Yongbin Li, and Cam-Tu Nguyen. 2023. Coarse-to-fine knowledge selection for document grounded dialogs. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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*.

Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022a. ConNER: Consistency training for cross-lingual named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8438–8449, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022b. MELM: Data augmentation with masked entity language modeling for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2251–2262, Dublin, Ireland. Association for Computational Linguistics.

A Experiments Hyperparameters

A.1 Hyper-parameters for retriever

train_batch_size=128
epochs=50
max_input_length=512
dropout=0.1
weight_decay=0.1
warmup_steps=1000
gradient_checkpoint_segments=32
optim=adam
learning_rate=4e-05
preKturns=2

A.2 Hyper-parameters for re-reanker

learning_rate=2e-5
dropout=0.1

epochs=20
train_batch_size=1
accumulation_steps=32
weight_decay=0.1
warmup_steps=1000
max_input_length=512
passages=20
preKturns=2

A.3 Hyper-parameters for generator

learning_rate=2e-4
dropout=0.1
epochs=20
accumulation_steps=16
max_grad_norm=1
train_batch_size=1
accumulation_steps=1
weight_decay=0.1
warmup_steps=1000
max_input_length=1024
max_output_length=128
beam_size=3
passages4gen=5
preKturns=2