Exploration of multilingual prompts in document-grounded dialogue

Xiaocheng Zhang[1], **Xuelin Fu**[2], **Yongqing Huang**[3], **Xiaohong Su**[†]
[1,†]Harbin Institute of Technology, Harbin, Heilongjiang, China
[2]Guilin University of Technology, Guilin, Guangxi, China
[3]Guangdong University of Technology, Guangzhou, Guangdong, China
22s136029@stu.hit.edu.cn 1735573894@qq.com 1486590231@qq.com
sxh@hit.edu.cn

Abstract

Transferring DGD models from high-resource languages to low-resource languages is a meaningful but challenging task. Being able to provide multilingual responses to multilingual documents further complicates the task. This paper describes our method at DialDoc23 Shared Task (Document-Grounded Dialogue and Conversational Question Answering) for generate responses based on the most relevant passage retrieved. We divide it into three steps of retrieval, re-ranking and generation. Our methods include negative sample augmentation, prompt learning, pseudo-labeling and ensemble. On the submission page, we rank 2nd based on the sum of token-level F1, SacreBleu and Rouge-L scores used for the final evaluation, and get the total score of 210.25.

1 Introduction

Our team fanjuanju participates in the Third DialDoc Workshop Shared Task co-located with ACL 2023. The goal of this task is to query document knowledge through a multilingual dialogue system. The dataset contains 797 dialogues in Vietnamese (3,446 turns), 816 dialogues in French (3,510 turns), and a corpus of 17272 paragraphs, that each dialogue turn is grounded in a paragraph from the corpus. We need to use the dialogue history and the current query to retrieval the paragraph that supports the answer to the current question, and generate corresponding responses based on the knowledge in the paragraph. The score is calculated based on the sum of token-level F1(Rajpurkar et al., 2016), SacreBleu(Post, 2018) and Rouge-L metrics, hereinafter referred to as F1, Bleu, Rouge respectively.

2 Related Work

2.1 Document-grounded Dialogue (DGD)

When we have a conversation, we usually refer to the document information we know. DGD refers to the technology that uses the document as a reference in the conversation to support the conversation interaction. In practical applications, such as customer service conversation system, smart home control, etc, the document can be a product description, user manual or an article, in this case, documentation is external knowledge provided to the model, and document-based conversations can help people find answers and solve problems faster. Doc2dial(Feng et al., 2020), a doc-based dialogue data set, consists of two tasks: 1. Seeking sentences related to questions from documents (information-seeking); 2. Use the results of the previous step to generate a reasonable response; In Chinese, there are movie-chats published by Tsinghua University, in which both parties are chatting about one or more movies in a dataset; Existing document dialogue data sets mainly focus on the plain text content in documents, while ignoring the importance of common structural information such as title, serial number and table in documents to machine understanding of document content. Therefore, Doc2Bot(Fu et al., 2022), a large-scale multi-domain document dialogue data set in Chinese, was proposed by Fu et al.

2.2 Pre-trained language models

The representation of natural language is to represent human language in a way that is easier for computer to understand. Methods such as word2vec(Mikolov et al., 2013) and glove(Pennington et al., 2014) based on deep learning can represent words with similar semantics, but they can't solve the polysemy problem well. The subsequent Elmo(Peters et al., 2018), which takes into account contextual information, can better solve the polysemy representation problem. And Elmo started the pre-training and fine-tuning paradigm. Since Elmo, transformer(Vaswani et al., 2017), as a more powerful feature extractor than lstm, has been applied to various subsequent pre-

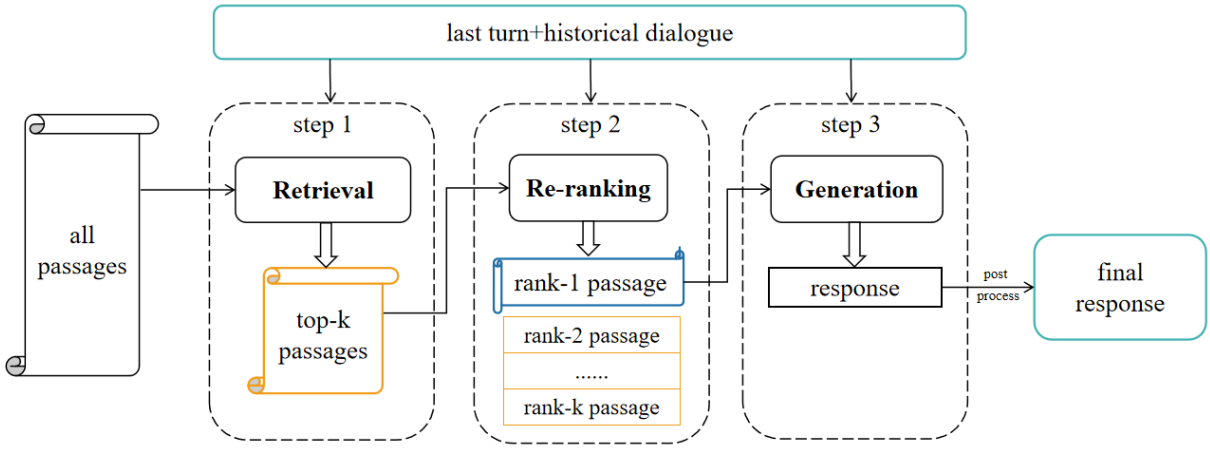

Figure 1: The framework we used in this competition.

training language models (such as GPT(Radford et al., 2018) and BERT(Devlin et al., 2019)), constantly updating the existing optimal results in various tasks of natural language processing. The most classic pre-training language model is BERT, which designed two pre-training tasks to dynamically learn word vectors. There are many subsequent improved versions of BERT, such as roberta dynamic masking(Liu et al., 2021), Bert-wwm(Cui et al., 2021) implementation of full word masking in Chinese, and ernie(Zhang et al., 2019) introducing entity information, etc. GPT series models are more representative of autoregressive models, which learn word vectors by predicting the next word by the current statement. These pre-training models pretrain and learn on large-scale corpus, and then fine-tune downstream tasks to fit the current data. For this competition, we also relied on the "shoulders of giants" of the pre-training language model, and since the data set was geared towards French and Vietnamese, we used a multi-language version of the pre-training language model for this competition.

3 Method

According to baseline(Zhang et al., 2023), we divide the tasks into three steps: retrieval, re-ranking and generation. Firstly, we use the method of contrast learning to train the retrieval model, and expand the negative example in the training process to improve the performance of the retrieval model. In the re-ranking step, we fine-tune the XLM-RoBERTa(Conneau et al., 2020) and InfoXLM(Chi et al., 2021) models, then ensemble the two models to predict the scores of the retrieved paragraphs. In the generation step, we use the prompt learning method to fine-tune MT5(Xue et al., 2021) to generate the corresponding language response, and finally add the pseudo-tag retraining to get the final response. The framework we used in this competition is illustrated in Figure 1.

3.1 Retrieval

Based on the conventional comparative learning training method, the original data set is divided into n small batches of data, and the n mini-batches of data are stored in advance. When training begins, each training batch is constructed with a normal In-Batch(IB) negative sample. At the same time, for n mini-batches of data stored in advance, if i >= 1, the previous batch of data is taken to construct incremental negative samples. We use e_{Query} and $e_{Passage}$ to represent the vectors of query and passage respectively, and use the cosine similarity function to calculate the correlation score between them.

$$cos(e_{Query}, e_{Passage}) = \frac{e_{Query} \cdot e_{Passage}}{\| e_{Query} \| \| e_{Passage} \|} \quad (1)$$

When prescribed to Addictive Margin InfoNCE Loss(Chen et al., 2020)(Yang et al., 2019) and a learnable temperature parameter τ, it's the following:

$$\mathcal{L} = -log \frac{e^{(\varphi(h,r,t)-\gamma)/\tau}}{e^{(\varphi(h,r,t)-\gamma)/\tau} + \sum_{i=1}^{|N|} e^{(\varphi(h,r,t_i))/\tau}} \quad (2)$$

Margin $\gamma > 0$, usually 1.0. is the score of the triplet, which is in the range of -1 to 1. The temperature t is adjustable and $\tau = log\frac{1}{\tau}$ is defined as a learnable parameter(Wang et al., 2022).

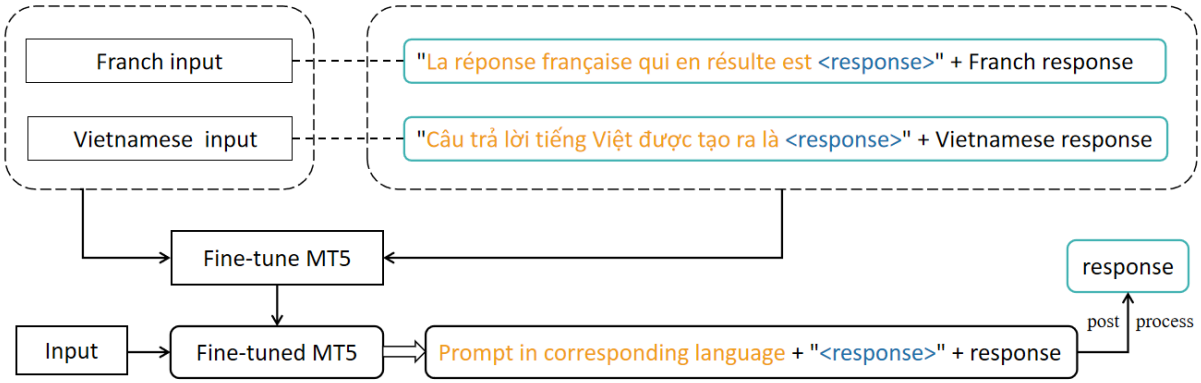

Figure 2: Concrete example of adding a prompt to a response.

3.2 Re-ranking

We fine-tune the XLM-RoBERTa and InfoXLM models on the FrDoc2BotRerank and ViDoc2BotRerank datasets. In the training process, FGM is used for adversarial training to increase the generalization performance of the model. We load the fine-tuned XLM-RoBERTa and InfoXLM models for inference. The query and the passages retrieved from the previous step are spliced separately as the input of the model, and the logits output by the model are weighted average. We use the softmax function to get the probability, and sort according to the probability to get the final result of the re-ranking model.

3.3 Response Generation

We fine-tune MT5 on the FrDoc2BotGeneration and ViDoc2BotGeneration datasets. The input of the model is a simple concatenation of the query and the passage most relevant to the query, and the output is a response. Given the input $x = \{x_i\}_{i=1}^{M}$ and its response $y = \{y_i\}_{i=1}^{N}$, we minimize the following negative log likelihood (NLL) loss:

$$\mathcal{L}_{NLL} = -\sum_{i=1}^{N} logp_\theta(y_i|x, y_{<i}) \qquad (3)$$

We add prompt and pseudo-labels to increase model performance, we also add FGM(Miyato et al., 2017) and AWP(Wu et al., 2020) for confrontation training to improve the generalization ability of the model.

3.3.1 Multilingual prompts

We employ a simple but effective prompt strategy: Add the prefix of the corresponding language to the label to guide the model to generate the response of the corresponding language. The biggest challenge in multilingual generation tasks is the problem of multilingual performance degradation(Zhu et al., 2021), the essence of its performance degradation is the interference between languages. Most of today's multilingual translation models tell the model which language to translate to by adding language tags. Inspired by this, we add corresponding prefixes to French response and Vietnamese response as prompt when fine-tuning MT5. This guides the model to generate responses for the corresponding languages. Then we use post-processing to remove the corresponding prompt in the generated text. See the Figure 2 for specific practices.

3.3.2 Pseudo label

Because the competition does not restrict pseudo-label, we use the fine-tuned model to infer the test set to obtain pseudo-label. We add it to the training set to fine-tune the model again, and load this model for inference to get the result of final test set.

4 Experiments

4.1 Experimental Settings

Our implementations of XLM-RoBERTa, InfoXLM and MT5 are based on the public Pytorch implementation from Transformers[1]. The query encoder and context encoder of the retrieval model both use XLM-RoBERTa-base, and other models are in large size. In the search task, we set the maximum input length of both query and context to 512 tokens, and set to top-48 on the dev-test set and top-100 on the final-test set. The input to both the re-ranking model and the generation model is a concatenation of query and passage. When fine-tuning the re-ranking model and generate model,

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

Table 1: The results of comparative experiments on retrieval model. "pre-batch-neg" means "use the data of the previous batch to expand the negative example" and "top-48" means "the number of retrieval model recalls is set to 48".

Methods	On dev-test set		
	F1	Bleu	Rouge
Baseline	58.39	40.12	55.64
pre-batch-neg	**59.54**	**46.56**	**57.37**
pre-batch-neg/top-48	59.06	46.29	56.79

we truncate the length of the query to 195 tokens and maximum input length to 512 tokens. We fine-tune these models on a single Tesla A100s GPU with 80gb memory, and the three steps of retrieval, re-ranking, and generation take about 10 hours, 24 hours, and 8 hours respectively.

4.2 Experimental Results and Analysis

Since the organizer is not provide the labels of the final-test set, we only did comparative experiments on the dev-test set. We conduct experiments on retrieval, reranking, and generation in sequence, and the current experiment is based on the results of the previous step. Table 1 shows the retrieval contrast experimental results on dev-test set of our method. We fine-tune baseline on the three steps corresponding data sets and get the F1 of 58.39, Bleu of 40.12 and Rouge of 55.64 on the dev-test set. We extend the negative example when fine-tuning the retrieval model, and get the F1 of 59.54, Bleu of 46.56 and Rouge of 57.37. This result proves that the expansion of negative examples in training can improve the performance of retrieval. The top-k of baseline is set to 20. When we expand it to 48 (our setting of the best score on the dev-test set submission page), the performance will slight drop. This is because the re-ranking model at this time is underperforming, and wrong predictions cause the generation model to receive mismatched input.

Table 2 shows the contrast experimental results on dev-test set of re-ranking step. Experiment under top-48, we replace the initial pre-training weight of the re-rank model with XLM-RoBERTa and InfoXLM, both of which are large size(baseline use base size). We get the F1 of 63.14, Bleu of 49.23 and Rouge of 60.78 on XLM-RoBERTa. By comparing top-48 and top-20, it can be seen that after the performance of the re-ranking model is improved, increasing the number of recalls of the retrieval model can improve the score. We add adversarial perturbations during training, and performance has been improved on both models. Our ensemble of the two models shows a slight drop in performance on the dev-test set, but a 2.05-point improvement on the final-test set.

Table 2: The results of comparative experiments on re-ranking model. "top-20" means "the number of retrieval model recalls is set to 20" and "Adv" means “adversarial”.

Methods	On dev-test set		
	F1	Bleu	Rouge
RoBERTa(top-20)	62.74	48.76	60.35
RoBERTa	63.14	49.23	60.78
InfoXLM	62.83	48.75	59.46
RoBERTa(Adv)	63.59	**50.47**	**61.43**
InfoXLM(Adv)	62.77	49.21	60.38
RoBETA(adv)+ InfoXLM(adv)	**63.62**	50.41	61.40

Table 3: The results of comparative experiments on generation model. "GS/Adv/Prompt/PL" in the table respectively represents "greedy search/adversarial/prompt learning/pseudo label".

Methods	On dev-test set		
	F1	Bleu	Rouge
GS	65.76	50.58	64.44
GS/Adv	67.83	58.42	66.59
GS/Adv/prompt	69.56	60.23	67.51
GS/Adv/prompt/PL	**70.14**	**60.98**	**68.26**

Table 3 shows the contrast experimental results of generation on dev-test set. We conduct experiments on the improvement of the generation model under the highest score combination of the current retrieval model and the re-ranking model. We replace the generation strategy from beam search to greedy search and get the F1 of 65.76, Bleu of 50.58 and Rouge of 64.44. During training, we add adversarial perturbations using fgm and awp. The F1, Bleu and Rouge increase to 67.83, 58.42 and 66.59. Then we add a prompt to the response and get the F1 of 69.56, Bleu of 60.23 and Rouge of 67.51. It proves that prompt-learning on task data can further improve performance. At last, we add pseudo-labeled data for training and achieve 70.14 F1, 60.98 Bleu and 68.26 Rouge on the dev-test set. The last method(use ensemble on re-ranking step)achieves 210.25 score on the final-test set.

5 Conclusion

We have introduced our submission for the Third DialDoc Workshop Shared Task. Our team ranks 2nd on the final submission page. We have made improvements on the baseline and tried programs such as negative sample augmentation, ensemble, prompt learning, adversarial training, and pseudo-tagging. There are other methods that could further improve the performance of our model. Try to translate the official Chinese and English data into Vietnamese and French, and then use all available Vietnamese and French data for pre-training. You can try to combine the retrieved top k with the re-ordered relevancy for weighted ranking. Try combining a reorder task with a build task for training. You can try training a dichotomous model to score the generated statements to pick out the responses with the highest scores. Because of time and equipment constraints, we didn't try everything during the competition. We hope the above methods can be helpful to future contestants.

Acknowledgements

We thank the thoughtful suggestions from the reviewers. This work is supported by the National Natural Science Foundation of China (Grant Nos.62272132).

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3576–3588. 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. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese BERT. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3504–3514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Song Feng, Hui Wan, R. Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, and Luis A. Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8118–8128. 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, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1820–1836. Association for Computational Linguistics.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021. A robustly optimized BERT pre-training approach with post-training. In *Chinese Computational Linguistics - 20th China National Conference, CCL 2021, Hohhot, China, August 13-15, 2021, Proceedings*, volume 12869 of *Lecture Notes in Computer Science*, pages 471–484. Springer.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. 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, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022. Simkgc: Simple contrastive knowledge graph completion with pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4281–4294. Association for Computational Linguistics.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. 2020. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Yinfei Yang, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5370–5378. ijcai.org.

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 *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics.

Yaoming Zhu, Jiangtao Feng, Chengqi Zhao, Mingxuan Wang, and Lei Li. 2021. Counter-interference adapter for multilingual machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2812–2823. Association for Computational Linguistics.