# Enhanced Training Methods for Multiple Languages

**Hai Li**[1*], **Yang Li**[2]
[1]Tencent, China
[2]Shanghai Jiaotong University, China
mikehli@tencent.com liyangber@sjtu.edu.cn

## Abstract

Document-grounded dialogue generation based on multilingual is a challenging and realistic task. Unlike previous tasks, it need to tackle with multiple high-resource languages facilitating low-resource languages. This paper summarizes our research based on a three-stage pipeline that includes retrieval, re-rank and generation where each component is individually optimized. In different languages with limited data scenarios, we mainly improve the robustness of the pipeline through data augmentation and embedding perturbation with purpose of improving the performance designing three training methods: cross-language enhancement training, weighted training with neighborhood distribution augmentation, and ensemble adversarial training, all of that can be used as plug and play modules. Through experiments with different settings, it has been shown that our methods can effectively improve the generalization performance of pipeline with score ranking 6th among the public submissions on leaderboards.

## 1 Introduction

Question Answering (QA) system has received extensive attention in recent researches. The QA system aims to provide precise answers in response to the user's questions in natural language. An essential task in the QA system is conversational question answering and document-grounded dialogue modeling. Lack of data is one of the main challenges (Zhang et al., 2020).

Retrieval-augmented Generation (RAG) (Lewis et al., 2020) proposes a two-stage generation method with retriever extracting multiple documents related to the query and feeding them into answer generator. A survey of document-grounded dialogue systems (Ma et al., 2020) points that it is a mainstream method to indirectly search for key text before directly generating replies. There have been various works for knowledge-grounded dialogue systems (Zhan et al., 2021; Wen et al., 2022; Ma et al., 2020) to address this problem. A new framework UniGDD (Gao et al., 2022) use prompt learning for context guidance and design multitask learning. PPTOD (Su et al., 2022) proposes a dialogue pre-trained model that implements the current SOTA.

As a more realistic task, MultiDoc2Dial (Feng et al., 2021) faces challenges of identifying useful pieces of text from documents and generating response simultaneously which is goal-oriented dialogues generation based on multiple documents. Unlike former task, Doc2dial 2023 upgrades the difficulty level by introducing multiple languages.

To alleviate the problem of limited datasets in low-resource languages, on the one hand, it is necessary to effectively utilize datasets in the other high-resource languages. On the other hand, we design three training methods. These designs are all aimed at enhancing the generalization ability of the model. Our model is based on a three-stage framework: retriever, re-ranker and generator, the aims of first and second step are obtaining the most relevant paragraphs to the question, and then generating answer text. The first stage is responsible for the coverage of relevant texts that is the comprehensiveness of input texts; in the second stage, it is necessary to filter out the most relevant text that is the accuracy of the input text; the third stage generates answers based on the input text, which is clearly the most important part. Our contributions are as follows:

- A cross language enhancement training method is designed which can effectively improve generalization ability by replacing the high-frequency tokens of high-resource languages with that of low-resource languages in pre-trained model.

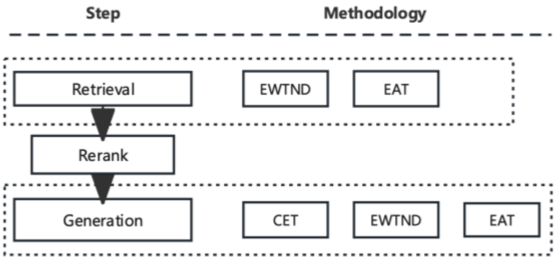

Figure 1: Training process of our pipeline.

- Enhanced weighted training approach based on neighborhood distribution is presented, the diversity of input texts can be increased through data augmentation, and the problem of semantic inaccuracy can be alleviated through weight.

- Ensemble adversarial training method is proposed including two classic adversarial training methods to improve the model's anti-interference ability and reduce text generation bias.

The above three enhancement training methods can be easily applied to other languages models as plug and play modules. Based on the published dataset, sufficient experiments are conducted confirming the method can effectively improve the generalization performance of the model.

## 2 Task Definition

Given dialogue history $\{q_1, \cdots, q_{t-1}\}$ and current user's query $q_t$, DialDoc task need to produce the response based on knowledge from a set of relevant documents $D_0 \subseteq D$ , where $D$ denotes all knowledge documents. Besides, the task provides similar format dataset of four languages including two high-resource languages (English and Chinese) and two low-resource languages ( French and Vietnamese), and the latter one is evaluated.

## 3 Methodology

To start with design，our pipeline is based on the three-stage baseline (Zhang et al., 2023). The three training augmentation methods that we propose can be applied to retrieval and generation. The specific framework process is as Figure 1.

### 3.1 Cross-Language Enhancement Training (CET)

From perspective of tokenizer, we designed a enhancement training method with token exchange between various languages. In different languages pairs, words with high frequency may have similar semantics, so that transfer learning can be used to facilitate low-resource languages training with embedding layers of high-resource languages. The basic idea is that as for pre-training model's tokenizer , replace high-resource languages' tokens with that of low-resource languages according to the rank of tokens' frequency which should follow four principles: (i) the total number of tokens of the high-resource languages need to be larger than that of the low-resource languages. (ii) select every similar language pairs, replace the high-resource tokens with low-resource tokens according to the rank order of frequency separately. In this paper, it should replace Chinese with Vietnamese and English with French. (iii) if the tokens of a language pair are insufficient, they can be mapped to the remaining unaligned tokens of another language. In this paper, there does not need to do it as the number of tokens in English higher than that of French, so do Chinese and Vietnamese. (iv) punctuation marks, [UNK] and other special marks remain unchanged.

After obtaining the mapping relationship of the tokenizer, we replace low-resource languages' datasets into high-resource languages' datasets as additional data, setting training weight $w$ for the new one.

### 3.2 Enhanced Weighted Training of Neighborhood Distribution (EWTND)

To alleviate the limited datasets about low-resource languages, we propose enhanced weighted training of neighborhood distribution method. By enhancing the texts from semantic neighborhood distribution, the diversity of input text increases, and the problem of semantic inaccuracy of neighborhood distribution is alleviated through weighted training. The steps of the method are as follows: (i) in top $n$ words $\{w_1, \cdots, w_n\}$ with the highest frequency, using the last layer of pre-trained mT5 (Xue et al., 2021; Raffel et al., 2020; Zhang et al., 2020) encoder to produce 512 dimensional vectors $\{v_1, \cdots, v_n\}$ for each token (except for punctuation mark). (ii) for every $v$, find the $k$ words with the largest similarity through vector retrieval by Faiss (Johnson et al., 2019) vector retrieval library, and record their similarities. So we get the text neighborhood matrix $t_{\mathrm{ij}}$ and similarity matrix $s_{\mathrm{ij}}$, where $1 \le \mathrm{i} \le \mathrm{n}, 1 \le \mathrm{j} \le \mathrm{k}$. (iii) during training, each sentence

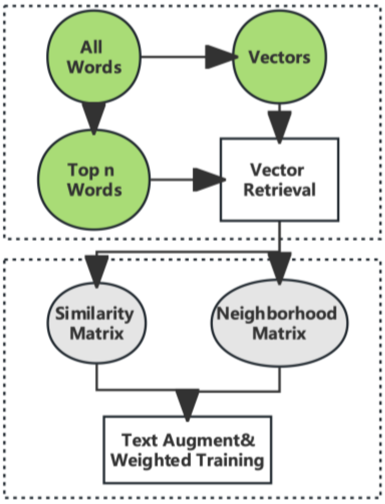

Figure 2: The key parts of EWTND.

has a $p\%$ probability to apply replacing that is words in $w$ are replaced by one of its neighborhood from $t$ with equal probability, and the calculation weight of sample loss is updated to the mean of similarity from $s$ in every sentence.

### 3.3 Ensemble Adversarial Training (EAT)

As a regularization method, adversarial training can improve the robustness of the model by introducing perturbations in embedding (Tramèr et al., 2020; Miyato et al., 2021). We propose an ensemble adversarial training method that blend two classic adversarial training methods to improve the model's anti-interference ability and reduce text generation bias. Adversarial training can be described by a general formula as follows: (Madry et al., 2019)

$$\overset{min}{\theta}\mathbb{E}_{(x,y)\sim D}[\underset{\Delta x\in\Omega}{\overset{max}{}}L(x+\Delta x,y;\theta)]$$

where $D$ is training dataset, x is input, y is target, $\theta$ is model parameter, $L(x+\Delta x,y;\theta)$ is loss of single sample, $\Omega$ is disturbance space, $\Delta x$ is perturbation. What's more, the main changes in different adversarial training methods are $\Delta x$ and $\Omega$. FGM method (Ian et al., 2015; Wong et al., 2020) raise the gradient with parameter $\epsilon$ and standardize it getting new $\Delta x$:

$$\Delta x=\epsilon\frac{\nabla L(x,y;\theta)}{\|\nabla L(x,y;\theta)\|}$$

While PGD method (Madry et al., 2019) split $\Delta x$ into multiple steps, set the constraint space to a sphere:

$$\Delta x_{t+1}=\prod\nolimits_{x+S}(\Delta x_t+\alpha\frac{\nabla L(x_t,y;\theta)}{\|\nabla L(x_t,y;\theta)\|})$$

where $S=r\in\mathbb{R}^d,\|r\|_2<\epsilon$, $\alpha$ is step size.

We add the FGM and PGD into training. For each batch in training process, we set the probabilities of the different training methods, there is $p_1\%$ probability of PGD, $p_2\%$ probability of FGM, and $p_3\%$ probability of not changing. The proportion can be determined by the ordinal of the model's convergence effect. In this paper, the rank of PGD, FGM, and non enhancement are 3:2:1 respectively, which means the probabilities are 50%, 33%, 17%.

After multiple experiments, we believe that there is a correlation between the final convergence loss of the method and the dataset, so the all possibilities should cannot be directly set and need to be determined based on the training results.

## 4 Experiments

We evaluate our methods using datasets provided by shared task which include four languages. As for generator, EWTND uses French and Vietnamese dialogue generation dataset, while CET also requires English and Chinese dialogue dataset. Besides, the score is calculated based on the sum of token-level F1, SacreBleu and Rouge-L metrics.

The experiments are mainly conducted on fine-tuning the retriever and generator based on the open-source baseline in three-stage framework. All the performances of methods can be evaluated by score of generator.

| $w$ | F1 | Sarcebleu | Rouge-L | Score |
|---|---|---|---|---|
| 0 | 58.55 | 42.03 | 55.83 | 156.42 |
| 0.2 | 60.74 | 43.30 | 57.92 | 161.96 |
| 0.25 | 61.85 | 43.72 | 59.21 | 164.78 |
| 0.3 | 61.97 | 44.38 | 59.31 | **165.66** |
| 0.35 | 61.71 | 43.63 | 59.08 | 164.42 |
| $0^{\text{half bz}}$ | 61.13 | 43.36 | 58.18 | 162.67 |

Table 1: The results of CET on Doc2dial validation dataset.

**Implementation** As for CET and EWTND, when they are used in generator, we change the "passages" and "re-rank" corresponding text in dataset; when they are used in retriever, we change the "positive" and "negative" corresponding text in dataset; while "query" text and "target" text won't be changed. As for EWTND, we use the cosine similarity. Faiss vector retrieval use product quantization to divide vector into 8 sub vectors, with 100 k-means clustering for each sub vector.

There is no threshold set to limit the number of synonyms $k$ which facilitates parallelization acceleration. We also set no limit to training epochs with early stopping epochs as 5, as EAT will need at least double training time.

**Results** Table 1 reports the performance of generator by using CET. When the weight is small, there can be a significant improvement. As weight increases to a certain extent, there will be score jitter. It proves that the CET can utilize the embedding of high-resource languages to improve low-resource languages. Meanwhile, this may also be due to more training batches. By reducing the batch size to half, it can be observed that score still improves, but under nearly equal training time, CET still achieves better results.

| $\boldsymbol{n}$ | $\boldsymbol{k}$ | $p$ | **Score** |
|---|---|---|---|
| 500 | 1 | 0.2 | 170.23 |
| 500 | 2 | 0.2 | **172.45** |
| 500 | 3 | 0.2 | 166.38 |
| 500 | 2 | 0.3 | 171.81 |
| 1000 | 2 | 0.2 | 170.75 |

Table 2: The results of EWTND on Doc2dial validation dataset.

Table 2 shows the effect of generator by using EWTND, it still use CET and EWTND but only strengthen the origin data. When $\boldsymbol{k}$ increases from 2 to 3, the reason why score drops might be uncertainty of the neighborhood's semantic meaning, the same reason can explain the time when $\boldsymbol{n}$ increases.

| $\boldsymbol{p_1}$ | $\boldsymbol{p_2}$ | $\boldsymbol{p_3}$ | **Score** |
|---|---|---|---|
| 100% | 0% | 0% | 175.05 |
| 0% | 100% | 0% | 172.45 |
| 50% | 33% | 17% | **175.39** |
| 60% | 25% | 15% | 174.48 |
| 45% | 35% | 20% | 173.60 |

Table 3: The results of EAT on Doc2dial validation dataset.

Table 3 shows the ensemble effect of adversarial training, it proves that such training method will provide stable improving although not much.

| **Method** | **EWTND** | **EAT** | **CET** | **Score** |
|---|---|---|---|---|
| Retriever | ✓ | | | 181.57 |
| Retriever | ✓ | ✓ | | 181.60 |
| mT5 | | | | 173.42 |
| mT5 | | | ✓ | 183.05 |
| mT5 | ✓ | | ✓ | 186.71 |
| mT5 | ✓ | ✓ | ✓ | 188.62 |

Table 4: The results of adding training methods into other models on Doc2dial validation dataset.

Table 4 shows effectiveness of three training methods as plug and play modules. By enhancing the retriever, the generator still improves but disadvantage is that it increases training time around 1.5 times. Besides, the improved performance is not as good as methods applied to the generator. With the best retriever and origin re-ranker, we replace the generator with origin mT5 (Xue et al., 2021) model which shows that it is better than generator in baseline. Finally, we achieve best performance by adding three enhanced training methods into mT5.

The above experiments have shown that our methods have significant advantages: (i) three training methods can effectively increase model's performance without affecting prediction speed. (ii) almost all language models with token as input can apply these methods. (iii) the methods can have more potentials in future work, especially in cross language scenarios, EWTND can be extended to more similar language pairs; EAT can use more complex sampling methods based on the neighborhood distribution of different languages.

## 5 Conclusion

In this paper, we propose three training methods to improve model's performance from perspective of embedding enhancement and data augmentation. CET Introduces cross language learning through high-frequency words; EWTND use weighted augmentation from the neighborhood distribution of high-frequency words; EAT strengthen the robustness of the model through embedding perturbation. Compared to the baseline mode, our methods achieve the stable rise in score.

## References

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

Zhang, Yeqin and Fu, Haomin and Fu, Cheng and Yu, Haiyang and Li, Yongbin and Nguyen, Cam-Tu. 2023. Coarse-to-Fine Knowledge Selection for Document Grounded Dialogs. *2023 IEEE International Conference on Acoustics, Speech and Signal Processing.*

Patrick S. H. Lewis, Ethan Perez, Aleksandra Pik- tus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *In Advances in Neu- ral Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Haolan Zhan, Lei Shen, Hongshen Chen, and Hainan Zhang. 2021. CoLV: A collaborative latent variable model for knowledge-grounded dialogue generation. *In Proceedings of the 2021 Conference on Empiri- cal Methods in Natural Language Processing, pages 2250–2261, Online and Punta Cana, Dominican Re- public. Association for Computational Linguistics.*

Xiaofei Wen, Wei Wei and Xian-Ling Mao. 2022. Sequential Topic Selection Model with Latent Variable for Topic-Grounded Dialogue. *In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP'2022), Abu Dhabi.*

Ian J. Goodfellow, Jonathon Shlens and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv: 1412.6572.*

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras and Adrian Vladu. 2019. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv preprint arXiv: 1706.06083.*

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. *arXiv preprint arXiv: 2010.11934.*

Johnson, Jeff Douze, Matthijs Jegou, Herve. 2019. Billion-scale similarity search with {GPUs}. *IEEE Transactions on Big Data.*

Chang Gao, Wenxuan Zhang and Wai Lam. 2022. UniGDD: A Unified Generative Framework for Goal-Oriented Document-Grounded Dialogue. *arXiv preprint arXiv: 2204.07770.*

Longxuan Ma, Wei-Nan Zhang,Mingda Li and Ting Liu. 2020. A Survey of Document Grounded Dialogue Systems (DGDS). *arXiv preprint arXiv: 2004.13818.*

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint arXiv: 1910.10683.*

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh and Patrick McDaniel. 2020. Ensemble Adversarial Training: Attacks and Defenses. *arXiv preprint arXiv: 1705.07204.*

Takeru Miyato, Andrew M. Dai, Ian Goodfellow. 2021. Adversarial Training Methods for Semi-Supervised Text Classification. *arXiv preprint arXiv: 1605.07725.*

Longxuan Ma, Weinan Zhang, Runxin Sun, Ting Liu. 2020. A Compare Aggregate Transformer for Understanding Document-grounded Dialogue. *arXiv preprint arXiv: 2010.00190.*

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, Minlie Huang and Xiaoyan Zhu. 2020. Recent Advances and Challenges in Task-oriented Dialog System. *arXiv preprint arXiv: 2003.07490.*

Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *arXiv preprint arXiv: 1912.08777.*

Eric Wong, Leslie Rice, J. Zico Kolter. 2020. Fast is better than free: Revisiting adversarial training. *ICLR 2020.*

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, YiAn Lai, Yi Zhang. 2022. Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics.*