Achieving State-of-the-Art Multilingual Translation Model with Minimal Data and Parameters

Hui Zeng
LanguageX AI Lab
felix_zeng_ai@aliyun.com

Abstract

This is LanguageX (ZengHuiMT)'s submission to WMT 2023 General Machine Translation task for 13 language directions. We initially employ an encoder-decoder model to train on all 13 competition translation directions as our baseline system. Subsequently, we adopt a decoder-only architecture and fine-tune a multilingual language model by partially sampling data from diverse multilingual datasets such as CC100 and WuDaoCorpora. This is further refined using carefully curated high-quality parallel corpora across multiple translation directions to enable the model to perform translation tasks. As per automated evaluation metrics, our model ranks first in the translation directions from English to Russian, English to German, and English to Ukrainian. It secures the second position in the directions from English to Czech, English to Hebrew, Hebrew to English, and Ukrainian to English, and ranks third in German to English, Japanese to English, and Russian to English among all participating teams. Our best-performing model, covering 13 translation directions, stands on par with GPT-4. Among all 13 translation directions, our multilingual model surpasses GPT-4 in bleu scores for 7 translation directions.

1 Introduction

Since 2023, large language models like ChatGPT (Brockman et al., 2023) have had a profound impact on the field of machine translation, characterized by an ever-increasing scale in terms of parameters and data requirements. Many research institutions and language service providers struggle to keep pace with this computational arms race. For smaller teams, the only viable strategy is to maximize model performance under constrained hardware resources. We participated in the WMT 2023 General Machine Translation task, covering 13 translation directions. Given our limited computational power and time constraints, it was infeasible to craft dedicated models for each translation direction, making a large-scale multilingual translation model our optimal choice. We utilized Fairseq (Ott et al., 2019) to train our baseline multilingual translation model and further employed the Hugging Face Transformers Toolkit (Wolf et al., 2020) to train a multilingual language model. Subsequent fine-tuning with task-specific instructions enabled it to perform multilingual translation tasks effectively.

2 Data Filtering and Selection

We participated in the WMT 2023 General MT task, competing in 13 language pairs including Chinese to/from English, German to/from English (at the document level), Hebrew to/from English (in a low-resource setting), Japanese to/from English, Russian to/from English, Ukrainian to/from English, and English to Czech.

Given the challenge of concurrently training for 13 translation directions, it was imperative for us to judiciously regulate the size of parallel corpora specific to each direction, as well as the parameter count of the multilingual translation model. This was crucial to ensure training completion within a constrained timeline. For the Chinese-English bi-directional translation, our primary sources for parallel corpora were the CCMT Corpus (which can be found at:

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 181–186

http://mteval.cipsc.org.cn:81/agreement/description), genuine internal translation project data, in addition to content extracted and curated from websites and e-books. This rigorous process resulted in a refined collection of approximately 5 million parallel sentence pairs. For other translation orientations, we used the English segments from the derived Chinese-English parallel corpus as foundational data. This seed data enabled the retrieval of analogous language pairs from our comprehensive in-house multilingual parallel corpus, with each translation direction maintaining a parallel sentence count in the ballpark of 5 million.

Given the need to train a decoder-only multilingual model, we primarily utilized public datasets such as Book Corpus (Zhu et al., 2015), CC100 (Conneau et al., 2020), and WuDaoCorpora (Yuan et al., 2021). It was also imperative for us to regulate the data volume for each language and the parameter count of the multilingual model. Table 1 delineates the sources and the respective quantities for monolingual data across different languages. Owing to the extended length of text segments in the WuDaoCorpora (Yuan et al., 2021), the number of extracted text passages is fewer compared to other languages. However, the character count remains substantial.

Language	Data Source	Size in GB	Paragraph Count
Chinese	WuDaoCorpora	6.4	3,980,000
Czech	CC100	4.5	29,630,985
German	CC100	5.1	24,958,540
English	BookCorpus	4.3	20,000,000
English	CC100	3.3	20,000,000
Hebrew	CC100	5.4	30,877,445
Japanese	CC100	5.8	30,985,700
Russian	CC100	5.7	13,928,244
Ukrainian	CC100	4.6	16,818,862

Table 1: Sources and Quantities of Monolingual Data for Each Language.

2.1 Monolingual Data Filtering

The following rules are used to filter parallel corpus.

• Remove duplicated sentence pairs.
• Remove the sentence pairs containing special characters.
• Remove the sentence pairs containing html addresses or tags.

2.2 Parallel Data Filtering Using Rules

The following rules are used to filter parallel corpus.

a. Remove duplicated sentence pairs.
b. Remove the lines having identical source and target sentences.
c. Remove the sentence pairs containing special characters.
d. Remove the sentence pairs containing html addresses or tags.
e. Remove the sentence pairs with empty source or target side.

2.3 Parallel Data Filtering Using Multilingual Language Model

We used a multilingual model - sentence-transformers/paraphrase-multilingual-mpnet-base-v2 (Reimers et al., 2019) that generates embeddings for sentences or paragraphs in various languages. Using these embeddings, we calculated semantic similarity scores for parallel sentence pairs. Based on these scores, we filtered out low-quality parallel sentence pairs.

3 System Description

This section illustrates how the model is trained step by step.

3.1 Data pre-processing

Data pre-processing of multilingual translation model. We utilized the NLLB (NLLB Team, 2022) tokenizer from Hugging Face as the foundation and incorporated additional Chinese tokens to create an enhanced tokenizer specifically for Chinese language processing. This resulted in a final vocabulary size of 266,786 tokens.

To ensure synchronized training across all translation directions and to prevent the model from mastering one translation direction at the expense of another, we evenly blended the multilingual parallel corpora. This involved sequentially placing a fixed number of parallel sentence pairs from different translation directions into the training set, typically set to 100 pairs per direction.

To facilitate the simultaneous training of multiple translation directions within a single large model, we shared the embeddings and vocabulary for both source and target languages. Furthermore, we prefixed the source part of the parallel sentence pairs with specific prompt tokens.

The structure of the parallel sentence pairs is as follows: {engine name} engine. Translation from {source language} to {target language}: {source line} 🐯 {target line} <eos>. 🐯 is the delimiter used for parallel corpora.

To better accommodate the German to/from English (at the document level) translation task, we combined conventional sentence-level German to/from English parallel corpora into paragraph-level corpora based on a specified number of sentences. We then mixed this with the regular sentence-level parallel corpora, ensuring the resultant model is trained to handle a broader range of sequence lengths.

Data pre-processing of multilingual language model. We employed the same tokenizer as used in the multilingual translation model.

Due to the vast size of the CC100 dataset (Conneau et al., 2020), we performed sampling on the data for all languages, with 1,000 lines as the sampling unit. Multiple units were extracted from various parts of the entire dataset to cover it as comprehensively as possible, while keeping the individual language data size at around 5GB.

To mitigate catastrophic forgetting, we uniformly mixed the monolingual data of each language. This ensured that the training process included synchronized training on data from all languages, rather than training on one language first and then training on another.

The structure of the supervised finetuning prompt for translation task is as follows: {engine name} engine. Text in {source language}: {source line} Translation of the previous text to {target language}: {target line} 🥰.

To prevent endless generation and excessive translation, the 🥰 emoji is placed at the end of the translation to signify its completion, signaling the model to cease generation.

3.2 Baseline Translation Model Training

The parallel data prepared in step 3.1 is used to train a multilingual translation model using transformer (Vaswani et al., 2017) architecture as the baseline. Training was conducted using Fairseq (Ott et al., 2019) over the entire dataset for four epochs. The crucial training parameters are as follows:

```
--encoder-layers 12 \
--encoder-attention-heads 16 \
--encoder-embed-dim 1024 \
--encoder-ffn-embed-dim 4096 \
--decoder-layers 6 \
--decoder-attention-heads 16 \
--decoder-embed-dim 1024 \
--decoder-ffn-embed-dim 4096 \
--share-decoder-input-output-embed \
--share-all-embeddings \
--max-source-positions 1024 \
--max-target-positions 1024 \
--lr 5e-4 \
--lr-scheduler inverse_sqrt \
--warmup-updates 4000 \
```

Parameter	Value
Trainable parameters	1,091,315,712
Vocabulary size	266,786
Max length	1024
Embedding Dimension	1536
Decoder layers	24
Attention heads	16
Learning rate	5e-5
Lr scheduler type	linear
Warmup steps	4,000

Table 1: Parameters for Training Multilingual Language Model.

Translation Direction	Baseline Translation Model	Multilingual Language Model
en-cs	41.20	**43.67**
en-de	40.20	**41.00**
en-he	35.00	**36.52**
en-ru	31.20	**32.07**
en-uk	26.60	**28.29**
en-zh	47.30	**53.01**
en-ja	17.00	**17.60**
de-en	26.70	**42.08**
he-en	56.00	**57.51**
ja-en	21.20	**23.54**
ru-en	30.90	**32.15**
uk-en	42.50	**44.28**
zh-en	25.2	**28.27**

Table 3: BLEU scores on Newstest 2023 for all directions and different training methodologies.

Translation Direction	GPT 4-5shot	Multilingual Language Model
en-cs	38.26	**43.67**
en-de	**44.08**	41.00
en-he	27.08	**36.52**
en-ru	31.09	**32.07**
en-uk	25.78	**28.29**
en-zh	49.65	**53.01**
en-ja	**20.55**	17.60
de-en	**49.54**	42.08
he-en	52.04	**57.51**
ja-en	**25.27**	23.54
ru-en	**35.31**	32.15
uk-en	**44.84**	44.28
zh-en	27.87	**28.27**

Table 4: BLEU score comparison between the multilingual model and GPT-4 across all language directions on Newstest 2023.

3.3 Multilingual Language Model Training

We utilized DeepSpeed (Rasley et al., 2020) and Hugging Face transformers (Vaswani et al., 2017) as our training tools and trained the models on the uniformly mixed monolingual data and SFT data prepared in step 3.1 after applying bf16 precision. The specific training parameters are presented in Table 2. The entire training process was completed using four RTX A6000 GPUs. After completing one full pass of the entire dataset, we

terminated the training of the multilingual language model.

3.4 Results

The BLEU (Papineni et al., 2002) scores on Newstest 2023 for all translation directions and different training methodologies are presented in Table 3.

Based on the automated assessment metrics, our system takes the lead in translation directions from English to Russian, English to German, and English to Ukrainian. It claims the runner-up spot for English to Czech, English to Hebrew, Hebrew to English, and Ukrainian to English directions, and occupies the third place for the German to English, Japanese to English, and Russian to English directions among the contenders.

4 Conclusion

This paper describes LanguageX (ZengHuiMT)'s translation system for the WMT2023 General MT task. Initially, we utilize a comprehensive encoder-decoder structure to establish our baseline system by training across all 13 contest translation directions. In the subsequent stages, we embrace a solely decoder-focused design and harness a multilingual language model, drawing samples from multilingual datasets like CC100 (Conneau et al., 2020) and WuDaoCorpora (Yuan et al., 2021). This model is then meticulously fine-tuned using select high-grade parallel corpora from various translation domains, empowering it to execute translation task.

Our best-performing model, covering 13 translation directions, boasts around 1 billion parameters. This is less than one percent of the parameter count of mammoth models like GPT-4 (OpenAI, 2023), which possess hundreds of billions of parameters. In translation evaluations across all languages, our system stands on par with GPT-4 (OpenAI, 2023). Among all 13 translation directions, our multilingual model surpasses GPT-4 (OpenAI, 2023) in bleu (Papineni et al., 2002) scores for 7 translation directions.

Acknowledgments

Thanks to my wife who spends most of her time to take care of our two kids, so that I am able to participate in the contest and complete this paper.

References

Greg Brockman, Atty Eleti, Elie Georges, Joanne Jang, Logan Kilpatrick, Rachel Lim, Luke Miller, and Michelle Pokrass. 2023. Introducing ChatGPT and Whisper APIs. *https://openai.com/blog/introducing-chatgpt-and-whisper-apis*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems,* pages 6000–6010.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wen-zek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov: Unsupervised Cross-lingual Representation Learning at Scale. *In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451. (2020)

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf and Lysandre Debut and Victor Sanh and Julien Chaumond and Clement Delangue and Anthony Moi and Pierric Cistac and Tim Rault and Rémi Louf and Morgan Funtowicz and Joe Davison and Sam Shleifer and Patrick von Platen and Clara Ma and Ya-cine Jernite and Julien Plu and Canwen Xu and Teven Le Scao and Sylvain Gugger and Ma-riama Drame and Quentin Lhoest and Alexander M. Rush: Transformers: State-of-the-Art Natural Language Processing. *In: Proceedings of the 2020 Conference on Empirical Meth-ods in Natural Language Processing: System Demonstrations,* pp. 38–45. (2020)

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, Jie Tang: WuDaoCorpora: A super large-scale Chinese corpora for pre-training language models. AI Open, 65-68 (2021)

Yukun Zhu and Ryan Kiros and Richard S. Zemel and Ruslan Salakhutdinov and Raquel Urtasun and Antonio Torralba and Sanja Fidle: Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. http://arxiv.org/abs/1506.06724. (2015)

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

NLLB Team and Marta R. Costa-jussà and James Cross and Onur Çelebi and Maha Elbayad and Kenneth Heafield and Kevin Heffernan and Elahe Kalbassi and Janice Lam and Daniel Licht and Jean Maillard and Anna Sun and Skyler Wang and Guillaume Wenzek and Al Youngblood and Bapi Akula and Loic Barrault and Gabriel Mejia Gonzalez and Prangthip Hansanti and John Hoffman and Semarley Jarrett and Kaushik Ram Sadagopan and Dirk Rowe and Shannon Spruit and Chau Tran and Pierre Andrews and Necip Fazil Ayan and Shruti Bhosale and Sergey Edunov and Angela Fan and Cynthia Gao and Vedanuj Gos-wami and Francisco Guzmán and Philipp Koehn and Alexandre Mourachko and Christophe Ropers and Safiyyah Saleem and Holger Schwenk and Jeff Wang: No Language Left Be-hind: Scaling Human-Centered Machine Translation. https://arxiv.org/abs/2207.04672. (2022)

Rasley, Jeff and Rajbhandari, Samyam and Ruwase, Olatunji and He, Yuxiong: DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. Association for Computing Machinery, 3505-3506 (2020)

OpenAI. GPT-4 technical report. arXiv preprint arXiv:2303.08774, 2023

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *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 3982–3992, Hong Kong, China. Association for Computational Linguistics.