Yishu: Yishu at WMT2023 Translation Task

Qiulin Chen
Dtranx AI, Chengdu, China
415626524@qq.com

Min Luo
Dtranx AI, Chengdu, China
1148859199@qq.com

Yixin Tan
Dtranx AI, Chengdu, China
420129823@qq.com

Abstract

This paper introduces the Dtranx AI translation system, developed for the WMT 2023 Universal Translation Shared Task. Our team participated in two language directions: English to Chinese and Chinese to English. Our primary focus was on enhancing the effectiveness of the Chinese-to-English model through the implementation of bilingual models. Our approach involved various techniques such as data corpus filtering, model size scaling, sparse expert models (especially the Transformer model with adapters), large-scale back-translation, and language model reordering. According to automatic evaluation, our system secured the first place in the English-to-Chinese category and the second place in the Chinese-to-English category.

1 Introduction

This year, the Dtranx AI team participated in the WMT2023 Universal Translation Sharing Task and focused on enhancing the performance in the zh-en and en-zh language directions. For data preprocessing, we employed various methods, including knowledge-based rules, language detection, and language modeling, to clean parallel, monolingual, and back-translated data. Our data primarily comprised large-scale data mining and back-translation. Additionally, we applied punctuation regularization, byte pair encoding (BPE) (Sennrich et al., 2015), and subword regularization methods (Provilkov et al., 2019) for processing the data, which yielded excellent results across all languages.

In the modeling section, we have made enhancements to Fairseq (Ott et al., 2019) by increasing the model's depth and width. Specifically, we augmented the Transformer model et al.(Vaswani et al., 2017) by significantly increasing the number of layers and widening the model architecture.This modification allows the model to capture more complex patterns and dependencies in the data. Additionally, we have embraced the concept introduced by Bapna et al. (Bapna et al., 2019) to expand the Transformer model by incorporating language-specific adapters, thus bridging the gap between diverse languages. Lastly, we integrated the dense Transformer model with the sparse Adapter model and leveraged the language model to reranking the final results, leading to further improvements in system performance.

For both English and Chinese translation tasks, we have developed separate systems. We have enhanced the model capacity and applied Adapter fine-tuning techniques, while also incorporating additional proprietary data for system training purposes.During the model inference phase, we have implemented reordering techniques to select more optimal translations. Based on automatic evaluation, our system achieved first place in English to Chinese translation and second place in Chinese to English translation.

2 Method

2.1 Data

In this section, we will present our primary dataset, which consists of bitext data and monolingual data sources, along with the preprocessing methods employed to prepare this initial data. Additionally, we will provide details about the setup utilized for training our baseline model.

2.1.1 Bitext Data

For the Chinese-English-English-Chinese language pairs, we utilize all bitext data in the shared task and include additional data sources for English-Chinese conversion. During data processing, we implemented the following knowledge-based rules for enhancement:

- Remove empty sentences.
- Eliminate escaped HTML characters.

- Standardize different punctuation variations.
- Normalize spacing.
- Remove sentences with repetition marks, including single characters repeated more than four times, two characters repeated more than three times, and three characters repeated more than twice.
- Delete sentence pairs with inconsistent punctuation at the end of the original and translated texts.
- Remove sentence pairs with a source/target token ratio exceeding 1:3 (or 3:1).
- Delete segments that exceed 150 tokens in length.
- Remove sentence pairs with fewer than 5 tokens in the source text or translation.
- Convert traditional Chinese characters to simplified Chinese characters.
- Delete corpora with an unaligned number of parentheses.
- Delete corpora with an unaligned number of Arabic numerals.
- Remove corpora with non-native character ratios greater than 0.4.

We employed Moses (Koehn et al., 2007) for normalizing spacing and punctuation. We utilized all accessible data sources to train our model.

Considering the aforementioned concerns regarding corpus quality, we implemented additional filtering steps to ensure data availability. Initially, we attempted to filter out low-quality sentence pairs using the word alignment method of fast-align (Dyer et al., 2013). We retained the top 80% of sentence pairs based on the alignment score(a score generated by the word alignment model that measures the quality of word alignment between source and target sentences), encompassing all directions. Subsequently, we trained the Transformer model for all languages using Fairseq, following a similar approach as outlined in the study conducted by Bei et al. (Bei et al., 2019). The scores were calculated as follows:

$$Score_{sentence} = PPL \tag{1}$$

Language Pair	Data
zh-en	50M

Table 1: Ultimate bitext training data

Language Pair	Data
zh	72M
en	10M

Table 2: Ultimate monolingual data

$$Score_{com} = \lambda * Score_{src} + (1 - \lambda) * Score_{tgt} \tag{2}$$

Here, we employ PPL as an abbreviation for perplexity, which represents the perplexity of the sentence language model. The value of λ, on the other hand, is determined empirically based on language pairs and ranges from 0.2 to 0.8. For instance, if our source language is English and the target language is Chinese, we would set λ to 0.7.

Finally, the training data, as presented in Table 1, was carefully curated to serve as the foundational resource for our model training. This bilingual training dataset consists of 50 million sentence pairs in the Chinese-English (zh-en) language pair,and it can be utilized bidirectionally.

2.1.2 Monolingual Data

To ensure the quality of our data and to create synthetic parallel texts, we harnessed the capabilities of a well-trained bilingual model. We compiled high-quality monolingual corpora in various languages from reputable sources, including news commentaries, europarl, and news crawls. The monolingual data, after undergoing a rigorous filtering process, is presented in Table 2.

2.1.3 Tokenizer

We opted for SentencePiece (Kudo and Richardson, 2018) as the training tool for our subword tagger. To enhance subwording efficiency, we adopted the approach of Tran et al. (Tran et al., 2021) by employing sampled text with a temperature of 5. For the bilingual model, we utilized a vocabulary of 32,000 words.

In addition, we integrated the subword regularization method (Provilkov et al., 2019) (Raffel et al., 2020) into the tagged text. This technique was exclusively applied to the source side, as it has the potential to enhance the model's robustness by allowing different subword tokenizations.

2.2 Model Architectures

We have developed a dedicated model for bidirectional translation between Chinese and English, capitalizing on our proficiency as native Chinese speakers. To enhance the quality of our training data, we integrated a private corpus comprising approximately 20 million high-quality sentence pairs spanning various domains, including general text, technology, medicine, law, finance, and more.

Regarding "basic fine-tuning," our approach involves parameter adjustments, including fine-tuning the learning rate, the number of training epochs, and batch sizes. These adjustments are made to optimize the model's adaptation to the specific translation task at hand.

In the realm of back-translation, we harness English and Chinese monolingual corpora. This approach leverages monolingual text data in both languages, enriching the model's translations by back-translating them into English. This technique seamlessly augments the diversity of our training dataset.

As for the specifications of our Transformer model, it features 12 encoder layers and 6 decoder layers, each equipped with 8 attention heads. The embedding size is set to 512, and the width of the feed-forward neural network (FFN) is 4096. Additionally, we have incorporated techniques like Layer Normalization and residual connections to stabilize the model training process.

2.2.1 Language Specific Adapter

In essence, a language-specific adapter layer is a dense layer that incorporates residual connections and nonlinear projections. The hyperparameter "b" represents the dimension of the internal dense layer. These adapter layers consist of a multitude of globally shared parameters, along with several task-specific layers. This unique design allows us to train and optimize individual models for multiple languages.

Bapna et al. (Bapna et al., 2019) demonstrated the improved translation performance achieved by machine translation models employing adapters. Therefore, following the training of our bilingual models, we integrated adapter layers into them and subsequently conducted additional training and fine-tuning on these adapter layers. To be specific, for the Chinese-English and English-Chinese models, we introduced a language-specific adapter with a dense layer dimension of 4096.

Regarding the incorporation of adapters in the bilingual models, we seamlessly integrate the adapter layers into the existing architecture, where they operate alongside the standard layers. The globally shared parameters refer to the model parameters that are common across various languages and tasks, which are shared among different adapter layers in the model.

For the fine-tuning of the adapters, we utilized additional bilingual data specific to the translation tasks. These adapters were fine-tuned with the same data used for training the main translation model, allowing them to adapt to the particular translation requirements of our task.

2.2.2 Finetune

To enhance the model's performance, we implemented in-domain fine-tuning, a proven effective technique in previous news translation tasks. We generated various types of fine-tuned data using the following approach. According to the studies conducted by Li et al. (Li et al., 2020) and Wang et al.(Wang et al., 2021), low-frequency and high-frequency words often pertain to domain-specific nouns and other related terms that directly reflect the topic at hand. However, this year's shared task has transitioned from the news domain to a more generalized translation task. Recognizing that previous fine-tuning using news domain data could potentially have a detrimental effect on the model, we adopted the strategy outlined by Li et al. (Li et al., 2020) and Wang et al. (Wang et al., 2021), which involves selecting topic-related data based on a test set. Subsequently, we identified specific data for further fine-tuning and conducted experiments on the 2022 news development set, subsequently applying the refined model directly to the 2022 test set.We fine-tuned the full model, not just the adapters, to ensure it was well-suited for the task at hand.

2.2.3 Model Ensemble

Model integration has been widely adopted as a technique in previous WMT sharing tasks. To mitigate bias towards more recent training data, it is common practice to average multiple checkpoint parameters of the model. Specifically, during training, we consistently take the average of the last five checkpoints. In the fine-tuning phase, we fine-tune the hyperparameters (e.g., num epoch and num average checkpoints) based on the performance on the development set and directly apply them to the

Team	Bleu	Chrf	Comet
HW-TSC	33.6	57.5	82.8
Yishu	33.4	57.4	82.7
GPT4-5shot	26.8	53.1	81.6
ZengHuiMT	27.0	54.6	79.6

Table 3: Submission results for zh-en in WMT23

Team	Bleu	Chrf	Comet
HW-TSC	58.6	53.8	87.3
Yishu	57.6	53.0	88.1
GPT4-5shot	49.6	46.5	87.1
ZengHuiMT	52.9	47.0	84.3

Table 4: Submission results for en-zh in WMT23

test set of WMT23.

3 Results

3.1 Experimental setup

Each model was trained on eight NVIDIA A100 GPUs, each equipped with 40 GB of memory. Additionally, we employed high-volume processing and higher learning rates, as mentioned in Ott et al. (Ott et al., 2018). The maximum learning rate was set to 0.0005, and we used 10,000 warm-up steps. All dropout probabilities were set to 0.1. To expedite training, we utilized half-precision floating-point numbers (FP16). In the context of multilingual training, we incorporated source language labels and target language labels to leverage the distinctions between languages. Following the approach proposed by Tran et al. (Tran et al., 2021), we segmented the data into multiple parts and downsized the data in both the high-resource direction and in synthetic backtranslation for each training cycle.

3.2 Results

We trained bilingual models for English to Chinese (en-zh) and Chinese to English (zh-en). In our model, we enhanced the model capacity by introducing specific adapter layers for each translation direction, addressing the unique linguistic challenges of each language pair. These adapter layers do not induce sparsity; instead, they add more trainable parameters to the model. Each model has dedicated adapters for en-zh and zh-en, as they are not shared between the bilingual models. To refine our training set, we extracted additional relevant corpus from the raw text in the test set. This extracted data was structured using the language model and augmented through reverse translation. The results demonstrate the effectiveness of our systematic approach, and we achieved the highest scores on the COMET evaluation metric. These outcomes are detailed in Table 3 and Table 4

4 Conclusion

In this paper, we present Dtranx AI's submission to the WMT2023 Universal Translation Shared Task. For the Chinese-English and English-Chinese language pairs, we adopt a bilingual model as the fundamental structure and enhance it through various strategies. These include increasing the model capacity, fine-tuning with Adapters, incorporating private relevant corpus, and optimizing the translation output by reordering. Our experimental results demonstrate the effectiveness of these optimization techniques.

References

Ankur Bapna, Naveen Arivazhagan, and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. *arXiv preprint arXiv:1909.08478.*

Chao Bei, Hao Zong, Conghu Yuan, Qingming Liu, and Baoyong Fan. 2019. Gtcom neural machine translation systems for wmt19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 116–121.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226.*

Zuchao Li, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2020. Sjtu-nict's supervised and unsupervised neural machine translation systems for the wmt20 news translation task. *arXiv preprint arXiv:2010.05122*.

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. *arXiv preprint arXiv:1904.01038*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2019. Bpe-dropout: Simple and effective subword regularization. *arXiv preprint arXiv:1910.13267*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

C Tran, S Bhosale, J Cross, P Koehn, S Edunov, and A Fan. 2021. Facebook ai wmt21 news translation task submission. arxiv.

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.

Longyue Wang, Mu Li, Fangxu Liu, Shuming Shi, Zhaopeng Tu, Xing Wang, Shuangzhi Wu, Jiali Zeng, and Wen Zhang. 2021. Tencent translation system for the wmt21 news translation task. In *Proceedings of the sixth conference on machine translation*, pages 216–224.