# Lan-Bridge MT's Participation in the WMT 2022 General Translation Shared Task

**Bing Han**
Lan-Bridge / Sichuan (China)
hanbing@lan-bridge.com

**Yangjian Wu**
Lan-Bridge / Sichuan (China)
wuyangjian@lan-bridge.com

**Gang Hu**
Lan-Bridge / Sichuan (China)
hugang@lan-bridge.com

**Qiulin Chen**
Lan-Bridge / Sichuan (China)
chenqiulin@lan-bridge.com

## Abstract

This paper describes Lan-Bridge Translation systems for the WMT 2022 General Translation shared task. We participate in 18 language directions: English to and from Czech, German, Ukrainian, Japanese, Russian, Chinese, English to Croatian, French to German, Yakut to and from Russian, and Ukrainian to and from Czech. We mainly focus on multilingual models to develop systems covering all these directions. In general, we apply data corpus filtering, scaling model size, sparse expert model (in particular, Transformer with adapters), large-scale backtranslation, and language model reranking techniques. Our system ranks first in 6 directions based on the automatic evaluation.

## 1 Introduction

Our Lan-Bridge MT team participate in the WMT 2022 General Translation shared task. As machine translation expands into more and more languages, multilingual machine translation has attracted more and more attention in both academia and industry. It can not only avoid training a separate model for each language pair but also transfer knowledge from high-resource languages to low-resource ones. Many systems such as Tran et al. (2021) submitted in previous years have proved this point and achieved a state of the art results in some language directions.

For data preprocessing, knowledge-based rules, language detection, and language model are involved to clean parallel data, monolingual data, and synthetic data (mainly from large-scale data mining and backtranslation). Punctuation normalization and BPE (byte pair encoding) (Sennrich et al., 2015) with subword regularization method (Provilkov et al., 2019) are applied for all languages. As for models, we fork Fairseq (Ott et al., 2019) as our development tool and use Transformer (Vaswani et al., 2017) as the main architecture. In addition, we follow Bapna et al. (2019) to extend Transformer by adding language-specific adapters to bridge the gap between different language pairs. Finally, we ensemble dense Transformer models and sparse adapter models, and the final result are re-ranked by language models. For English to and from Chinese, we develop a separate system. In addition to optimization techniques similar to multilingual models, We also use additional private data. And for Yakut to and from Russian, due to a smaller corpus, we simply apply fine-tuning and backtranslation on our multilingual models.

We win the first place in Russian ↔ Yakut, Russian → English, English → Croatian, Czech → English and Ukrainian → English based on BLEU (Papineni et al., 2002) score. [1]

## 2 System Overview

### 2.1 Data

Here we describe our base datasets, including bitext and monolingual data sources, and the preprocessing methods we apply to prepare these initial data sets to train our baseline models.

### 2.1.1 Bitext Data

We use all available bitext data from the shared task for all language pairs, besides, for English to and from Chinese, we add extra data from ai-challenger. For high-resource language pairs such as English to Chinese or English to German, which provides millions of high-quality bitext, we only choose those high-quality resources, and simply apply language identification using fasttext (Joulin et al., 2017) with an ID threshold of 0.8 and knowledge-based rules shows below as data process:

- Remove empty sentences
- De-escaping HTML characters

[1] This result is based on the submission website https://ocelot-wmt22.mteval.org/, not the official final result.

*Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 268–274
Abu Dhabi, December 7–8, 2022. ©2022 Association for Computational Linguistics

- Normalization of different languages of punctuation

- Normalization of spacing

- Remove sentences with repeated tokens, including single character that repeat more than four times, two characters that repeat more than three times, and more than three characters that repeat more than twice.

- Delete the corpus with inconsistent punctuation marks at the end of the original text and the translation

- Deletion of segments where source/target token ratio exceeds 1:3 (or 3:1)

- Deletion of segments longer than 150 tokens

- Deletion of segments shorter than 5 tokens

- Transfer traditional Chinese characters to simplified Chinese characters

- Delete corpus with misaligned number of parentheses

- Delete corpus with misaligned number of Arabic numerals

- Delete corpus with a proportion of non-native language characters exceeding 0.4

The normalization of spacing and punctuation is applied using Moses (Polykovskiy et al., 2020).

For medium- and low-resource language pairs, we incorporate additional sources of data from OPUS (Tiedemann, 2012), ccAligned (El-Kishky et al., 2020), and ccMatrix (Schwenk et al., 2019). All available data sources are utilized to train our models.

Due to the low-quality issue of corpora mentioned above, we add a few filter steps to make them usable. First, we try the word alignment method using fast_align (Dyer et al., 2013) to filter low-quality sentence pairs and keep top 80% for all directions ranked by alignment score. Then we use Fairseq to train the transformer multilingual language model for all languages, similar to Bei et al. (2019), the score is calculated as follows:

$$Score_{sentence} = PPL$$

$$Score_{combine} = \lambda * Score_{src} + (1 - \lambda) * Score_{tgt}$$

| Language Pair | Data |
|---|---|
| cs-en | 100M |
| de-en | 250M |
| fr-de | 20M |
| hr-en | 70M |
| sah-ru | 0.1M |
| uk-cs | 6M |
| uk-en | 20M |
| zh-en | 50M |
| ru-en | 80M |
| ja-en | 20M |

Table 1: Ultimate bitext training data

| Language | Data |
|---|---|
| cs | 64M |
| en | 72M |
| de | 63M |
| fr | 79M |
| ja | 81M |
| sah | 0.2M |
| ru | 70M |
| uk | 5M |
| zh | 10M |
| hr | 14M |

Table 2: Ultimate monolingual data

Here $PPL$ is the perplexity of a language model for sentence, $\lambda$ is an empirical value between 0.2–0.8 depending on the language pair, such as the source language is English, and the target language is Croatian, then our empirical value of $\lambda$ is 0.7. Finally, we consult Parallel Corpus Filtering Zhang et al. (2020) for finetuning a multilingual high-resource corpus classifier using mBERT (Gonen et al., 2020) to get our ultimate training data described in Table 1.

### 2.1.2 Monolingual Data

As we need a multilingual language model to filter low-quality corpus and create synthetic parallel text, we collect all high-quality monolingual corpus from News-Commentary, europarl, and news-crawl for all languages if available. For medium and low-quality resources, we use all available monolingual data from the shared task, and filter according to the above steps (where applicable). The ultimate monolingual data is described in Table 2.

| Module | Big | Large |
|---|---|---|
| Layers | 12 | 24 |
| Attention Heads | 16 | 16 |
| Embedding Size | 1024 | 1024 |
| FFN Size | 2048 | 4096 |
| Shared Vocab | True | True |

Table 3: Hyper-parameters and model sizes of different models used in our systems.

## 2.2 Tokenizer

We use sentencepiece (Kudo and Richardson, 2018) to train a multilingual subword tokenizer. To represent the low-resource languages better, we follow Tran et al. (2021)'s settings, sampling text with temperature 5. Especially, for Yakut, we take monolingual data into account, since it's an extremely low-resource. Finally, For bilingual models, we used a vocabulary size of 32,000, and for multilingual models, we used 100,000.

We also apply subword regularization methods (Provilkov et al., 2019; Raffel et al., 2020) when tokenizing text. For low- and medium-resource directions, we apply BPE dropout on both the source and target sides and double the corpus size. And for high-resource directions, we only use it on the source side and don't do data augmentation stuff.

## 2.3 Model Architectures

Similar to Tran et al. (2021), we train two separate models: Many to English, or one system encompassing every language translated into English, and Many to Many directions, or one for English into every language and other non-English directions. Due to the very late release of the Yakut to the Russian corpus, we apply simple finetuning and backtranslation in this direction. For Chinese to and from English, we train a separate model. Because we are native speakers of Chinese and good at English, we introduce about 20 million high-quality private corpus [2]

**Dense Multilingual Model** Our model settings are empirically designed based on Transformer (Vaswani et al., 2017). We introduce two model architectures seen in Table 3. All models are implemented on top of the open-source toolkit Fairseq (Ott et al., 2019).

We also train three bilingual models: English to/from Chinese, and French to Germany. The aim is to compare how similarities among different languages will influence multilingual model. Due to the limitation of computing resources, we do not test in other language directions.

**Language Specific Adapter** In brief, an adapter layer is a dense layer with residual connection and non-linear projection. The hyperparameter b is the dim size of the inner dense layers. With a large set of globally shared parameters and small interspersed task-specific layers, adapters allow us to train and adapt a single model for a huge number of languages. Bapna et al. (2019) shows translation performance improvement in multilingual models with residual adapters. So after training the dense multilingual model, we add adapters for each language direction and apply further training and finetuning on these adapter layers. In detail, for high-resource directions, we add a larger adapter (b=4096). As for medium-resource, we set b=2048 and for the low-resource, we set b=1024.

## 2.4 Optimization Tricks

**Backtranslation** As shown in previous news task submissions, such as Tran et al. (2021) and Wang et al. (2021), backtranslation can significantly improve the BLEU score in low- and medium-resource language directions. We find no significant improvement in high-resource directions. And for some "X-en" high-resource directions, like zh-en shows in Table 4, backtranslation even lower the BLEU score. For this reason, we collect monolingual data for low- and medium-resource directions. All backtranslation data are generated by our well-trained multilingual model with Transformer Big settings. We use this generated data to train models with "Large" settings.

**Finetune** We use in-domain finetuning to further improve the model performance, which has proven effective on previous news translation tasks. We construct different types of finetuning data with the following approaches. Li et al. (2020); Wang et al. (2021) shows that low-frequency words frequency words are mostly domain-specific nouns, etc., which may indicate the topic directly. On the other hand, this year the shared task has changed from a news domain to a general translation task. We think finetuning our model by previous in-domain news data may be harmful to our model.

[2] We have a data group and a translation review team. First, we collect public monolingual data to make it multilingual. Second, we have a cooperative corpus or terminology base with our clients. With the consent of our clients, some non-public corpora and terminology are used for training.

| | Test Set | Big Model | Large Model | +BT | +Adapter | +Finetune | +LM Rerank |
|---|---|---|---|---|---|---|---|
| cs-en | wmt21 | 23.0 | 23.9 | — | 24.2 | 24.8 | 25.2 |
| uk-en | flore101 | 35.7 | 35.9 | 36.1 | 37.0 | 37.0 | 37.5 |
| ja-en | wmt21 | 21.5 | 21.8 | 24.0 | 27.2 | 28.0 | 28.0 |
| de-en | wmt21 | 29.4 | 29.9 | — | 30.0 | 32.1 | 32.3 |
| ru-en | wmt21 | 30.1 | 31.3 | 32.5 | 34.0 | 37.5 | 37.9 |
| en-cs | wmt21 | 15.7 | 17.0 | — | 20.4 | 20.2 | 21.3 |
| en-uk | flore101 | 24.1 | 24.5 | 27.1 | 28.0 | 28.9 | 29.0 |
| en-ja | wmt21 | 16.9 | 18.0 | 22.5 | 22.8 | 25.0 | 25.1 |
| en-de | wmt21 | 24.4 | 24.8 | 25.0 | 25.0 | 27.1 | 27.3 |
| en-ru | wmt21 | 20.6 | 21.0 | 21.1 | 23.4 | 24.2 | 25.6 |
| en-hr | flore101 | 25.8 | 26.4 | 28.9 | 29.3 | 30.0 | 30.3 |
| cs-uk | flore101 | 19.8 | 21.3 | 25.0 | 25.9 | 26.2 | 26.6 |
| uk-cs | flore101 | 20.8 | 22.0 | 24.1 | 24.3 | 24.3 | 24.5 |
| fr-de | wmt21 | 35.8 | 36.1 | 37.3 | 39.1 | 39.0 | 39.1 |
| zh-en | wmt21 | 31.4 | 32.0 | 31.7 | — | 34.0 | 34.1 |
| en-zh | wmt21 | 33.0 | 33.4 | 35.1 | — | 35.5 | 35.7 |
| Avg Incremental | — | — | 0.70 | 2.99 | 2.40 | 4.11 | 4.47 |

Table 4: Evaluation result on dev dataset. The inside of the dividing line represents the same model. We train X-en, en-X X-X, zh-en, and en-zh models separately. All translations are generated by beam search with beam size 5. All the models are the average of the final 5 checkpoints.

So we follow Li et al. (2020); Wang et al. (2021)'s strategies to select topic-related data based on a testset. We use the selected data for further finetuning. We experimented with the 2022 news development set and apply it directly to the 2022 test set.

**Language Model Reranking** Following Yee et al. (2019); Tran et al. (2021), we train language models and apply noisy channel reranking to the outputs of our final system. Unlike Tran et al. (2021), which trains a separate language model for each language, we train a multilingual language model for all languages to evaluate whether the multilingual language models can also improve the quality of translations.

**Model Ensemble** Model ensemble is a widely used technique in previous WMT shared tasks. To deal with biases toward recent training data, it is common to average parameters across multiple checkpoints of a model. We always average the last 5 checkpoints during training. During finetuning, we tune this hyperparameters (num epoch and num average checkpoints) on the development set and use it directly on the test set of wmt22.

## 3 Experiment

We conduct experiments to quantify the impact of each component in our system. The evaluation conduct on newstest2021 or development set on wmt22 using SacreBLEU (Post, 2018).

### 3.1 Settings

Every single model is trained on 8 NVIDIA A100 GPUs, each of which has 40 GB of memory. We also employ large batching with larger learning rates (Ott et al., 2018). We set the max learning rate to 0.0005 and warmup steps to 10000. All the dropout probabilities are set to 0.1. To speed up the training process, we conduct training with a half-precision floating point (FP16). During training multilingual, we add both source-side language tags and target-side language tags to leverage the gap between different language pairs. Following Tran et al. (2021), we divide data into multiple shards and downsample data from both high-resource directions and synthetic backtranslated with each training epoch using one shard.

### 3.2 Multilingual Models Result

We mainly evaluate our model and method on the wmt21 test set and flore101 dataset (Goyal et al., 2021). We analyze each aspect in our final submission and the cumulative effect. The effect of each component is shown in Table 4.

According to our experimental results, increasing the model capacity, increasing the sparsity of the model (adding a specific set of Adapter Layer

| | en-zh | zh-en | fr-de |
|---|---|---|---|
| Bilingual Model Big | 33.0 | 31.4 | 32.6 |
| Multilingual Model Big | 31.9 | 30.2 | 35.8 |

Table 5: BLEU score on wmt21 test set. All result is based on Big Model without any optimization

for each speech direction), fine-tuning the training set by extracting more relevant corpus based on the original text of the test set, and using the language model for reranking is effective on all language directions and the test set. Backtranslation is particularly effective in low- and medium-resource languages. Although the improvement of BLEU value by backtranslation on high-resource languages is not obvious or even worse, the average improvement by backtranslation is as obvious in a comprehensive view, with an average improvement of 1.68 BLEU per language direction.

Because this year's task is a general-purpose machine translation, rather than the news domain machine translation task of previous years, we are not submitting translation results that validate the optimal model on the development set, but rather the results of model fine-tuning on selected domain data after the release of the test set.

### 3.3 Distant Language Pairs Analysis

As shown in Tran et al. (2021), the multilingual model can significantly improve the BLEU score of medium- and low-resource language directions. For high-resource language directions, there are no significant enhancements. For high-resource languages, such as en-de, the BLEU score decreases slightly, and this is even more severe for distant language directions, en-ja, and en-zh for example. To compare the influence of the distance of the language family on the multilingual model. We train bilingual models for en-zh, zh-en, and fr-de. The test result is shown in Table 5. Since most of the language directions of wmt22 are Indo-European, distant languages, Chinese and Japanese for example, cannot benefit from the knowledge transfer additive of other languages, while the parameter capacity of the multilingual model is limited. These factors lead to poor results. Overall, when training multilingual models, languages with similar language families should be trained together, instead of putting all the languages together.

| Task | BLEU | Task | BLEU |
|---|---|---|---|
| cs-en | 25.3⋆ | en-cs | 26.3 |
| de-en | 33.4 | en-de | 36.1 |
| ja-en | 22.8 | en-ja | 39.4 |
| ru-en | 45.2* | en-ru | 32.6 |
| uk-en | 44.6⋆ | en-uk | 29.5 |
| zh-en | 28.1 | en-zh | 48.3 |
| fr-de | 41.8 | en-hr | 18.2⋆ |
| uk-cs | 36.5 | cs-uk | 38.3 |
| ru-sah | 15.3⋆ | sah-ru | 7.1⋆ |

Table 6: Our final submission results in 18 tasks. ⋆ represents the best score in the automatic evaluation. Note that the result is based on the submission website https://ocelot-wmt22.mteval.org/, not the official final result.

### 3.4 Submission Results

The results we finally submitted are shown in Table 6. We participate in 18 tasks this year. On the whole, all of our systems performed competitively, especially for Many-to-English directions. Yakut to/from Russian tasks is added bonus. Few teams participate in these two tasks.

## 4 Conclusion

In this paper, we described Lan-Bridge's submission to the WMT2022 General Translation shared task. Our main exploration was using a multilingual model to train different language pairs. It shows that the multilingual model can achieve state of art results in both high- and low-resource language directions. Meanwhile, we found that the multilingual model worked better for languages from the same or close language families than languages from distant language families. Finally, for extremely low-resource languages, even a multilingual model can boost their performance of them, but the translation is still far from usable.

## 5 Acknowledgments

We would like to thank our translator colleagues (Ruicheng Liu, and Qingtong Yang) for their assistance with linguistic knowledge and partners development teams (Jun Zhang et al.) for their foundational contributions to this project. We would also like to thank Yafu Li, and Yulong Chen from Westlake University for helpful discussions early on in the project.

## 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.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not greek to mbert: Inducing word-level translations from multilingual bert. *arXiv preprint arXiv:2010.08275*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. 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. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

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 of the Association for Computational Linguistics*, pages 311–318.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alan Aspuru-Guzik, and Alex Zhavoronkov. 2020. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

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, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

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

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook ai wmt21 news translation task submission. *arXiv preprint arXiv:2108.03265*.

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.

Kyra Yee, Yann Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*.

Boliang Zhang, Ajay Nagesh, and Kevin Knight. 2020. Parallel corpus filtering via pre-trained language models. *arXiv preprint arXiv:2005.06166*.