No Domain Left Behind

Hui Zeng
LanguageX AI Lab
felix_zeng_ai@aliyun.com

Abstract

We participated in the WMT General MT task and focus on four high resource language pairs: English to Chinese, Chinese to English, English to Japanese and Japanese to English). The submitted systems (LanguageX) focus on data cleaning, data selection, data mixing and TM-augmented NMT. Rules and multilingual language model are used for data filtering and data selection. In the automatic evaluation, our best submitted English to Chinese system achieved 54.3 BLEU score and 63.8 COMET score, which is the highest among all the submissions.

1 Introduction

Training neural machine translation models for a specific domain is a well-studied task. However, maximizing the performance of a single NMT model for multiple domains remains difficult. As a former translator and a current machine translation engineer, I always dream about building a versatile machine translation system – no domain left behind. Our neural machine translation system is developed using big transformer (Vaswani et al., 2017) architecture and the toolkit I used is fairseq (Ott et al., 2020). Rules, multilingual language model and faiss (Johnson et al., 2021) are used to align, clean and select parallel data. The following techniques are used in model training: a. Data mixing is used to mix general domain corpus with specific domain corpus; b. Back translation (Sennrich et al., 2016) is not applied because it is time-consuming. Instead, Neural Machine Translation with Monolingual Translation Memory (Cai et al., 2021) is used to fully utilize the monolingual corpus.

2 Data Filtering and Selection

The Chinese-English parallel data is mainly from CCMT Corpus [1], inhouse domain data from translation projects, as well as parallel data aligned from multilingual websites and e-books. The monolingual data for multiple domains is collected from the internet and e-books. WMT newstest2021 is used to evaluate the model's general domain performance. Multiple domain-specific test sets are created to evaluate the model's specific domain performance. Each domain has a test set of 1,000 sentences.

In order to build a versatile machine translation system, a total of 15 domains are covered in preparing the parallel and monolingual corpus. The primary domains and subdomains are listed as follows:

Literature
- Web novel
- Famous literary work
- Literature/Poetry
- Idioms/maxims/sayings
- Slang
- Conversation
- Names (personal, company)
- Symbols / Abbreviations / Acronyms

Art, History and Philosophy
- Arts/crafts/painting
- Cooking/culinary/gastronomy
- Folklore
- History
- Philosophy
- Graphic arts/photo/imaging
- Music
- Religion

[1] http://mteval.cipsc.org.cn:81/agreement/description

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 423–427
Abu Dhabi, December 7–8, 2022. ©2022 Association for Computational Linguistics

Social Science, Sociology, Ethics, etc.

Economy, Finance and Business

Business/commerce
Accounting
Finance (general)
Investment / Securities
Insurance
Economics
Real Estate

Fashion and Marketing

Advertising / Public Relations
Marketing / Market Research
Cosmetics / Beauty
Fashion
Textiles / Clothing / Fashion
Clothing/textiles

Politics and National Defense

Government/politics
International org/Dev/coop
Military / Defense

Law

Law (general)
Law: Contract(s)
Law: patents/trademarks/copyrights
Law: Taxation & Customs

Computers and IT (Information Technology)

Computers (general)
Computers: Systems, Networks
Computers: Hardware
IT (Information Technology)
Telecommunications
Internet, e-Commerce
SAP System Applications and Products
Media / Multimedia

Films and Television

Cinema/film/TV/drama

Games, Sports and Entertainment

Games / Video Games / Gaming / Casino
Sports / Fitness / Recreation
Tourism & Travel

Medical

Medical (general)
Medical: Cardiology
Medical: Dentistry
Medical: Health Care
Medical: Instruments
Medical: Pharmaceuticals
Dentistry
Veterinary
Genetics
Nutrition

Industry and Engineering

Engineering
Nuclear Eng/Sci
Automation & Robotics
Automotive / Cars & Trucks
Mechanics / Mech Engineering
Construction / Civil Engineering
Transport / Transportation / Shipping
Electronics / Elect Eng
Petroleum Eng/Sci
Surveying
Metallurgy / Casting
Mining & Minerals / Gems
Energy / Power Generation
Maritime / Sailing / Ships
Industrial
Food/drink
Paper / Paper Manufacturing
Printing & Publishing
Nuclear
Manufacturing
Furniture/household/appliance
Materials (Plastics, Ceramics, Rubber, Glass, Wood etc.)

Science

Astronomy/space
Aerospace/aviation/space
Mathematics & Statistics
Physics
Chemistry
Geography/geology
Architecture
Zoology
Biology
Botany
Meteorology
Metrology
Psychology
Education/pedagogy
Linguistics
Environment & Ecology
Anthropology
Archaeology
Genealogy

Agriculture and Animal Husbandry

Agriculture
Fisheries
Forestry wood timber
Wood Industry = Forestry
Wine / Oenology / Viticulture
Animal husbandry/livestock

Management and Training

Management

Human Resources
Safety

News and Journalism

2.1 Monolingual Data Filtering

The monolingual data for 15 primary domains are mainly collected from websites and e-books. The following rules are used for a simple cleaning:

•Remove duplicated sentences.

•Remove the sentences containing special characters.

•Remove the sentences containing html addresses or tags.

2.2 Parallel Corpus Aligning

There are a large number of multilingual websites and multilingual e-books, which are easily accessible. However, these data need to be aligned to create sentence level parallel corpus. To this end, a corpus aligner is created using Sentence-BERT (Reimers et al., 2019) and faiss (Johnson et al., 2021).

Regardless of order, thousands of source sentences and target sentences are first encoded into sentence embeddings using Sentence-BERT, and then faiss is used to retrieve the target sentences which is most similar in meaning to the source sentences. The aligning of thousands of parallel sentences could be finished within a few seconds.

2.3 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.

2.4 Parallel Data Filtering Using Multilingual Language Model

As mentioned in section 2.2, a corpus aligner is created using Sentence-BERT (Reimers et al., 2019) and faiss (Johnson et al., 2021). This can also be used to filter parallel data.

Apart from the corpus aligned from websites and e-books, in-house data from translation projects and public corpus like CCMT are also used.

The aforesaid corpus aligner can be used to score each parallel sentence pair so that the pairs with extremely low scores can be removed.

3 System Description

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

3.1 Data pre-processing

For data preprocessing, we use the tokenizer developed on my own to process both Chinese and English. Chinese text (including punctuations and numbers) is split to single character level. We keep the upper- and lower-case letters of English as they are, since we believe they are also important features for the model. Numbers in English text are also split into single digits. We use byte pair encoding (BPE) (Sennrich et al., 2016) to create a shared vocabulary, so that the vocabulary size is reduced to 45467. We also wrote a post-processor to restore the Chinese and English text to normal form.

3.2 Baseline Model Training

WMT newstest2021 is used to evaluate the model's general domain performance. Multiple domain-specific test sets are created to evaluate the model's specific domain performance. Each domain has a test set of 1,000 sentences.

The CCMT parallel Corpus filtered by rules and corpus aligner is used to train big transformer (Vaswani et al., 2017) English to Chinese and Chinese to English translation models as the general domain baselines.

Validation is performed every 2000 steps. The training is terminated if there is no gain in BLEU (Papineni et al., 2002) for 20 consecutive validations.

The BLEU scores on specific domains are also calculated as baselines.

3.3 Training on Mixed Data

Data mixing (Hasler et al., 2021) is used to improve translation quality for multiple new domains represented by small amounts of parallel data while maintaining the performance of a high-quality, general-purpose NMT model.

The importance of the training data sample can be increased by increasing its size, thereby

Model + Corpus	Literature EN2ZH	Law EN2ZH	Medical EN2ZH	Newstest2021 EN2ZH
filtered CCMT Corpus big transformer	21.5	23.2	19.8	28.7
filtered CCMT Corpus data mixing (general data, domain specific data) big transformer	28.7	38.6	36.3	35.9
filtered CCMT Corpus data mixing (general data, domain specific data) NMT with domain specific monolingual translation memory	31.2	43.5	40.1	39.2

Table 1: Different systems and their BLEU scores (only three typical domains are listed)

changing the ratio of training data and domain data to influence the trade-off between generic and domain performance.

3.4 NMT with Monolingual Translation Memory

Prior work has proved that Translation memory (TM) can boost the performance of Neural Machine Translation (NMT). In contrast to existing work that uses bilingual corpus as TM and employs source-side similarity search for memory retrieval, Cai (Cai et al., 2021) proposed a new framework that uses monolingual memory and performs learnable memory retrieval in a crosslingual manner.

This framework has unique advantages. First, the cross-lingual memory retriever allows abundant monolingual data to be TM. Second, the memory retriever and NMT model can be jointly optimized for the ultimate translation goal. The "plug and play" property of TM is useful for domain adaptation, where a single general-domain model can be adapted to a specific domain by using domain-specific monolingual TM.

3.5 Results

The BLEU scores on general test sets and some domain specific test sets for each corpus plus model combination are shown in Table 1.

In the automatic evaluation, our best submitted English to Chinese system achieved 54.3 BLEU score and 63.8 COMET score, which is the highest among all the submissions.

4 Conclusion

This paper describes LanguageX's translation system for the WMT2022 General MT task. The potential of a single translation model for all domains is explored. We are pleased to argue that, with data mixing and TM-augmented NMT, a versatile machine translation system with all-round translation performance could be built.

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

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.

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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-Scale Similarity Search with GPUs. *In IEEE Transactions on Big Data*, pp. 535-547, vol. 7.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural Machine Translation with Monolingual Translation Memory. *In*

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 7307–7318 August 1–6, 2021. Association for Computational Linguistics.

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, pages 3982–3992, Hong Kong, China, November 3–7, 2019*. Association for Computational Linguistics

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* pages 86–96. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Eva Hasler, Tobias Domhan, Jonay Trenous, Ke Tran, Bill Byrne, and Felix Hieber. 2021. Improving the Quality Trade-Off for Neural Machine Translation Multi-Domain Adaptation. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* pages 8470–8477, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.