Huawei BabelTar NMT at WMT22 Biomedical Translation Task: How we further improve domain-specific NMT

Weixuan Wang , Xupeng Meng, Suqing Yan, Ye Tian, Wei Peng[*]
Artificial Intelligence Application Research Center, Huawei Technologies
peng.wei1@huawei.com

Abstract

This paper describes Huawei Artificial Intelligence Application Research Center's neural machine translation system ("BabelTar"). Our submission to the WMT22 biomedical translation shared task covers language directions between English and the other seven languages (French, German, Italian, Spanish, Portuguese, Russian, and Chinese). During the past four years, our participation in this domain-specific track has witnessed a paradigm shift of methodology from a purely data-driven focus to embracing diversified techniques, including pre-trained multilingual NMT models, homograph disambiguation, ensemble learning, and preprocessing methods. We illustrate practical insights and measured performance improvements relating to how we further improve our domain-specific NMT system.

1 Introduction

The existing mainstream neural machine translation (NMT) system is predominantly data-driven. Our participation in WMT biomedical tasks traced back from 2019 has witnessed pursuits extending beyond this modality. In our WMT20 and WMT21 submissions, various domain adaption technologies (Bawden et al., 2020; Akhbardeh et al., 2021) have been applied including practical approaches fine-tuning on general-purpose models, back-translation (Sennrich et al., 2016a) and leveraging in-domain dictionaries (Peng et al., 2020; Wang et al., 2021). Despite achieving state-of-the-art (SOTA) BLEU scores for most of our submissions in the last two years, under-translation occurred in the "English ↔ Chinese" due to the models' incapability to handle long sentences (Wang et al., 2021). It was rectified by ensembling the affected model with the baseline, resulting in a decrease in BLEU scores. In addition, the models trained predominately with the general domain data still face challenges associated with domain adaptation.

In this paper, we present practical insights into how we further improve Huawei Artificial Intelligence Application Research Center's neural machine translation system ("BabelTar") in domain-specific machine translation. This year, our participation in the WMT22 biomedical translation task covers language directions between "English (EN)" and the other seven languages "German (DE)", "Spanish (ES)", "French (FR)", "Italian (IT)", "Portuguese (PT)", "Russian (RU)" and "Chinese (ZH)". More specifically, we adopt in-house general-purposed bilingual NMT models built upon the transformer-big architecture (Vaswani et al., 2017) and a pre-trained multilingual NMT model (M2M100) (Fan et al., 2021) with an M2M100-418M configuration as baseline models. Finetuned with the in-domain data provided by the organizer, the back-translated monolingual Medline data in English dating before July 2018, the in-domain dictionaries enhanced with terminologies, the models can be improved significantly over the last year's submissions, for example, +1.18 BLEU on "EN → IT" and +1.24 BLEU on "EN → DE". Leveraging the knowledge learned in addressing the ambiguities caused by homographs, we can further boost +0.65 BLEU in the language direction of "EN → ZH". By optimizing the sequence length during decoding, we successfully solve the issue of under-translation in the language pair of "EN ↔ ZH".

2 The Data

In this section we detail the bilingual and monolingual corpora used in this shared task (Table 1).

- **OOD**: The general domain data (OOD) are in-house data used to train the baseline models.
- **IND**: In all directions, we use the in-domain

[*] Corresponding author

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 930 - 935
December 7-8, 2022 ©2022 Association for Computational Linguistics

Directions	Train					Dev.	Test	Vocab.
	OOD	IND	IND-Dict.	IND-Aug.	IND-BT.			
EN→DE	6M	2.4M	62.5K	-	5.5M	1.1K	340	42K
DE→EN	6M	2.4M	62.5K	-	53M	1.1K	370	42K
EN→ES	3.3M	1.1M	131K	-	-	1K	410	40K
ES→EN	3.3M	1.1M	131K	-	52.5M	1K	382	40K
EN→FR	3M	2.8M	62.5K	-	-	1.6K	342	40K
FR→EN	3M	2.8M	62.5K	889K	53M	1.6K	314	40K
EN→IT	6M	139K	60.6k	235K	695k	0.8K	339	40K
IT→EN	6M	139K	60.6k	235K	55M	0.8K	327	40K
EN→PT	3M	7.1M	60.3K	-	-	1k	403	32K
PT→EN	3M	7.1M	60.3K	-	52.5M	1k	423	32K
EN→RU	3M	32K	60.4K	-	-	792	161	40K
RU→EN	3M	32K	60.4K	-	52.5M	792	210	40K
EN→ZH	3M	-	60.1K	847K	-	5K	347	50K
ZH→EN	3M	-	60.1K	847K	-	5K	311	50K

Table 1: Data used for training and evaluating the system. "M" is the acronym for "million", and K stands for "thousand", indicating the records of sentences, lexicon pairs or vocabularies. The Dev. datasets are extracted from the training datasets, and we use WMT21 shared task test data to evaluate our submission this year.

data (IND) provided by the shared task organizers to finetune the baseline models. [1] The IND data consists of WMT-released bitexts from Pubmed, UFAL, [2] Medline, [3] MeSpEn, [4] Scielo [5] and Brazilian Thesis and Dissertations.[6]

- **IND-dict.**: The lexicon pairs are collected from SNOMED-CT, [7] DOPPS[8] and WFOT.[9] Other terminologies are from Babel linguistics, [10] with COVID-19 related terms obtained from Neulab. [11]

- **IND-Aug.**: We augment the in-domain data using parallel corpora collected from TAUS [12] for the English ↔ Spanish, English ↔ French, English ↔ Italian, and English ↔ Chinese language pairs.

- **IND-BT.**: A batch of monolingual Medline data in English (IND-BT.) dated before July 2018 has been collected and back-translated for data augmentation. The official released IND data from WMT is also back-translated. The models used for back-translation are from our last year's shared task (Wang et al., 2021).

It is noted that OOD, IND, IND-dict. and IND-Aug. are combined and subsequently partitioned for training and evaluation.

3 The Approaches

The proposed systems are finetuned using the following methods. It is noted that bilingual models are trained on one Tesla V100 GPU, taking approximately 8-20 hours. All multilingual models are trained on eight Tesla V100 GPUs, taking 6-50 hours, depending on the volumes of data involved.

3.1 Multilingual NMT Models

Unlike our previous submissions focusing merely on bilingual NMT models, we leverage pre-trained multilingual NMT models (M2M-100) in the shared task this year.

[1] http://www.statmt.org/wmt21/biomedical-translation-task.html

[2] https://ufal.mff.cuni.cz/ufal_medical_corpus

[3] https://github.com/biomedical-translation-corpora/corpora

[4] https://temu.bsc.es/mespen/

[5] https://figshare.com/articles/dataset/A_Large_Parallel_Corpus_of_Full-Text_Scientific_Articles/5382757

[6] https://figshare.com/articles/A_Parallel_Corpus_of_Thesis_and_Dissertations_Abstracts/5995519

[7] https://www.nlm.nih.gov/healthit/snomedct/index.html

[8] https://static.lexicool.com/dictionary/XJ9XO98314.pdf

[9] https://static.lexicool.com/dictionary/HY1TK12777.pdf

[10] https://babel-linguistics.com/resources/glossaries/

[11] https://github.com/neulab/covid19-datashare/tree/master/parallel/terminologies

[12] https://md.taus.net/corona

System	**EN→DE**	**EN→ES**	**EN→FR**	**EN→IT**	**EN→PT**	**EN→RU**	**EN→ZH**
Bi-baseline	31.25	51.01	**47.27**	43.92	48.94	32.26	39.98
Bi-best	**32.49**	**51.81**	47.27	**45.10**	53.87	34.41	**42.23**
Multi-baseline	21.46	42.13	36.31	33.53	38.73	25.25	24.04
Multi-best	30.5	51.48	45.5	43.46	**53.98**	**37.14**	38.69
WMT22 Submission	33.42	44.75	37.85	48.48	52.55	37.03	47.68
Official Best	39.14	52.35	40.17	48.48	52.55	41.27	55.71

System	**DE→EN**	**ES→EN**	**FR→EN**	**IT→EN**	**PT→EN**	**RU→EN**	**ZH→EN**
Bi-baseline	40.46	50.79	48.82	44.73	47.36	44.69	**39.62**
Bi-best	**41.57**	**53.47**	**48.86**	44.73	**59.41**	47.69	39.62
Multi-baseline	33.67	43.23	35.73	36.43	41.84	39.76	21.57
Multi-best	40.68	52.02	46.37	**45.67**	58.08	**48.48**	34.96
WMT22 Submission	43.75	59.02	49.36	49.89	56.03	46.75	46.12
Official Best	46.95	60.45	50.95	49.89	56.03	50.01	46.17

Table 2: BLEU scores on related submissions. The Bi-baseline models represent the best bilingual models in our WMT21 participation (Wang et al., 2021) for language pairs in EN ↔ DE, EN ↔ FR, EN ↔ IT and EN ↔ ZH with others are out-of-domain bilingual NMT models newly trained for EN ↔ ES, EN ↔ PT and EN ↔ RU. The results of the Multi-baseline are the pre-trained multilingual NMT models from M2M100-418M on related language directions. The Bi-best and Multi-best are the bilingual and multilingual NMT models trained using the depicted methods achieving the best results.

Data	**EN→IT**	**IT→EN**	**EN→PT**	**PT→EN**	**EN→RU**	**RU→EN**
Baseline	33.53	36.43	38.73	41.84	25.25	39.76
+IND	42.17	43.72	50.12	54.74	36.25	47.09
+IND-all + IND	**43.46 (+1.29)**	**45.67 (+1.95)**	**53.98 (+3.86)**	**58.08 (+3.34)**	**37.14 (+0.89)**	**48.48 (+1.39)**

Table 3: Effects of applying different finetuning order to train English⇔Italian, English⇔Portuguese, English⇔Russian M2M-100 models on WMT21.

3.2 Domain-specific Dictionaries

Leveraging domain-specific dictionaries is proved a viable solution for domain adaptation in NMT (Peng et al., 2020; Wang et al., 2021) to enhance IND data coverage. A terminology dictionary is generated from the collected lexicons and attached to the end of the parallel corpus for each language direction to train the models.

3.3 Ensemble Learning

Ensemble learning is a representative method aggregating several models' predictions to obtain more accurate predictions. We average the probabilities of NMT output layers at each time step as depicted in Garmash and Monz (2016). In these experiments, we choose the top 3 best bilingual NMT models to participate in ensemble learning.

3.4 Homograph Disambiguation

Homographs may confuse an NMT model in selecting an inaccurate prediction due to conflicting word sense meanings in different domains. We design a novel approach to tackle homographic issues of NMT in the latent space to handle cross-domain ambiguities. The method is under review and will appear in another venue.

3.5 Preprocessing and Postprocessing

The under-translation problem presented in Wang et al. (2021) is associated with the inability of an NMT model to handle long sentences. The presence of noisy training data may cause under-translation. We optimize the preprocessing pipeline to include techniques like sentence segmentation, punctuation normalization, special tokens replacement, etc., leading to a resolution of the under-

Models	EN→DE	DE→EN	EN→FR	FR→EN	EN→IT	IT→EN	EN→ZH	ZH→EN
Model-1	31.25	40.46	**47.27**	48.82	43.92	**44.73**	**42.23**	**39.62**
Model-2	31.65	40.42	47.21	48.34	43.92	44.22	41.58	39.14
Model-3	31.01	40.17	47.25	48.55	45.04	44.05	41.29	38.92
Ensemble	**32.49**	**41.57**	46.79	**48.86**	**45.10**	44.71	41.36	38.50

Table 4: Results from the ensemble learning of the top three models on WMT21.

translation problem. More specifically, we first perform punctuation normalization to standardize data formats using Moses library (Koehn et al., 2007). Sentencepiece approach (Sennrich et al., 2016b) is subsequently used to tokenize the sentences into a series of subwords. Sentences with a length longer than a threshold (i.e., 80 subwords) are segmented to handle issues wrt under-translation. Preprocessing also replaces some unique tokens with placeholders, such as roman numbers, to avoid the out-of-vocabulary (OOV) problem. Postprocessing strategies are used to recover the previously segmented sentences. The detokenization is performed to convert subwords into words. Finally, we apply specific rules to handle punctuations and remove undesirable spaces.

4 Experimental Results and Analysis

As OOD data also contribute to the domain-specific NMT (Wang et al., 2021), both OOD data and IND data are used to finetune the NMT bilingual and multilingual NMT models. OK-aligned WMT21 test data are used for evaluation in the experiments. The BLEU scores are evaluated using the MTEVAL script from Moses (Koehn et al., 2007) with results shown in Table 2.

4.1 Multilingual NMT

It is challenging to finetune a pre-trained multilingual NMT model with hundreds of millions of parameters (i.e., 418 millions parameters for M2M-100-418M) with limited numbers of in-domain data. We design a two-stage training procedure in which a multilingual baseline initially finetuned on IND data of all available language pairs ("IND-all") is subsequently trained on data from a specific language pair ("IND"). As depicted in Table 3, such a two-stage training method ("IND-all + IND") is more effective than a simple finetuning step, achieving a significant improvement to the BLEU score (up to +3.86). Multilingual NMT models outperform bilingual NMT models, particularly for low-resource language pairs, such as EN ↔ RU and IT → EN (shown in Table 2).

4.2 Ensemble Decoding

We choose the three best models to ensemble in all experiments, including our best model submitted in the WMT21 shared task and the other two models trained following the methods depicted in this paper. Unlike the way mentioned in Wang et al. (2021) in averaging the logarithmic probabilities of a decoded token, we average the outputs of the output layer. This proves to be a more effective approach than the one used in previous years' submissions. The results are shown in Table 4. We have not investigated means to ensemble a pre-trained multilingual NMT model with our SOTA bilingual NMT models due to time and resource constraints in this year's shared task.

4.3 The Effect of Homograph Disambiguation

Table 6 demonstrates the effectiveness of applying a method designated for homographic disambiguation. It can be observed that resolving homographic issues in domain-specific NMT can significantly improve the BLEU score to up to +0.65.

4.4 Preprocessing to Solve Under-translation

To handle issues relating to under-translation, we design a segmentation strategy to break sentences longer than 80 subwords. Combined with other preprocessing techniques, we can further improve the performance of our domain-specific NMT system. Table 7 shows a +0.89 BLEU enhancement. A comparison of translated examples is shown in Table 5 to aid our understanding.

5 Discussion

It is the fourth year we have participated in this shared task, and we have made significant progress in our submissions measured against officially released test data from previous years. But the improvements for some language directions are not always accompanied by a consistent uplift of BLEU

Sentence	Example
Input	The disease duration ranged from 2 weeks to 60 months (median, 4 months), and the affected segment was C All the patients were followed up 3 to 42 months (median, 12 months).
Wang et al. (2021)	病程2周
This year	病程2周-60个月（中位，4个月），累及节段为C。随访3-42个月（中位，12个月）。
Input	The median age of the 30 patients was 56.5 (28-80) years old, among them, 25 patients were primary plasma cell leukemia, and 5 patients were secondary plasma cell leukemia.
Wang et al. (2021)	30例患者的中位年龄为56.5（28
This year	30例患者中位年龄为56.5（28-80）岁，其中原发性浆细胞白血病25例，继发性浆细胞白血病5例。

Table 5: A comparison of examples produced by Wang et al. (2021) and by models submitted this year in the translation task for EN → ZH.

Model	EN→ZH
Baseline	41.58
Homographic Disambiguation	**42.23 (+0.65)**

Table 6: The effect of applying an approach designed for homograph disambiguation to domain-specific NMT. The baseline is the NMT model for EN ⇔ ZH, without the assistance of the homograph disambiguation technique.

Model	EN→ZH
Baseline	40.69
Preprocessing + Baseline	**41.58 (+0.89)**

Table 7: Compared results between models with or without preprocessing when training EN → ZH translation model on WMT21.

for the contest year. The learned NMT models still suffer from "out of distribution" issues many deep learning models have encountered. Apart from maintaining the NMT models with a large amount of the latest IND data, we need to design deep learning systems to adapt to changes in distributions (Bengio et al., 2021).

On another point, we realized that the reference data sometimes do not reflect the ground truth of the translation during our manual evaluation process. It raises a related question about the rationale of using BLEU as an exclusive automatic evaluation criterion. Although BLEU may remain the default metric for evaluating machine translation quality, we strongly suggest the community investigate complementary metrics capable of accommodating good translation results with semantics variations in this shared task.

6 Conclusion

This paper depicts Huawei's neural machine translation system ("BebelTar") and the submission to the WMT22 biomedical shared task. The submission consists of fourteen models covering language directions between English and all seven other languages available in this track. We can improve the domain-specific NMT significantly by leveraging a broad range of techniques, which includes pretrained multilingual NMT models, lexicon-based enhancement, homograph disambiguation, ensemble learning, preprocessing and postprocessing, etc. In the meantime, we share practical insights on achieving the measured performance, hoping to contribute to the machine translation community in this shared task. Our future work will focus on investigating mechanisms to adapt a domain-specific NMT model to different distributions.

Acknowledgements

We express our gratitude to colleagues from HUAWEI Artificial Intelligence Application Center for their continuous support. We also appreciate the WMT22 organizers for hosting this shared task and anonymous reviewers' insightful comments, which enable us to further improve domain-specific NMT.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno-Yepes, Nancy Mah, David Martínez, Aurélie Névéol, Mariana L. Neves, Maite Oronoz, Olatz Perez-de-Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, italian and russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 660–687. Association for Computational Linguistics.

Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. 2021. Deep learning for ai. *Communications of the ACM*, 64(7):58–65.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1409–1418. ACL.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej 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.

Wei Peng, Jianfeng Liu, Minghan Wang, Liangyou Li, Xupeng Meng, Hao Yang, and Qun Liu. 2020. Huawei's submissions to the WMT20 biomedical translation task. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 857–861. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Weixuan Wang, Wei Peng, Xupeng Meng, and Qun Liu. 2021. Huawei aarc's submissions to the wmt21 biomedical translation task: Domain adaption from a practical perspective. In *Proceedings of the Sixth Conference on Machine Translation*, pages 868–873, Online. Association for Computational Linguistics.