Huawei AARC's Submissions to the WMT21 Biomedical Translation Task: Domain Adaption from a Practical Perspective

Weixuan Wang[1] *, **Wei Peng**[1] *, **Xupeng Meng**[1], **Qun Liu**[2]
[1]Artificial Intelligence Application Research Center, Huawei Technologies
{weixuanwang2,peng.wei1,mengxupeng}@huawei.com
[2]Noah's Ark Lab, Huawei Technologies
{qun.liu}@huawei.com

Abstract

This paper describes Huawei Artificial Intelligence Application Research Center's neural machine translation systems and submissions to the WMT21 biomedical translation shared task. Four of the submissions achieve state-of-the-art BLEU scores based on the official-released automatic evaluation results (EN→FR, EN↔IT and ZH→EN). We perform experiments to unveil the practical insights of the involved domain adaptation techniques, including finetuning order, terminology dictionaries, and ensemble decoding. Issues associated with overfitting and under-translation are also discussed.

1 Introduction

General-purpose machine translation systems have limited capability in addressing domain-specific tasks (Koehn and Knowles, 2017), for example, the WMT biomedical translation shared task, due to the low availability for high-quality in-domain data. In our WMT20 submission, various domain adaption technologies (Bawden et al., 2019, 2020) have been applied including practical approaches finetuning on general-purpose models, back-translation (Sennrich et al., 2016) and leveraging in-domain dictionaries (Peng et al., 2020b). Despite achieving state-of-the-art (SOTA) BLEU scores for most of the submissions, few efforts were put in place to disclose the practical insights associated with these techniques.

This year, the Artificial Intelligence Application Research Center (AARC) participate in the WMT21 biomedical translation task for eight language directions between English and other four languages (French, German, Italian, and Chinese). The baseline model is an in-house general-purpose NMT model built upon the transformer-big architecture (Vaswani et al., 2017). Apart from presenting an overview of the proposed biomedical Neural Machine Translation (NMT) system, we investigate the practical insights of the involved domain adaptation techniques, including finetuning order, terminology dictionaries, and ensemble decoding. Issues associated with overfitting to in-domain data and under-translation are also discussed.

2 The Data

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

2.1 Bilingual Data

In-domain bilingual data In all directions, we use the in-domain data (IND) provided by the shared task organizers to finetune the base model. [1] The IND data consists of WMT-released bitexts from Pubmed, UFAL [2] and Medline. [3]

We notice that the official release of IND data suffers from issues of misalignment between source and target sentences, and missing target sentences. The translation of a source sentence may be misplaced in a different line or even appeared in multiple lines on the target side. Moreover, a source sentence may have not been translated into in a target sentence. A data processing pipeline is developed to address the issues mentioned above (depicted in 3.4). The test data is the official release of the WMT19 shared task.

Augmented Bilingual Data We collect in-domain data from TAUS [4] for the English-French, English-Italian and English-Chinese language pairs (depicted in Table 1 as IND-Aug.) to address the in-domain data scarcity issue. For English-Chinese data, after collecting a portion of abstracts of China

*Co-first authors.

[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://md.taus.net/corona

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 868–873
November 10–11, 2021.

Directions	Train					Dev.	Test	Vocab.
	OOD	IND	IND-Dict.	IND-Aug.	IND-BT.			
EN→FR	3M	2.8M	62.5K	-	-	1.6K	1147	40K
FR→EN	3M	2.8M	62.5K	889K	53M	1.6K	952	40K
EN→DE	6M	2.4M	62.5K	-	5.5M	1.1K	963	42K
DE→EN	6M	2.4M	62.5K	-	53M	1.1K	794	42K
EN→IT	6M	139K	60.6k	235K	695k	0.8K	708	40K
IT→EN	6M	139K	60.6k	235K	55M	0.8K	760	40K
EN→ZH	3M	-	60.1K	847K	-	5K	774	50K
ZH→EN	3M	-	60.1K	847K	-	5K	418	50K

Table 1: Data used for training and evaluating the system. Note that "OOD" is short for the general domain data. "IND" is the in-domain data provided by the WMT organizers. "IND-Dict." refers to the in-domain dictionary. "IND-Aug." is the augmented IND data collected manually (not from MEDLINE, as depicted in 2.1). "IND-BT." is the IND monolingual data used for the back-translation. M is the acronym for "million," and K stands for "thousand".

Master's and Doctoral Dissertations, we align the data on the sentence level by using a model proposed by Açarçiçek et al. (2020). This is done by finetuning a RoBERTa (Liu et al., 2019) filter model on the TAUS dataset and selecting the source-target sentence pairs above a normalized log-probability threshold of 90%.

General-domain bilingual data We observe that finetuning the base model with IND data alone may incur sub-optimal BLEU scores. A conjecture is that the test data has a different distribution to that of the IND data. We present a case to show that finetuning the base model on a mixture of general domain data (OOD) and IND data can produce minor improvements (shown in 4.2).

2.2 Monolingual Data

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 competition (Peng et al., 2020b).

3 The Approaches

The proposed systems are finetuned using the following methods. All models are trained on one Tesla V100 GPU, taking approximately 8-20 hours depending on the volumes of data involved.

3.1 Leveraging In-domain Dictionary

Leveraging domain-specific dictionaries is a viable solution for domain adaptation in NMT (Peng et al., 2020a,b) to enhance IND data coverage. We collect lexicons from SNOMED-CT [5], DOPPS[6], WFOT [7] and generate a terminology dictionary which is subsequently attached to the end of training data. Terminology is further entended to cover COVID-19 related terms obtained from Neulab.[8]

3.2 Ensemble

Ensembling methods is a machine learning technique that aggregates several base models to generate one optimal predictive model (Garmash and Monz, 2016). We choose the top two models to ensemble in an attempt to produce a more general NMT model.

3.3 Architecture

To train the in-domain NMT model, we choose the in-house NMT system trained on general domain data as a baseline built upon the transformer-big architecture. LazyAdam optimizer is used during the training phase with a learning rate of $1e^{-5}$ and a warm-up period of 16,000 steps. The dropout ratio is set to 0.1, and the batch size for training and validation is 6,144 and 32 tokens, respectively. The width of the beam search is 4. Early stopping is applied to the training.

[5]https://www.nlm.nih.gov/healthit/snomedct/index.html
[6]https://static.lexicool.com/dictionary/XJ9XO98314.pdf
[7]https://static.lexicool.com/dictionary/HY1TK12777.pdf
[8]https://github.com/neulab/covid19-datashare/tree/master/parallel/terminologies

System I	EN→FR	FR→EN	EN→DE	DE→EN	EN→IT	IT→EN	EN→ZH	ZH→EN
baseline	42.94	42.10	31.05	38.24	40.54	49.19	34.41	33.41
+ IND	45.03	44.81	31.90	33.81	36.35	42.28	-	-
+ IND, IND-Dict.	45.93	45.05	32.68	38.98	36.69	45.13	-	-
+ IND, IND-Dict., OOD	45.65	-	32.45	39.26	41.77	48.88	-	-
+ IND, IND-Dict., OOD, IND-BT	-	44.56	33.79	40.25	42.69	50.80	-	-
+ IND, IND-Dict., OOD, IND-Aug.	-	-	-	-	40.83	-	36.08	35.35
+ IND, IND-Dict., OOD, IND-Aug., IND-BT	-	45.15	-	-	41.39	50.91	-	-
WMT21 Submission (Huawei_AGI)	**45.31**	**48.71**	**31.98**	**41.32**	**44.25**	**45.70**	**44.40**	**39.43**
WMT21 Best Official	**45.31**	**49.28**	**32.59**	**45.01**	**44.25**	**45.70**	**46.50**	**39.43**

Table 2: BLEU scores on all related submissions. The baseline models are finetuned in various configurations, including mixed finetuning on general-domain data (aka "OOD"), IND bitexts ("IND"), "IND-Dict." and the augmented IND data ("IND-Aug.").

3.4 Data Processing

Several pre-processing techniques are introduced to ensure the quality of the data.

- First, we perform punctuation normalization to standardize their formats using Moses library (Koehn et al., 2007).
- Then we carry out a primary data cleaning process to remove nonstandard sentences, including those with special characters, weblinks, extra spaces, and other bad cases.
- According to the length of the sentence after segmentation and the proportion of rare words, we remove bitexts with more rare words in the sentences. We further clean the data by skipping those sentence pairs with more than 100 subwords or less than one subword. The bitexts with a source and target sentence length ratio of more than 2.5 are excluded. A language detection tool [9] is used to filter out bitexts with abnormal language patterns, i.e., sentences with undesirable *langid*.
- An alignment model trained by fast-align (Dyer et al., 2013) [10] is used to score the corpus to remove misaligned parallel sentences.

After decoding, post-processing is performed to detokenize subwords and remove undesirable spaces between special characters and numbers, i.e., converting "rs = 0.9148" into "rs=0.9148".

4 Experimental Results and Analysis

The base systems are trained with OOD data and finetuned using IND data enhanced with monolingual data to produce the submitted results. We extract the OK-aligned data from the last two years (WMT19 and WMT20) and produce the test data to evaluate the NMT models. The BLEU scores are calculated using the MTEVAL script from Moses (Koehn et al., 2007). Results are shown in Table 2. The final two rows demonstrate the results of our submissions this year and the best official records released by the organizers.

4.1 Finetuning Order Does Matter

We identify the order of training is crucial in the experiment. We perform the experiment under the following three training strategies:

1. Strategy 1 (S1): the baseline is finetuned on the back-translation (BT) pseudo parallel corpus, followed by another finetuning using IND data.
2. Strategy 2 (S2): the baseline is finetuned using the IND data, followed by another finetuning using the BT data.
3. Strategy 3 (S3): the baseline is finetuned using a mixture of BT and IND data.

Table 5 presents the results of this comparative study for French→English translation direction. It can be observed that finetuning order generates significantly different BLEU scores, with Strategy 1 achieving a BLEU score +8.89 higher than that from Strategy 2. We follow the training strategy 1 in WMT21 shared task to this end.

4.2 OOD Data Mixed Finetuning

We observe that finetuning the base model with IND data alone (particularly with a limited amount of IND data) may result in sub-optimal BLEU scores. This may indicate overfitting to the training data, which has a different distribution to the test data. We perform a series of experiments to

[9] https://github.com/aboSamoor/polyglot
[10] https://github.com/clab/fast_align

Data	EN→FR	FR→EN	EN→DE	DE→EN	EN→IT	IT→EN
baseline	42.94	42.1	31.05	38.24	40.54	49.19
+IND	45.03	44.81	31.9	38.81	36.35	42.28
+IND + IND-Dict.	**45.93 (+0.90)**	**45.05 (+1.24)**	**32.68 (+0.78)**	**38.98 (+0.17)**	**36.69 (+0.34)**	**45.13 (+2.85)**

Table 3: Effects of applying terminology dictionaries to train English⇔French, English⇔German, English⇔Italian models on WMT20.

models	EN→FR	FR→EN	EN→DE	DE→EN	EN→IT	IT→EN	EN→ZH	ZH→EN
baseline	42.94	42.10	31.05	38.24	40.54	49.19	34.41	33.41
model-1	45.93	45.23	**33.37**	**40.15**	42.52	50.91	**36.05**	**35.31**
model-2	45.57	45.15	33.10	39.97	42.39	50.80	34.94	35.13
ensemble	**46.15**	**46.21**	33.27	40.12	**42.59**	**51.28**	35.78	35.11

Table 4: Results on the ensemble of three models on WMT20

Data	FR→EN	
	WMT19	WMT20
baseline	37.98	42.1
BT	30.06	34.19
IND	38.26	44.81
BT-IND (S1)	**39.26**	**45.10**
IND-BT (S2)	33.10	36.21
BT+IND (S3)	39.09	42.17

Table 5: The comparative study of finetuning order in French→English translation direction.

Data	EN→IT	IT→EN
baseline	40.54	49.19
IND	36.35	42.28
OOD-1M + IND + IND-Dict.	**41.77**	48.88
OOD-3M + IND + IND-Dict.	41.63	**49.10**
OOD-6M + IND + IND-Dict.	38.32	-

Table 6: Mixed finetuning OOD data creates improvements to address overfitting to IND when training English⇔Italian translation models on WMT20.

disclose this issue. As shown in Tables 6 and 7, finetuning with a mixture of OOD and IND data generates minor improvements. Interestingly, the experiment results are sensitive to the amount of OOD data involved. Future work is planned to look into this issue in detail.

4.3 The Effect of Terminology Dictionaries

In this section, we perform an ablation study to show the effectiveness of terminology dictionaries. The IND dictionaries are appended to bitexts as a part of the corpus to train NMT models. Table 3 presents consistent improvements for all six models in the experiment.

Data	EN→FR	
	WMT19	WMT20
baseline	39.06	42.94
IND	43.56	45.03
OOD-3M + IND + IND-Dict.	**43.65**	**45.65**
OOD-9M + IND + IND-Dict.	39.70	43.50

Table 7: The effects of mixed finetuning OOD data in improving the potential overfitting issue with IND data when training English→French translation models.

4.4 Ensemble Decoding

Ensemble decoding is applied to improve the generality of the NMT model by averaging the logarithmic probabilities of a decoded token. It can be observed from Table 4 that ensemble decoding is marginally effective compared to well-learned NMT models. This finding is consistent with that obtained from Wang et al. (2020).

4.5 Under-translation with Overfitting

Under-translation occurs when the NMT model fails to decode a portion of the input sentence. One of Chinese→English models under-translates a particular sentence of the WMT21 test data. For example, as shown in Table 8, "无危险器官受累患者的预后显著优于有危险器官受累的患者" of the input has been left untranslated. After increasing the width of the beam search, under-translation can be avoided. In our opinion, under-translation may be caused by noisy IND data, in which the learned self-attentions are not differentiable during decoding. By ensembling the affected model with the baseline, we successfully rectify the problem.

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).
prediction	病程2周
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.
prediction	30例患者的中位年龄为56.5（28
input	无危险器官受累患者的预后显著优于有危险器官受累的患者，患者10年OS率分别为100%和60.6%（P=0.0007）。
prediction	The 10-year os rate was 100% and 60.6% respectively (p=0.0007).

Table 8: Under-translated examples of English⇔Chinese. The portion of the sentence marked in red is under-translated.

5 Conclusion

This paper depicts Huawei's neural machine translation systems and submissions to the WMT21 biomedical shared task. We have achieved state-of-the-art BLEU scores for four of eight language pairs (EN→FR, EN↔IT and ZH→EN) based on the official-released results. We also explore practical issues for the involved domain adaptation techniques, including the effects of finetuning order, terminology dictionaries, and ensemble decoding on enhancing the performances of cross-domain NMT. We have discussed issues associated with overfitting and under-translation.

Acknowledgements

We express our gratitude to colleagues from HUAWEI AARC and Noah's Ark Lab for their continuous support. We also appreciate the WMT 21 organizers for hosting this shared task and the anonymous reviewers' insightful comments.

References

Haluk Açarçiçek, Talha Çolakoglu, Pinar Ece Aktan Hatipoglu, Chong Hsuan Huang, and Wei Peng. 2020. Filtering noisy parallel corpus using transformers with proxy task learning. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 940–946. Association for Computational Linguistics.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno-Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurélie Névéol, Mariana L. Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 3: Shared Task Papers, Day 2*, pages 29–53. 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.

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, Atlanta, Georgia. Association for Computational Linguistics.

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.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 28–39. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Wei Peng, Chongxuan Huang, Tianhao Li, Yun Chen, and Qun Liu. 2020a. Dictionary-based data augmentation for cross-domain neural machine translation. *CoRR*, abs/2004.02577.

Wei Peng, Jianfeng Liu, Minghan Wang, Liangyou Li, Xupeng Meng, Hao Yang, and Qun Liu. 2020b. 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. 2016. 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.

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.

Yiren Wang, Lijun Wu, Yingce Xia, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2020. Transductive ensemble learning for neural machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 6291–6298. AAAI Press.