Huawei's NMT Systems for the WMT 2019 Biomedical Translation Task

Wei Peng*, **Jianfeng Liu**
Artificial Intelligence Application Research Center
Huawei Technologies
Shenzhen, PRC
peng.wei1@huawei.com
liujianfeng@huawei.com

Liangyou Li*, **Qun Liu**
Noah's Ark Lab
Huawei Technologies
Hong Kong, PRC
liliangyou@huawei.com
qun.liu@huawei.com

Abstract

This paper describes Huawei's neural machine translation systems for the WMT 2019 biomedical translation shared task. We trained and fine-tuned our systems on a combination of out-of-domain and in-domain parallel corpora for six translation directions covering English–Chinese, English–French and English–German language pairs. Our submitted systems achieve the best BLEU scores on English–French and English–German language pairs according to the official evaluation results. In the English–Chinese translation task, our systems are in the second place. The enhanced performance is attributed to more in-domain training and more sophisticated models developed. Development of translation models and transfer learning (or domain adaptation) methods has significantly contributed to the progress of the task.

1 Introduction

In recent years, neural machine translation (NMT) has achieved substantial progress and outperforms statistical machine translation (SMT), especially when large volumes of parallel corpora are available. However, compared to out-of-domain (OOD) data, in-domain data is typically in a small volume and hard to obtain. Therefore, a lot of research focuses on how to make use of OOD data to improvement in-domain NMT systems. Among them, a well-accepted method for domain adaptation is to fine-tune a pre-trained baseline model using in-domain data (Koehn and Knowles, 2017; Luong and Manning, 2015; Freitag and Al-Onaizan, 2016).

In this paper, we present Huawei's practices on adapting our NMT systems from general-domain to in-domain. In addition to fine-tuning our OOD systems on in-domain data, we also resort to a broader spectrum of domain adaptation settings (Chu and Wang, 2018), including training models from scratch on a mixture of shuffled OOD and in-domain data and ensemble various models at the decoding stage. Final systems are submitted to the biomedical shared task of WMT 2019 on six translation directions for English–Chinese, English–French and English–German language pairs.

This paper is organized as below: Section 2 illustrates the system architecture followed by details of parallel corpora for training in Section 3. Section 4 presents our experimental settings. Results are presented and discussed in Section 5. In Section 6, we conclude the paper and unveil future work.

2 System Architecture

Our systems are implemented in TensorFlow 1.8 platform with the Transformer architecture (Vaswani et al., 2017) which consists of an encoder stack and a decoder stack with multi-head attention mechanisms. Each encoder layer consists of two sub-layers: a multi-head self-attention layer and a feed-forward layer with `relu` as the activation function. Compared to the encoder, each decoder layer includes an additional sub-layer to attend to outputs of the encoder. The hyperparameters used in our systems are defined in Table 1 which follow the transformer-big settings in Vaswani et al. (2017).

Hyperparameters	Values
Encoder Layers	6
Decoder Layers	6
Embedding Units	1,024
Attention Heads	16
Feed-forward Hidden Units	4,096

Table 1: Hyperparameters of our systems.

*Co-first author

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 164–168

3 Parallel Corpora

In this section, we present the parallel corpora used to train and evaluate translation models. The statistics of the data used is shown in Table 2. The OOD parallel corpora are collected from a number of sources. In addition to WMT parallel corpora for the news translation task, we also gather data from OPUS.[1] For English–Chinese tasks, we also include in-house data. The data generated by back-translating WMT monolingual corpus is named as "BT" data. Data from other sources such as the UM-Corpus (Tian et al., 2014) and Wikipedia are also included.

The in-domain data is from WMT biomedical translation shared task website.[2] More specifically, the in-domain data are gathered from the following sources (shown in Table 2):

- The EMEA corpus (Tiedemann, 2012). The EMEA corpus encompasses biomedical documents from the European Medicines Agency (EMEA). This corpus is a major component of in-domain training data.
- The UFAL medical corpus collection.[3] The extracted EN–FR parallel corpus contains data predominantly from PatTR Medical data whilst EMEA (OpenSubtitles and crawled) contributing to approximately one-third of the EN–DE data. PathTR is a parallel EN–DE and EN–FR corpus extracted from the MAREC patent collection and it has been used for this task since 2014, containing aligned sentence segments from patent titles, abstracts, and claims.[4]
- A small portion of in-domain data are from Medline and Pubmed.[5] This source of data is provided by the WMT Biomedical task organizers.

4 Experiments

The data depicted in Table 2 are mixed, pre-processed and split into training and development sets. The development data is created by random selection 1% from the mixed data sets. We also pre-processed the WMT 2018 test data and treated it as test data to benchmark the models trained under various settings.

Corpus	EN–ZH	EN–FR	EN–DE
OOD Parallel	48.94M	66.33M	22.28M
BT	6.12M	-	24.19M
UM-Corpus	875K	-	-
Wikipedia	-	818K	2.46M
UFAL	-	2.81M	3.04M
EMEA	-	1.09M	1.11M
Medline [6]	-	55K	29K
Pubmed	-	613K	-
Total	55.93M	71.72M	53.11M

Table 2: Corpora statistics in the numbers of sentence pairs after cleaning.

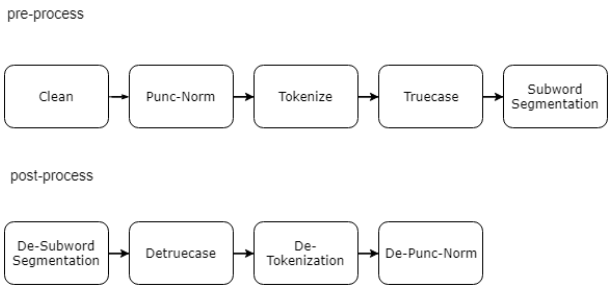

Figure 1: Data Processing Pipeline

4.1 Pre-processing and Post-processing

We noticed that the data processing procedure is an important factor in enhancing the quality of training data and thus the performance of trained models. Our pre-processing pipeline is composed of a number of steps (depicted in Figure 1). The data is undergone data cleaning, puncture normalization (Punc-Norm), tokenization, truecasing and subword segmentation:

- Data cleaning addresses the issues of noisy training data. For example, we remove sentence pairs which are potentially misaligned according to scores from fast-align. We also remove sentence pairs if the ratio of language-specific characters is lower than a threshold. As we found parallel corpora of a language pair may contain sentences pairs which are in a third language, we apply language detection[7] and filtering as well.

[1] http://opus.nlpl.eu/

[2] http://www.statmt.org/wmt19/biomedical-translation-task.html

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

[4] https://www.cl.uni-heidelberg.de/statnlpgroup/pattr/

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

[7] https://github.com/aboSamoor/polyglot

- After cleaning data, a few common steps used for machine translation are applied by using scripts from Moses (Koehn et al., 2007). Punc-Norm deals with variations of punctuation in different languages (i.e., French, German) by normalizing them into a standard form. Tokenization is a language-dependent process of splitting a sentence into a sequence of tokens. Truecasing models are trained for each language and applied appropriate case forms on words.

- In order to alleviate the out-of-vocabulary problem, subword segmentation (Sennrich et al., 2016) is used as well. Instead of training an individual segmentation model for each language independently, we directly use subsets of the multilingual vocabularies[8] from the BERT (Devlin et al., 2018) project.[9] It is generated by the WordPiece (Schuster and Nakajima, 2012) model trained on Wikipedia dump. A greedy algorithm is then applied to segment a word in our corpus into a sequence of subwords according to the vocabulary if applicable. For example, "Bitstream" is segmented into "Bit" and "##stream".

After decoding, the outputs are post-processed by combining subwords, de-truecasing and detokenization. Punctuation is also converted back to their original form in a specific language when translating to Chinese, French and German.

4.2 Training and Decoding Details

The models are trained in two different ways: (1) Mixed: the model is simply trained on a mixture of data without differentiating OOD and in-domain data. The data is shuffled randomly and there is no oversampling technique applied; (2) Fine-tuned: the baseline model is first pre-trained on the OOD parallel corpus and then fine-tuned on the in-domain data.

All systems are trained for 400K steps, except that, in the Fine-tuned setting, we further fine-tune base systems for 300K unless early stopped. The training was performed on GPU clusters with 4 or 8 Tesla V100. Follow Transformer, we use Adam as a optimizer and a dynamic learning rate with a linear warmup and root-squared decay. The batch size is set to be 3K source or target words on each GPU card.

We average top 10 checkpoints (Vaswani et al., 2017) evaluated against the development set as the final model for decoding. The beam size is set to 4 and a length penalty weight factor with a value 1 is used (Wu et al., 2016).

We further optionally apply ensemble decoding to combine best models trained in the two settings mentioned above. Ensemble decoding (or prediction) is an approach combining multiple predictors to reduce the errors. It has been widely used in improving NMT performance.

5 Experimental Results

We experimented with more than twenty models in total trained on different combinations of various data and under different settings. `sacrebleu.py` (Post, 2018) and `multi-bleu.perl` from Moses[10] are used to evaluate translations on the development and test data. Table 3 shows BLEU scores on WMT 2018 test set under different settings. We found that models from fine-tuning on in-domain data outperform models trained on the mixed data set when reasonable volumes of in-domain data are available (e.g., on EN–FR and EN–DE). By contrast, the mixed method performs the best on EN–ZH where we do not have genuine in-domain data for fine-tuning. Another interesting finding is that the ensemble decoding consistently takes the middle place when we simply combine the best two models under the three settings. We presume this is caused by domain issues as at least one of the two models used was not well trained on in-domain data.

The results in terms of official BLEU scores of our submissions for WMT 2019 are presented in Table 4 and Table 5. Our final systems achieve the best BLEU scores on English–French and English–German language pairs according to the official evaluation results. In the English–Chinese translation task, our systems are in the second place. We can also find from the tables that training with the mixed data, fine-tuning on in-domain data have contributed to a number of winning models on different language pairs. While the mixed method works better than the Fine-tuned method on English–Chinese and English–

[8] Vocabulary size for EN–ZH: 42K (ZH), 46K (EN); Vocabulary size for EN–DE: 58K (DE), 58K (EN); Vocabulary size for EN–FR: 59K (FR), 58K (EN).

[9] https://github.com/google-research/bert

[10] https://github.com/moses-smt/mosesdecoder

BLEU Scores on WMT 18 Data						
Models	**EN2ZH**	**ZH2EN**	**EN2DE**	**DE2EN**	**EN2FR**	**FR2EN**
Baseline	33.49	19.46	24.4	27.98	30.57	35.40
Fine-tuned	31.65	21.46	**26.56**	**32.8**	**34.38**	**40.56**
Mixed	**34.36**	**24.37**	24.54	29.18	31.88	36.46
Ensemble (top 2)	34.27	23.41	25.28	32.36	34.30	38.77

Table 3: BLEU scores of the trained models measured against a subset of the test data for WMT 18 biomedical task (bold fonts show the best scores).

German, the fine-tuned method outperforms on French–English (EN2FR Run1) due to a reasonable volume of high-quality in-domain data included. It is noted that the submission (EN2FR Run3) based on the ensemble decoding method has resulted in much lower performance.

According to our experiments and experiences, we reached the same conclusion as that from the WMT biomedical task organizers (Neves et al., 2018): the enhanced performance is attributed to more in-domain training and more sophisticated models developed (i.e., Transformers). The development of translation models and transfer learning (or domain adaptation) methods have significantly contributed to the progress of the task.

6 Conclusions

In this paper, we present Huawei's neural machine translation systems for the WMT 2019 biomedical translation shared task. More than twenty models have been trained and tested under different training settings on three language pairs (six translation directions), i.e., English–Chinese, English–French and English–German. A number of pre-processing and post-processing techniques have been employed to enhance the quality of the data. Our final systems rank the best BLEU scores on English–French and English–German language pairs and the second on English–Chinese according to the official evaluation results in terms of BLEU scores.

Acknowledgments

We would like to express our gratitude to colleagues from HUAWEI Noah's Ark Lab and HUAWEI AARC for their continuous support. We also appreciate the organizers of WMT 19 Biomedical Translation Task for their prompt replies to our inquiries.

References

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation.

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*, pages 28–39, Vancouver, Canada. Association for Computational Linguistics.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Mariana L. Neves, Antonio Jimeno-Yepes, Aurélie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin M. Verspoor. 2018. Findings of the wmt 2018 biomedical translation shared task: Evaluation on medline test sets. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 324–339, Belgium, Brussels. Association for Computational Linguistics.

WMT 19 Submission	EN2ZH	ZH2EN	EN2DE	DE2EN	EN2FR	FR2EN
Best Official	**42.34**	**34.13**	**27.89**	**28.82**	**39.95**	**35.56**
ARC Run 1	35.47	30.07	**27.89**	28.71	**39.95**	35.51
ARC Run 2	35.47	30.05	27.86	28.79	36.67	35.51
ARC Run 3	35.47	30.05	27.85	**28.82**	36.19	**35.56**
ARC Best Model	Mixed	Mixed	Mixed	Mixed	Fine-tuned	Fine-tuned

Table 4: Official BLEU scores of ARC submission for WMT 19 biomedical task test sets with all sentences (bold fonts show the best official scores).

WMT 19 Submission	EN2ZH	ZH2EN	EN2DE	DE2EN	EN2FR	FR2EN
Best Official	**43.92**	**35.61**	**35.39**	**38.84**	**42.41**	**38.24**
ARC Run 1	37.09	32.15	**35.39**	38.66	**42.41**	38.18
ARC Run 2	37.09	32.16	35.28	38.80	38.89	38.18
ARC Run 3	37.09	32.16	35.26	**38.84**	38.29	**38.24**
ARC Best Model	Mixed	Mixed	Mixed	Mixed	Fine-tuned	Fine-tuned

Table 5: Official BLEU scores of our submissions for WMT 19 biomedical task with OK-aligned test sets (bold fonts show the best official scores).

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.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5149–5152.

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, Berlin, Germany. Association for Computational Linguistics.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, Yi Lu, Shuo Li, Yiming Wang, and Longyue Wang. 2014. Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1837–1842, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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 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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.