Sarah's Participation in WAT 2019

Raymond Hendy Susanto, Ohnmar Htun, Liling Tan
Rakuten Institute of Technology
Rakuten, Inc.
{first.last}@rakuten.com

Abstract

This paper describes our MT systems' participation in the WAT 2019. We participated in the (i) Patent, (ii) Timely Disclosure, (iii) Newswire and (iv) Mixed-domain tasks. Our main focus is to explore how similar Transformer models perform on various tasks. We observed that for tasks with smaller datasets, our best model setup are shallower models with lesser number of attention heads. We investigated practical issues in NMT that often appear in production settings, such as coping with multilinguality and simplifying pre- and post-processing pipeline in deployment.

1 Introduction

This paper describes our machine translation systems' participation in the 6th Workshop on Asian Translation (WAT-2019) translation task (Nakazawa et al., 2019). We participated in the (i) Patent, (ii) Timely Disclosure, (iii) Newswire, and (iv) Mixed-domain tasks. We trained our systems under a constrained setting, meaning that no additional resources were used other than those provided by the shared task organizer. We built all MT systems based on the Transformer architecture (Vaswani et al., 2017). Our main findings for each task are summarized in the following:

- **Patent task:** We built several translation systems for six translation directions. We also explored a multilingual approach and compared it with the unidirectional models.
- **Timely disclosure task:** We tried a simplified data processing such that the model is trained directly on raw texts without requiring language-specific pre/post-processing.
- **Newswire task:** We explored fine-tuning the hyperparameters of a Transformer model on a relatively small dataset and found that a compact model is able to achieve a competitive performance.
- **Mixed-domain task:** We explored low-resource translation approaches for Myanmar-English.

2 JPO Patent Task

2.1 Task Description

For the patent translation task, we used the JPO Patent Corpus (JPC) version 4.3, which is constructed by the Japan Patent Office (JPO). Similar to previous WAT tasks (Nakazawa et al., 2015, 2016, 2017, 2018), the task includes patent description translations for Chinese-Japanese, Korean-Japanese, and English-Japanese. Each language pair's training set consists of 1M parallel sentences. We used the official training, validation, and test split provided by the organizer without any external resources. We trained individual unidirectional models for each language pair. Additionally, we explored multilingual NMT approaches for this task.

2.2 Data Processing

We used SentencePiece (Kudo and Richardson, 2018) for training subword units based on byte-pair encoding (BPE). We pre-tokenized the data using the following tools:

- Juman version 7.01[1] for Japanese,
- Stanford Word Segmenter version 2014-06-16[2] with Peking University (PKU) model for Chinese,
- Mecab-ko[3] for Korean, and
- Moses tokenizer for English.

Source and target sentences are merged for training a joint vocabulary. We set the vocabulary size to 100,000 and removed subwords that occur less than 10 times from the vocabulary, following similar pre-processing steps for the baseline NMT system released by the organizer.[4]

[1] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN
[2] https://nlp.stanford.edu/software/segmenter.shtml
[3] https://bitbucket.org/eunjeon/mecab-ko/
[4] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2019/baseline/dataPreparationBPE.html

Proceedings of the 6th Workshop on Asian Translation, pages 152–158
Hong Kong, China, November 4, 2019.

Embedding dim.	1024
Tied embeddings	Yes
Transformer FFN dim.	4096
Attention heads	8
En/Decoder layers	6
Label smoothing	0.1
Dropout	0.3
Attention weight dropout	0.1
Transformer FFN dropout	0.1
Learning rate	0.001
Batch size in tokens	4000
Update frequency	1

Table 1: JPO model settings

2.3 Model

Our NMT model is based on the Transformer (Vaswani et al., 2017) implementation in the Fairseq toolkit (Ott et al., 2019). The details of the parameters used for our experiments are summarized in Table 1. Encoder's input embedding, decoder's input and output embedding layers were tied together (Press and Wolf, 2017), which saves significant amounts of parameters without impacting performance. The model was optimized with Adam (Kingma and Ba, 2015) using $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon =$ 1e-8. We used the same learning rate schedule as (Ott et al., 2018) and run the experiments on 4 Nvidia V100 GPUs, enabling mixed-precision training in Fairseq (`--fp16`). The best performing model on the validation set was chosen for decoding the test set. We trained 4 independent models with different random seeds to perform ensemble decoding.

2.4 Results

Table 2 shows our model performance for the patent task. For brevity, we only reported the results on the JPCN test set, which is a union of JPCN{1,2,3}, and the Expression Pattern task (JPCEP) for zh-ja. For the detailed breakdown for each test set, we would like to refer readers to the overview paper. Since human evaluation result is not available as the time of this writing, we only present the results in terms of BLEU scores. It is clear that ensemble decoding significantly outperforms single model decoding. Under the constrained settings, our best submissions obtain the first place in the WAT leaderboard[5] for zh-ja, ja-zh, and ja-en. Interestingly, our model did not perform well on ja-ko translation, where the performance was behind the organizer's baseline system which is based on a sequence-to-sequence LSTM. More careful investigation could help us understand which component in our training pipeline (e.g., data processing or tokenization) could possibly cause this difference.

[5] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html

Task	Model	BLEU
JPCN zh-ja	Unidirectional, single	46.77
JPCN zh-ja	Unidirectional, ensemble	48.68
JPCN zh-ja	Multilingual, single*	45.98
JPCN ja-zh	Unidirectional, single	40.78
JPCN ja-zh	Unidirectional, ensemble	42.22
JPCN ja-zh	Multilingual, single*	39.57
JPCN ko-ja	Unidirectional, single	71.41
JPCN ko-ja	Unidirectional, ensemble	72.55
JPCN ko-ja	Multilingual, single*	69.80
JPCN ja-ko	Unidirectional, single	69.81
JPCN ja-ko	Unidirectional, ensemble	70.94
JPCN ja-ko	Multilingual, single*	67.87
JPCN en-ja	Unidirectional, single	44.14
JPCN en-ja	Unidirectional, ensemble	44.97
JPCN en-ja	Multilingual, single*	43.82
JPCN ja-en	Unidirectional, single	41.74
JPCN ja-en	Unidirectional, ensemble	43.34
JPCN ja-en	Multilingual, single*	39.82
JPCEP zh-ja	Unidirectional, single	35.41
JPCEP zh-ja	Unidirectional, ensemble	36.73
JPCEP zh-ja	Multilingual, single*	34.45

Table 2: JPO task results. Note that we did not submit our multilingual model output (marked with *) and it serves as comparative purposes.

2.5 Multilingual Experiments

Given that multiple language pairs are involved for this task, we further experimented with multilingual NMT approaches after the submission period. We followed the approach in (Johnson et al., 2017), which adds an artificial token in each source sentence for indicating the target translation language (`--encoder-langtok tgt` in Fairseq). Encoder and decoder parameters are shared among all the language pairs. We merged all training data from all four languages for training a joint subword vocabulary of size 100,000 approximately. As a result, we can share the embedding layer in the encoder and decoder. Since the number of training examples for each direction is

the same, we iterate round-robin over batches from the six language pairs.

As shown in Table 2, our multilingual result did not show improvement in the NMT systems, falling behind the unidirectional model by not more than 2 BLEU points for single decoding. Nonetheless, parameter sharing in multilingual model reduces the total number of parameters to approximately the same as that of one unidirectional model. In practice, this can potentially simplify production deployment for multiple language translation. Effectively, the model is able to perform a zero-shot translation for language pairs not included in this task (such as Chinese-Korean), although we left this for future investigation.

3 JPX Timely Disclosure Task

3.1 Task Description

The timely disclosure task evaluated Japanese to English translations from the Timely Disclosure Document Corpus (TDDC), which is constructed by the Japan Exchange Group (JPX). The corpus consists of 1.4M parallel Japanese-English sentences made from past timely disclosure documents between 2016 and 2018. The validation and test sets are further split into two sub data sets: 1) nouns and phrases ("X_ITEMS") and 2) complete texts ("X_TEXTS"). We used the official data split given by the organizer with no additional external resources. For this task, we did a brief study on the effect of different pre-processing procedures on model performance.

3.2 Data Processing

MT systems typically include complicated pre/post-processing pipeline, which is often language-specific. This usually forms a long chain in the pipeline: *tokenization/segmentation* → *truecasing* → *translation* → *detruecasing* → *detokenization*.

While tools like Moses (Koehn et al., 2007), Experiment Management System[6] and SacreMoses[7] simplify the data processing pipeline, handling various languages produces significant technical debt in maintaining language specific resources and rules. Although there are language agnostic approaches to tokenization/truecasing, (e.g. Evang et al., 2013; Susanto et al., 2016), the errors from various components in the pipeline are propagated. Instead we propose to use a single step pre-processing using SentencePiece subword tokenizer.

SentencePiece is an unsupervised tokenizer that can learn directly on raw sentences, and pre-tokenization is an optional step. This greatly simplifies the training process as we can feed the data directly into SentencePiece to produce subword tokens based on BPE. We merged source and target sentences for training a shared vocabulary of 32,000 tokens with 100% character coverage and no further filtering. We removed empty lines and sentences exceeding 250 subword tokens from the training set. Both items and texts sub data sets were processed in the same manner. We concatenated the items and texts development data sets together for model validation.

Embedding dim.	1024
Tied embeddings	Yes
Transformer FFN dim.	4096
Attention heads	8
En/Decoder layers	6
Label smoothing	0.1
Dropout	0.1
Attention weight dropout	0.1
Transformer FFN dropout	0.1
Learning rate	0.001
Batch size in tokens	14336
Update frequency	2

Table 3: JPX model settings

3.3 Model

For the timely disclosure task, we used a 6-layer Transformer with 8 heads as shown in Table 3. The overall model is similar to the JPO model, except a couple differences: 1) Smaller dropout probability, 2) Larger number of tokens per batch, and 3) Delayed updates. Particularly, gradients for multiple sub-batches on each GPU were accumulated, which reduces variance in processing time and reduces communication time (Ott et al., 2019). With `--update-freq 2`, this effectively doubles the batch size. We trained 4 independent models with different random seeds to perform ensemble decoding on both the items and texts test sets. Every model was trained for 40 epochs and the best performing checkpoint on validation set was chosen.

[6] http://www.statmt.org/moses/?n=FactoredTraining.EMS

[7] https://github.com/alvations/sacremoses

Task	Model	Tokenization	BLEU	Human
TDDC Item ja-en	Single	None	52.77	29.25
TDDC Item ja-en	Ensemble	None	54.25	36.75
TDDC Item ja-en	Single	Juman*	52.83	-
TDDC Text ja-en	Single	None	54.84	37.75
TDDC Text ja-en	Ensemble	None	58.38	49.50
TDDC Text ja-en	Single	Juman*	57.34	-

Table 4: JPX task results. Note that we did not submit the output from our model that includes Japanese word segmentation (marked with *) and it serves as comparative purposes.

3.4 Results

Table 4 shows our model performance for the timely disclosure task. Human evaluation ranks our best submissions in the first place for the "Item" test set and second place for the "Text" test set. After the submission period has ended, we did a further study on the effect of including Japanese segmentation in data pre-processing. We tokenized the Japanese text using Juman and re-trained our model. Comparing their BLEU scores on single decoding, we observe that tokenization slightly improves on Item data, while it significantly improves on Text data by 2.5 BLEU points, which might have boosted our scores in the leaderboard. Nonetheless, a single step pre-processing greatly simplifies our training and translation pipeline. This is particularly helpful in deploying MT systems for several languages in production settings because it allows us to build an end-to-end system that does not rely on language-specific pre/post-processing.

4 JIJI Newswire Task

4.1 Task Description

The newswire task evaluated Japanese-English translations on the JIJI corpus. The corpus was created by Jiji Press in collaboration with National Institute of Information and Communications Technology (NICT). The data set contains 200,000 parallel sentences for training, 2,000 for validation and 2,000 for testing. We did not use any external resources other than the provided corpus. For this task, we investigated the importance of choosing a suitable Transformer network size with respect to the size of our training set.

4.2 Data Processing

We ran Juman version 7.01 for Japanese word segmentation but English sentences were not tokenized. After tokenization, both Japanese and English sentences were combined and fed into SentencePiece for training BPE subword units. The subword vocabulary size is 32,000 with 100% character coverage and no further filtering. We further removed empty lines and sentences exceeding 250 subword tokens from the training set. We also tried feeding the sentences directly into SentencePiece without pre-tokenization for Japanese but we observed a weaker performance on the JIJI task by doing so.

Embedding dim.	512
Tied embeddings	Yes
Transformer FFN dim.	2048
Attention heads	2
En/Decoder layers	5
Label smoothing	0.2
Dropout	0.4
Attention weight dropout	0.2
Transformer FFN dropout	0.2
Learning rate	0.001
Batch size in tokens	4000
Update frequency	1

Table 5: JIJI model settings

4.3 Model

Sennrich and Zhang (2019) adapted RNN-based NMT systems in low-resource settings by reducing vocabulary size and careful hyperparameter tuning. Similarly, we applied system adaptation techniques to Transformer-based NMT systems for this task, given that the JIJI corpus is a relatively small data set. As shown in Table 5, we chose to scale down our Transformer model so as to prevent overfitting. We made the following adjustments: (i) Halving embedding and hidden dimension, (ii) Reducing the number of attention heads and encoder/decoder layers, and (iii) Increasing regularization through

Task	Model	BLEU	Human
JIJI en-ja	BASE (Single)*	16.70	-
JIJI en-ja	MINI (Single)	21.80	55.25
JIJI en-ja	MINI (Ensemble)	22.65	63.25
JIJI ja-en	BASE (Single)*	15.91	-
JIJI ja-en	MINI (Single)	21.34	44.75
JIJI ja-en	MINI (Ensemble)	21.84	50.75

Table 6: JIJI task results. Note that we did not submit the BASE model output (marked with *) and it serves for comparative purposes.

dropout and weight decay (`--weight-decay 0.0001`). We used the same model set up for both directions. Considering the size of this data set, we were able to run longer epochs for JIJI tasks: 150 epochs for Japanese→English and 100 for English→Japanese. We compare the performance of this downsized model (MINI) to the previous model setup used for the JPX task (BASE).

4.4 Results

Table 6 shows our model performance on the newswire task. We can observe that the MINI model significantly outperforms the BASE model by around 5 BLEU points on single decoding. These results affirm our hypothesis that it is possible to improve NMT performance in low-resource settings by more careful hyperparameter tuning without relying too much on auxiliary resources. Overall, our submissions for both translation directions ranked the first in the leaderboard in terms of BLEU scores and under the constrained settings. Unfortunately, our system output are the only constrained submissions that were humanly evaluated and thus we are not able to do a comparative evaluation.

5 Mixed-domain Task

5.1 Task Description

The mix-domain task evaluated Myanmar-English translations from the University of Computer Studies, Yangon (UCSY) (Ding et al., 2018) and the Asian Language Treebank (ALT) corpora (Ding et al., 2019). The models were trained on the UCSY corpus, then validated and tested on the ALT corpus. The UCSY corpus contains approximately 200,000 sentences, ALT validation and test sets had 1,000 sentences each. No other resources were used to train our models for the task participation.

5.2 Data Processing

For the mix-domain task, no special preprocessing steps were taken to handle the data; sentences were fed directly to the SentencePiece to produce subwords tokens. We experimented with two Transformer models of varying sizes using the Marian[8] toolkit (Junczys-Dowmunt et al., 2018).

5.3 Model

Using similar models settings as (i) JPX model in Table 3 with 32,000 subwords tokens at 100% character coverage (BASE) and (ii) the JIJI model in Table 5 with 10,000 subwords tokens at 100% character coverage (MINI), we train one model each to compare (i) vs (ii) in the Mixed-domain Task. We only participated in the English to Myanmar task.

5.4 Results

Task	Model	BLEU
ALT2 en-my	BASE (Single)	12.55
ALT2 en-my	MINI (Single)	19.64
ALT2 en-my	MINI (Ensemble)	19.94

Table 7: Mixed-domain Task Results

Table 7 shows the result of our English to Myanmar models. Due to the low resource nature of the Myanmar-English language pair and the added difficulty of domain adaptation, for future work, we will explore extending language resources in the generic domain to further improve translation quality in this language pair.

We have compiled the *Myth Corpus*[9] with various Myanmar-English datasets that researchers can use to improve Myanmar-English models. The datasets created ranges from manually cleaned dictionaries to synthetically translated data using

[8] https://marian-nmt.github.io

[9] https://github.com/alvations/myth

commercial translation API and unsupervised machine translation algorithms.

6 Conclusion

In this paper we presented our submissions to the WAT 2019 translation shared task. We trained similar Transformer-based NMT systems across different tasks. We found that shallower Transformers with a small number of heads perform better on smaller data sets. We also found a trade-off between simplifying data processing pipeline and model performance. Finally, we attempted simple techniques for training a multilingual NMT system and we will continue our investigation along this direction in future work.

References

Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2019. Towards Burmese (Myanmar) morphological analysis: Syllable-based tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):5.

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2018. NOVA: A feasible and flexible annotation system for joint tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):17.

Kilian Evang, Valerio Basile, Grzegorz Chrupała, and Johan Bos. 2013. Elephant: Sequence labeling for word and sentence segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1426, Seattle, Washington, USA. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Toshiaki Nakazawa, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Nobushige Doi, Yusuke Oda, Anoop Kunchukuttan, Shantipriya Parida, Ondej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Hideya Mino, Isao Goto, Hideto Kazawa, Yusuke Oda, Graham Neubig, and Sadao Kurohashi. 2017. Overview of the 4th workshop on Asian translation. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 1–54, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2015. Overview of the 2nd workshop on Asian translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, Kyoto, Japan.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2016. Overview of the 3rd workshop on Asian translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, Kyoto, Japan.

Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi. 2018. Overview of the 5th workshop on Asian translation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

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.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Belgium, Brussels. Association for Computational Linguistics.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Raymond Hendy Susanto, Hai Leong Chieu, and Wei Lu. 2016. Learning to capitalize with character-level recurrent neural networks: An empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2090–2095, Austin, Texas. 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.