Osaka University MT Systems for WAT 2018: Rewarding, Preordering, and Domain Adaptation

Yuki Kawara[†] **Yuto Takebayashi**[†] **Chenhui Chu**[‡] **Yuki Arase**[†]
[†]Graduate School of Information Science and Technology, Osaka University
[‡]Institute for Datability Science, Osaka University
{kawara.yuki,takebayashi.yuto,arase}@ist.osaka-u.ac.jp
chu@ids.osaka-u.ac.jp

Abstract

In this paper, we present Osaka University MT systems submitted to WAT 2018 shared translation tasks and analysis of their performances. For the ASPEC Japanese-English task, we use our rewarding model on neural machine translation (NMT) and preordering model on phrase-based statistical machine translation (PBSMT). For the Myanmar-English task, we further apply our mixed fine tuning method for domain adaptation on NMT. We report the translation results on these two tasks, where the rewarding model performs the best.

1 Introduction

This paper describes our systems submitted to WAT 2018 shared translation task (Nakazawa et al., 2018) and analyzes these systems. This year, Osaka University participated in two tasks: the ASPEC Japanese-English and Myanmar-English tasks. We use three different methods that we have been proposed in the past.

For the first system, we use the rewarding model boosting target words in the decoder of NMT (Takebayashi et al., 2018). It predicts target words that are promising to be used in a correct translation and rewards them to give them better chances to be output. For the second system, we preorder source sentences before translation so that the word order becomes similar to target sentences, which is applied to PBSMT (Kawara et al., 2018). For the third system, we use our mixed fine tuning method (Chu et al., 2017). It is a domain adaptation method that uses out-of-domain data to leverage for in-domain translation. The rewarding and preordering models are applied to both the ASPEC Japanese-English and Myanmar-English tasks, while mixed fine tuning is only applied to the Myanmar-English task because it is designed for low-resource translation.

We first describe statistics of datasets provided in the translation tasks in Section 2. Then, we present the details of the rewarding model, preordering model, and mixed fine tuning, as well as our internal evaluation results in Sections 3, 4, and 5, respectively. Finally, we analyze the official results of the shared tasks in Section 6 and conclude this paper in Section 7.

2 Datasets

We conduct English-to-Japanese, Japanese-to-English, English-to-Myanmar, and Myanmar-to-English translation, referred to as *En-Ja*, *Ja-En*, *En-My*, and *My-En* for short, hereafter.

Table 1 shows statistics of the datasets provided in the ASPEC (Asian Scientific Paper Excerpt Corpus) (Nakazawa et al., 2016) Japanese-English and Myanmar-English tasks. The ASPEC Japanese-English task is of a scientific domain, providing 3M, 1,790, and 1,812 sentences for training, development, and test, respectively. The Myanmar-English task provides two corpora, namely, the ALT (Asian Language Treebank) and UCSY (NLP Lab, University of Computer Studies, Yangon). The ALT corpus extracted from the Wikinews, providing 18k, 993, and 1,007 sentences for training, development, and test, respectively. The UCSY is a mixed domain corpus, which is supplementary for this task and pro-

32nd Pacific Asia Conference on Language, Information and Computation
The 5th Workshop on Asian Translation
Hong Kong, 1-3 December 2018

Corpus name	ASPEC	ALT	UCSY
Language	En-Ja/Ja-En	En-My/My-En	
Train	3,008,500	17,965	208,638
Dev	1,790	993	N/A
Test	1,812	1,007	N/A

Table 1: Data statistics of the WAT 2018 ASPEC and Myanmar-English tasks.

vides 208k sentences for training only.

3 Rewarding Model

3.1 Model

We employed the rewarding model using bilingual dictionaries (Takebayashi et al., 2018) to address the adequacy problem in NMT. Our model *rewards* target words that are promising to be used in correct translations by boosting their probabilities to be output by a decoder as shown in Figure 1.

Specifically, it first predicts a set of target words D_{f2e} that are promising to be used in translations by looking up bilingual dictionaries. Then, it *rewards* a target word y_j if it is contained in D_{f2e} by adding weight to the logarithm of the posterior probability $p(\cdot)$ of the decoder given a source sentence X:

$$Q(y_j|y_{<j}, X) = \log p(y_j|y_{<j}, X) + \lambda r_{y_j}, \quad (1)$$

where λ is the weight of reward that will be tuned using a development set. This means that our model boosts the probabilities of predicted words that might have been slipped away during beam search in the conventional decoder. We use a simple binary rewarding that performed the best in Takebayashi et al. (2018):

$$r_{y_j} = \begin{cases} 1 & (y_j \in D_{f2e}), \\ 0 & (\text{otherwise}). \end{cases} \quad (2)$$

Finally, a target word is output as:

$$y_j = \arg\max_{y_j} Q(y_j|y_{<j}, X).$$

3.2 Experiments

For the ASPEC Japanese-English task, we used the first 2M parallel sentence pairs among the entire 3M pairs sentences for training following Morishita et al. (2017), because the remaining 1M sentences were noisy. As preprocessing, we segmented Japanese sentences using MeCab,[1] and tokenized and truecased the English sentences with the *truecase.perl* script in Moses[2]. We further split the words into sub-words using joint BPE (Sennrich et al., 2016) with 32,000 merge operations. The vocabulary sizes of the Japanese and English side were 28,852 and 22,340, respectively. For the Myanmar-English task, we simply concatenated the available ALT and UCSY corpora for training. We tokenized and truecased the English corpus, and used the tokenized and romanized Myanmar corpus released by the organizers.

We used the mlpnlp-nmt system[3] that is an LSTM based encoder-decoder NMT model with attention, which achieved the best translation performance in human evaluations for both the Ja-En, and En-Ja tasks at WAT 2017 (Nakazawa et al., 2017). We implemented our rewarding model on top of the mlpnlp-nmt system. We followed the hyperparameter settings of Morishita et al. (2017). We used 2-layer LSTMs for both the encoder and decoder with the beam size of 5. Stochastic gradient descent was used as the learning algorithm, with an initial learning rate of 1.0. The mini batch size was 128.

For the rewarding model, accurate prediction of D_{f2e} is crucial. We used the GIZA++ toolkit[4] on the training corpus to automatically create a bilingual dictionary. We applied the "grow-diag-final-and" heuristic and obtained lexical translation probabilities using Moses.[5] We then pruned translation pairs with low probabilities by δ. λ in Equation (1) was tuned on the development sets from 0.1 to 1.0 by 0.1 interval. The threshold δ was tuned on 0, 0.0001, 0.001, 0.01 and 0.1 for the Japanese-English task and 0 and 0.001 for the Myanmar-English task. We selected the best combination among all combinations of δ and λ on the development set for each model.

Table 2 shows the BLEU scores of each task on the test sets. We can see that the rewarding model improves the BLEU score for 0.57 points and 1.07

[1] https://github.com/taku910/mecab

[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/recaser/truecase.perl

[3] https://github.com/mlpnlp/mlpnlp-nmt/

[4] http://code.google.com/p/giza-pp

[5] http://www.statmt.org/moses/

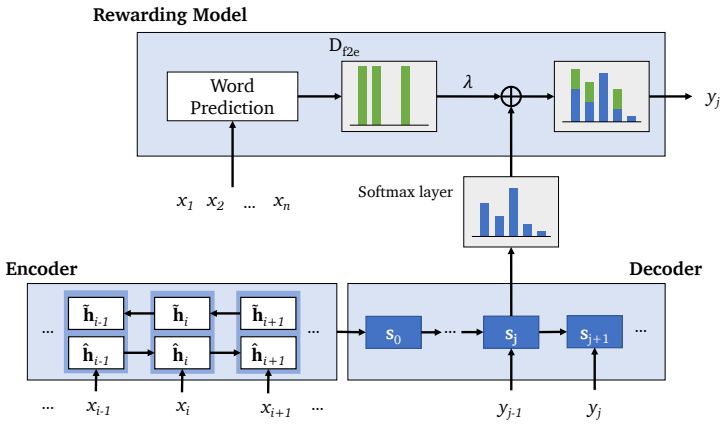

Figure 1: Rewarding model at decoding step j: predicted target words D_{f2e} are rewarded to have better chances to be output at each decoding time step. Note that the attention model is omitted for clarity.

	En-Ja	Ja-En	En-My	My-En
Mlpnlp-nmt	39.50	27.21	**22.34**	13.67
Rewarding	**40.07**	**28.28**	22.33	**13.79**

Table 2: Mlpnlp system and rewarding results (BLEU-4) on the WAT 2018 ASPEC and Myanmar-English tasks.

	En-Ja		Ja-En	
	pre.	rec.	pre.	rec.
Mlpnlp-nmt	73.90	69.03	66.42	61.66
Rewarding	72.87	70.22	66.06	63.09

Table 3: The precision and recall of unigram calculated by comparing the translation hypotheses against the reference translations on WAT 2018 ASPEC task.

points in the En-Ja and Ja-En tasks, respectively. However, there are no significant differences between the mlpnlp-nmt and the proposed model in the En-My and My-En tasks. We think the reason for this is that word alignments between English and Myanmar are not reliable because the size of the corpus is too small, which significantly degrades the word prediction quality. Hence, the rewarding model could not reward correct words. Table 3, 4 show that the precision and recall of unigram calculated by comparing the translation hypotheses against the reference translations on WAT 2018 ASPEC Japanese-English and Myanmar-English tasks, respectively. We can see that the recall increase 1.19 and 1.43 in exchange of decreasing the precision on En-Ja and Ja-En tasks, respectively. However, there are no significant differences between the mlpnlp-nmt and the proposed model in the En-My and My-En tasks.

	En-My		My-En	
	pre.	rec.	pre.	rec.
Mlpnlp-nmt	67.65	48.35	56.44	46.31
Rewarding	67.61	48.34	56.23	46.29

Table 4: The precision and recall of unigram calculated by comparing the translation hypotheses against the reference translations on WAT 2018 Myanmar-English task.

4 Preordering Model

4.1 Model

The word order between source and target languages significantly influences the translation quality in MT. Preordering, arranging words of source sentences so that the order is similar to that of the target language before translation, can effectively address this problem and significantly improves BLEU score of PBSMT (Nakagawa, 2015). Although NMT has been shown its strong performance in translation, it requires a large amount of training corpus, which is not the case for the Myanmar-English task. Hence, we use our preordering model with PBSMT for WAT submission.

We applied the preordering model based on recursive neural networks (RvNN) (Kawara et al., 2018) to En-Ja and Ja-En translation for ASPEC and En-My translation for the Myanmar-English tasks.[6] We first parse source sentences to obtain their syntax trees with a parser, then assign either *Inverted (I)* or *Straight (S)* labels at each node of the source syn-

[6]We could not conduct experiments on the My-En translation because parsers are unavailable for Myanmar.

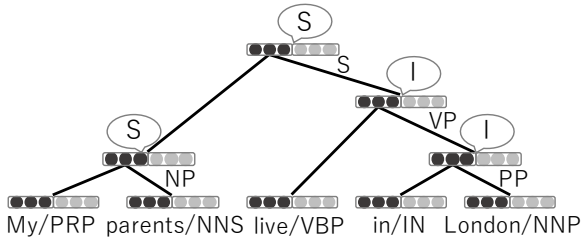

Figure 2: Preordering an English sentence "My parents live in London" with RvNN for Japanese. (*I* indicates to reorder the child nodes, and *S* indicates not to reorder the child nodes.)

tax tree to indicate whether the node should be reordered or not. Gold labels are automatically determined to achieve the highest Kendall's τ computed based on word alignment links. RvNN predicts labels at the node in test time and reorders source sentences. We then train a PBSMT system with reordered source sentences.

Figure 2 shows an example of the labeled parse tree of the English sentence "My parents live in London." RvNN learns to predict correct labels for nodes of a source syntax tree. For example, at the VP node of "live in London," its child nodes of "live" and "in London" are inverted to have the same word order with the Japanese counterpart.

4.2 Experiments

We used Stanford CoreNLP[7] for tokenization and POS tagging, Enju[8] for parsing of English, and MeCab[9] for tokenization and Ckylark for parsing[10] of Japanese. Myanmar corpus was tokenized and romanized by organizers. For the En-My translation, we concatenated the ALT and UCSY corpora for training. For word alignment, we used MGIZA.[11] Source-to-target and target-to-source word alignments were calculated using IBM model 1 and hidden Markov model, and they were combined with the intersection heuristic following Nakagawa (2015). We used 100k sentences sampled from training corpus for preordering. The embedding size and hidden size were set to 200. The vocabulary size was set to 50k. We used Adam (Kingma and Ba, 2015) with a weight decay (10^{-4}) and gradient clipping (5) for optimization. The mini batch size was set to 500.

For PBSMT, we used Moses.[12] We trained the 5-gram language model on the target side of the training corpus with KenLM.[13] Tuning was performed by minimum error rate training (Och, 2003). We repeated tuning and testing of each model 3 times and reported the average of scores. The distortion limit of PBSMT system trained by preordered sentences was set to 0, while that without preordering was set to 20.

	En-Ja	Ja-En	En-My
Moses PBSMT	24.54	15.31	19.71
Preordering	**29.16**	**17.30**	**20.93**

Table 5: PBSMT results (BLEU-4) with and without preordering on the WAT 2018 ASPEC and Myanmar-English tasks.

Table 5 shows the results. We can see that the preordering model improves the results on the PBSMT (4.62, 1.99, 1.22 for En-Ja, Ja-En, En-My, respectively). Translation quality of the En-My task is improved less than the En-Ja task (1.22 point and 4.62 point, respectively). We think this is caused by unbalanced corpus sizes of ALT and UCSY. The ALT corpus, from which the test set was derived, is significantly smaller than the UCSY corpus. This makes the English-Myanmar translation task difficult.

5 Domain Adaptation

5.1 Method

It has been known that vanilla NMT performs poorly for domain specific translation in low-resource scenarios (Chu and Wang, 2018). The WAT 2018 Myanmar-English task is a low-resource setting that only contains 18k in-domain training sentences for the ALT task. However, it also provides the UCSY out-of-domain corpus, containing 208k training sentences. This is a proper domain adaptation setting, where out-of-domain data can be leveraged for in-domain translation.

[7] http://stanfordnlp.github.io/CoreNLP/
[8] http://www.nactem.ac.uk/enju/
[9] http://taku910.github.io/mecab/
[10] http://odaemon.com/?page=tools_ckylark
[11] http://github.com/moses-smt/giza-pp
[12] https://github.com/moses-smt/mosesdecoder
[13] http://github.com/kpu/kenlm

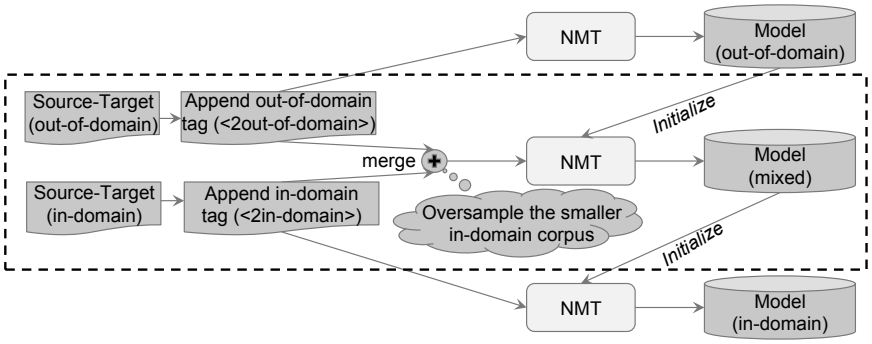

Figure 3: Mixed fine tuning with domain tags for domain adaptation.

In this work, we applied the domain adaptation method of mixed fine tuning (see Figure 3) for the WAT 2018 Myanmar-English task. Mixed fine tuning is a transfer learning based approach proposed by Chu et al. (2017). We first train an NMT model on the resource-rich out-of-domain (*i.e.*, UCSY) corpus till convergence. Then we resume training on the in-domain (*i.e.*, ALT) and out-of-domain (*i.e.*, UCSY) mixed corpus, which simply concatenates the corpora of two domains by appending artificial tokens that indicate the domains and by oversampling the corpus of the resource-poor domain (*i.e.*, ALT). This prevents over-fitting and enables smooth domain transition.

5.2 Experiments

For English, we tokenized and true-cased the sentences using the *tokenizer.perl* and *truecase.perl* scripts in Moses. For Myanmar, we used the transcribed and tokenized data released by the organizers. For the NMT system, we used the open source implementation of the Transformer model (Vaswani et al., 2017) in *tensor2tensor*[14]. We used the Transformer because it is the current state-of-the-art NMT model. For training, we used the default model settings corresponding to *transformer_base_single_gpu* in the implementation and to *base model* in (Vaswani et al., 2017). We compared the MT performance with vanilla NMT, which was trained on the in-domain data only using the Transformer. We trained 100k steps for the vanilla NMT system. For mixed fine tuning, we trained the out-of-domain and fine tuning models for 200k and 200k steps, respectively. As development and test data were not provided for the UCSY corpus, we randomly sampled $1,043$ and $1,043$ sentences from the corpus for development and test, respectively. Note that we removed the development and test sentences from the UCSY corpus for training.

	En-My	My-En
Transformer	**12.28**	0.45
Mixed fine tuning	9.45	**11.63**

Table 6: Transformer and mixed fine tuning results (BLEU-4) on the WAT 2018 Myanmar-English task.

Table 6 shows the results. We can see that mixed fine tuning significantly improves the results on the My-En direction (10.18 BLEU points higher) but performs worse than vanilla NMT on the En-My direction (2.83 BLEU points lower). We think the reason for this is that Myanmar sentences were tokenized into writing units and romanized and thus has a very small vocabulary. This makes the output word embeddings good enough for the En-My translation direction when training on the in-domain data only. Mixed fine tuning on the mixed data decreases the quality of word output word embeddings due to the mix of domains, leading to the drop in BLEU scores.

6 Official Results on WAT 2018

Table 7 shows the official results of our systems, organizer's systems, and the best systems on the WAT 2018 ASPEC and Myanmar-English tasks.[15] Translation qualities were evaluated with both automatic evaluation metrics (BLEU, RIBES, and AMFM) and human annotations. BLEU is calculated based on the proportion of matched n-gram between output

[14]https://github.com/tensorflow/tensor2tensor

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

		En-Ja	Ja-En	En-My	My-En
Rewarding	BLEU	**38.01**	**26.19**	**22.33**	**11.38**
	RIBES	**82.51**	**74.98**	**66.86**	**65.56**
	AMFM	**76.31**	**58.83**	74.08	51.09
	human	**4.50**	**−37.00**	**3.00**	**−57.00**
Preordering	BLEU	23.24	13.97	20.88	-
	RIBES	71.69	66.54	63.95	-
	AMFM	70.51	57.14	**77.48**	-
	human	−82.25	−95.75	−23.50	-
Mixed fine tuning	BLEU	-	-	9.45	9.99
	RIBES	-	-	58.19	64.89
	AMFM	-	-	66.54	**55.20**
	human	-	-	-	−99.50
Organizer baseline (PBSMT)	BLEU	27.48	18.45	-	-
	RIBES	68.37	64.51	-	-
	AMFM	73.64	59.10	-	-
Organizer baseline (NMT with attention)	BLEU	36.37	26.91	22.42	14.44
	RIBES	82.50	76.50	66.74	69.69
	AMFM	75.99	59.54	74.56	52.60
Organizer baseline (Transformer)	BLEU	40.79	28.06	-	-
	RIBES	84.49	76.76	-	-
	AMFM	76.86	59.56	-	-
Best system	BLEU	*43.43*♣	*30.59*♢	*32.30*♠	*29.14*♠
	RIBES	*85.03*♢	*77.79*♢	*74.65*♠	*79.40*♠
	AMFM	*78.10*♢	*61.94*♢	*81.65*♠	*65.59*♠
	human	*28.50*♡	*15.75*♭	*61.00*♮	*22.25*♯

Table 7: Official results of the WAT 2018 ASPEC and Myanmar-English tasks. Best systems are from different teams as indicated by the following symbols. ♣: Transformer with relative position, ensemble of 4 models, rerank, ♢: Transformer with relative position, ensemble of 3 models, ♠: many PBSMT and NMT n-best lists combined and reranked using Wikipedia data for back-translation and language model trainings, ♡: big bidirectional Transformer using 1.5M sentences only, ♭: Transformer vanilla model using 3M sentences, ♮: 4 models ensemble, ♯: NMT baseline, ensemble (system descriptions are borrowed from Nakazawa et al. (2018)).

and reference sentences. RIBES is calculated based on uni-gram precision and similarity the word order between system output and reference sentence. AMFM is calculated based on both adequacy and fluency, which is designed to decouple semantic and syntactic components of the translation process to provide a balanced view of translation quality. Because human evaluation was restricted to 2 systems of each team, we report human evaluation results of ASPEC En-Ja and Ja-En tasks for the rewarding and preordering systems, and a result of the My-En task for the rewarding and mixed fine tuning systems, and En-My tasks for all systems.

We can see that in terms of BLEU score and human evaluation, the rewarding model performed best among our three systems for all languages. In terms of AMFM, preordering and mixed fine tuning achieved 3.4 and 4.11 points higher scores than the rewarding model in the En-My and My-En tasks, respectively.

Our rewarding model outperformed the organizer's baseline of NMT with attention for 1.64 BLEU points on En-Ja task. However, the organizer's baseline of the Transformer achieved 2.78 and 1.87 higher BLEU points than the rewarding model on En-Ja and Ja-En, respectively. Because the rewarding model can be easily applied to different NMT decoders, we will apply it to the Trans-

former for further improvement. For En-My and My-En translations, our rewarding model is comparable to the organizer's baseline (En-My) and 3.06 BLEU points lower (My-En) due to the poor word alignment quality as discussed in Section 3.2.

There are significant gaps between our results and those of the best systems. These best systems ensemble multiple systems, while all of our results are from a single system. Ensembling multiple systems would improve our results, which is the future work.

7 Conclusion

We have described our systems submitted to WAT 2018 shared translation tasks. Among which, the rewarding model showed the best performance. As future work, we first plan to conduct system combination of these three systems. Secondly, we will apply the rewarding model to the decoder of the Transformer in order to further improve its translation quality.

Acknowledgments

This work was supported by NTT communication science laboratories and Grant-in-Aid for Research Activity Start-up #17H06822, JSPS.

References

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1304–1319, Santa Fe, USA, August.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 385–391, Vancouver, Canada, July.

Yuki Kawara, Chenhui Chu, and Yuki Arase. 2018. Recursive neural network based preordering for english-to-japanese machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Student Research Workshop*, pages 21–27, Melbourne, Australia, July.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference for Learning Representations (ICLR)*, San Diego, USA, December.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2017. NTT neural machine translation systems at WAT 2017. In *Proceedings of the Workshop on Asian Translation (WAT)*, pages 89–94, Taipei, Taiwan, November.

Tetsuji Nakagawa. 2015. Efficient top-down BTG parsing for machine translation preordering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 208–218, Beijing, China, July.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 2204–2208, Portoro, Slovenia, May.

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 Workshop on Asian Translation (WAT)*, pages 1–54, Taipei, Taiwan, November.

Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Win Pa Pa, Isao Goto, Hideya Mino, Katsuhito Sudoh, and Sadao Kurohashi. 2018. Overview of the 5th workshop on asian translation. In *Proceedings of the 5th Workshop on Asian Translation (WAT)*, Hong Kong, China, December.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1715–1725, Berlin, Germany, August.

Yuto Takebayashi, Chenhui Chu, Arase Yuki, and Masaaki Nagata. 2018. Word rewarding for adequate neural machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 14–22, Bruges, Belgium, October.

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 *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008. Long Beach, USA, December.