Ensemble and Reranking: Using Multiple Models in the NICT-2 Neural Machine Translation System at WAT2017

Kenji Imamura and **Eiichiro Sumita**
National Institute of Information and Communications Technology,
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{kenji.imamura, eiichiro.sumita}@nict.go.jp

Abstract

In this paper, we describe the NICT-2 neural machine translation system evaluated at WAT2017. This system uses multiple models as an ensemble and combines models with opposite decoding directions by reranking (called bi-directional reranking).

In our experimental results on small data sets, the translation quality improved when the number of models was increased to 32 in total and did not saturate. In the experiments on large data sets, improvements of 1.59–3.32 BLEU points were achieved when six-model ensembles were combined by the bi-directional reranking.

1 Introduction

This paper presents the NICT-2 machine translation system evaluated at WAT2017 (Nakazawa et al., 2017). This system is a basic encoder-decoder with an attention mechanism (Sutskever et al., 2014; Bahdanau et al., 2015). This methodology is known to achieve high translation quality, even when using a single model. It is also known that better quality can be achieved by utilizing multiple models. In this paper, we use as many models as possible and attempt to improve the translation quality.

There are two major approaches that use multiple models: ensemble (Hansen and Salamon, 1990) and reranking (e.g., (Och et al., 2004)). The ensemble approach independently encodes and decodes input sentences by multiple models and averages the word distributions output from the decoder (c.f., Sec. 2.1). The reranking approach first creates an n-best list of translations using a model A, rescores it using another model B, and selects the highest scoring translation (c.f., Sec. 2.2).

	Pros	Cons
Ensemble		
	• All hypotheses in the search space are candidates for translation.	• Models that have different output layers in the decoders cannot be incorporated (from the viewpoints of vocabulary and decoding direction).
	• It is possible to speed up the computations by parallel processing.	• All models should be loaded on graphics processing units (GPUs) at the same time.
Reranking		
	• Arbitrary models can be combined if the language pairs are the same.	• The system cannot select candidates that are not in the n-best list.
	• The models for the generation and rescoring of the n-best candidates have to be loaded separately on GPUs.	• The n-best generation and rescoring processes are sequential.

Table 1: Pros and Cons of Ensemble and Reranking

Both methods have pros and cons, as shown in Table 1. The aim of this paper is to use as many models as possible, based on these characteristics.

In this paper, we first obtain the following information on small data sets and then apply the ensemble and reranking methods on large data sets.

- How many models contribute to the translation quality?
- If both methods use the same number of models, which method is better? In this paper, we only evaluate the translation quality and ignore the translation speed.

The rest of this paper is organized as follows. Sec. 2 describes in detail the ensemble and reranking methods and their combination used at WAT2017. Sec. 3 evaluates characteristics of the

Proceedings of the 4th Workshop on Asian Translation, pages 127–134,
Taipei, Taiwan, November 27, 2017. ©2017 AFNLP

ensemble and reranking methods using small data sets (the JIJI Corpus and the MED Corpus, which was developed in-house). In Sec. 4, we evaluate the NICT-2 system using ASPEC (Asian Scientific Paper Excepts Corpus; (Nakazawa et al., 2016b)) data sets, and the paper is concluded in Sec. 5.

Note that we only evaluate Japanese-English (Ja-En) and Japanese-Chinese (Ja-Zh) pairs. Thus, additional investigation of whether the conclusions are valid for other language pairs is necessary. However, we believe that the results in this paper are valuable as a case study.

2 Ensemble and Reranking

2.1 Ensemble

The ensemble approach is a method for neural networks that trains multiple models using the same data sets and applies them to test data while averaging the outputs (Hansen and Salamon, 1990). In the case of neural machine translation, an input sentence is encoded and decoded using multiple models. Then, the word distributions output from the decoder (i.e., vectors of the target vocabulary size) are averaged. A beam search is applied to this averaged distribution. Note that each model is independently trained in the same way as the training of a single model.

If we represent the output word selection for a single model by Eq. (1), the selection for an ensemble is represented by Eq. (2). In this case, we use the geometric mean.

$$\hat{y}_t = \operatorname{argmax} \log Pr(y_t | y_1^{t-1}, \mathbf{x}; M) \quad (1)$$

$$\hat{y}_t = \operatorname{argmax} \frac{1}{J} \sum_{j=1}^{J} \log Pr(y_t | y_1^{t-1}, \mathbf{x}; M_j) \quad (2)$$

where y_t denotes the tth output word, y_1^{t-1} denotes the history of the output words from the beginning of the sentence to the $(t-1)$th position, $\mathbf{x}$ denotes the input word sequence, M denotes the model (M_j denotes the jth model), and J denotes the number of ensemble models.

The ensemble approach has some restrictions. Firstly, it has to use identical target vocabularies for all models because it averages the output vectors. Secondly, the decoding direction (from the beginning to the end of a sentence or from the end to the beginning) has to be consistent over all models because the beam search is applied after averaging. In this paper, we call the directions from the beginning to the end and from the end to the beginning the left-to-right and right-to-left directions, respectively.

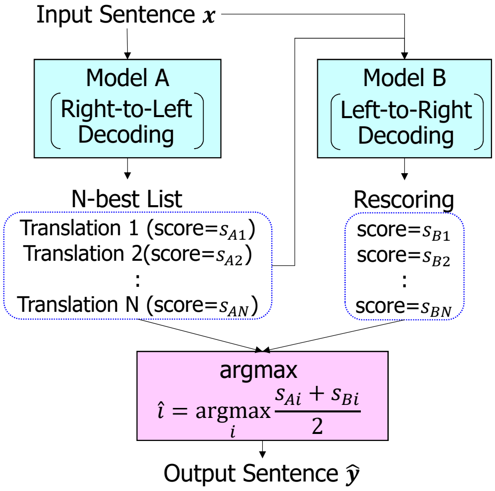

Input Sentence $\boldsymbol{x}$
Model A
Right-to-Left Decoding
Model B
Left-to-Right Decoding
N-best List
Translation 1 (score=s_{A1})
Translation 2(score=s_{A2})
:
Translation N (score=s_{AN})
Rescoring
score=s_{B1}
score=s_{B2}
:
score=s_{BN}
argmax
$\hat{\imath} = \operatorname{argmax}_i \frac{s_{Ai} + s_{Bi}}{2}$
Output Sentence $\hat{\boldsymbol{y}}$

Figure 1: Structure of Bi-directional Reranking

2.2 Reranking

The reranking method for machine translation (Och et al., 2004) comprises two steps. Firstly, an input sentence is translated using a model A, and an n-best list is generated. Then, the translations in the n-best list are rescored using another model B. Finally, the translation that has the highest score is selected/output (Figure 1). The models A and B are independently trained as single models.

The final translations are influenced by the rescoring method. In this paper, we use the arithmetic mean of the log-likelihoods of the models A and B.

The reranking method has the advantage that arbitrary models can be used if the target languages are the same. In addition, the reranking method consumes less memory than the ensemble method because only one model is used at each step in the reranking method, even though it uses two models in total. However, it has the disadvantage that the translation quality cannot be improved if good translation hypotheses are not included in the n-best list.

2.3 Combination of Ensemble and Reranking

The pros and cons of the ensemble and reranking methods are shown in Table 1. To combine both methodologies while retaining as many advantages as possible, we employ reranking as the general methodology. The ensemble method is used for the n-best list generation and rescoring

Corpus	Language Pair	#Sentences	#Sub-word Types	Remarks
ASPEC	Ja ↔ En	Train: 2,977,320 Dev.: 1,790 Test.: 1,812	Ja: 49,656 En: 49,776	Scientific paper excerpts Sentences in which the number of sub-words is equal to or less than 80
	Ja ↔ Zh	Train: 656,635 Dev.: 2,090 Test.: 2,107	Ja: 49,654 Zh: 49,385	Scientific paper excerpts Sentences in which the number of sub-words is equal to or less than 80
JIJI	Ja → En	Train: 199,905 Dev.: 2,000 DevTest.: 2,000 Test.: 2,000	Ja: 35,009 En: 33,934	Newswire Sentences in which the number of sub-words is equal to or less than 80
MED	Ja → En	Train: 238,214 Dev.: 1,000 Test.: 1,000	Ja: 20,327 En: 21,043	Pseudo-dialogues at hospitals Sentences in which the number of sub-words is equal to or less than 80

Table 2: Corpus Statistics

to combine multiple models. We can combine many models using this architecture because the reranking method can combine twice the number of models without consuming extra memory.

We use an identical vocabulary set among all models so that the ensemble method can be applied. In addition, the models used here have the same structure for simplicity. The only difference is that each model is learned using a different random seed.

For the generation and rescoring of the n-best translations in the reranking, we use models with opposite decoding directions, which are impossible to combine with the ensemble method. In this paper, we call this bi-directional reranking. More precisely, the n-best list is generated by right-to-left decoding (i.e., from the end to the beginning of a sentence). Then, the hypotheses in the list are rescored by left-to-right decoding (i.e., from the beginning to the end of the sentence). Finally, the translation likelihoods for both directions are averaged, and the hypothesis with the highest likelihood is output.

The bi-directional reranking approach realizes Liu et al. (2016)'s method, which uses bi-directional decoding, by reranking. In the bi-directional reranking approach, the target word sequence is inverted during training and translation. Therefore, small changes are required in the training and translation programs.

3 Experiments Using Small Data Sets

We perform Japanese-English translation experiments using small data (with approximately 200k sentences) to clarify characteristics of the ensemble and the bi-directional reranking approaches.

3.1 Experimental Settings

Corpora: Table 2 shows the list of corpora that were used here. We used two corpora as small data sets. The first is the JIJI Corpus, which consists of newswires. Japanese and English articles were automatically aligned sentence by sentence. Note that the translations are sometimes not literal because the original articles were not translated sentence by sentence.

The second is the corpus of pseudo-dialogues at hospitals (MED Corpus). This corpus is a collection of conversations between patients and hospital staffs, which were created by writers (developed in-house). The pseudo-dialogues were first written in Japanese and then translated into English.

The byte-pair encoding (Sennrich et al., 2016) rules were acquired from a training set of each corpus, and they were applied to the training, development, and test sets. The number of sub-word types is 34–35k in the JIJI Corpus and 20–21k in the MED Corpus. We used sentences with 80 or fewer sub-words for training.

Preprocessing, Postprocessing: Table 3 shows a summary of our system. As shown in the table, we used the same preprocessing and postprocessing steps as the WAT baseline systems (Nakazawa et al., 2016a).

Translation System: We used OpenNMT (Klein et al., 2017)[1] as the base translation system. The encoder comprises a two-layer bi-directional LSTM (long short-term memory), in which the number of units is 500 each. The decoder comprises a two-layer LSTM (1000

[1] http://opennmt.net/

		Japanese	English	Chinese
Preprocessing	Character Normalization	NFKC Normalization of Unicode		
	Tokenizer	MeCab (Kudo et al., 2004)	Moses Toolkit	Stanford Segmenter (CTB)
	TrueCaser	–	Moses Toolkit	–
	Byte Pair Encoding	In-house Encoder		
Training and Translation	System	OpenNMT (modified for right-to-left decoding and the ensemble method)		
	Encoder	Word embedding: 500 units, two-layer Bi-LSTM (500 + 500 units)		
	Decoder	Word embedding: 500 units, two-layer LSTM (1,000 units)		
	Attention	Global Attention		
	Training	Mini Batch Size:64, SGD Optimization (10+6 epochs), Dropout:0.3		
	Translation	Beam Width:5 (c.f., Sec. 3.2)		
Postprocessing	DeTrueCaser	–	Moses Toolkit	–
	DeTokenizer	WAT Official's	Moses Toolkit	WAT Official's

Table 3: Summary of the NICT-2 NMT System

units). Global Attention (Luong et al., 2015) was utilized.

We used the stochastic gradient descent (SGD) method for the optimization. The learning rate was 1.0 for the first ten epochs, and then annealing was performed for six epochs while decreasing the learning rate by half.

To implement the methods described in Section 2.3, we modified OpenNMT as follows.

- We enabled the ensemble in the translator.
- We enabled right-to-left decoding in the trainer and translator.

The n-best size for the reranking was determined by the experiment in Section 3.2.

Evaluation: Of the WAT official evaluation metrics, we employ BLEU (Papineni et al., 2002) for the evaluation. WAT official scores are changed by word segmenters. In this paper, we use JUMAN (Kurohashi et al., 1994) for Japanese, Moses tokenizer (Koehn et al., 2007) for English, and Stanford Word Segmenter (Chinese Penn Treebank Model) (Chang et al., 2008) for Chinese evaluation.

3.2 Optimal Size of N-best List

To output n-best translations using the beam search, beam width is better to set equal or more than n. In our experiments, we set the beam width equal to the size of the n-best list.

Figure 2 shows the BLEU scores of various n-best sizes on the DevTest set of the JIJI Corpus. It contains the scores obtained by left-to-right and right-to-left decoding and bi-directional reranking. A single model is used here, i.e., an ensemble is not used in this experiment.

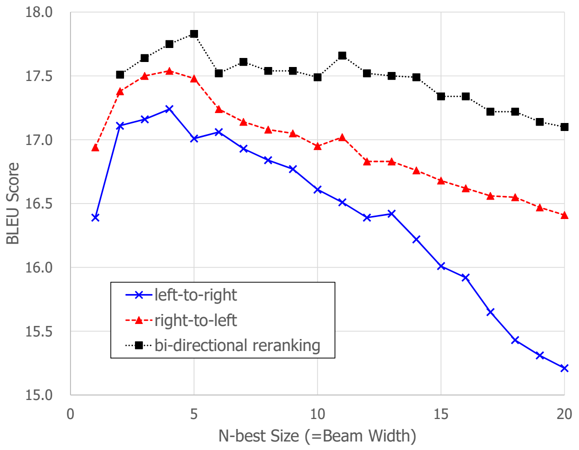
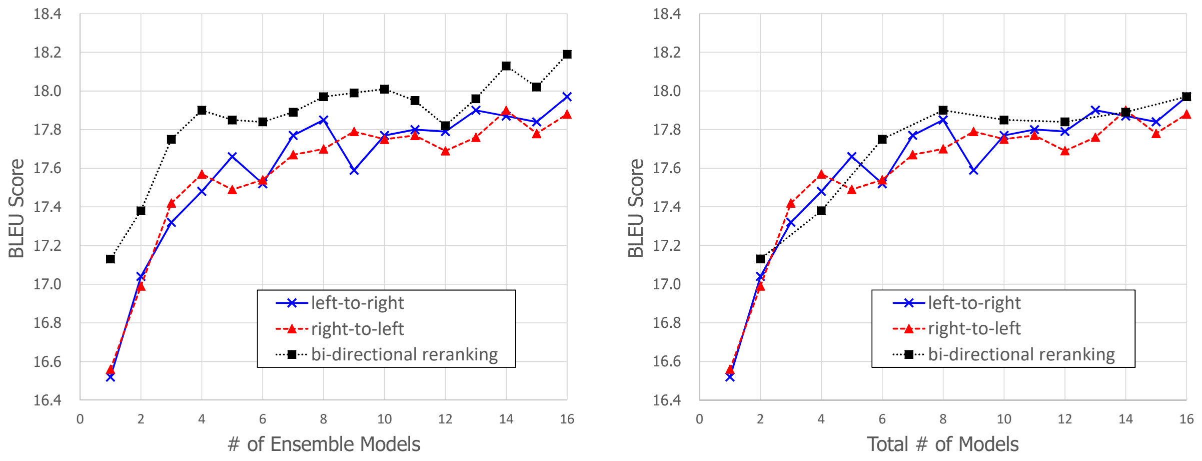


Figure 2: BLEU Scores According to N-best Size

In all methods, the BLEU scores changed according to the size of the n-best list. For left-to-right and right-to-left decoding, the BLEU scores were highest when the n-best size was 4, and the scores decreased when the n-best size increased above 4. After the bi-directional reranking, the BLEU score was the highest when the n-best size was 5, and slowly decreased when the size increased above 5.

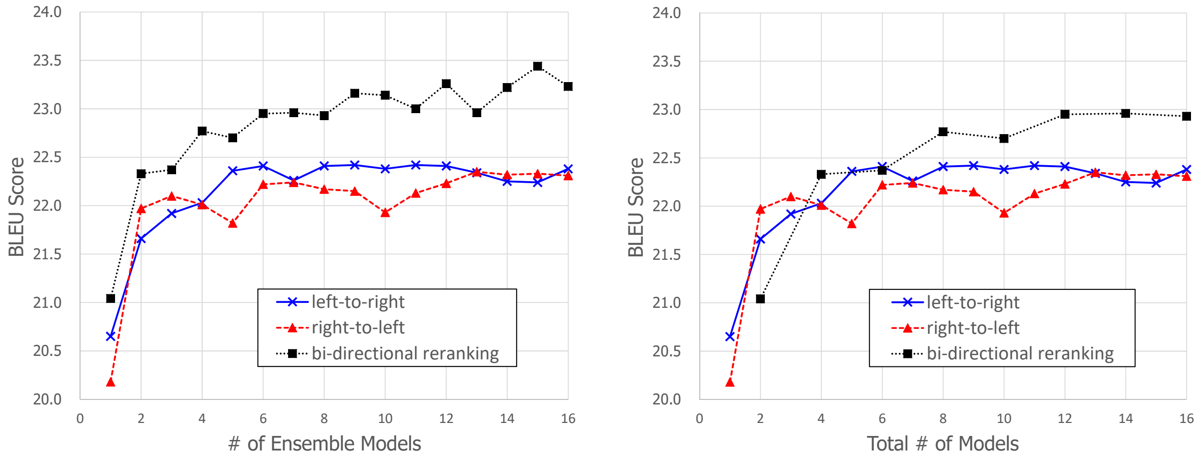
In general, large n-best size is expected in reranking to include good hypotheses. However, in our NMT system, the peak score was achieved with a small n-best size when a single model was used, and similarly, a small n-best size was the best in the reranking. This is because decreasing the accuracy of the single model had greater influence than improving the coverage of n-best sizes. Based on the above observation, we use 5 as the n-best size hereafter.



(a) Number of Ensemble Models vs. BLEU Scores

(b) Number of Total Models vs. BLEU Scores

Figure 3: Results of Multiple Model Combination on the JIJI Corpus
In the bi-directional reranking, the total number of models in (b) is equal to twice the number of ensemble models in (a).



(a) Number of Ensemble Models vs. BLEU Scores

(b) Number of Total Models vs. BLEU Scores

Figure 4: Results of Multiple Model Combination on the MED Corpus
In the bi-directional reranking, the total number of models in (b) is equal to twice the number of ensemble models in (a).

3.3 Effects of Multiple Models

Figures 3 and 4 show the results of the left-to-right and right-to-left decodings, which use the simple ensemble, and the bi-directional reranking on the JIJI and MED corpora, respectively. Note that the number of models used in the reranking is double of that used by the ensemble model. Therefore, we show two graphs: (a) a graph based on the number of ensemble models and (b) a graph based on the total number of models. We increased the number of models incrementally, i.e., models are added one at a time. Therefore, the settings for many models must be compatible with the settings of fewer models.

We firstly focus on the number of models and the translation quality. The BLEU scores tend to increase with the number of models for the all methods in the graphs. However, the rates of increase become slower as the number of models increases. On the JIJI Corpus, the BLEU scores are still increasing slightly with the 16-model ensemble. On the MED Corpus, the BLEU scores almost saturate with two- to six-model ensembles but do not saturate in the bi-directional reranking.

Zhou et al. (2002) indicated that the ensemble is more effective when models are selected, rather than using all models. However, in our experiment, degradation of the translation quality was not observed when all the models were used. The BLEU score improved by 1.67 points in the bi-directional reranking with 16 ensembles (32 models in total) on the JIJI Corpus compared with a single model in the left-to-right decoding. On the MED Corpus, the BLEU score improved by 2.58 points.

Secondly, we focus on the left-to-right and right-to-left decodings of the ensemble. On MED Corpus, the BLEU scores of the right-to-left decoding are higher than those of the left-to-right decoding. In contrast, the BLEU scores of the both decoding directions are almost the same on JIJI Corpus. We expected that the results would depend on the data sets and language pairs. However, these results show that the translation quality changed according to the decoding direction.

Thirdly, focusing on the graphs in (a), the scores of the reranking almost always surpass those of the simple ensembles (left-to-right and right-to-left decodings). From these results, we make the following observations.

- The model combination using the reranking favorably affects the translation quality.
- Bi-directional reranking can improve the translation quality from different aspects than the ensemble.

Since we combined models with opposite decoding directions, effects similar to those of bi-directional decoding (Liu et al., 2016) were realized.

The graphs in (b) show that the total number of models is double the number of ensemble models in the reranking. As shown in the graphs, the BLEU scores of the reranking almost always surpass those of the ensembles. In our experiments, bi-directional reranking was more effective than the ensembles if the number of models was the same.

4 Experiments Using Large Data Sets

In this section, we show the results of Ja-En, En-Ja, Ja-Zh, and Zh-Ja translation of the ASPEC task.

4.1 Experimental Settings

Corpora: The corpora used here are the ASPEC data sets listed in Table 2. From these training sets, we acquired the byte-pair encoding rules, which generate approximately 50k sub-word types per language, and used sentences in which the number of sub-words is equal to or less than 80.

Translation System: The other settings such as the translation system, preprocessing, and post-processing are the same as those in Section 3. Table 3 shows a summary of the settings.

4.2 Results

The results of the Ja-En and En-Ja translations are shown in Table 4, and those of the Ja-Zh and Zh-Ja translations are shown in Table 5.

We tested up to six ensembles due to resource limitations; however, the results have the same tendency as those of the small data sets. Namely, the BLEU scores increased with the number of models in both the cases, ensembles and reranking. The best BLEU scores were obtained in the bi-directional reranking with six-model ensembles in all language pairs, except En-Ja.

The improvements from the left-to-right single model to the bi-directional reranking with six-model ensemble were +1.97, +3.32, +1.59, and +2.58 points in the Ja-En, En-Ja, Ja-Zh, and Zh-Ja translations, respectively.

5 Conclusion

In this paper, we presented the NICT-2 neural machine translation system evaluated at WAT2017. The main characteristics of this system are that multiple models are used by the ensemble, and moreover, double models are used by the bi-directional reranking.

In the experiments on small data sets, we increased the number of models in the ensemble to 16. However, the translation quality did not saturate and can be further improved on some data sets.

We confirmed that the decoding direction influences the translation quality. In addition, the reranking can combine models with different properties from the ensemble. Using this feature, we combined models with opposite decoding directions in the reranking. By incorporating the ensemble and bi-directional reranking, we achieved higher translation quality than with the ensemble alone. In our experiments using ASPEC data

	Ja-En			En-Ja		
# of Ensemble Models	Ensemble (left-to-right)	Ensemble (right-to-left)	Reranking	Ensemble (left-to-right)	Ensemble (right-to-left)	Reranking
1	24.79	24.72	25.34	36.85	38.20	39.10
2	25.60	25.40	25.89	38.37	38.69	39.41
3	26.17	25.62	26.08	38.95	39.23	39.87
4	25.89	25.77	26.26	38.97	39.37	40.03
5	25.94	26.06	26.37	39.19	39.55	**40.23**
6	26.21	26.29	**26.76**	39.13	39.26	40.17

Table 4: WAT2017 Official Scores (Ja-En Pair of ASPEC).
Note: The Japanese scores are based on the JUMAN segmenter.

	Ja-Zh			Zh-Ja		
# of Ensemble Models	Ensemble (left-to-right)	Ensemble (right-to-left)	Reranking	Ensemble (left-to-right)	Ensemble (right-to-left)	Reranking
1	33.64	33.60	34.10	44.26	44.13	45.10
2	34.67	34.22	34.77	45.59	45.52	46.20
3	34.75	34.64	34.98	45.88	45.93	46.53
4	34.75	34.64	34.98	46.13	46.10	46.55
5	35.02	34.81	35.18	46.27	46.36	46.69
6	35.27	34.95	**35.23**	46.55	46.31	**46.84**

Table 5: WAT2017 Official Scores (Ja-Zh Pair of ASPEC).
Note: The Japanese and Chinese scores are based on the JUMAN and Stanford (CTB Model) segmenters, respectively.

sets, the BLEU scores improved from 1.59 to 3.32 points compared with the single model.

Both the ensemble and reranking can further improve the translation quality if the quality of a single model can be improved. Therefore, we will tackle the improvement of single models. At the time, we should evaluate the qualities of single and multiple models separately.

Currently, the ensemble approach might not be practical due to restrictions on the number and memory of GPUs. However, we assume that advances in hardware will decrease these restrictions.

Acknowledgments

This work was supported by "Promotion of Global Communications Plan — Research and Development and Social Demonstration of Multilingual Speech Translation Technology," a program of Ministry of Internal Affairs and Communications, Japan.

References

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of International Conference on Learning Representations (ICLR 2015)*.

Pi-Chuan Chang, Michel Galley, and Christopher D . Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Columbus, Ohio.

Lars Kai Hansen and Peter Salamon. 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada.

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.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of EMNLP 2004*, pages 230–237, Barcelona, Spain.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of

Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language*, pages 22–28.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Agreement on target-bidirectional neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 411–416, San Diego, California.

Thang Luong, Hieu Pham, and D. Christopher Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal.

Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Hideya Mino, Isao Goto, Graham Neubig, Hideto Kazawa, Yusuke Oda, Jun Harashima, and Sadao Kurohashi. 2017. Overview of the 4th Workshop on Asian Translation. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, Taipei, Taiwan.

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

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016b. ASEPC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth Edition of the Language Resources and Evaluation Conference (LREC-2016)*, Portoroz, Slovenia.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *HLT-NAACL 2004: Main Proceedings*, pages 161–168, Boston, Massachusetts, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pages 3104–3112.

Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. 2002. Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 137(1-2):239–263.