Investigating Phrase-Based and Neural-Based Machine Translation on Low-Resource Settings

Hai-Long Trieu, Duc-Vu Tran, Le-Minh Nguyen
Japan Advanced Institute of Science and Technology
{trieulh,vu.tran,nguyenml}@jaist.ac.jp

Abstract

Neural-based and phrase-based methods have shown the effectiveness and promising results in the development of current machine translation. The two methods are compared on some European languages, which show the advantages of the neural machine translation. Nevertheless, there are few work of comparing the two methods on low-resource languages, which there are only small bilingual corpora. The problem of unavailable large bilingual corpora causes a bottleneck for machine translation for such language pairs. In this paper, we present a comparison of the phrase-based and neural-based machine translation methods on several Asian language pairs: Japanese-English, Indonesian-Vietnamese, and English-Vietnamese. Additionally, we extracted a bilingual corpus from Wikipedia to enhance machine translation performance. Experimental results showed that when using the extracted corpus to enlarge the training data, neural machine translation models achieved the higher improvement and outperformed the phrase-based models. This work can be useful as a basis for further development of machine translation on the low-resource languages.

1 Introduction

Recent approaches have shown the promising results in the development of machine translation. During a long period from statistical models (Brown et al., 1990; Brown et al., 1993) to phrase-based models (Och et al., 1999; Koehn et al., 2003; Chiang, 2005) to recent neural-based methods (Sutskever et al., 2014; Cho et al., 2014), the phrase-based and neural-based become dominant methods in current machine translation. Statistical machine translation (SMT) systems achieve a high performance in many typologically diverse language pairs (Bojar et al., 2013). SMT can be applied to any pair of languages with minimal engineering effort (Bisazza and Federico, 2016). Meanwhile, neural machine translation (NMT) has obtained the state-of-the-art performance in machine translation for several languages including Czech-English, German-English, English-Romanian (Sennrich et al., 2016a). NMT has been proposed recently as a promising framework for machine translation, which learns sequence-to-sequence mapping based on two recurrent neural networks (Sutskever et al., 2014; Cho et al., 2014), called encoder-decoder networks. In a basic encoder-decoder network, the dimension of the context vector in the encoder is fixed, which leads to a low performance when translating for long sentences. In order to overcome the problem, (Bahdanau et al., 2015) proposed a method called attention mechanism, in which the model encodes the most relevant information in an input sentence rather than a whole input sentence into the fixed length context vector. NMT models with the attention mechanism have achieved significantly improvement in many language pairs (Jean et al., 2015; Gulcehre et al., 2015; Luong et al., 2015).

SMT and NMT models have shown successfully in language pairs in which large bilingual corpora are available such as English-German, English-French, Chinese-English, and English-Arabic. There are some work that evaluated the phrase-based versus neural-based methods such as

31st Pacific Asia Conference on Language, Information and Computation (PACLIC 31), pages 384–391
Cebu City, Philippines, November 16-18, 2017

the comparison of the two methods on English-German (Bentivogli et al., 2016), the comparison on 30 translation directions on the United Nations Parallel Corpus (Junczys-Dowmunt et al., 2016). Nevertheless, for low-resource settings like Asian language pairs which contain only small bilingual corpora, there are few work of the comparison of the two methods on such language pairs. Additionally, the problem of unavailable large bilingual corpora causes a bottleneck for machine translation on such languages.

In this work, we compared the SMT and NMT methods on several low-resource language pairs. The standard phrase-based SMT was used based on the work of (Koehn et al., 2007). The NMT model was used based on the state-of-the-art model (Sennrich et al., 2016a) in the WMT 2016,[1] which used encoder-decoder networks with attention mechanism and open-vocabulary translation. Experiments were conducted on Asian language pairs: Japanese-English, Indonesian-Vietnamese, and English-Vietnamese with only small bilingual corpora. Furthermore, in order to overcome the problem of unavailable large bilingual corpora, we extracted a bilingual corpus from Wikipedia to enhance machine translation on both SMT and NMT models. Moreover, we aim to evaluate the effects of enlarging training data to the two different machine translation methods and to the overall performance. Experimental results showed meaningful findings in the comparison of the two machine translation methods on the low-resource settings. This work can be useful as a basis for further development of NMT as well as machine translation in general on the low-resource languages. The scripts, corpora, and trained models used in this research can be found at the repository.[2]

2 Approaches

In this section, we discuss the two powerful approaches in machine translation currently: SMT and NMT. Additionally, we discuss one of the main factors that affects translation quality using both of the two machine translation approaches: bilingual corpora. For most language pairs in the world, large bilingual corpora are unavailable (Wang et al., 2016), which causes a bottleneck for machine translation on such language pairs. We extracted a parallel corpus from comparable data to enhance machine translation.

2.1 Phrase-based Machine Translation

In phrase-based SMT models (Koehn et al., 2003; Och and Ney, 2004), phrases are used as atomic units for translation. An input sentence is separated into phrases. Then, each phrase is translated to target phrases, which can be reordered to produce the translation output.

Given a source sentence s, the goal is to find the best translation t, which maximizes both the adequacy and fluency. Assume that the source sentence s can be segmented into a sequence of phrases $s_1^I = s_1 s_2 ... s_I$, which can be decoded into a sequence of target phrases $t_1^J = t_1 t_2 ... t_J$. The best translation $\hat{t}$ can be modeled as follows.

$$\hat{t}_1^J = argmax\ P(t_1^J | s_1^I) \tag{1}$$

The translation probability $P(t_1^J | s_1^I)$ can be computed using the Bayes theorem.

$$P(t_1^J | s_1^I) = \frac{P(s_1^I | t_1^J) P(t_1^J)}{P(s_1^I)} \tag{2}$$

Since the objective is to find the best translation $\hat{t}$, it can be computed based on the two components as follows.

$$\hat{t}_1^J = argmax\ P(s_1^I | t_1^J) P(t_1^J) \tag{3}$$

Where: the component $P(s_1^I | t_1^J)$ is called *translation model*; $P(t_1^J)$ is called *language model*.

2.2 Neural Machine Translation

For neural machine translation, one of the basis frameworks is the encoder-decoder (Cho et al., 2014; Sutskever et al., 2014). The basis framework can be improved by several components such as attention mechanism, open-vocabulary. We discuss the basis framework and the components in this section.

[1] http://www.statmt.org/wmt16/

[2] https://github.com/nguyenlab/MT-LowRes

NMT Models Given a source sentence $s = (s_1, ..., s_m)$, and a target sentence $t = (t_1, ..., t_n)$, the goal of a NMT is to model the conditional probability $p(t|s)$. This process bases on the encoder-decoder framework as proposed in (Cho et al., 2014; Sutskever et al., 2014).

$$logp(t|s) = \sum_{j=1}^{n} logp(t_i|\{t_1, ..., t_{i-1}\}, s, c) \quad (4)$$

in which, the source sentence s is represented by the context vector c using the encoder. For each time, a target word is translated based on the context vector using the decoder.

For the decoding, the probability of each target word t_i can be computed as follows.

$$p(t_i|\{t_1, ..., t_{i-1}\}, s, c) = softmax(h_i) \quad (5)$$

where h_i is the current target hidden state as in Equation 6.

$$h_i = f(h_{i-1}, t_{i-1}, c) \quad (6)$$

Finally, for the bilingual corpus B, the training objective is computed as in Equation 7.

$$I = \sum_{(s,t)\in B} -logp(t|s) \quad (7)$$

Attention Mechanism As shown in (Bahdanau et al., 2015), the translation performance decreases when translating long sentences. Instead of encoding entire the input sentence into the context vector, the most relevant information of the input sentence is encoded into the single, fixed-length vector. The representation c for the source sentences is set as follows.

$$c = [\bar{h_1}, ..., \bar{h_m}] \quad (8)$$

There are two stages in the function f in Equation 6: attention context and extended recurrent neural network (RNN). In the attention context, an alignment vector a_i is learned by comparing the previous hidden h_{i-1} with individual source hidden states in the context vector c; then the model derives a weighted average (c_i) of the source hidden states based on the alignment vector a_i. For the second stage, extended RNN, the RNN unit is expanded for the context vector c_i in addition to the previous hidden state h_{i-1} and the current input t_{i-1} to compute the next hidden state h_i.

Byte-pair Encoding In order to overcome the problem of out-of-vocabulary, (Sennrich et al., 2016b) proposed a method for open-vocabulary translation by encoding rare and unknown words as sequences of subword units. This is because various word classes can be translated by smaller units like compositional translation for compounds, phonological and morphological transformations for cognates and loanwords. In order to do that, words are segmented using byte-pair encoding that originally devised as a compression algorithm (Gage, 1994).

2.3 Bilingual Corpus: An Essential Resource in Machine Translation

Current Status Both of the two approaches: SMT and NMT require large bilingual corpora to train machine translation models. There are several large bilingual corpora which contain up to millions of parallel sentences such as European languages (Europarl corpus (Koehn, 2005), JRC-Acquis corpus (Steinberger et al., 2006)), English-French (the Canadian Hansard[3], the Giga-FrEn corpus[4]), and English-Chinese (the UM-Corpus (Tian et al., 2014)). Nevertheless, such large bilingual corpora are unavailable for most language pairs in the world (Irvine, 2013; Wang et al., 2016), which causes a bottleneck for both of the SMT and NMT machine translation methods. We extracted a bilingual corpus from comparable data in order to: i) investigate how the extracted bilingual corpus affects the two SMT and NMT approaches, and ii) enhance machine translation using SMT and NMT methods.

Extracting Bilingual Sentences from Wikipedia We extracted a bilingual corpus from Wikipedia, a large comparable data that contains a number of articles in the same domain in many languages. First, we extracted parallel titles of Wikipedia's articles based on the Wikipedia database dumps.[5] For a

[3]http://www.isi.edu/naturallanguage/download/hansard/
[4]http://www.statmt.org/wmt14/translation-task.html
[5]https://dumps.wikimedia.org/backup-index.html

language pair, the two resources were used to extract the parallel titles: the articles' titles and IDs in a particular language (ending with *-page.sql.gz*) and the interlanguage link records (file ends with *-langlinks.sql.gz*). Then, the title pairs were used to collect parallel articles using a crawler that we implemented on Java. After article pairs were collected, we preprocessed the data including: removing noisy characters, splitting sentences from paragraphs, word tokenization using the Moses scripts.[6] Finally, for each parallel article pair, sentences were aligned using the Microsoft sentence aligner (Moore, 2002), a powerful sentence alignment algorithm. The extracted bilingual corpus was used to improve SMT and NMT models.

3 Experiments

We conducted experiments on Asian language pairs: Japanese-English, Indonesian-Vietnamese, and English-Vietnamese using the two machine translation methods: SMT and NMT. Additionally, we extracted a bilingual corpus from Wikipedia to enhance the machine translation on both of the two methods.

3.1 Setup

For SMT models, we used the Moses toolkit (Koehn et al., 2007). The word alignment was trained using GIZA++ (Och and Ney, 2003) with the configuration *grow-diag-final-and*. A 5-gram language model of the target language was trained using KenLM (Heafield, 2011). For tuning, we used the batch MIRA (Cherry and Foster, 2012). For evaluation, we used the BLEU scores (Papineni et al., 2002).

For NMT models, we adapted the attentional encoder-decoder networks combined with byte-pair encoding (Sennrich et al., 2016a). In our experiments, we set the word embedding size 500, and hidden layers size of 1024. Sentences are filtered with the maximum length of 50 words. The minibatches size is set to 60. The models were trained with the optimizer Adadelta (Zeiler, 2012). The models were validated each 3000 minibatches based on the BLEU scores on development sets. We saved the models for each 6000 minibatches. For decoding, we used beam search with the beam size of 12. We trained NMT models on an Nvidia GRID K520 GPU.

[6]https://github.com/moses-smt/mosesdecoder/tree/master/scripts/tokenizer

3.2 SMT vs. NMT on Low-Resource Settings

Experiments on Japanese-English We conducted experiments on Japanese-English using the Kyoto bilingual corpora (Neubig, 2011). The training data includes 329,882 parallel sentences. For the development and the test data, there are 1,235 parallel sentences in the development set and 1,160 parallel sentences in the test set (see Table 1 for the data sets).

	Train	**Dev**	**Test**
Sentences	329,882	1,235	1,160
ja Words	6,085,131	34,403	28,501
en Words	5,911,486	30,822	26,734
ja Vocabs	114,284	4,909	4,574
en Vocabs	161,655	5,470	4,912

Table 1: Bilingual data set of Japanese-English of the training set (Train), development set (Dev), and test set (Test), (ja: Japanese, en: English).

Experimental results of Japanese-English translation are showed in Table 2. The NMT model obtained 11.91 BLEU point on the development set. For the test set, the model achieved 14.91 BLEU point after training 20 epochs. Meanwhile, the SMT model obtained the higher performance: +1.18 BLEU point on the development set, and +2.86 BLEU point on the test set. The experimental results indicated that for a small bilingual corpus (329k parallel sentences of the Japanese-English Kyoto corpus), the SMT model showed the higher performance than the NMT model.

Model	**Dev**	**Test**
SMT	13.09	17.75
NMT	11.91	14.91

Table 2: Experimental results in Japanese-English translation (BLEU)

Experiments on Indonesian-Vietnamese We conducted experiments on the Indonesian-Vietnamese language pairs, which has yet investigated on machine translation to our best knowledge.

For training data, we used two resources: TED data (Cettolo et al., 2012) and the ALT corpus (Asian Language Treebank Parallel Corpus) (Thu et al., 2016). We extracted Indonesian-Vietnamese parallel sentences from the TED data. For the ALT corpus, we dived the Indonesian-Vietnamese bilingual corpus into three parts: 16,000 sentences for training, 1,000 sentences for the development set, and 1,084 sentences for the test set. We combined the Indonesian-Vietnamese TED data with the training set extracted from the ALT corpus to create 226,239 training sentence pairs. The data sets are described in Table 3.

	Train	**Dev**	**Test**
Sentences	226,239	1,000	1,084
id Words	1,932,460	22,736	25,423
vi Words	2,822,894	32,891	36,026
id Vocabs	52,935	4,974	5,425
vi Vocabs	29,896	3,517	3,751

Table 3: Bilingual data sets of Indonesian-Vietnamese translations (id:Indonesian, vi: Vietnamese).

We showed the experimental results of the Indonesian-Vietnamese translations in Table 4. The NMT model achieved 14.48 BLEU point on the development set and 14.98 BLEU point on the test set after training 22 epochs. Meanwhile, the SMT model obtained the much higher performance: 27.37 BLEU point on the development set and 30.17 BLEU point on the test set.

Model	**Dev**	**Test**
SMT	27.37	30.17
NMT	14.48	14.98

Table 4: Experimental results on Indonesian-Vietnamese translation (BLEU)

Experiments on English-Vietnamese We conducted experiments on English-Vietnamese using the data sets of the IWSLT 2015 machine translation shared task (Cettolo et al., 2015). The *constrained* training data contained 130k parallel sentences from the TED talks.[7] We used the *tst2012* for the development set, *tst2013* and *tst2015* for the test sets. The data set are presented in Table 5.

Data	**Sent.**	**Src Vocab.**	**Trg Vocab.**
constr	131,019	50,118	54,565
unconstr	456,350	114,161	124,846
tst2012	1,581	3,713	3,958
tst2013	1,304	3,918	4,316
tst2015	1,080	3,175	3,528

Table 5: Data sets on the IWSLT 2015 experiments; **constr, unconstr**: the constrained, unconstrained training data set; **Src Vocab. (Trg Vocab.)**: the vocabulary size in the source (target) side of the corpus

In addition, we used two other data sets to enlarge the training data from the two resources: the corpus of National project VLSP (Vietnamese Language and Speech Processing)[8] and the EVBCorpus (Ngo et al., 2013). The two data sets were merged with the *constrained* data to create a large training data called *unconstrained* data. This aims to investigate how the large training data affects the SMT and NMT models.

System	**tst2013**	**tst2015**
constr (SMT)	26.54	24.42
constr (NMT)	23.59	17.27
unconstr(SMT)	27.19	25.41
unconstr(NMT)	26.71	22.30

Table 6: Experimental results English-Vietnamese translations (BLEU); **constr (SMT)**: the model trained on the constrained data using SMT; **unconstr (NMT)**: the model trained on the unconstrained data using NMT

Experimental results of English-Vietnamese are presented in Table 6. In overall, the SMT model obtained the higher performance than the NMT model (26.54 vs. 23.59 BLEU points on the *tst2013* using the *constrained* data, 25.41 vs. 22.30 BLEU points on the *tst2015* using the *unconstrained* data). Another point is the effect of enlarging the training data using the *unconstrained* data set. Enlarging the training data (increasing from 130k to 456k parallel sentences) improved both SMT and NMT models. Specifically, the SMT model achived +0.65

[7] https://www.ted.com/talks

[8] http://vlsp.vietlp.org:8080/demo/?page=home

BLEU point on the *tst2013* and +0.99 BLEU point on the *tst2015*. The interesting point is that the NMT model showed the higher improvement than the SMT model when using the *unconstrained* data: +3.12 BLEU point on the *tst2013* and +5.03 BLEU point on the *tst2015*.

3.3 Improving SMT and NMT Using Comparable Data

Building An English-Vietnamese Bilingual Corpus from Wikipedia As presented in Section 2.3, we used the Wikpedia database dumps to extract parallel titles, which were updated on *2017-01-20*. After collecting, processing, and aligning sentences in parallel articles using the Microsoft sentence aligner (Moore, 2002), we obtained 408,552 parallel sentences for English-Vietnamese. The extracted corpus are available at the repository of this work.

Improving SMT and NMT models We evaluated the extracted bilingual corpus in improving SMT and NMT models. Experimental results are shown in Table 7. There are several interesting findings from this experiment. First, although using only the Wikipedia corpus to train SMT and NMT models, we obtained promising results: 20.34 BLEU point using SMT and 17.58 BLEU point using NMT on the *tst2015*. Second, when the Wikipedia corpus was merged with the *unconstrained* for the training data, both SMT and NMT models achieved the improvement. For the SMT model, the improvement was +0.09 BLEU point on the *tst2013* and +0.95 BLEU point on the *tst2015*. Meanwhile, the NMT model showed the higher improvement with +2.22 BLEU point on the *tst2013* and up to +4.51 BLEU point on the *tst2015*. The next interesting point is that when using the large training data (more than 800k parallel sentences of merging 456k sentences the *unconstrained* with 408k sentences of the Wikipedia corpus), the NMT model outperformed the SMT model: 28.93 BLEU point vs. 27.28 BLEU point on the *tst2013*, 26.81 BLEU point vs. 26.36 BLEU point on the *tst2015*.

System	tst2013	tst2015
wiki (SMT)	22.06	20.34
wiki (NMT)	18.43	17.58
unconstr(SMT)	27.19	25.41
unconstr(NMT)	26.71	22.30
unconstr+wiki(SMT)	27.28	26.36
unconstr+wiki(NMT)	28.93	26.81

Table 7: Experimental results of English-Vietnamese using the corpus extracted from Wikipedia (BLEU); **wiki (NMT)**: the model trained on the extracted corpus from Wikipedia using NMT models; **unconstr+wiki**: the unconstrained data was merged with the Wikipedia corpus for the training data

4 Conclusion

Recent methods of phrase-based and neural-based have showed the promising directions in the development of machine translation. Neural machine translation models have been applied successfully on several language pairs with large bilingual corpora available. The phrase-based and neural-based methods are also compared and evaluated on some European language pairs. Nevertheless, there is still a bottleneck in SMT and NMT on low-resource language pairs when large bilingual corpora are unavailable. In this work, we conducted a comparison of SMT and NMT methods on several Asian language pairs which contain small bilingual corpora: Japanese-English, Indonesian-Vietnamese, and English-Vietnamese. In addition, a bilingual corpus was extracted from Wikipedia to enhance the machine translation performance and investigate the effects of the extracted corpus on the two machine translation methods. Experimental results showed meaningful findings. For a small bilingual corpus, SMT models showed the better performance than NMT models. Nevertheless, when enlarging the training data with the extracted corpus, both SMT and NMT models were improved, in which NMT models showed the higher improvement and outperformed the SMT models. This work can be useful for further improvement for machine translation on the low-resource languages.

References

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

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

Arianna Bisazza and Marcello Federico. 2016. A survey of word reordering in statistical machine translation: computational models and language phenomena. *Computational Linguistics*, 42(2):163–205, June.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44. Association for Computational Linguistics, August.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The iwslt 2015 evaluation campaign. *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of HLT/NAACL*, pages 427–436. Association for Computational Linguistics.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. In *CoRR 2015*.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Ann Irvine. 2013. Statistical machine translation in low resource settings. In *Proceedings of HLT/NAACL*, pages 54–61. Association for Computational Linguistics.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal neural machine translation systems for wmt15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 134–140.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. *arXiv preprint arXiv:1610.01108*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

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 ACL*, pages 177–180. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

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

Robert C Moore. 2002. *Fast and accurate sentence alignment of bilingual corpora*. Springer.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Quoc Hung Ngo, Werner Winiwarter, and Bartholomäus Wloka. 2013. Evbcorpus-a multi-layer english-vietnamese bilingual corpus for studying tasks in comparative linguistics. In *Proceedings of the 11th Workshop on Asian Language Resources (11th ALR within the IJCNLP2013)*, pages 1–9.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational linguistics*, 30(4):417–449.

Franz Josef Och, Christoph Tillmann, Hermann Ney, et al. 1999. Improved alignment models for statistical machine translation. In *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation (WMT)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems (NIPS)*, pages 3104–3112.

Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Introducing the asian language treebank (alt). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 1574–1578.

Liang Tian, Derek F Wong, Lidia S Chao, Paulo Quaresma, Francisco Oliveira, and Lu Yi. 2014. Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In *LREC*, pages 1837–1842.

Pidong Wang, Preslav Nakov, and Hwee Tou Ng. 2016. Source language adaptation approaches for resource-poor machine translation. *Computational Linguistics*.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *CoRR*.