PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 108 JUNE 2017 171–182

Pre-Reordering for Neural Machine Translation: Helpful or Harmful?

Jinhua Du, Andy Way

ADAPT Centre, School of Computing, Dublin City University

Abstract

Pre-reordering, a preprocessing to make the source-side word orders close to those of the target side, has been proven very helpful for statistical machine translation (SMT) in improving translation quality. However, is it the case in neural machine translation (NMT)? In this paper, we firstly investigate the impact of pre-reordered source-side data on NMT, and then propose to incorporate features for the pre-reordering model in SMT as input factors into NMT (factored NMT). The features, namely parts-of-speech (POS), word class and reordered index, are encoded as feature vectors and concatenated to the word embeddings to provide extra knowledge for NMT. Pre-reordering experiments conducted on Japanese↔English and Chinese↔English show that pre-reordering the source-side data for NMT is redundant and NMT models trained on pre-reordered data deteriorate translation performance. However, factored NMT using SMT-based pre-reordering features on Japanese→English and Chinese→English is beneficial and can further improve by 4.48 and 5.89 relative BLEU points, respectively, compared to the baseline NMT system.

1. Introduction

In recent years, NMT has achieved impressive progress (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015). The state-of-the-art NMT model employs an encoder–decoder architecture with an attention mechanism, in which the encoder summarizes the source sentence into a vector representation, and the decoder produces the target string word by word from vector representations, and the attention mechanism learns the soft alignment of a target word against source words (Bahdanau et al., 2015). NMT systems have outperformed the state-of-the-art SMT model on various language pairs in terms of translation qual-

 Corresponding author: andy.way@adaptcentre.ie
Cite as: Jinhua Du, Andy Way. Pre-Reordering for Neural Machine Translation: Helpful or Harmful?. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 171–182. doi: 10.1515/pralin-2017-0018.

ity (Luong et al., 2015; Bentivogli et al., 2016; Junczys-Dowmunt et al., 2016; Wu et al., 2016; Toral and Sánchez-Cartagena, 2017). However, due to some deficiencies of NMT systems such as the limited vocabulary size, low adequacy for some translations, much research work has involved incorporating extra knowledge such as SMT features or linguistic features into NMT to improve translation performance (He et al., 2016; Sennrich and Haddow, 2016; Nadejde et al., 2017; Wang et al., 2017).

Pre-reordering, a preprocessing step in SMT, modifies the word order of a source-side sentence to be more similar to the word order in a target language, and has proven very helpful in improving translation quality for SMT systems (Xia and McCord, 2004; Collins et al., 2005; Neubig et al., 2012; Miceli-Barone and Attardi, 2013; Nakagawa, 2015).[1] NMT has a strong capability to learn word orders or word alignment from sequential lexical information using the soft alignment (attention) mechanism, and NMT systems introduce more changes in word order than pure phrase-based SMT (PB-SMT) systems. Furthermore, NMT's reorderings are closer to the reorderings in the reference than those of PB-SMT (Toral and Sánchez-Cartagena, 2017). Thus, in this paper, we ask the question whether pre-reordering is necessary and helpful for NMT.

The intuition behind pre-reordering for NMT is contradictory: on the one hand, if the word order of a source-side sentence is close to that of the target language, then the attention mechanism can easily learn a diagonal alignment, so pre-reordering might be helpful to the learning process; on the other hand, compared to the weak global reordering capability of PB-SMT, the attention mechanism in NMT can globally learn the word alignment, so pre-reordering might be redundant.

Zhu (2015) firstly reported the observation that performing pre-reordering on NMT hurts the model performance. In his experiment, the pre-reordered NMT system using long-short term memory (LSTM) degrades by 1.22 BLEU (Papineni et al., 2002) points compared to the baseline NMT system. However, he only empirically performed experiments on English→Japanese, and did not have a general verification on other language pairs and analyse the reason behind the result.

In this paper we investigate the impact and generality of pre-reordering on NMT, and verify whether pre-reordering is redundant for NMT by comprehensively experimenting on two language pairs, four translation directions in total, and then propose an indirect method of utilizing the pre-reordering features as factors in NMT to enhance the attention model to learn more accurate word alignments. The main contributions of this work include:

- We examine the effect of pre-reordered training data on NMT models on a number of translation directions, which shows that pre-reordering is not helpful to the current NMT architecture. The pre-reordering operation is like a hard constraint which deteriorates the learning capability of neural networks from the natural word order.

[1] A huge of amount of work has been done on this topic. Here we only list some example papers.

- We propose a new feature and incorporate it with SMT-based pre-reordering features as factors to NMT to verify their impact on translation quality.
- We provide a qualitative analysis on the translation results.

2. Related Work

To the best of our knowledge, there is limited work published on the issue of pre-reordering for NMT. Zhu (2015) is the first work to report that the NMT system trained on the pre-reordered data hurts translation quality compared to the NMT system trained on the naturally ordered data. In his experiments on English→Japanese task, the pre-reordered NMT system decreases by 1.22 BLEU points compared to the normal LSTM NMT system. However, he did not examine the reasons behind the result and verify on other language pairs.

Niehues et al. (2016) proposed a pre-translation strategy to combine SMT and NMT, in which the SMT system is used to pre-translate the input and then an NMT system generates the final hypothesis using the pre-translation. In this framework, they only use the pre-reordered data to train SMT systems rather than NMT systems. In their experiments, the pre-translation system using the pre-reordered SMT system can improve translation quality compared to that trained on naturally ordered data.

Toral and Sánchez-Cartagena (2017) carried out a multifaceted evaluation of NMT versus PB-SMT for 9 language directions. One evaluation is the reordering. However, their work is not to perform reordering in the source-side sentences to train the NMT systems, but to measure the amount of reordering performed by NMT and PB-SMT systems, i.e. whether NMT systems produce more changes in the word order of a sentence than the PB-SMT systems, and whether NMT systems make the word order of the translation closer to that of the reference.

A number of works on integrating extra knowledge or different features into NMT have been carried out recently. He et al. (2016) incorporate SMT features, such as a translation model and an n-gram language model, with the NMT model under the log-linear framework. Their experiments show that the proposed method significantly improves translation quality of the baseline NMT system on Chinese→English translation tasks.

Wang et al. (2017) propose to incorporate an SMT model into the NMT framework in which at each decoding step, SMT offers additional recommendations of generated words based on the decoding information from NMT, and then an auxiliary classifier is employed to score the SMT recommendations and a gating function is used to combine the SMT recommendations with NMT generations, both of which are jointly trained within the NMT architecture in an end-to-end manner. Experimental results on Chinese–English translation show that the proposed approach achieves significant and consistent improvements over state-of-the-art NMT and SMT systems.

Different from the above work, Sennrich and Haddow (2016) integrate linguistic features such as morphological features, POS tags, and syntactic dependency labels

as input features to NMT system by generalising the embedding layer of the encoder. In experiments on WMT16 training and test sets, linguistic input features improve model quality. García-Martínez et al. (2016) propose the concept of factored NMT, and they use the linguistic decomposition of the words in the output side rather than in the input.

Similar to the work in Sennrich and Haddow (2016), we propose to incorporate features such as SMT-based pre-reordering features and a new reordered index feature as inputs to NMT to verify their effectiveness in improving translation quality.

3. Neural Machine Translation

The basic principle of an NMT system is that it can map a source-side sentence $\mathbf{x} = (x_1, \ldots, x_m)$ to a target sentence $\mathbf{y} = (y_1, \ldots, y_n)$ in a continuous vector space, where all sentences are assumed to terminate with a special "end-of-sentence" token $< eos >$. Conceptually, an NMT system employs neural networks to solve the conditional distributions as in (1):

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} p(y_i|y_{<i}, x_{\leq m}) \tag{1}$$

We utilise the NMT architecture in Bahdanau et al. (2015), which is implemented as an attentional encoder-decoder network with recurrent neural networks (RNN).

In this framework, the encoder is a bidirectional neural network (Sutskever et al., 2014) with gated recurrent units (Cho et al., 2014) where a source-side sequence $\mathbf{x}$ is converted to a one-hot vector and fed in as the input, and then a forward sequence of hidden states $(\overrightarrow{h}_1, \ldots, \overrightarrow{h}_m)$ and a backward sequence of hidden states $(\overleftarrow{h}_1, \ldots, \overleftarrow{h}_m)$ are calculated and concatenated to form the annotation vector h_j. The decoder is also an RNN that predicts a target sequence $\mathbf{y}$ word by word where each word y_i is generated conditioned on the decoder hidden state s_i, the previous target word y_{i-1}, and the source-side context vector c_i as in (2):

$$p(y_i|y_{<i}, \mathbf{x}) = g(y_{i-1}, s_i, c_i) \tag{2}$$

where g is the activation function that outputs the probability of y_i, and c_i is calculated as a weighted sum of the annotations h_j. The weight α_{ij} is computed as in (3):

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{m} exp(e_{ik})} \tag{3}$$

where

$$e_{ij} = a(s_{i-1}, h_j)$$

is an alignment model which models the probability that the inputs around position j are aligned to the output at position i. The alignment model is a single-layer feed-forward neural network that is learned jointly through backpropagation.

4. Top-Down BTG-based Pre-reordering

In PB-SMT, the difference in word order between source and target languages is one of the major problems. Pre-reordering source-side word order closes to that of the target language is one of many approaches to deal with this issue. In this paper, we investigate a pre-reordering method based on Bracketing Transduction Grammar (BTG) (Neubig et al., 2012) for NMT systems.[2]

The BTG-based pre-reordering method reorders source sentences by handling sentence structures as latent variables. Nakagawa (2015) proposed an incremental top-down parsing method to improve the computational efficiency of the original BTG-based pre-reordering where model parameters can be learned using latent variable Perceptron with the early update technique. His experiments show that pre-ordering using the top-down parsing algorithm was faster and achieved higher BLEU scores than the original BTG-based pre-ordering method.

The advantage of the top-down BTG-based pre-reordering method is that it can be easily applied to any languages using only parallel text. Given a word x_i in a source-side sentence $\mathbf{x}$, three features are used to pre-reorder $\mathbf{x}$, namely the word surface form x_i^w, POS tag x_i^p and word class x_i^c. To train the pre-ordering model, the word alignment links between words in the source and target sentences of the parallel training data are also provided. The trained pre-reordering model is then employed to prereorder the training data and test data annotated by the above three features.

5. Factored NMT Using Pre-reordering Features

Factored NMT, introduced in Sennrich and Haddow (2016), represents the encoder input as a combination of features as in (4):

$$\overrightarrow{h}_j = g(\overrightarrow{W}(\Big\|_{k=1}^{|F|} E_k x_{jk}) + \overrightarrow{U}\overrightarrow{h}_{j-1}) \tag{4}$$

where $\|$ is the vector concatenation, $E_k \in \mathbb{R}^{m_k \times K_k}$ are the feature embedding matrices, with $\sum_{k=1}^{|F|} m_k = m$, and K_k is the vocabulary size of the k_{th} feature, and $|F|$ is the number of features in the feature set F (Sennrich and Haddow, 2016).

In factored NMT, the features can be any form of knowledge which might be useful to NMT systems, such as POS tags, lemmas, morphological features and dependency labels used in Sennrich and Haddow (2016). In our work, besides the the pre-reordering features, namely the POS tag and word class, we propose another feature

[2] In future work, we will examine the impact of different pre-reordering methods on NMT.

to verify how these features affect the performance of NMT systems. The new feature is defined as "Reordered Index" which is illustrated in Table 1.

Source:	Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi .
Original Index:	0 1 2 3 4 5 6 7 8 9 10
Reference:	Australia is one of the few countries that have diplomatic relations with North Korea .
Pre-reordered:	Aozhou shi zhiyi shaoshu guojia de you bangjiao yu Beihan .
Absolute Reordered Index:	0 1 9 7 8 6 4 5 2 3 10
Source:	Aozhou shi yu **Beihan** you bangjiao de shaoshu guojia zhiyi .
Relative Reordered Index:	0 0 6 **6** 2 2 -1 -4 -4 -7 0

Table 1. An example of reordered index as an input feature for NMT

In Table 1, the source language is Chinese (shown as Chinese Pinyin) and the reference is English. "Pre-reordered" indicates the reordered Chinese sentence by the BTG-based pre-reordering model. "Original Index" is the sequence of word position in the original source-side sentence, and "Absolute Reordered Index" is the reordered sequence of word positions where the number represents the word position in the original source-side sequence.

In order to reduce data sparseness, we convert the absolute word positions in "Absolute Reordered Index" to relative word positions in "Relative Reordered Index", which is calculated as in (5):

$$\mathrm{relative\_p} = \mathrm{p\_in\_reordered\_sequence} - \mathrm{p\_in\_original\_sequence} \quad (5)$$

For example, the word "**Beihan**" in Table 1 has the absolute position "3" in the original source sentence, while it moves to position "9" in the pre-reordered source sentence. Then we have $\{\mathrm{relative\_p} = 9 - 3 = \mathbf{6}\}$ as shown in the last row of Table 1.

6. Experiments

As Japanese and Chinese languages differ drastically from English in terms of word order and grammatical structure, we select Japanese–English and Chinese–English translations[3] to verify the impact of pre-reordering on NMT.

Two sets of experiments are set up as follows:

- Pre-reordering for NMT: four translation directions (JP↔EN and ZH↔EN) are evaluated on non-prereordered and pre-reordered data for NMT.

[3]In the rest of the paper, we use JP, ZH and EN to denote Japanese, Chinese and English, respectively.

- Factored NMT: SMT-based pre-reordering features are encoded as input factors for NMT systems.

In the following sections, we will report our experimental setup and results in terms of these two experiments.

6.1. Experimental Settings

For JP–EN translation tasks, the training data is the first part (train-1) of the JP–EN Scientific Paper Abstract Corpus (ASPEC-JE) that contains 1M sentence pairs, the development/validation set contains 1,790 sentence pairs, and the test set contains 1,812 sentence pairs (Nakazawa et al., 2016). There is only one reference for each source-side sentence in the validation and test sets.

For ZH–EN tasks, we use 1.4M sentence pairs extracted from LDC ZH–EN corpora as the training data, and NIST 2004 current set as the development/validation set that contains 1,597 sentences, and NIST 2005 current set as the test set that contains 1,082 sentences. There are four references for each Chinese sentence and there is only one reference for each English sentence in the validation and test sets. For EN→ZH, we use the first reference out of four references for Chinese as the input (English).

The pre-reordering factors, namely the POS tag, word class and reordered index are obtained by:

- POS tag: the Japanese data are segmented and tagged using KyTea (Neubig et al., 2011), and the Chinese data are segmented and tagged using the ICTCLAS toolkit (Zhang et al., 2003).
- Word Class (WoC): the word classes of the training data are obtained using "mkcls" by setting the number of classes to 50. For an Out-of-Vocabulary word in the validation and test sets, we randomly allocate a class between (1, 50) to it.
- Reordered Index (ReIdx): we generate two different kinds of reordered indices, namely the "Absolute Reordered Index" (AbsReIdx) and "Relative Reordered Index" (RelaReIdx) which are described in Section 5.

Chinese and Japanese are not suitable for using the Byte Pair Encoding (BPE) method (Sennrich et al., 2016) to encode words as subword units. Thus, we keep the words as translation units. We use Moses (Koehn et al., 2007) with default settings as the standard PB-SMT system, and use KenLM (Heafield et al., 2013) to train a 5-gram language model with the target side of the parallel data. We use Nematus (Sennrich et al., 2017) as the baseline NMT system, and set minibatches of size 80, a maximum sentence length of 60, word embeddings of size 600, and hidden layers of size 1024. The vocabulary size for input and output is set to 45K. Models are trained with the Adadelta optimizer (Zeiler, 2012), reshuffling the training corpus between epochs. We validate the model every 5,000 minibatches via BLEU scores on the validation set.

As in Sennrich and Haddow (2016), for factored NMT systems, in order to ensure that performance improvements are not simply due to an increase in the number of model parameters, we keep the total size of the embedding layer fixed to 600. Table 2

shows the vocabulary size and embedding size for pre-reordering features and the word as the input for the JP→EN NMT system. The total embedding size is fixed to 600. "Varied" indicates that for each single feature, the word embedding size will be different which is obtained by $[600 - embedding\_size(feature)]$. For example, the word embedding size will be $600 - 10 = 590$ for using POS tags as the input feature. Similar settings and parameters are for Chinese. We add 'UNK' to the vocabulary of each feature.

	Input Voc. Size		Input Voc. Size		Embedding Size	
Feature	JP	Model	ZH	Model	All	Single
POS tags	21	21	37	37	10	10
Word Class	51	51	51	51	15	15
AbsReIdx	61	61	61	61	15	15
RelaReIDX	117	117	117	117	20	20
Word	161,390	45,000	185,029	45,000	540	Varied

Table 2. Vocabulary size, and size of embedding layer of each feature.

In order to verify the impact of pre-reordered data on NMT systems and how pre-reordering features affects NMT systems, we only use the single NMT model rather than an ensemble model. The beam size for NMT decoding is 12. All results are reported by case-insensitive BLEU scores and carried out a bootstrap resampling significance test (Koehn, 2004).

6.2. Results and Analysis

Tables 3 and 4 show our main results for JP↔EN and ZH↔EN with and without pre-reordered data, respectively. The baseline system is a standard PB-SMT system trained on non-reordered and pre-reordered data, respectively.

	JP→EN				EN→JP			
	Non-reordered		Pre-reordered		Non-reordered		Pre-reordered	
SYS	Validation	Test	Validation	Test	Validation	Test	Validation	Test
SMT	18.25	17.64	**21.79***	**21.71***	27.03	26.32	**33.67***	**33.75***
NMT	**24.16***	**24.55***	20.42	21.43	**35.25***	**35.23***	32.75	32.98
Gain	+5.91	+ 6.91	-1.37	-0.31	+ 8.22	+ 8.91	-0.92	-0.77

Table 3. Results on JP–EN pre-reordering experiments. "" indicates translation performance is significantly better.*

	ZH→EN				EN→ZH			
	Non-reordered		Pre-reordered		Non-reordered		Pre-reordered	
SYS	Validation	Test	Validation	Test	Validation	Test	Validation	Test
SMT	33.13	29.24	**34.63***	**30.59***	14.50	12.77	**16.12***	**13.77***
NMT	**35.49***	**31.76***	33.95	30.23	**15.97***	**15.62***	14.14	13.53
Gain	+2.46	+2.52	-0.68	-0.36	+1.47	+2.85	-1.98	-0.22

Table 4. Results on ZH–EN pre-reordering experiments

NMT systems trained on the non-reordered data significantly improve on the validation set by 5.91 (18.25→24.16) and on the test set by 6.91 (17.64→24.55) absolute points for JP→EN, respectively; and by 8.22 (27.03→35.25) absolute points on the validation set and 8.91 (26.32→35.23) absolute points on the test set for EN→JP, respectively, compared to SMT systems.

Non-reordered NMT systems significantly improve on the validation set by 2.46 (33.13→35.49) and on the test set by 2.52 (29.24→31.76) absolute points for ZH→EN, respectively; and by 1.47 (14.50→15.97) on the validation set and 2.85 (12.77→15.62) absolute points on the test set for EN→ZH, respectively, compared to SMT systems.

However, for NMT systems trained on the pre-reordered data, translation performance decreases both on the validation set and test set compared to the SMT systems trained on the pre-reordered data. We also observe that 1) pre-reordered SMT systems achieve significant improvement compared to baseline SMT systems; 2) pre-reordered NMT systems perform worse than the non-reordered NMT systems.

From the results we can see that the pre-reordering has a negative impact on the learning capability of NMT systems. We infer that the pre-reordering is like a hard constraint for NMT and introduces more noise in terms of word order, which appears to make the learning process more difficult.

We also evaluate pre-reordering features as input factors for the NMT system against the baseline NMT system. The results are shown in Table 5.

	JP→EN		ZH→EN	
SYS	Validation	Test	Validation	Test
NMT	24.16	24.55	35.49	31.76
AbsReIdx	24.40	24.61	36.42*	31.90
RelaReIdx	24.52*	24.90*	36.87*	31.96*
POS+WoC	25.08*	25.17*	37.42*	33.15*
POS+WoC+RelaReIdx	**25.26***	**25.65***	**37.83***	**33.63***
Gain	1.1	1.1	2.34	1.87

Table 5. Results on JP→EN and ZH→EN factored NMT Experiments

We observe that the proposed "Reordered Index" features, namely the AbsReIdx and RelaReIdx can improve translation quality, but the former is not significant while the latter is significant, which shows that the relative reordering positions can provide more extra useful information to the words. The features of the pre-reordering model for SMT, namely the POS tags and word class, improve by 0.92 (24.16→25.08) and 1.93 (35.49→37.42) BLEU points on the validation set, respectively, and 0.62 (24.55→25.17) and 1.39 (31.76→33.15) BLEU points on the test set, respectively, compared to the baseline NMT system. In addition, adding the RelaReIdx further improves by 0.48 (25.17→25.65) and 0.48 (33.15→33.63) BLEU points on the test set, respectively. The incremental improvements in Table 5 show that the POS tags, word class and Reordered Index features contribute different information to the learning process of the NMT system to improve translation performance.

7. Conclusion

In this paper we investigate whether pre-reordering is beneficial to NMT and our empirical results show that it is not the case, i.e. pre-reordering the source-side data deteriorates translation performance. Linguistic knowledge has been verified to be useful in improving translation quality by resolving the reordering problem, so we propose to integrate SMT-based pre-reordering features, namely POS tags, word class and reordered index as input factors into the JP–EN and ZH–EN NMT systems. Our experiments show that these pre-reordering features yield improvements over the baseline NMT system, resulting in improvements on the test set of 1.1 and 1.87 BLEU points, respectively, on the test sets.

As to future work, we expect more experiments on different language pairs and different pre-reordering methods to verify the impact of pre-reordering on NMT, and we will explore the inclusion of novel and different reordering features for NMT to improve reordering in translations further.

Acknowledgements

We would like to thank the reviewers for their valuable and constructive comments. Thanks Dr. Jian Zhang for his initial idea and work on pre-reordered SMT. This research is supported under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106).

Bibliography

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proc. of the 3rd International Conference on Learning Representations*, pages 1–15, San Diego, USA, 2015.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus phrase-based machine translation quality: a case study. In *Proc. of the EMNLP*, 2016.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proc. of the EMNLP*, 2014.

Collins, Michael, Philipp Koehn, and Ivona Kucerova. Clause Restructuring for Statistical Machine Translation. In *Proc. of the ACL*, pages 531–540, Ann Arbor, Michigan, USA, 2005.

García-Martínez, Mercedes, Loïc Barrault, and Fethi Bougares. Factored Neural Machine Translation. In *arXiv:1609.04621*, 2016.

He, Wei, Zhongjun He, Hua Wu, and Haifeng Wang. Improved Neural Machine Translation with SMT Features. In *Proc. of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 151–157, Phoenix, Arizona, USA, 2016.

Heafield, Kenneth, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable Modified Kneser-Ney Language Model Estimation. In *Proc. of the ACL*, pages 690–696, Sofia, Bulgaria, 2013.

Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. In *Proc. of the IWSLT*, Tokyo, Japan, 2016.

Kalchbrenner, Nal and Phil Blunsom. Recurrent continuous translation models. In *Proc. of the EMNLP*, pages 1700–1709, Seattle, Washington, USA, 2013.

Koehn, Philipp. Statistical Significance Tests for Machine Translation Evaluation. In *Proc. of the EMNLP*, pages 388–395, Barcelona, Spain, 2004.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of the ACL*, pages 177–180, Prague, Czech Republic, 2007.

Luong, Thang, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proc. of the EMNLP*, pages 1412–1421, Lisbon, Portugal, 2015.

Miceli-Barone, Antonio Valerio and Giuseppe Attardi. Pre-Reordering for Machine Translation Using Transition-Based Walks on Dependency Parse Trees. In *Proc. of the Eighth Workshop on Statistical Machine Translation*, pages 164–169, Sofia, Bulgaria, 2013.

Nadejde, Maria, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. Syntax-aware Neural Machine Translation Using CCG. In *arXiv:1702.01147*, 2017.

Nakagawa, Tetsuji. Efficient Top-Down BTG Parsing for Machine Translation Preordering. In *Proc. of the ACL-IJCNLP*, pages 208–218, Beijing, China, 2015.

Nakazawa, Toshiaki, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian Scientific Paper Excerpt Corpus. In *Proc. of the Tenth LREC*, Portorož, Slovenia, 2016.

Neubig, Graham, Yosuke Nakata, and Shinsuke Mori. Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis. In *Proc. of the ACL-HLT*, pages 529–533, Portland, Oregon, USA, 2011.

Neubig, Graham, Taro Watanabe, and Shinsuke Mori. Inducing a Discriminative Parser to Optimize Machine Translation Reordering. In *Proc. of the EMNLP-CoNLL*, pages 843–853, Jeju Island, Korea, 2012.

Niehues, Jan, Eunah Cho, Thanh-Le Ha, and Alex Waibel. Pre-Translation for Neural Machine Translation. In *Proc. of the COLING*, pages 1828–1836, Osaka, Japan, 2016.

Papineni, Kishore, Salim Roukos, Todd Ward, and WeiJing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proc. of the ACL*, pages 311–318, 2002.

Sennrich, Rico and Barry Haddow. Linguistic Input Features Improve Neural Machine Translation. In *Proc. of the 1st Conference on Machine Translation*, pages 83–91, Berlin, 2016.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proc. of the ACL*, pages 1715–1725, Berlin, Germany, 2016.

Sennrich, Rico, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nădejde. Nematus: a Toolkit for Neural Machine Translation. In *arXiv:1703.04357*, 2017.

Sutskever, Ilya, Oriol Vinyals, , and Quoc V Le. Sequence to sequence learning with neural networks. In *Proc. of the 2014 Neural Information Processing Systems*, pages 3104–3112, Montreal, Canada, 2014.

Toral, Antonio and Víctor M. Sánchez-Cartagena. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. In *Proc. of the EACL*, Valencia, Spain, 2017.

Wang, Xing, Zhengdong Lu, Zhaopeng Tu, Hang Li, Deyi Xiong, and Min Zhang. Neural Machine Translation Advised by Statistical Machine Translation. In *Proc. of the AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, 2017.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, and Mohammad Norouzi et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. In *arXiv:1609.08144*, 2016.

Xia, Fei and Michael McCord. Improving a Statistical MT System with Automatically Learned Rewrite Patterns. In *Proc. of the COLING*, pages 508–514, IIT Bombay, India, 2004.

Zeiler, Matthew D. ADADELTA: An Adaptive Learning Rate Method. In *CoRR, abs/1212.5701*, 2012.

Zhang, Huaping, Hongkui Yu, Deyi Xiong, and Qun Liu. HHMM-based Chinese Lexical Analyzer ICTCLAS. In *Proc. of the Second SIGHAN Workshop on Chinese Language Processing*, pages 184–187, Sapporo, Japan, 2003.

Zhu, Zhongyuan. Evaluating Neural Machine Translation in English-Japanese Task. In *Proc. of the 2nd Workshop on Asian Translation*, pages 61–68, Kyoto, Japan, 2015.

Address for correspondence:
Andy Way
andy.way@adaptcentre.ie
ADAPT Centre, School of Computing, Dublin City University,
Glasnevin, Dublin 9, Dublin, Ireland