Improving Low-Resource NMT with Parser Generated Syntactic Phrases

Kamal Kumar Gupta, Sukanta Sen, Asif Ekbal, Pushpak Bhattacharyya

Department of Computer Science and Engineering,
Indian Institute of Technology Patna, India
{kamal.pcs17, sukanta.pcs15, asif, pb}iitp.ac.in

Abstract. Recently, neural machine translation (NMT) has become highly successful achieving state-of-the-art results on many resource-rich language pairs. However, it fails when there is a lack of sufficiently large amount of parallel corpora for a domain and/or language pair. In this paper, we propose an effective method for NMT under a low-resource scenario. The model operates by augmenting the original training data with the examples extracted from the parse trees of the target-side sentences. It provides important evidences to the model as these phrases are relatively smaller and linguistically correct. Our experiment on the benchmark WMT14 dataset shows an improvement of 3.28 BLEU and 3.41 METEOR score for Hindi to English translation. Evaluation on the same language pair with relatively much smaller datasets of judicial and health domains also show the similar trends with significant performance improvement in terms of BLEU (15.63 for judicial and 15.97 for health) and METEOR (14.30 for judicial and 15.93 for health).

Keywords: Neural Machine Translation · Low Resource Machine Translation · Low Resource NMT.

1 Introduction

Neural machine translation (NMT) has recently attracted a lot of attention to the translation community due to its promising results on several language pairs [4] and rapid adoption in the deployment services [22, 6, 10]. The advantages of an NMT system over the Statistical Machine Translation (SMT) are the followings: an entire machine translation (MT) system can be implemented with a single end-to-end architecture; it is better than SMT at generating the fluent outputs [14]. However, NMT requires a huge parallel corpus and the absence of which makes outputs suffer from adequacy.

Efficiency of any NMT model [14] greatly depends on the size of the parallel corpus. It performs well with a very large size of training data, but performs poorly when there is a scarcity of such a large corpus. In addition to that, SMT models are known to perform better when we do not have a sufficiently large amount of parallel data. However, [14] has shown that NMT based method makes a huge jump in BLEU score as we increase the training data size while SMT improves the BLEU score with a fixed rate.

Building NMT models under a low-resource scenario is a great challenge to the researchers. We consider the scenarios to be low-resource when we do not have a sufficient amount of parallel data for a certain language pair or do not have a sufficient amount of parallel data for a particular domain. In our case, we develop an NMT system for Hindi-English language pair which has relatively less parallel data as compared to some European language pairs. Our domains are judicial and health for which we do not have sufficient amount of parallel data, especially for language pair like Hindi-English. Translating documents related to health and judiciary are very crucial in a multilingual country like India. In general, the health information (electronic medical records, health tips available in social media etc.) are available in English and making these available in Hindi would be useful to the common people.

For both health and judicial domains, we have a dearth of parallel data, and we treat this situation as low-resource scenario. In this paper, we propose an approach for NMT that can effectively work for the purpose of improving low-resource NMT without using any additional monolingual data. With the help of a constituency parser, we extract the phrases from the parsed trees of the target sentences. Though the quality of these phrases depends on the robustness of the parser, we can consider these phrases[1] useful for providing important evidences during training. First, we extract noun, verb and prepositional phrases and then we *i.* back-translate these phrases to generate the source-target parallel phrases; and *ii.* make identical copies of these phrases in the source side to obtain the copied parallel phrases. These (translated and copied) parallel phrases are added with the original training data as training examples. Our method is inspired by the idea of adding additional monolingual data through back-translation [21] in order to reduce data sparsity. Our method does not add additional monolingual data because for some pair of languages or for the particular domains, sufficient data may not be present and adding out-of-domain data (to the parallel corpus) may create ambiguity.

Target side monolingual data helps to improve fluency. In encoder-decoder NMT architecture, decoder is indeed a Recurrent Neural Network (RNN) language model. So, it is important that target side data must be accurate in fluency. For our Hindi→English translation system extracting phrases from Hindi(source) and adding them to the English(target) side using back-translation and copied technique will affect the fluency because the augmented synthetic English phrases may not be very fluent. That is why we use phrases from English (target) side only. Our experiments with Hindi→English language for judicial, health domain and WMT14 dataset show impressive performance gains due to the inclusion of these phrases. The key points of our proposed work as follows:

- We augment the original training data using syntactic phrases extracted from the original training data with the help of a constituency parser. This augmented data is used for training of an NMT system.
- We empirically show that our proposed approach of augmenting training data by syntactic phrase pairs, which uses no additional monolingual data,

[1] Linguistically more accurate as the lengths are short

can improve the performance significantly over the baseline NMT model developed with only the original training data.

The remainder of the paper is organized as follows: in Section 2, we present the related works briefly. Section 3 describes in details our proposed approach. Section 4 and 5 provide the details of the datasets used and the experimental setup, respectively. We show the results and analysis in Section 6. Finally, in Section 7 we conclude our work with future work road-maps.

2 Related Works

Neural Machine Translation (NMT) requires a significantly large-scale parallel data for training. Many language pairs, especially under the low resource scenario, do not have this abundance of information. Hence researchers are currently focusing on exploring methods that could be effective in a low-resource scenario. The use of monolingual target language data with the available parallel corpus is one such method that researchers have attempted in recent times. [9] trained a language model using monolingual target language words and integrate it with NMT system trained on a low-resource language pair. They showed improvement in translation for Turkish-English language pair. [21] introduced the back-translation method in which instead of training language model on monolingual target data, they translated the monolingual target data into source language and use this synthetic parallel data along with the original parallel data to train our NMT models. Inspired by the back-translation method, [23] proposed a method by adding the translated monolingual source-side data to the target-side, and create the synthetic parallel data. [7] made an identical copy of monolingual target data at source side to make copied parallel data which they used along with the original training data for training NMT models.
Our approach is different from the above mentioned approaches as these make use of additional monolingual data with the original bilingual corpus while, in our case, we do not use any additional monolingual data with the original training corpus. Our approach aims at extracting additional information from the original target side training sentences, itself, in the form of syntactic phrases and use it with bilingual corpus to train the NMT model.

3 Proposed Method

Our proposed method is based on the standard attentional encoder-decoder approach [1], and augments the original training data with additional synthetic examples. This method of augmenting data does not force any changes in the NMT architecture. In our method, we are not just dividing the target sentence into small phrases, rather we extract NP, VP and PP phrases in a way so that we get the whole sentence in the form of small to large sequences. Adding instances to the training data in such a way helps the model to learn the sentences from smaller to larger sequences as mentioned in the example of Section 3.1. This,

in turn, helps to preserve adequacy as well as fluency. For generating the synthetic training examples, our proposed model leverages the information obtained from parsing and back translation. We extract the phrases by parsing the target side sentences, and generate their equivalents in the source sides. Thereafter, we combine these with the original parallel data to use it to train the NMT model for source→target. Figure 1 gives an idea about the steps to be followed in this approach.

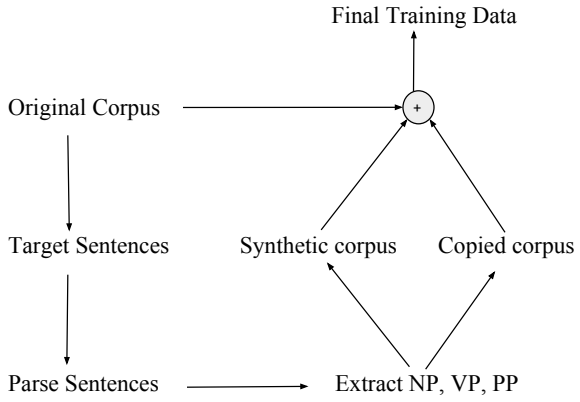

Fig. 1. Overall architecture. '+' means append.

3.1 Phrase Extraction from Parse Tree

We consider the target sentences from original training corpus, and parse it using a constituency parser. We perform experiments for the Hindi→English, and for parsing our target side (English) sentences, we use the Stanford parser[2]. Suppose in Hindi-English training data a source sentences is "इस अधिनियम को इंपीरियल लाइब्रेरी का नाम राष्ट्रीय पुस्तकालय में बदलने के लिए पारित किया गया था।" *is adhiniyam ko impeeriyal laibreree ka naam raashtreey pustakaalay mein badalane ke lie paarit kiya* and its aligned target sentences is "*This act was passed to change the name of the Imperial Library to National Library .*" After getting the parse tree of a target sentence, we extract the NP, VP and PP from the parse trees. For example, from the parse tree, as shown in Figure 2, we extract the syntactic phrases given in Table 2.

We collect the phrases in such a way that the phrases of the same category (e.g., NP, PP or VP) comes in the order of a small sequence to the large sequence.

[2] https://nlp.stanford.edu/software/lex-parser.shtml

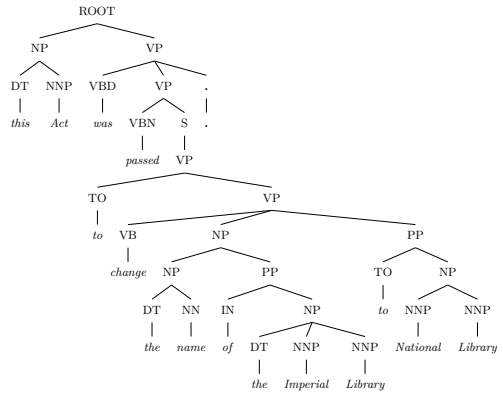

Fig. 2. Parsed tree

It means we just do not extract the single largest phrase from NP, VP and PP, rather we extract all the possible constituent phrases (in case of nested phrases). For example, first we obtain a NP "the Imperial Library" and then when we get the larger noun phrase "the name of the Imperial Library", we also consider its subset (i.e. constituent) NP as the potential candidate. We can see that all these phrases are linguistically sound.

Table 1. Number of English phrases generated from English sentences

Dataset	*Sentences*	*Phrases*
WMT_14	263,654	595,969
Judicial	5,000	81,308
Health	23,000	260,487

3.2 Phrase based SMT for Target→Source

We need a target to source translation system for generating synthetic source phrases by back translating the target phrases (c.f. Section 3.3). After extracting phrases from the target sentences of the original training corpus, we need their aligned parallel source translations. For this translation, we prefer PB-SMT [15] over NMT [1] because PB-SMT performs better in case of a small parallel corpus. For judicial domain, we have only 5000 parallel sentences, whereas, for the health domain, we have 23,000 parallel sentences. Table 1 shows the statistics of the extracted phrases from the target sentences.

3.3 Synthetic parallel corpus using back-translation

We take the target phrases obtained through parsing the target sentences of the training corpus (c.f. Section 3.1) and translate them into source language

Table 2. English phrases from parse tree with their Hindi translation

Phrase in English language	**Phrase in Hindi language**
National Library [NP]	राष्ट्रीय पुस्तकालय
to National Library [PP]	राष्ट्रीय पुस्तकालय के लिए
the Imperial Library [NP]	इंपीरियल पुस्तकालय
of the Imperial Library [PP]	इंपीरियल पुस्तकालय के
the name [NP]	नाम
the name of the Imperial Library [NP]	इंपीरियल पुस्तकालय का नाम
change the name of the Imperial Library to National Library [VP]	के नाम परिवर्तन करने के लिए राष्ट्रीय पुस्तकालय इंपीरियल पुस्तकालय
to change the name of the Imperial Library to National Library [VP]	के नाम में परिवर्तन करने के लिए राष्ट्रीय पुस्तकालय इंपीरियल पुस्तकालय
passed to change the name of the Imperial Library to National Library [VP]	के नाम में परिवर्तन करने के लिए पारित इंपीरियल पुस्तकालय के लिए राष्ट्रीय पुस्तकालय
was passed to change the name of the Imperial Library to National Library [VP]	के लिए किया गया था इंपीरियल पुस्तकालय का नाम को राष्ट्रीय पुस्तकालय
this Act [NP]	यह अधिनियम

using a phrase-based SMT system (descried in Section 3.2) to obtain the source-target parallel phrases. It has been shown in the literature that back-translated parallel corpus, when added to the original parallel corpus, helps in improving the performance of the system even though it may contain incorrect source translation [21].

3.4 Copied parallel corpus

[20] have used additional target side monolingual corpus with dummy source sentences. Following the work of [7], we take the additional target side monolingual data and convert it into a parallel corpus by making each source sentence identical to its target counterpart. We refer to this parallel corpus as the copied corpus. The decoder in encoder-decoder is essentially the RNN language model that also conditioned on source context. However, even though the source and target sentences are the same, NMT model performs better at predicting the next output word given the current output. Suppose, a target phrase is *the Imperial Library*, then using this method at training time, we feed *the Imperial Library* to the encoder and try to predict *the Imperial Library* at the decoder.

3.5 NMT training with synthetic and copied corpus

After obtaining the synthetic parallel data and copied parallel data, we mix these two with the original parallel corpus. It makes our final training data bigger. Statistics regarding the size of the available parallel corpus, number of extracted phrases and the size of synthetic and copied parallel corpus is described in Section 4. We take the system trained using the original training corpus as a baseline. Apart from the original training corpus, we create three new parallel

corpora by adding (i) only synthetic data; (ii) only copied corpus; and (iii) both synthetic and copied data with the original corpus. We shuffle these augmented corpora. Now we have four kinds of parallel corpora for the same language pair. We use these to train an attention based NMT model[1].

4 Datasets

We perform all our experiments for Hindi to English using the parallel corpora from WMT14 [5, 3] and two other domains: *judicial* and *health*.

Table 1 shows the number of phrases generated by the parse trees. These phrases are used to create synthetic and copied data used as additional inputs to the training. Table 4 shows the number of sentences and the number of tokens present in the training data before and after adding synthetic and copied phrases. Size of development set for WMT14, Judicial and Health dataset is 520, 1000 and 1000 respectively. Size of test set for WMT14, Judicial and Health dataset is 2507, 1561 and 1000 respectively.

Table 3. Vocabulary size of each dataset

Dataset	System	Hindi	English
WMT14	*Baseline*, *Baseline+BT*	112,344	104,016
	Baseline+Copied, *Baseline+BT+Copied*	216,360	104,016
Judicial	*Baseline*, *Baseline+BT*	8,357	9,324
	Baseline+Copied, *Baseline+BT+Copied*	17,681	9,324
Health	*Baseline*, *Baseline+BT*	19,996	17,255
	Baseline+Copied, *Baseline+BT+Copied*	37,251	17,255

5 Experimental Setup

We train two types of SMT models: one is English→Hindi (used for back translation) and anther is Hindi→English (used for checking the performance of phrase-based SMT on original training data). For both, we use the Moses [13] toolkit for training. We tokenize and true-case the sentences and remove the sentences with words more than 80 in the preprocessing step. We build 4-gram language model with modified Kneser-Ney smoothing [12] using IRSTLM [8]. For word alignment, we use GIZA++ [17] with grow-diag-final-and heuristics. For other parameters we use the default settings of Moses. The trained systems are tuned

Table 4. Evaluation results with BLEU and METEOR scores of different Hindi→English systems. #TrainingExamples, #SourceTokens and #TargetTokens columns show the training data size *(Increased dataset size is original training samples augmented with syntactic phrases).*

Domain	System	#TrainingExamples	#SourceTokens	#TargetTokens	BLEU	METEOR
WMT14	*PB-SMT*	263,654	3,330,273	3,033,689	8.24	22.63
	Baseline	263,654	3,330,273	3,033,689	7.52	15.39
	Baseline+Copied	859,623	9,663,437	9,366,853	**7.73**	**15.20**
	Baseline+BT	859,623	9,606,960	9,366,853	**10.80**	**18.80**
	Baseline+BT+Copied	1,455,592	15,940,124	15,700,017	**10.41**	**18.26**
Judicial	*PB-SMT*	5,000	129,971	121,430	23.13	29.94
	Baseline	5,000	129,971	121,430	1.87	6.47
	Baseline+Copied	86,308	627,578	619,020	**8.15**	**12.53**
	Baseline+BT	86,308	661,291	619,020	**13.81**	**18.09**
	Baseline+BT+Copied	167,616	1,158,898	1,116,627	**17.50**	**20.77**
Health	*PB-SMT*	23,000	418,853	391,943	20.02	30.10
	Baseline	23,000	418,853	391,943	2.19	8.90
	Baseline+Copied	283,487	1,670,053	1,643,143	**13.04**	**19.74**
	Baseline+BT	283,487	1,716,070	1,643,143	**16.53**	**23.39**
	Baseline+BT+Copied	543,974	2,967,270	2,894,343	**18.16**	**24.83**

using Minimum Error Rate Training (MERT) [16]. For getting synthetic parallel data using back translation, we use the English→Hindi PB-SMT system.

All the NMT models we train are based on attention-based encoder-decoder [1] approach. We train the models at word-level using the Nematus[3] toolkit [19] with the following settings: hidden layer size 512, word embedding dimension 256, max sentence length 80, batch size 40, and learning rate 0.0001. We use the default early-stopping[4] criteria of the Nematus. We use the Adam [11] optimizer.

We consider all the vocabulary words. The vocabulary size for each model is shown in Table 3. For decoding, we use beam width as 5. For other parameters of the Nematus, the default values are used.

6 Results and Analysis

We evaluate our proposed approach using BLEU [18] and METEOR [2] metrics. We summarize the results in Table 4. Some translation outputs produced using our proposed approach are shown in Table 5. Apart from the standard phrase-based SMT system (*PB-SMT*), we train the following NMT systems:

(i) *Baseline*: trained on the original training corpus only.
(ii) *Baseline+Copied*: trained on the original training corpus mixed with copied data.
(iii) *Baseline+BT*: trained on the original training corpus mixed with back-translated data.

[3] https://github.com/EdinburghNLP/nematus
[4] It is based on BLEU score with patience value = 10

Table 5. Some example outputs produced by different proposed systems. *We can observe that Baseline+BT+copied gives better results close to reference.*

	Output #1 (Judicial domain)
Source	यह शिकायतकर्ता द्वारा दायर अभियोजन से मुक्त होने का हकदार था .
Transliteration	yah shikaayatakarta dvaara daayar abhiyojan se mukt hone ka hakadaar tha .
Gloss	it complainant by filed prosecution from exempt be entitled was .
Reference	it was entitled to be exempt from prosecution filed by the complainant .
PB-SMT	it filed by the complainant of free from prosecution was entitled .
Baseline	The petitioner has been filed by the petitioner in this case .
Baseline+Copied	It is filed by the complainant passed by the prosecution .
Baseline+BT	It was entitled to the prosecution filed by the complainant .
Baseline+BT+copied	It was entitled to be free from the prosecution filed by the complainant .
	Output #2 (Health domain)
Source	अमेरिका के एक नामचीन प्रोफेसर के अनुसार एंटीएजिंग क्रीमों के इस्तेमाल से त्वचा विषैले पदार्थों के सम्पर्क में आ सकती है और सूरज की रोशनी से होनेवाले नुकसान की सम्भावना भी बढ़ जाती है ।
Transliteration	amerika ke ek naamacheen prophesar ke anusaar enteeejing kreemon ke istemaal se tvacha vishaile padaarthon ke sampark mein aa sakatee hai aur sooraj kee roshanee se honevaale nukasaan kee sambhaavana bhee badh jaatee hai .
Gloss	America of famous professor according anti-ageing creams of use skin poisonous substances of contact into come can and sun of light from harm of possibility also increase is .
Reference	According to a renowned American Professor, the skin may come in contact with harmful substances by using anti-ageing creams and the chance of skin damage from the sunlight increases .
PB-SMT	according to a famous professor of America skin by the use of anti - ageing creams toxic substances can come in contact with the and the possibility of the harm because of also increases .
Baseline	According to Dr . . . : This disease is used in the form of which there is a lot of benefit by which there is a lot of benefit in the skin and the patient gets destroyed .
Baseline+Copied	According to a famous appendix of America by using skin the skin of the skin becomes strong and the possibility of fear occurring due to the light there is also increase .
Baseline+BT	According to a famous professor of America skin can come in contact with ageing creams and the possibility of harm because of the light of Sun .
Baseline+BT+copied	According to a famous professor of America with the use of ageing creams and skin may come into contact with poisonous rashes and the possibility of the harm because of the accumulation of sun also increases .

(iv) *Baseline+BT+Copied*: trained on the original training corpus mixed with the back-translated and copied data.

BLEU scores for NMT baselines (*Baseline*) are very poor (7.52, 1.87 and 2.19) for *WMT14*, *judicial* and *health* data due to the size of training corpus. Table 4 shows the improvements in terms of BLEU and METEOR scores after adding the phrases in the baseline corpus. We observe that adding phrases using back translation (i.e. *(Baseline+BT)*) and back translation with copied data (i.e. *(Baseline+BT+copied)*) yield higher scores. Outputs produced by *Baseline+copied* are improved in terms of fluency over the baseline (*Baseline*). However, it lacks in adequacy because of the missing translations for some words. Adequacy increases when we use synthetic data with the original training data in *Baseline+BT* as it translates the missing terms that *Baseline+copied* fails to translate. But the lack of proper sequence of phrases reduces the fluency. Further, by adding the copied data to *Baseline+BT* increases the fluency and maintains adequacy of the outputs generated by the *Baseline+BT+copied*.

From the outputs #1 shown in Table 5, we can see that *Baseline* system translates the phrase *"that the petitioner has been filed by the petitioner"* wrongly, and generates some phrases (e.g., *the petitioner*) multiple times. Though fluency is improved by *Baseline+Copied*, some words (e.g., मुक्त-*Exempt,* हकदार-*Entitled*) are not translated. Further, the *Baseline+BT* system translates some missing words, but still lacks in ordering of the phrases and also drops some phrases (*be free from*). The *Baseline+BT+copied* system translates all the missing words and phrases and improves the fluency by maintaining the phrase order. Similar observation regarding the improvement in translation quality and behavior of our models can be seen in Output #2 where *Baseline+copied* is better at fluency compared to the *Baseline*. However, it drops the translation of few terms like *'professor', 'creams'* etc. and repeats the terms like *"the skin"* unnecessarily. The *Baseline+BT* improves adequacy by translating previously missing words (not all) but lacks in maintaining proper phrase ordering. *Baseline+BT+copied* maintains adequacy and fluency substantially except translating a word as *'ageing'* instead of *'anti-ageing'*. One more important observation should be made here that in both the examples PB-SMT translates the words correctly but badly fails in fluency because of the wrong ordering of phrases but because of its end to end nature NMT preserves the fluency and with sufficient amount of training data it improves the translation quality in context of adequacy too.

We have done significance tests and the results are significant with 95% confidence level (with p=0 which is < 0.05) and 99% confidence level (with p=0 which is < 0.01). Analysis shows that performance improvement in our proposed model is statistically significant over the baselines.

7 Conclusion

Training an NMT system requires large training set, which is not easily available for many languages. In this work, we proposed a technique to improve the translation quality of a NMT systems under the low-resource scenario by injecting

syntactic phrases extracted from the parse trees of target-side training data. We extract the noun, verb and prepositional phrases from target sentences of the training data, and perform back translation to generate phrases for the source side. We use these synthetic phrase pairs as additional training data. We empirically showed that our method of augmenting original training data, without using any additional monolingual data, can improve the baseline NMT system for Hindi→English translation in several domains. In future, we will investigate the effectiveness of this approach for the other low-resource Indian languages and domains. We will extract syntactic phrases using parser from both the source and target sides to analyze its impact on the translation quality.

8 Acknowledgement

We gratefully acknowledge TDIL, MeitY who supported this research work under development of the project "Hindi to English machine translation for judicial domain".

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. International Conference on Learning Representation(ICLR) (2015)
2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics (2005), http://www.aclweb.org/anthology/W05-0909
3. Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., et al.: Findings of the 2014 workshop on statistical machine translation. In: Proceedings of the ninth workshop on statistical machine translation. pp. 12–58 (2014)
4. Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A.J., Koehn, P., Logacheva, V., Monz, C., et al.: Findings of the 2016 conference on machine translation. In: ACL 2016 First Conference on Machine Translation (WMT16). pp. 131–198. The Association for Computational Linguistics (2016)
5. Bojar, O., Diatka, V., Rychlỳ, P., Stranák, P., Suchomel, V., Tamchyna, A., Zeman, D.: Hindencorp-hindi-english and hindi-only corpus for machine translation. In: LREC. pp. 3550–3555 (2014)
6. Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., Brunelle, P., Coquard, A., Deng, Y., et al.: Systran's pure neural machine translation systems. arXiv preprint arXiv:1610.05540 (2016)
7. Currey, A., Barone, A.V.M., Heafield, K.: Copied monolingual data improves low-resource neural machine translation. In: Proceedings of the Second Conference on Machine Translation. pp. 148–156 (2017)
8. Federico, M., Bertoldi, N., Cettolo, M.: Irstlm: an open source toolkit for handling large scale language models. In: Ninth Annual Conference of the International Speech Communication Association (2008)

9. Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.C., Bougares, F., Schwenk, H., Bengio, Y.: On using monolingual corpora in neural machine translation. arXiv preprint arXiv:1503.03535 (2015)
10. Junczys-Dowmunt, M., Dwojak, T., Hoang, H.: Is neural machine translation ready for deployment? a case study on 30 translation directions. In: In Proceedings of the International Workshop on Spoken Language Translation (IWSLT) (2016)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representation(ICLR) (2015)
12. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on. vol. 1, pp. 181–184. IEEE (1995)
13. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. pp. 177–180. Association for Computational Linguistics (2007)
14. Koehn, P., Knowles, R.: Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation. pp. 28–39. Association for Computational Linguistics, Vancouver (August 2017), http://www.aclweb.org/anthology/W17-3204
15. Koehn, P., Och, F.J., Marcu, D.: 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. pp. 48–54. Association for Computational Linguistics (2003)
16. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. pp. 160–167. Association for Computational Linguistics (2003)
17. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational linguistics **29**(1), 19–51 (2003)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Philadelphia, Pennsylvania (2002)
19. Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Barone, A.V.M., Mokry, J., et al.: Nematus: a toolkit for neural machine translation. arXiv preprint arXiv:1703.04357 (2017)
20. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709 (2015)
21. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany (2016)
22. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR **abs/1609.08144** (2016), http://arxiv.org/abs/1609.08144

23. Zhang, J., Zong, C.: Exploiting source-side monolingual data in neural machine translation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1535–1545 (2016)