Data Augmentation for Low-Resource Neural Machine Translation

Marzieh Fadaee **Arianna Bisazza** **Christof Monz**
Informatics Institute, University of Amsterdam
Science Park 904, 1098 XH Amsterdam, The Netherlands
{m.fadaee,a.bisazza,c.monz}@uva.nl

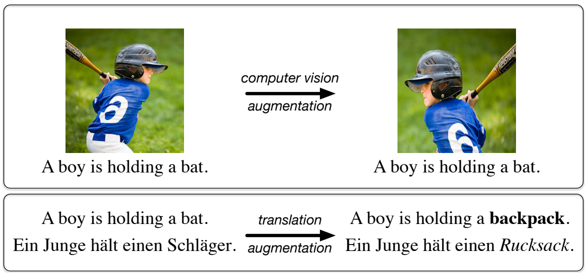
Abstract

The quality of a Neural Machine Translation system depends substantially on the availability of sizable parallel corpora. For low-resource language pairs this is not the case, resulting in poor translation quality. Inspired by work in computer vision, we propose a novel data augmentation approach that targets low-frequency words by generating new sentence pairs containing rare words in new, synthetically created contexts. Experimental results on simulated low-resource settings show that our method improves translation quality by up to 2.9 BLEU points over the baseline and up to 3.2 BLEU over back-translation.

1 Introduction

In computer vision, data augmentation techniques are widely used to increase robustness and improve learning of objects with a limited number of training examples. In image processing the training data is augmented by, for instance, horizontally flipping, random cropping, tilting, and altering the RGB channels of the original images (Krizhevsky et al., 2012; Chatfield et al., 2014). Since the content of the new image is still the same, the label of the original image is preserved (see top of Figure 1). While data augmentation has become a standard technique to train deep networks for image processing, it is not a common practice in training networks for NLP tasks such as Machine Translation.

Neural Machine Translation (NMT) (Bahdanau et al., 2015; Sutskever et al., 2014; Cho et al., 2014) is a sequence-to-sequence architecture where an encoder builds up a representation of the source sentence and a decoder, using the previous LSTM hidden states and an attention mechanism, generates the target translation.



A boy is holding a bat.
Ein Junge hält einen Schläger.

A boy is holding a **backpack**.
Ein Junge hält einen *Rucksack*.

Figure 1: Top: flip and crop, two label-preserving data augmentation techniques in computer vision. Bottom: Altering one sentence in a parallel corpus requires changing its translation.

To train a model with reliable parameter estimations, these networks require numerous instances of sentence translation pairs with words occurring in diverse contexts, which is typically not available in low-resource language pairs. As a result NMT falls short of reaching state-of-the-art performances for these language pairs (Zoph et al., 2016). The solution is to either manually annotate more data or perform unsupervised data augmentation. Since manual annotation of data is time-consuming, data augmentation for low-resource language pairs is a more viable approach. Recently Sennrich et al. (2016a) proposed a method to back-translate sentences from monolingual data and augment the bitext with the resulting pseudo parallel corpora.

In this paper, we propose a simple yet effective approach, translation data augmentation (TDA), that augments the training data by altering existing sentences in the parallel corpus, similar in spirit to the data augmentation approaches in computer vision (see Figure 1). In order for the augmentation process in this scenario to be label-preserving, any change to a sentence in one language must pre-

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 567–573
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-2090

serve the meaning of the sentence, requiring sentential paraphrasing systems which are not available for many language pairs. Instead, we propose a weaker notion of label preservation that allows to alter both source and target sentences at the same time as long as they remain translations of each other.

While our approach allows us to augment data in numerous ways, we focus on augmenting instances involving low-frequency words, because the parameter estimation of rare words is challenging, and further exacerbated in a low-resource setting. We simulate a low-resource setting as done in the literature (Marton et al., 2009; Duong et al., 2015) and obtain substantial improvements for translating English→German and German→English.

2 Translation Data Augmentation

Given a source and target sentence pair (S,T), we want to alter it in a way that preserves the semantic equivalence between S and T while diversifying as much as possible the training examples. A number of ways to do this can be envisaged, as for example paraphrasing (parts of) S or T. Paraphrasing, however, is by itself a difficult task and is not guaranteed to bring useful new information into the training data. We choose instead to focus on a subset of the vocabulary that we know to be poorly modeled by our baseline NMT system, namely words that occur rarely in the parallel corpus. Thus, the goal of our data augmentation technique is to provide novel contexts for rare words. To achieve this we search for contexts where a common word can be replaced by a rare word and consequently replace its corresponding word in the other language by that rare word's translation:

original pair	augmented pair
$S: s_1, ..., s_i, ..., s_n$	$S': s_1, ..., s'_i, ..., s_n$
$T: t_1, ..., t_j, ..., t_m$	$T': t_1, ..., t'_j, ..., t_m$

where t_j is a translation of s_i and word-aligned to s_i. Plausible substitutions are those that result in a fluent and grammatical sentence but do not necessarily maintain its semantic content. As an example, the rare word *motorbike* can be substituted in different contexts:

Sentence [original / **substituted**]	Plausible
My sister drives a [car / **motorbike**]	yes
My uncle sold his [house / **motorbike**]	yes
Alice waters the [plant / **motorbike**]	no (semantics)
John bought two [shirts / **motorbike**]	no (syntax)

Implausible substitutions need to be ruled out during data augmentation. To this end, rather than relying on linguistic resources which are not available for many languages, we rely on LSTM language models (LM) (Hochreiter and Schmidhuber, 1997; Jozefowicz et al., 2015) trained on large amounts of monolingual data in both forward and backward directions.

Our data augmentation method involves the following steps:

Targeted words selection: Following common practice, our NMT system limits its vocabulary V to the v most common words observed in the training corpus. We select the words in V that have fewer than R occurrences and use this as our targeted rare word list V_R.

Rare word substitution: If the LM suggests a rare substitution in a particular context, we replace that word and add the new sentence to the training data. Formally, given a sentence pair (S, T) and a position i in S we compute the probability distribution over V by the forward and backward LMs and select rare word substitutions $\mathcal{C}$ as follows:

$$\begin{aligned}\overrightarrow{\mathcal{C}} &= \{s'_i \in V_R : \text{topK}\, P_{\mathit{ForwardLM}_S}(s'_i \mid s_1^{i-1})\}\\ \overleftarrow{\mathcal{C}} &= \{s'_i \in V_R : \text{topK}\, P_{\mathit{BackwardLM}_S}(s'_i \mid s_n^{i+1})\}\\ \mathcal{C} &= \{s'_i \mid s'_i \in \overrightarrow{\mathcal{C}} \wedge s'_i \in \overleftarrow{\mathcal{C}}\}\end{aligned}$$

where topK returns the K words with highest conditional probability according to the context. The selected substitutions s'_i, are used to replace the original word and generate a new sentence.

Translation selection: Using automatic word alignments[1] trained over the bitext, we replace the translation of word s_i in T by the translation of its substitution s'_i. Following a common practice in statistical MT, the optimal translation t'_j is chosen by multiplying direct and inverse lexical translation probabilities with the LM probability of the translation in context:

$$t'_j = \underset{t \in \mathit{trans}(s'_i)}{\arg\max}\, P(s'_i \mid t) P(t \mid s'_i) P_{\mathit{LM}_T}(t \mid t_1^{j-1})$$

If no translation candidate is found because the word is unaligned or because the LM probability

[1] We use fast-align (Dyer et al., 2013) to extract word alignments and a bilingual lexicon with lexical translation probabilities from the low-resource bitext.

is less than a certain threshold, the augmented sentence is discarded. This reduces the risk of generating sentence pairs that are semantically or syntactically incorrect.

Sampling: We loop over the original parallel corpus multiple times, sampling substitution positions, i, in each sentence and making sure that each rare word gets augmented at most N times so that a large number of rare words can be affected. We stop when no new sentences are generated in one pass of the training data.

Table 1 provides some examples resulting from our augmentation procedure. While using a large LM to substitute words with rare words mostly results in grammatical sentences, this does not mean that the meaning of the original sentence is preserved. Note that meaning preservation is not an objective of our approach.

Two translation data augmentation (TDA) setups are considered: only one word per sentence can be replaced (TDA$_{r=1}$), or multiple words per sentence can be replaced, with the condition that any two replaced words are at least five positions apart (TDA$_{r\geqslant 1}$). The latter incurs a higher risk of introducing noisy sentences but has the potential to positively affect more rare words within the same amount of augmented data. We evaluate both setups in the following section.

En:	I had been told that you would [not / **voluntarily**] be speaking today.
De:	mir wurde signalisiert, sie würden heute [nicht / *freiwillig*] sprechen.
En:	the present situation is [indefensible / **confusing**] and completely unacceptable to the commission.
De:	die situation sei [unhaltbar / *verwirrend*] und für die kommission gänzlich unannehmbar.
En:	... agree wholeheartedly with the institution of an ad hoc delegation of parliament on the turkish [prison / **missile**] system.
De:	... ad-hoc delegation des parlaments für das regime in den türkischen [gefängnissen / *flugwaffen*] voll und ganz zustimmen.

Table 1: Examples of augmented data with highlighted [original / **substituted**] and [original / *translated*] words.

3 Evaluation

In this section we evaluate the utility of our approach in a simulated low-resource NMT scenario.

3.1 Data and experimental setup

To simulate a low-resource setting we randomly sample 10% of the English↔German WMT15 training data and report results on newstest 2014, 2015, and 2016 (Bojar et al., 2016). For reference we also provide the result of our baseline system on the full data.

As NMT system we use a 4-layer attention-based encoder-decoder model as described in (Luong et al., 2015) trained with hidden dimension 1000, batch size 80 for 20 epochs. In all experiments the NMT vocabulary is limited to the most common 30K words in both languages. Note that data augmentation does not introduce new words to the vocabulary. In all experiments we preprocess source and target language data with Byte-pair encoding (BPE) (Sennrich et al., 2016b) using 30K merge operations. In the augmentation experiments BPE is performed after data augmentation.

For the LMs needed for data augmentation, we train 2-layer LSTM networks in forward and backward directions on the monolingual data provided for the same task (3.5B and 0.9B tokens in English and German respectively) with embedding size 64 and hidden size 128. We set the rare word threshold R to 100, top K words to 1000 and maximum number N of augmentations per rare word to 500. In all experiments we use the English LM for the rare word substitutions, and the German LM to choose the optimal word translation in context. Since our approach is not label preserving we only perform augmentation during training and do not alter source sentences during testing.

We also compare our approach to Sennrich et al. (2016a) by back-translating monolingual data and adding it to the parallel training data. Specifically, we back-translate sentences from the target side of WMT'15 that are not included in our low-resource baseline with two settings: keeping a one-to-one ratio of back-translated versus original data (1 : 1) following the authors' suggestion, or using three times more back-translated data (3 : 1).

We measure translation quality by single-reference case-insensitive BLEU (Papineni et al., 2002) computed with the `multi-bleu.perl` script from Moses.

3.2 Results

All translation results are displayed in Table 2. As expected, the low-resource baseline performs much worse than the full data system, re-iterating

		De-En			En-De		
Model	Data	test2014	test2015	test2016	test2014	test2015	test2016
Full data (ceiling)	3.9M	21.1	22.0	26.9	17.0	18.5	21.7
Baseline	371K	10.6	11.3	13.1	8.2	9.2	11.0
Back-translation$_{1:1}$	731K	11.4 (+0.8)$^{\blacktriangle}$	12.2 (+0.9)$^{\blacktriangle}$	14.6 (+1.5)$^{\blacktriangle}$	9.0 (+0.8)$^{\blacktriangle}$	10.4 (+1.2)$^{\blacktriangle}$	12.0 (+1.0)$^{\blacktriangle}$
Back-translation$_{3:1}$	1.5M	11.2 (+0.6)	11.2 (–0.1)	13.3 (+0.2)	7.8 (–0.4)	9.4 (+0.2)	10.7 (–0.3)
TDA$_{r=1}$	4.5M	11.9 (+1.3)$^{\blacktriangle,-}$	13.4 (+2.1)$^{\blacktriangle,\blacktriangle}$	15.2 (+2.1)$^{\blacktriangle,\blacktriangle}$	10.4 (+2.2)$^{\blacktriangle,\blacktriangle}$	11.2 (+2.0)$^{\blacktriangle,\blacktriangle}$	13.5 (+2.5)$^{\blacktriangle,\blacktriangle}$
TDA$_{r\geqslant1}$	6M	**12.6** (+2.0)$^{\blacktriangle,\blacktriangle}$	**13.7** (+2.4)$^{\blacktriangle,\blacktriangle}$	**15.4** (+2.3)$^{\blacktriangle,\blacktriangle}$	**10.7** (+2.5)$^{\blacktriangle,\blacktriangle}$	**11.5** (+2.3)$^{\blacktriangle,\blacktriangle}$	**13.9** (+2.9)$^{\blacktriangle,\blacktriangle}$
Oversampling	6M	11.9 (+1.3)$^{\blacktriangle,-}$	12.9 (+1.6)$^{\blacktriangle,\vartriangle}$	15.0 (+1.9)$^{\blacktriangle,-}$	9.7 (+1.5)$^{\blacktriangle,\vartriangle}$	10.7 (+1.5)$^{\blacktriangle,-}$	12.6 (+1.6)$^{\blacktriangle,-}$

Table 2: Translation performance (BLEU) on German-English and English-German WMT test sets (newstest2014, 2015, and 2016) in a simulated low-resource setting. Back-translation refers to the work of Sennrich et al. (2016a). Statistically significant improvements are marked $^{\blacktriangle}$ at the $p < .01$ and $^{\vartriangle}$ at the $p < .05$ level, with the first superscript referring to baseline and the second to back-translation$_{1:1}$.

the importance of sizable training data for NMT. Next we observe that both back-translation and our proposed TDA method significantly improve translation quality. However TDA obtains the best results overall and significantly outperforms back-translation in all test sets. This is an important finding considering that our method involves only minor modifications to the original training sentences and does not involve any costly translation process. Improvements are consistent across both translation directions, regardless of whether rare word substitutions are first applied to the source or to the target side.

We also observe that altering multiple words in a sentence performs slightly better than altering only one. This indicates that addressing more rare words is preferable even though the augmented sentences are likely to be noisier.

To verify that the gains are actually due to the rare word substitutions and not just to the repetition of part of the training data, we perform a final experiment where each sentence pair selected for augmentation is added to the training data *unchanged* (Oversampling in Table 2). Surprisingly, we find that this simple form of sampled data replication outperforms both baseline and back-translation systems,[2] while TDA$_{r\geqslant1}$ remains the best performing system overall.

We also observe that the system trained on augmented data tends to generate longer translations. Averaging on all test sets, the length of translations generated by the baseline is 0.88 of the average reference length, while for TDA$_{r=1}$ and TDA$_{r\geqslant1}$ it is 0.95 and 0.94, respectively. We attribute this effect to the ability of the TDA-trained system to generate translations for rare words that were left untranslated by the baseline system.

4 Analysis of the Results

A desired effect of our method is to increase the number of correct rare words generated by the NMT system at test time.

To examine the impact of augmenting the training data by creating contexts for rare words on the target side, Table 3 provides an example for German→English translation. We see that the baseline model is not able to generate the rare word *centimetres* as a correct translation of the German word *zentimeter* . However, this word is not rare in the training data of the TDA$_{r\geqslant1}$ model after augmentation and is generated during translation. Table 3 also provides several instances of augmented training sentences targeting the word *centimetres*. Note that even though some augmented sentences are nonsensical (e.g. *the speed limit is five centimetres per hour*), the NMT system still benefits from the new context for the rare word and is able to generate it during testing.

Figure 2 demonstrates that this is indeed the case for many words: the number of rare words occurring in the reference translation ($V_R \cap V_{ref}$) is three times larger in the TDA system output than in the baseline output. One can also see that this increase is a direct effect of TDA as most of the rare words are not 'rare' anymore in the augmented data, i.e., they were augmented sufficiently many times to occur more than 100 times (see hatched pattern in Figure 2). Note that during the experiments we did not use any information from the evaluation sets.

To gauge the impact of augmenting the contexts for rare words on the source side, we examine normalized attention scores of these words before and after augmentation. When translating

[2] Note that this effect cannot be achieved by simply continuing the baseline training for up to 50 epochs.

Source	der tunnel hat einen querschnitt von 1,20 meter höhe und 90 zentimeter breite .
Baseline translation	the wine consists of about 1,20 m and 90 of the canal .
$\text{TDA}_{r\geqslant 1}$ translation	the tunnel has a UNK measuring meters 1.20 metres high and 90 **centimetres** wide .
Reference	the tunnel has a cross - section measuring 1.20 metres high and 90 centimetres across .
Examples of augmented data for the word *centimetres*	• the average speed of cars and buses is therefore around 20 [kilometres / **centimetres**] per hour .
	• grab crane in special terminals for handling capacities of up to 1,800 [tonnes / **centimetres**] per hour .
	• all suites and rooms are very spacious and measure between 50 and 70 [m / **centimetres**]
	• all we have to do is lower the speed limit everywhere to five [kilometers / **centimetres**] per hour .

Table 3: An example from newstest2014 illustrating the effect of augmenting rare words on generation during test time. The translation of the baseline does not include the rare word *centimetres*, however, the translation of our TDA model generates the rare word and produces a more fluent sentence. Instances of the augmentation of the word *centimetres* in training data are also provided.

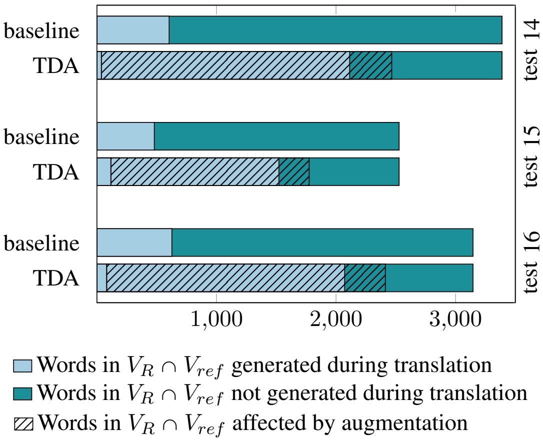


Figure 2: Effect of TDA on the number of unique rare words generated during De→En translation. V_R is the set of rare words targeted by $\text{TDA}_{r\geqslant 1}$ and V_{ref} the reference translation vocabulary.

English→German with our TDA model, the attention scores for rare words on the source side are on average 8.8% higher than when translating with the baseline model. This suggests that having more accurate representations of rare words increases the model's confidence to attend to these words when encountered during test time.

En:	registered users will receive the UNK newsletter free [of / **yearly**] charge.
De:	registrierte user erhalten zudem regelmäßig [den / *jährlich*] markenticker newsletter.
En:	the personal contact is [essential / **entrusted**] to us
De:	persönliche kontakt ist uns sehr [wichtig / *betraut*]

Table 4: Examples of incorrectly augmented data with highlighted [original / **substituted**] and [original / *translated*] words.

Finally Table 4 provides examples of cases where augmentation results in incorrect sentences. In the first example, the sentence is ungrammatical after substitution (*of / yearly*), which can be the result of choosing substitutions with low probabilities from the English LM topK suggestions.

Errors can also occur during translation selection, as in the second example where *betraut* is an acceptable translation of *entrusted* but would require a rephrasing of the German sentence to be grammatically correct. Problems of this kind can be attributed to the German LM, but also to the lack of a more suitable translation in the lexicon extracted from the bitext. Interestingly, this noise seems to affect NMT only to a limited extent.

5 Conclusion

We have proposed a simple but effective approach to augment the training data of Neural Machine Translation for low-resource language pairs. By leveraging language models trained on large amounts of monolingual data, we generate new sentence pairs containing rare words in new, synthetically created contexts. We show that this approach leads to generating more rare words during translation and, consequently, to higher translation quality. In particular we report substantial improvements in simulated low-resource English→German and German→English settings, outperforming another recently proposed data augmentation technique.

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project numbers 639.022.213 and 639.021.646, and a Google Faculty Research Award. We also thank NVIDIA for their hardware support, Ke Tran for providing the neural machine translation baseline system, and the anonymous reviewers for their helpful comments.

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)*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 131–198. http://www.aclweb.org/anthology/W/W16/W16-2301.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference*. BMVA Press. https://doi.org/http://dx.doi.org/10.5244/C.28.6.

Kyunghyun Cho, B van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 2014*.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. A neural network model for low-resource universal dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 339–348. http://aclweb.org/anthology/D15-1040.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, pages 644–648. http://www.aclweb.org/anthology/N13-1073.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. 2015. An empirical exploration of recurrent network architectures. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, Lille, France, volume 37 of *Proceedings of Machine Learning Research*, pages 2342–2350. http://proceedings.mlr.press/v37/jozefowicz15.pdf.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc., pages 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1412–1421. http://aclweb.org/anthology/D15-1166.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pages 381–390. http://www.aclweb.org/anthology/D/D09/D09-1040.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 86–96. http://www.aclweb.org/anthology/P16-1009.

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 (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1715–1725. http://www.aclweb.org/anthology/P16-1162.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pages 3104–3112. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1568–1575. https://aclweb.org/anthology/D16-1163.