

# Revisiting Low-Resource Neural Machine Translation: A Case Study

Rico Sennrich<sup>1,2</sup> Biao Zhang<sup>1</sup>

<sup>1</sup>School of Informatics, University of Edinburgh  
rico.sennrich@ed.ac.uk, B.Zhang@ed.ac.uk

<sup>2</sup>Institute of Computational Linguistics, University of Zurich

## Abstract

It has been shown that the performance of neural machine translation (NMT) drops starkly in low-resource conditions, underperforming phrase-based statistical machine translation (PBSMT) and requiring large amounts of auxiliary data to achieve competitive results. In this paper, we re-assess the validity of these results, arguing that they are the result of lack of system adaptation to low-resource settings. We discuss some pitfalls to be aware of when training low-resource NMT systems, and recent techniques that have shown to be especially helpful in low-resource settings, resulting in a set of best practices for low-resource NMT. In our experiments on German–English with different amounts of IWSLT14 training data, we show that, without the use of any auxiliary monolingual or multilingual data, an optimized NMT system can outperform PBSMT with far less data than previously claimed. We also apply these techniques to a low-resource Korean–English dataset, surpassing previously reported results by 4 BLEU.

## 1 Introduction

While neural machine translation (NMT) has achieved impressive performance in high-resource data conditions, becoming dominant in the field (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017), recent research has argued that these models are highly data-inefficient, and underperform phrase-based statistical machine translation (PBSMT) or unsupervised methods in low-data conditions (Koehn and Knowles, 2017; Lample et al., 2018b). In this paper, we re-assess the validity of these results, arguing that they are the result of lack of system adaptation to low-resource settings. Our main contributions are as follows:

- we explore best practices for low-resource

BLEU Scores with Varying Amounts of Training Data

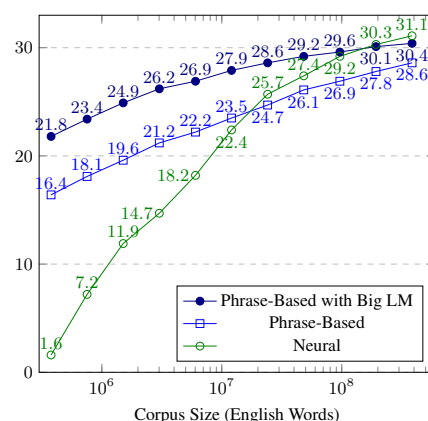


Figure 1: quality of PBSMT and NMT in low-resource conditions according to (Koehn and Knowles, 2017).

NMT, evaluating their importance with ablation studies.

- we reproduce a comparison of NMT and PBSMT in different data conditions, showing that when following our best practices, NMT outperforms PBSMT with as little as 100 000 words of parallel training data.

## 2 Related Work

### 2.1 Low-Resource Translation Quality Compared Across Systems

Figure 1 reproduces a plot by Koehn and Knowles (2017) which shows that their NMT system only outperforms their PBSMT system when more than 100 million words (approx. 5 million sentences) of parallel training data are available. Results shown by Lample et al. (2018b) are similar, showing that unsupervised NMT outperforms supervised systems if few parallel resources are available. In both papers, NMT systems are trained with hyperparameters that are typical for high-resource set-

tings, and the authors did not tune hyperparameters, or change network architectures, to optimize NMT for low-resource conditions.

## 2.2 Improving Low-Resource Neural Machine Translation

The bulk of research on low-resource NMT has focused on exploiting monolingual data, or parallel data involving other language pairs. Methods to improve NMT with monolingual data range from the integration of a separately trained language model (Gülçehre et al., 2015) to the training of parts of the NMT model with additional objectives, including a language modelling objective (Gülçehre et al., 2015; Sennrich et al., 2016b; Ramachandran et al., 2017), an autoencoding objective (Luong et al., 2016; Currey et al., 2017), or a round-trip objective, where the model is trained to predict monolingual (target-side) training data that has been back-translated into the source language (Sennrich et al., 2016b; He et al., 2016; Cheng et al., 2016). As an extreme case, models that rely exclusively on monolingual data have been shown to work (Artetxe et al., 2018b; Lample et al., 2018a; Artetxe et al., 2018a; Lample et al., 2018b). Similarly, parallel data from other language pairs can be used to pre-train the network or jointly learn representations (Zoph et al., 2016; Chen et al., 2017; Nguyen and Chiang, 2017; Neubig and Hu, 2018; Gu et al., 2018a,b; Kocmi and Bojar, 2018).

While semi-supervised and unsupervised approaches have been shown to be very effective for some language pairs, their effectiveness depends on the availability of large amounts of suitable auxiliary data, and other conditions being met. For example, the effectiveness of unsupervised methods is impaired when languages are morphologically different, or when training domains do not match (Søgaard et al., 2018).

More broadly, this line of research still accepts the premise that NMT models are data-inefficient and require large amounts of auxiliary data to train. In this work, we want to re-visit this point, and will focus on techniques to make more efficient use of small amounts of parallel training data. Low-resource NMT without auxiliary data has received less attention; work in this direction includes (Östling and Tiedemann, 2017; Nguyen and Chiang, 2018).

## 3 Methods for Low-Resource Neural Machine Translation

### 3.1 Mainstream Improvements

We consider the hyperparameters used by Koehn and Knowles (2017) to be our baseline. This baseline does not make use of various advances in NMT architectures and training tricks. In contrast to the baseline, we use a BiDeep RNN architecture (Miceli Barone et al., 2017), label smoothing (Szegedy et al., 2016), dropout (Srivastava et al., 2014), word dropout (Sennrich et al., 2016a), layer normalization (Ba et al., 2016) and tied embeddings (Press and Wolf, 2017).

### 3.2 Language Representation

Subword representations such as BPE (Sennrich et al., 2016c) have become a popular choice to achieve open-vocabulary translation. BPE has one hyperparameter, the number of merge operations, which determines the size of the final vocabulary. For high-resource settings, the effect of vocabulary size on translation quality is relatively small; Haddow et al. (2018) report mixed results when comparing vocabularies of 30k and 90k subwords.

In low-resource settings, large vocabularies result in low-frequency (sub)words being represented as atomic units at training time, and the ability to learn good high-dimensional representations of these is doubtful. Sennrich et al. (2017a) propose a minimum frequency threshold for subword units, and splitting any less frequent subword into smaller units or characters. We expect that such a threshold reduces the need to carefully tune the vocabulary size to the dataset, leading to more aggressive segmentation on smaller datasets.<sup>1</sup>

### 3.3 Hyperparameter Tuning

Due to long training times, hyperparameters are hard to optimize by grid search, and are often re-used across experiments. However, best practices differ between high-resource and low-resource settings. While the trend in high-resource settings is towards using larger and deeper models, Nguyen and Chiang (2018) use smaller and fewer layers for smaller datasets. Previous work has argued for larger batch sizes in NMT (Morphita et al., 2017; Neishi et al., 2017), but we

<sup>1</sup>In related work, Cherry et al. (2018) have shown that, given deep encoders and decoders, character-level models can outperform other subword segmentations. In preliminary experiments, a character-level model performed poorly in our low-resource setting.

find that using smaller batches is beneficial in low-resource settings. More aggressive dropout, including dropping whole words at random (Gal and Ghahramani, 2016), is also likely to be more important. We report results on a narrow hyperparameter search guided by previous work and our own intuition.

### 3.4 Lexical Model

Finally, we implement and test the lexical model by Nguyen and Chiang (2018), which has been shown to be beneficial in low-data conditions. The core idea is to train a simple feed-forward network, the lexical model, jointly with the original attentional NMT model. The input of the lexical model at time step  $t$  is the weighted average of source embeddings  $f$  (the attention weights  $a$  are shared with the main model). After a feedforward layer (with skip connection), the lexical model’s output  $h_t^l$  is combined with the original model’s hidden state  $h_t^o$  before softmax computation.

$$f_t^l = \tanh \sum_s a_t(s) f_s$$

$$h_t^l = \tanh(W f_t^l) + f_t^l$$

$$p(y_t | y_{<t}, x) = \text{softmax}(W^o h_t^o + b^o + W^l h_t^l + b^l)$$

Our implementation adds dropout and layer normalization to the lexical model.<sup>2</sup>

## 4 Experiments

### 4.1 Data and Preprocessing

We use the TED data from the IWSLT 2014 German→English shared translation task (Cettolo et al., 2014). We use the same data cleanup and train/dev split as Ranzato et al. (2016), resulting in 159 000 parallel sentences of training data, and 7584 for development.

As a second language pair, we evaluate our systems on a Korean–English dataset<sup>3</sup> with around 90 000 parallel sentences of training data, 1000 for development, and 2000 for testing.

For both PBSMT and NMT, we apply the same tokenization and truecasing using Moses scripts. For NMT, we also learn BPE subword segmentation with 30 000 merge operations, shared between German and English, and independently for Korean→English.

<sup>2</sup>Implementation released in Nematus:  
<https://github.com/EdinburghNLP/nematus>

<sup>3</sup><https://sites.google.com/site/koreanparalleldata/>

		subword vocabulary	
sentences	words (EN)	DE/KO	EN
DE→EN			
159 000	3 220 000	18 870	13 830
80 000	1 610 000	9850	7740
40 000	810 000	7470	5950
20 000	400 000	5640	4530
10 000	200 000	3760	3110
5000	100 000	2380	1990
KO→EN			
94 000	2 300 000	32 082	16 006

Table 1: Training corpus size and subword vocabulary size for different subsets of IWSLT14 DE→EN data, and for KO→EN data.

To simulate different amounts of training resources, we randomly subsample the IWSLT training corpus 5 times, discarding half of the data at each step. Truecaser and BPE segmentation are learned on the full training corpus; as one of our experiments, we set the frequency threshold for subword units to 10 in each subcorpus (see 3.2). Table 1 shows statistics for each subcorpus, including the subword vocabulary.

Translation outputs are detruccased, detokenized, and compared against the reference with cased BLEU using sacreBLEU (Papineni et al., 2002; Post, 2018).<sup>4</sup> Like Ranzato et al. (2016), we report BLEU on the concatenated dev sets for IWSLT 2014 (tst2010, tst2011, tst2012, dev2010, dev2012).

### 4.2 PBSMT Baseline

We use Moses (Koehn et al., 2007) to train a PBSMT system. We use MGIZA (Gao and Vogel, 2008) to train word alignments, and Implz (Heafield et al., 2013) for a 5-gram LM. Feature weights are optimized on the dev set to maximize BLEU with batch MIRA (Cherry and Foster, 2012) – we perform multiple runs where indicated. Unlike Koehn and Knowles (2017), we do not use extra data for the LM. Both PBSMT and NMT can benefit from monolingual data, so the availability of monolingual data is no longer an exclusive advantage of PBSMT (see 2.2).

ID	system	BLEU	
		100k	3.2M
1	phrase-based SMT	$15.87 \pm 0.19$	$26.60 \pm 0.00$
2	NMT baseline	$0.00 \pm 0.00$	$25.70 \pm 0.33$
3	2 + "mainstream improvements" (dropout, tied embeddings, layer normalization, bideep RNN, label smoothing)	$7.20 \pm 0.62$	$31.93 \pm 0.05$
4	3 + reduce BPE vocabulary (14k $\rightarrow$ 2k symbols)	$12.10 \pm 0.16$	-
5	4 + reduce batch size (4k $\rightarrow$ 1k tokens)	$12.40 \pm 0.08$	$31.97 \pm 0.26$
6	5 + lexical model	$13.03 \pm 0.49$	$31.80 \pm 0.22$
7	5 + aggressive (word) dropout	$15.87 \pm 0.09$	<b><math>33.60 \pm 0.14</math></b>
8	7 + other hyperparameter tuning (learning rate, model depth, label smoothing rate)	<b><math>16.57 \pm 0.26</math></b>	$32.80 \pm 0.08$
9	8 + lexical model	$16.10 \pm 0.29$	$33.30 \pm 0.08$

Table 2: German $\rightarrow$ English IWSLT results for training corpus size of 100k words and 3.2M words (full corpus). Mean and standard deviation of three training runs reported.

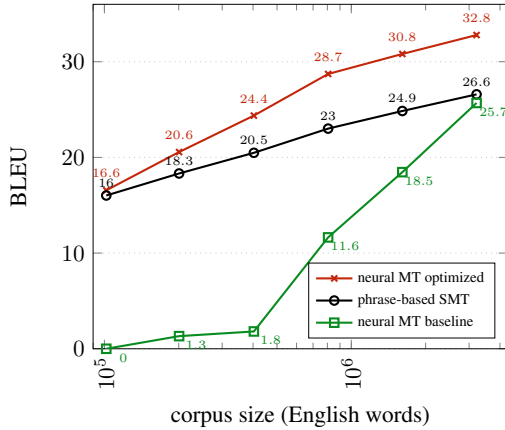


Figure 2: German $\rightarrow$ English learning curve, showing BLEU as a function of the amount of parallel training data, for PBSMT and NMT.

### 4.3 NMT Systems

We train neural systems with Nematus (Sennrich et al., 2017b). Our baseline mostly follows the settings in (Koehn and Knowles, 2017); we use adam (Kingma and Ba, 2015) and perform early stopping based on dev set BLEU. We express our batch size in number of tokens, and set it to 4000 in the baseline (comparable to a batch size of 80 sentences used in previous work).

We subsequently add the methods described in section 3, namely the bideep RNN, label smoothing, dropout, tied embeddings, layer normalization, changes to the BPE vocabulary size, batch

size, model depth, regularization parameters and learning rate. Detailed hyperparameters are reported in Appendix A.

## 5 Results

Table 2 shows the effect of adding different methods to the baseline NMT system, on the ultra-low data condition (100k words of training data) and the full IWSLT 14 training corpus (3.2M words). Our "mainstream improvements" add around 6–7 BLEU in both data conditions.

In the ultra-low data condition, reducing the BPE vocabulary size is very effective (+4.9 BLEU). Reducing the batch size to 1000 token results in a BLEU gain of 0.3, and the lexical model yields an additional +0.6 BLEU. However, aggressive (word) dropout<sup>6</sup> (+3.4 BLEU) and tuning other hyperparameters (+0.7 BLEU) has a stronger effect than the lexical model, and adding the lexical model (9) on top of the optimized configuration (8) does not improve performance. Together, the adaptations to the ultra-low data setting yield 9.4 BLEU (7.2 $\rightarrow$ 16.6). The model trained on full IWSLT data is less sensitive to our changes (31.9 $\rightarrow$ 32.8 BLEU), and optimal hyperparameters differ depending on the data condition. Subsequently, we still apply the hyperparameters that were optimized to the ultra-low data condition (8)

<sup>5</sup>beam search results reported by Wiseman and Rush (2016).

<sup>6</sup> $p = 0.3$  for dropping words;  $p = 0.5$  for other dropout.

<sup>4</sup>Signature BLEU+c.mixed+#.l+s.exp+tok.l3a+v.l.3.2.

system	BLEU
MIXER (Ranzato et al., 2016) <sup>5</sup>	21.8
BSO (Wiseman and Rush, 2016)	25.5
NPMT+LM (Huang et al., 2018)	30.1
MRT (Edunov et al., 2018)	32.84 $\pm$ 0.08
Pervasive Attention (Elbayad et al., 2018)	33.8
Transformer Baseline (Wu et al., 2019)	34.4
Dynamic Convolution (Wu et al., 2019)	35.2
our PBSMT (1)	28.19 $\pm$ 0.01
our NMT baseline (2)	27.16 $\pm$ 0.38
our NMT best (7)	35.27 $\pm$ 0.14

Table 3: Results on full IWSLT14 German→English data on tokenized and lowercased test set with *multi-bleu.perl*.

system	BLEU
(Gu et al., 2018b)	5.97
(supervised Transformer)	
phrase-based SMT	6.57 $\pm$ 0.17
NMT baseline (2)	2.93 $\pm$ 0.34
NMT optimized (8)	<b>10.37</b> $\pm$ 0.29

Table 4: Korean→English results. Mean and standard deviation of three training runs reported.

to other data conditions, and Korean→English, for simplicity.

For a comparison with PBSMT, and across different data settings, consider Figure 2, which shows the result of PBSMT, our NMT baseline, and our optimized NMT system. Our NMT baseline still performs worse than the PBSMT system for 3.2M words of training data, which is consistent with the results by Koehn and Knowles (2017). However, our optimized NMT system shows strong improvements, and outperforms the PBSMT system across all data settings. Some sample translations are shown in Appendix B.

For comparison to previous work, we report lowercased and tokenized results on the full IWSLT 14 training set in Table 3. Our results far outperform the RNN-based results reported by Wiseman and Rush (2016), and are on par with the best reported results on this dataset.

Table 4 shows results for Korean→English, using the same configurations (1, 2 and 8) as for German→English. Our results confirm that the techniques we apply are successful across datasets, and result in stronger systems than previously reported on this dataset, achieving 10.37

BLEU as compared to 5.97 BLEU reported by Gu et al. (2018b).

## 6 Conclusions

Our results demonstrate that NMT is in fact a suitable choice in low-data settings, and can outperform PBSMT with far less parallel training data than previously claimed. Recently, the main trend in low-resource MT research has been the better exploitation of monolingual and multilingual resources. Our results show that low-resource NMT is very sensitive to hyperparameters such as BPE vocabulary size, word dropout, and others, and by following a set of best practices, we can train competitive NMT systems without relying on auxiliary resources. This has practical relevance for languages where large amounts of monolingual data, or multilingual data involving related languages, are not available. Even though we focused on only using parallel data, our results are also relevant for work on using auxiliary data to improve low-resource MT. Supervised systems serve as an important baseline to judge the effectiveness of semisupervised or unsupervised approaches, and the quality of supervised systems trained on little data can directly impact semisupervised workflows, for instance for the back-translation of monolingual data.

## Acknowledgments

Rico Sennrich has received funding from the Swiss National Science Foundation in the project CoNTra (grant number 105212.169888). Biao Zhang acknowledges the support of the Baidu Scholarship.



## References

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Unsupervised Statistical Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium.
- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised Neural Machine Translation. In *International Conference on Learning Representations*.
- Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. *CoRR*, abs/1607.06450.
- 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)*.
- Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014. In *Proceedings of the 11th Workshop on Spoken Language Translation*, pages 2–16, Lake Tahoe, CA, USA.
- Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A Teacher-Student Framework for Zero-Resource Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935, Vancouver, Canada.
- Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-Supervised Learning for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany.
- Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 427–436, Montreal, Canada.
- Colin Cherry, George Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey. 2018. Revisiting Character-Based Neural Machine Translation with Capacity and Compression. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4295–4305, Brussels, Belgium.
- Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied Monolingual Data Improves Low-Resource Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark.
- Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc’Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 355–364, New Orleans, Louisiana.
- Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2018. Pervasive Attention: 2D Convolutional Neural Networks for Sequence-to-Sequence Prediction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 97–107, Brussels, Belgium.
- Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29*, pages 1019–1027.
- Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio.
- Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018a. Universal Neural Machine Translation for Extremely Low Resource Languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana.
- Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018b. Meta-Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium.
- Çağlar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On Using Monolingual Corpora in Neural Machine Translation. *CoRR*, abs/1503.03535.
- Barry Haddow, Nikolay Bogoychev, Denis Emelin, Ulrich Germann, Roman Grundkiewicz, Kenneth Heafield, Antonio Valerio Miceli Barone, and Rico Sennrich. 2018. The University of Edinburgh’s Submissions to the WMT18 News Translation Task. In *Proceedings of the Third Conference on Machine Translation*, pages 403–413, Belgium, Brussels.
- Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual Learning for Machine Translation. In *Advances in Neural Information Processing Systems 29*, pages 820–828.
- Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified

- Kneser-Ney Language Model Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.
- Po-Sen Huang, Chong Wang, Sitao Huang, Dengyong Zhou, and Li Deng. 2018. Towards Neural Phrase-based Machine Translation. In *International Conference on Learning Representations*.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *The International Conference on Learning Representations*, San Diego, California, USA.
- Tom Kocmi and Ondřej Bojar. 2018. Trivial Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation*, pages 244–252, Belgium, Brussels.
- Philipp Koehn, 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. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL-2007 Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.
- Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver.
- Guillaume Lample, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018a. Unsupervised Machine Translation Using Monolingual Corpora Only. In *International Conference on Learning Representations*.
- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018b. Phrase-Based & Neural Unsupervised Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5039–5049, Brussels, Belgium.
- Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task Sequence to Sequence Learning. In *The International Conference on Learning Representations*.
- Antonio Valerio Miceli Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2017. Deep Architectures for Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers*, Copenhagen, Denmark.
- Makoto Morishita, Yusuke Oda, Graham Neubig, Koichiro Yoshino, Katsuhito Sudoh, and Satoshi Nakamura. 2017. An Empirical Study of Mini-Batch Creation Strategies for Neural Machine Translation. In *The First Workshop on Neural Machine Translation (NMT)*, pages 61–68, Vancouver, Canada.
- Masato Neishi, Jin Sakuma, Satoshi Tohda, Shonosuke Ishiwatari, Naoki Yoshinaga, and Masashi Toyoda. 2017. A Bag of Useful Tricks for Practical Neural Machine Translation: Embedding Layer Initialization and Large Batch Size. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 99–109, Taipei, Taiwan.
- Graham Neubig and Junjie Hu. 2018. Rapid Adaptation of Neural Machine Translation to New Languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium.
- Toan Nguyen and David Chiang. 2018. Improving Lexical Choice in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 334–343, New Orleans, Louisiana.
- Toan Q. Nguyen and David Chiang. 2017. Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan.
- Robert Östling and Jörg Tiedemann. 2017. Neural machine translation for low-resource languages. *CoRR*, abs/1708.05729.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels.
- Ofir Press and Lior Wolf. 2017. Using the Output Embedding to Improve Language Models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Valencia, Spain.
- Prajit Ramachandran, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Copenhagen, Denmark.
- Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence Level Training with Recurrent Neural Networks. In *The*

- International Conference on Learning Representations*.
- Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017a. The University of Edinburgh’s Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, Copenhagen, Denmark.
- Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017b. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain.
- 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, Volume 2: Shared Task Papers*, pages 368–373, Berlin, Germany.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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)*, pages 86–96, Berlin, Germany.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.
- Anders Søgaard, Sebastian Ruder, and Ivan Vulic. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 3104–3112, Montreal, Quebec, Canada.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Z. Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.
- Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas.
- Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.
- 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*, pages 1568–1575, Austin, Texas.



## A Hyperparameters

Table 5 lists hyperparameters used for the different experiments in the ablation study (Table 2). Hyperparameters were kept constant across different data settings, except for the validation interval and subword vocabulary size (see Table 1).

## B Sample Translations

Table 6 shows some sample translations that represent typical errors of our PBSMT and NMT systems, trained with ultra-low (100k words) and low (3.2M words) amounts of data. For unknown words such as *blutbeflecken* ('bloodstained') or *Spaniern* ('Spaniards', 'Spanish'), PBSMT systems default to copying, while NMT systems produce translations on a subword-level, with varying success (*blue-flect*, *bleed*; *spaniers*, *Spanians*). NMT systems learn some syntactic disambiguation even with very little data, for example the translation of *das* and *die* as relative pronouns ('that', 'which', 'who'), while PBSMT produces less grammatical translation. On the flip side, the ultra low-resource NMT system ignores some unknown words in favour of a more-or-less fluent, but semantically inadequate translation: *erobert* ('conquered') is translated into *doing*, and *richtig aufgezeichnet* ('registered correctly', 'recorded correctly') into *really the first thing*.

hyperparameter	system						
	2	3	5	6	7	8	9
hidden layer size	1024						
embedding size	512						
encoder depth	1	2				1	
encoder recurrence transition depth	1	2					
decoder depth	1	2				1	
dec. recurrence transition depth (base)	2	4				2	
dec. recurrence transition depth (high)	-	2				-	
tie decoder embeddings	-	yes					
layer normalization	-	yes					
lexical model	-			yes	-		yes
hidden dropout	-	0.2			0.5		
embedding dropout	-	0.2			0.5		
source word dropout	-	0.1			0.3		
target word dropout	-				0.3		
label smoothing	-	0.1				0.2	
maximum sentence length	200						
minibatch size (# tokens)	4000		1000				
learning rate	0.0001					0.0005	
optimizer	adam						
early stopping patience	10						
validation interval:							
IWSLT 100k / 200k / 400k	50	100	400				
IWSLT $\geq$ 800k / KO-EN 2.3M	1000	2000	8000				
beam size	5						

Table 5: Configurations of NMT systems reported in Table 2. Empty fields indicate that hyperparameter was unchanged compared to previous systems.

source	In einem blutbefleckten Kontinent, waren diese Menschen die einzigen, die nie von den Spaniern erobert wurden.
reference	In a bloodstained continent, these people alone were never conquered by the Spanish.
PBSMT 100k PBSMT 3.2M	In a blutbefleckten continent, were these people the only, the never of the Spaniern erobert were. In a blutbefleckten continent, these people were the only ones that were never of the Spaniern conquered.
NMT 3.2M (baseline)	In a blinging tree continent, these people were the only ones that never had been conquered by the Spanians.
NMT 100k (optimized)	In a blue-flect continent, these people were the only one that has never been doing by the spaniers.
NMT 3.2M (optimized)	In a bleed continent, these people were the only ones who had never been conquered by the Spanians.
source	Dies ist tatschlich ein Poster von Notre Dame, das richtig aufgezeichnet wurde.
reference	This is actually a poster of Notre Dame that registered correctly.
PBSMT 100k PBSMT 3.2M	This is actually poster of Notre lady, the right aufgezeichnet was. This is actually a poster of Notre Dame, the right recorded.
NMT 3.2M (baseline)	This is actually a poster of emergency lady who was just recorded properly.
NMT 100k (optimized)	This is actually a poster of Notre Dame, that was really the first thing.
NMT 3.2M (optimized)	This is actually a poster from Notre Dame, which has been recorded right.

Table 6: German→English translation examples with phrase-based SMT and NMT systems trained on 100k/3.2M words of parallel data.