The University of Maryland's Kazakh–English Neural Machine Translation System at WMT19

Eleftheria Briakou and **Marine Carpuat**
Department of Computer Science
University of Maryland
College Park, MD 20742, USA
ebriakou@cs.umd.edu, marine@cs.umd.edu

Abstract

This paper describes the University of Maryland's submission to the WMT 2019 Kazakh to English news translation task. We study the impact of transfer learning from another low-resource but related language. We experiment with different ways of encoding lexical units to maximize lexical overlap between the two language pairs, as well as back-translation and ensembling. The submitted system improves over a Kazakh–only baseline by $+5.45$ BLEU on *newstest*2019.

1 Introduction

Neural Machine Translation (NMT) outperforms traditional phrase-based statistical machine translation provided that large amounts of parallel data are available (Bahdanau et al., 2014; Sennrich et al., 2017; Vaswani et al., 2017). However, it performs poorly under low-resource conditions (Koehn and Knowles, 2017).

While much work addresses this problem via semi-supervised learning from monolingual text (Sennrich et al., 2016; He et al., 2016), we focus on transfer learning from another language pair (Zoph et al., 2016; Nguyen and Chiang, 2017; Lakew et al., 2018). In this setting, an NMT system is firstly trained using auxiliary parallel data from a so-called "parent" language pair and then the trained model is used to initialize a "child" model which is further trained on a low-resource language pair. Similar approaches that support cross-lingual transfer learning for Multi-lingual NMT train a model on the concatenation of all data instead of employing sequential training (Gu et al., 2018; Zhou et al., 2018; Wang et al., 2019).

Transfer learning has been found effective in submissions to WMT in previous years: Kocmi et al. (2018) reported improvements of $+2.4$ BLEU on the low-resource Estonian→English translation task by transfer learning from Finnish→English. Interestingly, Kocmi and Bojar (2018) observed that the transfer learning approach is still effective when there is no relatedness between the "child" and "parent" language-pairs and also hypothesize that the size of the parent training set is the most important factor leading to translation quality improvements. However, previous work has also empirically validated that transfer learning benefits most when "child"-"parent" languages belong to the same or linguistically similar language family (Dabre et al., 2017). Specifically, Nguyen and Chiang (2017) showed consistent improvements in two Turkic languages via transfering from another related, low-resource language.

Taking those recent results into consideration, our main focus at WMT19 is to examine transfer learning for the Kazakh–English language pair using additional parallel data from Turkish–English. While using distinct writing systems, both source languages belong to the Turkic language family and preserve many morphological and syntactic features common for that group (Kessikbayeva and Cicekli, 2014). As a result, they constitute a suitable "child"-"parent" language-pair choice for exploring transfer learning between related low-resource languages. In this direction, we conduct experiments to address the following questions:

- How can we represent lexical units to exploit vocabulary overlap between languages? We compare bilingual and monolingual byte-pair encoding models with the recently proposed soft decoupled encoding model.

- How can we leverage both "child" and "parent" parallel data to obtain synthetic back-translated data from monolingual resources?

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 134–140
Florence, Italy, August 1-2, 2019.

2 Approach

Our method follows a simple strategy used in Wang et al. (2019) for multilingual training: we directly train NMT models on the concatenation of parallel data covering both the "child" and "parent" languages with no metadata to distinguish between them.[1]

Within this framework, we study the impact of (a) different lexical representations that attempt to maximize parameter sharing across related languages, (b) romanization to increase overlap between Turkish and Kazakh which are originally written in distinct scripts, (c) synthetic training data obtained by back-translation.

2.1 Lexical Units

How can we define lexical units to maximize information sharing across related source languages? We compare different configurations of sub-word segmentations using different variants of the standard Byte-Pair Encoding (BPE) framework (Sennrich et al., 2016), and compare them with the Soft Decoupled Encoding framework that exploits character n-gram representations of words instead of sub-words (Wang et al., 2019).

Joint BPEs (JBPEs) BPEs are learned jointly from the concatenation of "child" and "parent" parallel data. The advantage of this strategy is that the sub-word segmentations of related words in the two languages are encouraged to be more aligned; thus enabling the sharing of their representations on the source side due to a larger vocabulary overlap. Although, the "child" language might be "overwhelmed" by the "parent" language when there is a significant difference in the amount of their data (Neubig and Hu, 2018). This could lead to over-segmentation of the "child" language and subsequently limit the expressive power of the NMT system.

Separate BPEs (SBPEs) BPEs are learned separately for each language. This framework was found to be effective in the multilingual setting, especially for translation from extremely low-resource languages (Neubig and Hu, 2018). However, learning the merging operations separately might lead to unaligned sub-units between the two languages that fail to exploit relationships between their lexical representations.

Soft Decoupled Encoding (SDE) Small discrepancies in the spelling of words that share the same semantics across the two languages could lead to different segmented sub-units and hinder the lexical-level sharing between them. To take into account those spelling differences, we further experiment with the SDE encoding framework that is not based on any pre-processing segmentation. Specifically, SDE represents a word as a decomposition of two components: a character encoding that models the language-specific spelling of the word and a latent semantic embedding that captures its language-agnostic semantics. Following, we briefly summarize the main SDE components as proposed in Wang et al. (2019):

Lexical embedding Each word w is first decomposed to its bag of character n-grams (BoN(w)). Let C be the number number of character n-grams in the vocabulary and D be the dimension of the corresponding character n-gram embeddings. To acquire a lexical representation $c(w)$, the word is looked up to an embedding matrix $W_c \in \mathbb{R}^{C \times D}$ as shown below:

$$c(w) = \tanh(\text{BoN}(w) \cdot W_c) \quad (1)$$

Language Specific Transformation Next each word is passed through a language dependent transformation. For each language L_i a matrix $W_{L_i} \in \mathbb{R}^{D \times D}$ is learned and the transformed embeddings $c_i(w)$ is computed:

$$c_i(w) = \tanh(c(w) \cdot W_{L_i}) \quad (2)$$

Latent Semantic Embedding The shared semantic concepts among languages are represented by a matrix $W_s \in \mathbb{R}^{S \times D}$, where S corresponds to the number of semantic concepts a language can express. The latent embeddings of a word w is then given as:

$$e_{\text{latent}}(w) = \text{Softmax}(c_i(w) \cdot W_s^T) \cdot W_s \quad (3)$$

Finally, the SDE embedding of word w is extracted as a combination of the language-dependent lexical encoding and the latent embedding:

$$e_{\text{SDE}}(w) = e_{\text{latent}}(w) + c_i(w) \quad (4)$$

[1] We did not experiment with sequential training of the "parent" and "child" language pairs to establish a fair comparison between our BPE-based models and the SDE model that opts for joint training.

Encoding	Original		Romanized	
Word	molekül	молекула	molekuel	molekula
SBPEs	m_ol_ek_ül	мол_ек_ул_а	m_ol_ek_uel	mol_ek_ul_a
JBPEs	mol_ek_ül	мол_ек_ ул_а	mol_ek_uel	mol_ek_ula
Word	fosfor	фосфор	fosfor	fosfor
SBPEs	f_os_for	ф_ос_фор	f_os_for	f_os_for
JBPEs	fos_for	ф_ос_фор	fos_for	fos_for
Word	kalamar	кальмар	kalamar	kalmar
SBPEs	kal_am_ar	к_аль_ мар	kal_am_ar	kalm_ar
JBPEs	kal_am_ar	к_аль_ мар	kalam_ar	kal_mar

Table 1: Examples of words sharing significant lexical overlap in Kazakh and Turkish among with their corresponding sub-words segmentations.

2.2 Romanization

Given that the provided Kazakh and Turkish data are written in the Cyrillic and Latin scripts respectively, we investigate the impact of mapping text in the two languages into a common orthography. We transliterate both the "child" and the "parent" data using a transliteration tool[2] that applies the same romanization rules to encourage more overlap between child and parent data. Table 1 illustrates how romanization makes shared vocabulary and similarity between the two languages more explicit than using the original scripts.

Table 2 summarizes the statistical overlap on the source side vocabularies between the two languages for different lexical encodings with and without romanization. This analysis indicates that using the original script can be seen as an attempt to explore transfer learning when the lexical-level sharing between the two languages is limited. On the other hand, the vocabulary overlap between them is significantly increased once we romanize the data.

2.3 Synthetic Data

We further explore different ways to incorporate target-side English monolingual data provided by the competition into low-resource NMT. Following the widely used back-translation approach (Sennrich et al., 2016), we create synthetic parallel data and then train new NMT models on the mixture of real and synthetic parallel data.

Empty source baseline The source side of each monolingual example sentence is linked to an empty sentence (denoted by an artificial <null> token).

[2] https://www.isi.edu/~ulf/uroman.html

Method	Romanization	# Merge op.	Overlap
JBPEs	✓	32K	0.44
	✗		0.13
	✓	64K	0.33
	✗		0.11
SBPEs	✓	32K	0.18
	✗		0.04
	✓	64K	0.13
	✗		0.04
		n-gram	**Overlap**
SDE	✓	4	0.67
	✓	5	0.62

Table 2: Statistical overlap results between the vocabularies of the "child" and "parent" languages on the source sides for different encoding schemes. # Merge op. refers to the number of merge operations when BPEs are explored. For the SDE method we compute the overlap between the n-gram character vocabularies (e.g., n-gram=4 corresponds to n={1,2,3,4}).

Back-translation We create synthetic source sentences from automatically back-translating each target (English) sentence into the source language (Kazakh). Within this setting, we only use the original English-Kazakh parallel data to train a model that translates in the opposite direction.

Back-translation+transfer Given the data scarcity of the Kazakh parallel data we also attempt to incorporate both Kazakh and Turkish data to train a model that translates in the opposite direction. In order to produce output that is more similar to our main language of interest, we

introduce two artificial tokens (<2kk>, <2tr>) at the beginning of the input sentence to indicate the target language the model should translate to (Johnson et al., 2017). After the reversed system is trained we back-translate each target sentence to a Kazakh synthetic sentence.[3]

3 Model Configuration

Our NMT systems are built upon the publicly available code[4] of Wang et al. (2019) and are sequence-to-sequence 1-layer attentional long-short term memory units (LSTMs) with a hidden dimension of 512 for both the encoder and the decoder. The word embedding dimension is kept at 128, and all other layer dimensions are set to 512. We use a dropout rate of 0.3 for the word embedding and the output vector before the decoder Softmax layer. The batch size is set to be 1500 words. We evaluate by development set BLEU score (Papineni et al., 2002) for every 2500 training batches. For training, we use the Adam optimizer with a learning rate of 0.001. We use learning rate decay of 0.8, and stop training if the model performance on development set doesn't improve for 5 evaluation steps. We run each experiment with 3 different random seeds.

4 Data and Pre-processing

Parallel Data We use all the parallel data available for the Kazakh–English shared task except for the Wikipedia Titles as they consist of very short sentences (approximately 3 words each). Specifically, the "child" training data consist of about 7.5K sentence pairs from the News Commentary Corpus, and 98K sentence pairs from the English-Kazakh crawled corpus[5]. Additionally, we used approximately 200K Turkish–English sentence-pairs from the Setimes2 Corpus that are provided by the WMT18 competition.

Monolingual For the *Empty source* and *Back-translation* methods of creating synthetic data we used the target-side of the Turkish–English parallel corpus as monolingual data. For the *Back-Translation+transfer* experiment we used 100K randomly selected sentences from the News Commentary corpus, excluding sentences with less than 5 words and more than 50 words.

[3]Each English sentence of the monolingual corpus is augmented with a <2kk> token at the beginning.

[4]https://github.com/cindyxinyiwang/SDE

[5]We didn't filter out any sentence pairs from this corpus.

Pre-processing We process all corpora consistently. We tokenize the sentences and perform truecasing with the Moses scripts (Koehn et al., 2007). For all the experiments we consistenly use 8K BPEs on the English target side. We experiment with $\{32, 64\}$K merge operations for the models using BPE encoding and $\{4, 5\}$ n-grams for the SDE framework. To establish a fair comparison between the source language representations, we consistently use the same encoding for English words (target side) using BPEs learned on the concatenation of all the English data.

Tuning and Testing Data The official newsdev2019 is used as the validation set, and newstest2019 is used as the test set.

5 Experiments

Starting from *Baseline* BPE-based NMT systems trained using only the Kazakh data provided by the competition, we conduct the following experiments.

5.1 Byte Pair Encoding

Table 3 presents our results of 3 runs using $\{32, 64\}$K merge operations in total for each experiment. Generally, both Joint and Separate BPE segmentation strategies, with and without romanization improve BLEU over the *Baseline*. Previous empirical results on transfer learning for extremely low-resource languages indicated that training the BPE operations separately for the "child" and "parent" languages has a large positive effect on the performance of the model (Wang et al., 2019). By contrast, JBPEs and SBPEs perform comparably well in almost all configurations here. This could be attributed to our less imbalanced setting where the ratio of "child"-"parent" data is $1:2$, and the child language therefore contributes more to sub-word segmentation rules.

The best BLEU score is achieved using 32K JBPEs on the romanized data which is consistent with the configuration with the largest vocabulary overlap, according to Table 2. However, using $\{32, 64\}$K SBPEs on the original data only hurts BLEU by 0.5 and 1.24, despite the lack of lexical overlap. This suggests that most of the improvement does not come from the shared encoder vocabulary.

	32K BPEs		64K BPEs	
Method	Original	Romanized	Original	Romanized
Baseline	4.33 ± 0.16	4.49 ± 0.02	4.35 ± 0.13	4.21 ± 0.28
JBPEs	$\mathbf{9.35 \pm 0.10}$	$\mathbf{9.89 \pm 0.14}$	$\mathbf{8.65 \pm 0.27}$	8.77 ± 0.09
SBPEs	7.10 ± 0.26	9.70 ± 0.28	8.41 ± 0.08	$\mathbf{8.85 \pm 0.34}$

Table 3: Kazakh → English BLEU score results on news-test2019 for different BPE configurations and versions of data.

N-gram	*Lexical*	*Latent*	*Specific*	**BLEU**
	✓			$\mathbf{9.12 \pm 0.27}$
4	✓	✓		8.76 ± 0.29
	✓	✓	✓	6.57 ± 0.20
	✓			$\mathbf{9.17 \pm 0.21}$
5	✓	✓		8.69 ± 0.21
	✓	✓	✓	6.21 ± 0.18
Baseline-BPE				8.65 ± 0.27

Table 4: SDE Experiments using 64K n-grams of the concatenated corpora. The last line refers to the best BLEU score using 64K BPEs for comparison.

5.2 Soft-Decoupled Encoding

We compare the BPE results with different configurations of the SDE model. Table 4 presents average results of 3 runs with different random seeds, where we use 64K character n-grams as our vocabulary. The *Language Specific Transformation* consistently harms the BLEU score for both $n = 4, 5$. This result validates the empirical observations of Wang et al. (2019); the separate projection does not help when the "child"-"parent" languages have a significant surface lexical overlap. We also observe comparable BLEU results when we use SDE embeddings or lexical embeddings (where the latent embedding is not taken into account) to encode the semantics of words. The best BLEU scores are achieved for the lexical encoding using either 4-grams or 5-grams of words.

In both cases we observe that the n-gram models perform sligthly better than the best BPE model that uses the same number of merge operations as the n-gram vocabulary size (we refer to that model as *Baseline-BPE* on Table 4). However, we do not adopt SDE in our submitted system as the small BLEU score improvement comes with higher computational cost when compared to the BPE models.

5.3 Synthetic Data

Finally we experiment with back-translation of monolingual English corpora. All experiments used romanized text segmented with 32K BPE merge operations. Table 5 compares 3 different ways of using the same English data extracted from the target side of the Turkish–English parallel corpus. Each target sentence is coupled with a synthetic Kazakh sentence (*Back-translation*), an empty source sentence as a control (*Empty*) or a real Turkish sentence (*Transfer*). The ratio of real to additional data is kept to $1:2$ in all cases.

NMT training does not benefit from the back-translated data as it achieves nearly the same BLEU as the baseline model. Suprisingly empty source sentences yield better results than back-translation, suggesting that the synthetic back-translations are of low quality. Translating into Kazakh is challenging given the small amount of data available, especially for translating from a morphologically poor to a morphologically rich language. Finally, using real Turkish data on the source side achieves the best improvement over the baseline system (+4.4 BLEU).

Method	**Synthetic**	**BLEU**
Baseline		4.49
Empty	✓	5.26
Back-Translation	✓	4.64
Transfer		**9.89**

Table 5: Experiments using 200K monolingual data extracted from the target side of Turkish–English parallel corpus. The *Baseline* system is trained only on Kazakh data.

Given that in all these 3 experiments the decoder model was trained on the exact same English data, these results suggest that the transfer learning benefits both the encoder and decoder models.

Method	Synthetic	BLEU
Baseline-Transfer		**9.89**
Empty	✓	9.17
Back-Translation	✓	9.38
+ ensemble(4)*	✓	**9.94**

Table 6: Experiments using additional 100K News Commentary monolingual data. The Baseline system is trained on the concatenation of Kazakh–Turkish parallel data. The $\star$ symbol denotes our primary submission for human evaluation.

Finally, we attempt to combine Kazakh and Turkish parallel data to back-translate 100K additional monolingual data to Kazakh via training a NMT model that has control over the output language, as can be seen in Table 6. In this experiment our *Baseline-Transfer* system refers to the best model trained on the concatenation of "child" and "parent" data. In contrast to the previous experiment we now combine Kazakh, Turkish and synthetic data with a ratio $1:2:1$. We observe that in both cases (*Back-translation*, *Empty*) the BLEU score of the system trained on the augmented data fails to outperform the *Baseline-Transfer* performance, possibly due to the fact that the real Kazakh data have been "overwhelmed" by the auxiliary ones (Poncelas et al., 2018). However, we could assume that the quality of the back-translated data is slightly better once we utilized the Turkish data (given that it performs better than the *Empty* experiment).

Finally, the last row of Table 6 reports the BLEU score of our **primary submission**.[6] Specifically, the submitted model is an ensemble obtained by averaging the output distributions of 4 models trained on Kazakh, Turkish and Back-Translated using different random seeds.

6 Conclusion

This paper presents the University of Maryland's NMT system for WMT 2019 Kazakh $\rightarrow$ English news translation task. Specifically, we explored how to improve neural machine translation of a low-resource language by incorporating parallel data from a related, also low-resource language. Our empirical results validate that transfer learning benefits BLEU even when transfering from a low-resource language pair. Furthermore, our results suggest that translation quality (in terms of BLEU score) of the language-pair of focus is most benefited when the surface-level parameter sharing between the lexical representations of the two related languages is maximized. Finally, we observed that NMT training with synthetic data is sensitive to the quality of the back-translation.

[6]The *Baseline-Transfer* model slightly under-performed the *Baseline-Transfer+Back-Translation* model on the development set. Given that we did not have access to test data during evaluation time, our primary submission was based on evaluation on the development set.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv e-prints*, abs/1409.0473.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286. The National University (Phillippines).

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. 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. Association for Computational Linguistics.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 820–828, USA. Curran Associates Inc.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Gulshat Kessikbayeva and Ilyas Cicekli. 2014. Rule based morphological analyzer of kazakh language. In *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM*, pages 46–54, Baltimore, Maryland. Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Belgium, Brussels. Association for Computational Linguistics.

Tom Kocmi, Roman Sudarikov, and Ondřej Bojar. 2018. CUNI submissions in WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 431–437, Belgium, Brussels. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Surafel Melaku Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)*, pages 54–62, Brussels, Belgium.

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. Association for Computational Linguistics.

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. Asian Federation of Natural Language Processing.

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*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. *CoRR*, abs/1804.06189.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The university of Edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 389–399, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019. Multilingual neural machine translation with soft decoupled encoding. *CoRR*, abs/1902.03499.

Zhong Zhou, Matthias Sperber, and Alexander Waibel. 2018. Massively parallel cross-lingual learning in low-resource target language translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 232–243, Belgium, Brussels. Association for Computational Linguistics.

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. Association for Computational Linguistics.