CUNI Submissions in WMT18

Tom Kocmi **Roman Sudarikov** **Ondřej Bojar**

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
<surname>@ufal.mff.cuni.cz

Abstract

We participated in the WMT 2018 shared news translation task in three language pairs: English-Estonian, English-Finnish, and English-Czech. Our main focus was the low-resource language pair of Estonian and English for which we utilized Finnish parallel data in a simple method. We first train a "parent model" for the high-resource language pair followed by adaptation on the related low-resource language pair. This approach brings a substantial performance boost over the baseline system trained only on Estonian-English parallel data. Our systems are based on the Transformer architecture. For the English to Czech translation, we have evaluated our last year models of hybrid phrase-based approach and neural machine translation mainly for comparison purposes.

1 Introduction

This paper describes the Charles University's submission to WMT 2018 Shared Task: Machine Translation of News.

We have experimented with three language pairs: Czech *(CS)*, Estonian *(ET)* and Finnish *(FI)* paired with English *(EN)*. Altogether, we covered five directions: both direction for English-Estonian, both directions for English-Finnish and English to Czech translation.

Our main focus is improving the low-resource language translation and therefore we concentrate on the English and Estonian language pair with the help of Finnish-English parallel data. The Finnish is a good candidate since it is closely related to the Estonian language but considerably more training data are available.

For the Finnish and English language pair, we use standard Neural Machine translation *(NMT)* system Transformer (Vaswani et al., 2017) with model averaging.

Our last language pair of interest is English to Czech translation, where we use our last year's model Sudarikov et al. (2017) for comparison purposes. The system is based on a hybrid combination of phrase-based, transfer-based and NMT approaches.

The structure of the paper is the following. In Section 2, we describe the setup of our main systems for Estonian and Finnish. Section 3 presents the English-Czech model. Section 4 is devoted to the description of our datasets. Section 5 details the results achieved by our systems. Section 6 discusses other works in the area of multi-lingual translation systems. And finally Section 7 concludes the paper.

2 Estonian and Finnish Setup

The main focus of our participation is improving low-resource language Estonian with the use of Finnish data. Our method consists of first training a "parent" high-resource model and continue the training on the "child" (low-resource) parallel data as a means of model adaptation.

2.1 Low-Resource Language Adaptation

We present a method that uses related high-resource language pair as a boost in performance for a low-resource language pair. The method needs relies on only one condition and that is a vocabulary shared across all the languages in the parent as well as child language pairs.

The shared vocabulary is obtained by combining all training data when the vocabulary is generated. To avoid bias in the vocabulary towards the high-resource language pair, we use only as many sentence pairs from the high-resource pair as are available for the low-resource pair, calling this approach "balanced vocabulary". We did not experiment with other proportions of data.

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 431–437

https://doi.org/10.18653/v1/W18-64043

Our method is based on transfer learning (also called "adaptation" or "finetuning"). It starts with training of the parent high-resource language pair (English-Finnish in our case) until it reaches its best performance or is trained for sufficiently long. Then, the training corpus is switched to the low-resource language pair (English-Estonian) for the rest of the training, without resetting any of the training hyperparameters. Note that we are not resetting even the state of the adaptive learning rate. As mentioned in Kocmi and Bojar (2018), if the learning rate is reset, this approach stops working.

As such, this method is very similar to the transfer learning proposed by Zoph et al. (2016) and improved by the using the shared vocabulary as in Nguyen and Chiang (2017). Moreover, in contrast to those two papers, we show that this simple style of transfer learning can be used on both sides (i.e. either the source or the target language), not only with the target language common to both parent and child model. More details of our method are described in Kocmi and Bojar (2018).

This method does not need any modification of existing NMT frameworks. The only requirement is to use the shared vocabulary across both language pairs (we use vocabulary of wordpieces, Johnson et al., 2017). This is achieved by learning the wordpiece segmentation from the concatenated source and target sides of both the parent and child language pair.

All other parameters of the model can stay the same as for the standard NMT training.

2.2 Model Description

We use the Transformer model (Vaswani et al., 2017) which translates through an encoder-decoder framework, with each layer involving an attention network followed by a feed-forward network. The architecture is much faster than other NMT due to the absence of recurrent and convolutional layers.

The Transformer model seems superior to other NMT approaches as documented in e.g. Popel and Bojar (2018) and also several language pairs in the manual evaluation of WMT18 (Bojar et al., 2018).[1]

We use the Transformer sequence-to-sequence model as implemented in Tensor2Tensor (Vaswani et al., 2018) version 1.4.2. Our models are based on the "big single GPU" configuration as defined in the paper. We set the batch size to 2300 and maximum sentence length to 100 wordpieces, in order to fit the model to our GPUs (NVIDIA GeForce GTX 1080 Ti with 11 GB RAM).

We use exponential learning rate decay with the starting learning rate of 0.2 and 32000 warm-up steps. Decoding uses the beam size of 8 and length normalization penalty is set to 1.

3 Chimera Description

For English-Czech translation task, we took the same system combination setup as described in Sudarikov et al. (2017). We used outputs of three different individual forward translation systems, trained on a synthetic backtranslated training dataset and combined them into the final output. These systems are Chimera2016 (Tamchyna et al., 2016; Bojar et al., 2016b), NeuralMonkey (Helcl et al., 2018)[2] and Marian (where the translation part was formerly known as AmuNMT) (Junczys-Dowmunt et al., 2016) with pretrained English-to-Czech Nematus models.[3] All the used datasets are described in Section 4.

The outputs of the two neural systems, consisting of translations of WMT15–18 test sets, were used to extract additional phrase tables for Moses. These tables were added to the Chimera2016 system, which already had one phrase table from genuine parallel data and one synthetic phrase table from TectoMT (Žabokrtský et al., 2008) output. After that, we used MERT (Och, 2003) to estimate the weights for Moses alternative decoding paths with multiple translation tables. MERT was run on the WMT16 test set. Further details on experiments with different combinations of phrase tables are available in Sudarikov et al. (2017).

4 Data Preparation

This section describes the data used for the training of our models. First, we describe training data for Estonian and Finnish.

There are many different sources for WMT18 News shared task that are allowed for the constrained task. We used most of the allowed data but decided to drop some sources.

For the Estonian-English, we use Europarl and Rapid corpora. We did not use Paracrawl because

[1] http://www.statmt.org/wmt18/translation-task.html

[2] http://ufal.mff.cuni.cz/neuralmonkey

[3] http://data.statmt.org/rsennrich/wmt16_systems

Language pair	Sentences
Estonian-English	0.8 M
Finnish-English	2.8 M
Czech-English	71.7 M
Estonian mono-news	2.6 M
Finnish mono-news	12.0 M
Czech mono-news	59.2 M

Table 1: Overview of training datasets. The top half lists sentence pair counts for parallel corpora and the bottom half the sentence counts of monolingual data.

we find it very noisy. The development set is from WMT News 2018.

The Finnish-English was prepared as in Östling et al. (2017), removing Wikipedia headlines. The dev set is from WMT News 2015.

We dropped sentence pairs shorter than 4 words or longer than 75 words on either source or target side to allow for a speedup of Transformer training by capping the maximal sentence length and increasing the batch size. Our experiments showed no translation performance change due to the reduction of the training data.

For English-Czech models, we used the same datasets as described in Sudarikov et al. (2017). First we took Czech monolingual news corpus, which was translated into English using Nematus (Sennrich et al., 2017) model, with 59 million sentences. We also used the genuine parallel data extracted from CzEng 1.6 (Bojar et al., 2016a) using the XenC toolkit (Rousseau, 2013) with Czech monolingual news corpus as the reference in-domain text. That part gave us additinal 12M sentences. The same monolingual news corpus was used for the language models.

The final data sizes are presented in Table 1.

4.1 Backtranslated Data

The organizers of WMT 2018 provide participants with vast amounts of monolingual data to use in translation systems, both in-domain and out-of-domain. We exploit the in-domain monolingual data for training as described by Sennrich et al. (2016) and previously suggested for PBMT e.g. by Bojar and Tamchyna (2011).

The idea is to translate the target side the monolingual data by an already trained machine translation system for the opposite translation direction and then use the synthetic data as a parallel corpus for the training of the main system. In this setup, the synthetic side is used as the input and the original monolingual sentences serve as the target.

Specifically, for the examined language pair EN→FI, we backtranslate monolingual Finnish data with the FI→EN model and mix the synthetic data with the available parallel EN→FI data to create the training corpus for EN→FI.

Sennrich et al. (2016) motivates the use of monolingual data with domain adaptation, due to the usage of in-domain monolingual data, reducing overfitting, and better modeling of fluency. Bojar and Tamchyna (2011) explain how backtranslation (with some fall-back for unknown words) allows to improve the vocabulary when targetting morphologically rich languages.

We get monolingual News Crawl data from all years of both Finnish and Estonian. We created the synthetic data from all monolingual data; we only drop sentences shorter than 6 words or longer than 75 words.

The monolingual data sizes are presented in Table 1.

It is important to stress that all the results in this paper are *without* the use of backtranslation. Only Table 4 presents the results with the use of backtranslated data.

5 Results and Discussion

In this section, we first present the results for Estonian-English and Finnish-English language pairs, focusing on transfer learning from the high-resource language pair to low-resource one. At the end, we compare the current NMT outputs to our last year's system for English to Czech translation.

The scores are evaluated by uncased SacreBLEU (Post, 2018).

We have computed statistical significance with pairwise bootstrap resampling with 1000 samples and alpha equal to 0.05 (Koehn, 2004).

Table 2 presents the effect of transfer learning from the parent model to the child model. The improvement is noticeable in both sides: the language unique to the child model can appear in the source or in the target.

Whenever the child language pair has more resources than the parent (Finnish-English in our case), the improvement is small or even (insignificantly) negative, as in ETEN-FIEN.

One could argue that the languages are too related and simply using the high-resource language pair model could work for the low-resource test sentences. The second column of Table 2 shows that this is not the case: the parent model without

Parent - Child	Baseline	Only Parent	Transfer
ENFI - ENET	17.03	2.32	19.74‡
FIEN - ETEN	21.74	2.44	24.18‡
ENET - ENFI	19.50	2.04	20.07‡
ETEN - FIEN	24.40	1.94	23.95
ETEN - ENET	17.03	1.41	17.46
ENET - ETEN	21.74	1.01	22.04‡

Table 2: Uncased BLEU scores for transfer learning of child models on various combinations of parent and child. The baseline is obtained by training only on the child parallel data. "Only Parent" represent result when no adaptation of parent model is done, i.e. running MT for the wrong language. The results are only comparable within each row. Results significantly better than the baseline are marked with ‡.

Child Training Sents	Child BLEU	Baseline BLEU
800k	19.74	17.03
400k	19.04	14.94
200k	17.95	11.96
100k	17.61	9.39
50k	15.95	5.74
10k	12.46	1.95

Table 3: The maximal score reached by the English-to-Estonian child models for decreasing sizes of child's training data, trained on an English to Finnish parent (all models build upon the same parent ENFI after 800k steps trained on the whole ENFI training set). The baselines use only the reduced English-Estonian data.

any transfer learning does not work for translation of the child test set.

With this result in mind, we also tested the effect of using only the low-resource language pair in both directions: first as a parent trained in the reverse direction, followed by training of the child on the same parallel corpus, now in the intended direction. The results of this can be seen in the bottom part of Table 2. It is an interesting result that only by using the low-resource data twice (in the reverse and then the correct direction), we could get a small boost in performance, significant when targetting ETEN.

In Table 3, we simulate extremely low-resource languages by downscaling the data for the child model. The smaller the child data, the bigger relative improvement is obtained. A reasonable performance is obtained even with as few as 10k sentence pairs in the child. This result suggests that when dealing with the very low-resource language, it is useful to utilize a related language pair as a pre-training parent step.

Language Pair	Only Parallel	Transfer learning	With Backtranslated Equal Size	With Backtranslated All
EN-ET	17.03	**19.74**	**21.43**	**22.73**‡
EN-FI	19.50	-	22.96	23.57‡

Table 4: Results with backtranslated data, either up to the size of the original parallel corpus ("Equal Size") or all available ("All"). The significance is computed between "Equal Size" and "All". The bold results are with additional use of transfer learning.

Language pair	Baseline	Submitted
FI-EN	21.52	21.52
EN-FI	15.13	15.13
ET-EN	20.68	23.50
EN-ET	16.54	19.49

Table 5: WMT18 newstest BLEU scores for the baseline runs and the runs submitted as "CUNI-Kocmi-*" for manual evaluation.

5.1 Effect of Backtranslation

The size of the training set can be extended also with the backtranslated data. We experiment with backtranslation only for two language directions: English to Estonian and English to Finnish.

First, we trained FI→EN and ET→EN models on parallel data for each of the language pairs. With those models, we translated all monolingual data. Finally, we mixed the synthetic and genuine parallel corpora for FI→EN and (separately) for ET→EN.

Table 4 presents our experiment with two setups. We either used only a subset of the synthetic corpus of the size equal to the genuine parallel data, or we use all available synthetic data. The former approach results in a training corpus with half of monolingual backtranslated data and half of original parallel texts. The latter approach results in parallel training set containing 76.5% monolingual data for Estonian and 81.1% for Finnish. In both cases, we report the score on the dev set after 600k steps of training.

The motivation for applying this upper bound is that the synthetic corpus could introduce more translation errors and damage translation quality. The results in Table 4 however document that this is not the case and more data is better.

5.2 Estonian and Finnish Submitted Models

Our submitted models for Finnish and Estonian are presented in Table 5, with the baseline of no transfer. Unfortunately, we submitted models without backtranslation for manual evaluation.

Language pair	WMT17	WMT18
CUNI-Transformer	23.8	26.0
UEDIN-NMT	22.8	23.4
CUNI-Chimera2017	20.5	19.8

Table 6: Cased-BLEU results from matrix.statmt.org.

For Finnish, the submitted models did not include the transfer learning step so the FI→EN and EN→FI Baseline and Submitted scores are identical.

The Estonian-to-English model was trained from the Finnish-to-English model at its 800k training steps. The English-to-Estonian built upon the English-to-Finnish, trained also for 800k steps.

5.3 English-to-Czech Benchmark

Table 6 shows cased-BLEU scores for WMT17 and WMT18 test sets as presented at http://matrix.statmt.org.[4]

The Chimera setup remains the same in both years, so it can serve as a reference point, documenting the improvement of other systems. The gap between Chimera and the best neural systems considerably widened in terms of BLEU score (from +2.3 on WMT17 to +3.6 on WMT18 when comparing to UEDIN-NMT and from +3.3 to +6.2 when comparing to CUNI-Transformer).

6 Related Work

Firat et al. (2016) propose zero-resource multi-way multilingual systems, with the main goal of reducing the total number of parameters needed to train multiple source and target languages. To keep all the language pairs "active" in the model, a special training schedule is needed. Otherwise, catastrophic forgetting would remove the ability to translate between the languages trained earlier.

Johnson et al. (2017) test another multilingual approach: all translation pairs are simply used at once and the desired target language is indicated with a special token at the end of the source side. The model implicitly learns translation between many languages and it can even translate among language pairs never seen together.

The lack of parallel data can be tackled by unsupervised translation (Artetxe et al., 2018; Lample et al., 2018). The general idea is to mix monolingual training of autoencoders for the source and target languages with translation trained on data translated by the previous iteration of the system.

Aside from the common back-translation (Sennrich et al., 2016), simple copying of target monolingual data back to source (Currey et al., 2017) has been also shown to improve translation quality in low-data conditions.

Similar to transfer learning is also curriculum learning (Bengio et al., 2009; Kocmi and Bojar, 2017), where the training data are ordered from foreign out-of-domain to the in-domain training examples.

7 Conclusion

In this paper, we presented our systems for WMT 2018 shared news translation task in three language pairs: English-Estonian, English-Finnish, and English-Czech.

English-Estonian was the main focus of our research, with the English-Finnish used to improve the quality of the translations. Both Finnish and Estonian systems used the Transformer architecture. Our results show that a simple transfer learning is beneficial. Further gains (not in of our submitted systems) were obtained by including back-translated data.

Our English-Czech submission was prepared and used mainly for comparison purposes and it showed the widening gap between hybrid phrase-based and neural systems.

Acknowledgments

This study was supported in parts by the grants SVV 260 453, GAUK 8502/2016, and 18-24210S of the Czech Science Foundation. This work has been using language resources and tools stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (projects LM2015071 and OP VVV VI CZ.02.1.01/0.0/0.0/16 013/0001781).

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural ma-

[4] http://matrix.statmt.org/matrix/systems_list/1867 for 2017 and http://matrix.statmt.org/matrix/systems_list/1883 for 2018.

chine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016a. Czeng 1.6: enlarged czech-english parallel corpus with processing tools dockered. In *International Conference on Text, Speech, and Dialogue*, pages 231–238. Springer.

Ondrej Bojar, Roman Sudarikov, Tom Kocmi, Jindrich Helcl, and Ondrej Cıfka. 2016b. Ufal submissions to the iwslt 2016 mt track. *IWSLT. Seattle, WA*.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving Translation Model by Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

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.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Jindřich Helcl, Jindřich Libovickỳ, Tom Kocmi, Tomáš Musil, Ondřej Cífka, Dusan Varis, and Ondřej Bojar. 2018. Neural monkey: The current state and beyond. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, volume 1, pages 168–176.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a Vigas, 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.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. In *Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum Learning and Minibatch Bucketing in Neural Machine Translation. In *Recent Advances in Natural Language Processing 2017*.

Tom Kocmi and Ondřej Bojar. 2018. Trivial Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 3rd Conference on Machine Translation (WMT)*, Brussels, Belgium.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, volume 4, pages 388–395.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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. Asian Federation of Natural Language Processing.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

Robert Östling, Yves Scherrer, Jörg Tiedemann, Gongbo Tang, and Tommi Nieminen. 2017. The helsinki neural machine translation system. In *Proceedings of the Second Conference on Machine Translation*, pages 338–347, Copenhagen, Denmark. Association for Computational Linguistics.

Martin Popel and Ondej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. *arXiv preprint arXiv:1804.08771*.

Anthony Rousseau. 2013. Xenc: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, 100:73–82.

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. 2017. 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. 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.

Roman Sudarikov, David Mareček, Tom Kocmi, Dušan Variš, and Ondřej Bojar. 2017. CUNI Submission in WMT17: Chimera Goes Neural. In *Proceedings of the 2nd Conference on Machine Translation (WMT)*, Copenhagen, Denmark.

Aleš Tamchyna, Roman Sudarikov, Ondřej Bojar, and Alexander Fraser. 2016. CUNI-LMU submissions in WMT2016: Chimera constrained and beaten. In *Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics*.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Lukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for Neural Machine Translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 193–199, Boston, MA. Association for Machine Translation in the Americas.

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 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 6000–6010. Curran Associates, Inc.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170. 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.