CUNI NMT System for WAT 2018 Translation Tasks

Tom Kocmi **Shantipriya Parida** **Ondřej Bojar**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
{kocmi,parida,bojar}@ufal.mff.cuni.cz

Abstract

This paper describes the CUNI submission to WAT 2018 for the English-Hindi translation task using a transfer learning techniques which has proven effective under low resource conditions. We have used the Transformer model and utilized an English-Czech parallel corpus as additional data source. Our simple transfer learning approach first trains a "parent" model for a high-resource language pair (English-Czech) and then continues the training on the low-resource (English-Hindi) pair by replacing the training corpus. This setup improves the performance compared with the baseline and in combination with back-translation of Hindi monolingual data, it allowed us to win the English-Hindi task. The automatic scoring by BLEU did not correlate well with human judgments.

1 Introduction

Neural Machine Translation (NMT) systems are superior to Phrase-Based Statistical Machine Translation (PBMT) in large data conditions but they suffer when parallel resources are limited (Bojar et al., 2017; Koehn and Knowles, 2017; Lakew et al., 2017). In the current situation, only few language pairs have such high quality parallel corpora of sufficient size (Chu and Wang, 2018).

Many approaches were proposed in the past few years to utilize additional data to improve machine translation for low-resource languages. Currey et al. (2017) copied the target side of monolingual data to the source to forge a parallel corpus creating a "copied corpus". After mixing with the bilingual corpus and training NMT systems, they got accuracy improvements. Zoph et al. (2016) proposed transfer learning which uses an additional large corpus of another language pair (parent model) for training and then transfer the learned parameters to the low resource pair (child model) to initialize and constrain training, resulting an increase of BLEU scores. Nguyen and Chiang (2017) proposed transfer learning for low resource language pairs starting from a low resource parent pair. They used sub-word units (BPE, Sennrich et al. (2016a); Shibata et al. (1999)) and focused on increasing vocabulary overlap during transfer of model parameter from the parent language pair to the child one. Closely related to the transfer learning is also curriculum learning (Bengio et al., 2009; Kocmi and Bojar, 2017), where the training data can be ordered from parent out-of-domain to the child in-domain training examples.

In this system description paper, we explain our approach of using Hindi monolingual data and applying transfer learning using additional English-Czech parallel corpus. Section 1 describes related work carried out by different researchers using domain adaptation techniques. Section 2 explains the techniques which we followed in our work. Section 3 describes the datasets used in our experiment. Section 4 presents the model and experimental setups used in our approach. Section 5 provides the official evaluation results of WAT 2018 followed by the conclusion in Section 6.

2 Method Description

We utilize transfer learning based on the work of Kocmi and Bojar (2018). The method of training is

		#Tokens		
Set	#Sentences	EN	CS	HI
Train (EN-CS)	40.1M	563.4M	490.5M	-
Train (EN-HI)	1.4M	20.6M	-	22.1M
TrainBack (EN-HI)	8.8M	161M	-	167M
Dev (EN-HI)	520	10656	-	10174
Test (EN-HI)	2507	49394	-	57037

Table 1: Statistics of our data.

similar to domain adaptation, where we first train a more general model later followed by training on a more domain-specific dataset. The domain in our case is the actual language pair. The method by Kocmi and Bojar (2018) does not require any of the languages to be linguistically related.

The method has only one constrain and that is a shared vocabulary between language pairs of parent and child. This is solved by generating wordpiece segmentation (Johnson et al., 2017) from the concatenated source and target sides of both the parent and the child language pair. To avoid bias in the vocabulary towards the high-resource language pair, Kocmi et al. (2018b) showed that best performance is obtained by using a "balanced vocabulary" approach which uses only as many sentence pairs from the high-resource pair as there are available for the low-resource pair.

We start with the parent model, in our case English to Czech translation, and keep training as long as it improves the results on the development set. Then the training corpus is switched to the child parallel corpus and the training continues without any hyperparameter modifications. We do not even reset the learning rate.

This transfer learning method does not need any modifications of existing NMT frameworks.

We have observed that a small number of outputs of some of our systems were not translated into the target language. For those cases identified by the Python language detection library "langdetect" (Thoma, 2018), we use the output with the output of another model with different settings instead.

3 Dataset

This section describes the dataset provided by WAT 2018 for the translation task and the dataset used for domain adaptation. We have used two language pairs: one as the high-resource one (parent model) and another as the low-resource one (child model).

Kocmi and Bojar (2018) showed, that relatedness of parent and child language is not the main criterion for better performance, but it is the sheer volume of parent training size. Therefore we have decided to use Czech-English as the parent model, since it is one of the most resourceful language pairs available and allowed for the WAT 2018 shared task. And it is reasonably clean since it does not contain dirty crawled data.

We use CzEng 1.7 (Bojar et al., 2016) as the parent language pair training set. We preprocessed the data in the same manner as in the work of Kocmi et al. (2018a) by dropping sentences shorter than 4 words and longer than 75 words. We use IITB English-Hindi parallel corpus[1] (Kunchukuttan et al., 2018) provided by WAT 2018 for the English-Hindi translation task as the child language pair. This is supposedly the largest publicly available English-Hindi parallel corpus. This corpus contain 1.49 million parallel segments and 45 million monolingual segments and it was found very effective for English-Hindi translation task (Parida and Bojar, 2018). Apart from the above language pairs, we have also used the Hindi monolingual dataset for generating synthetic data using back translation. Recently many researchers have shown that back translating monolingual data can be used to create synthetic parallel corpora which in combination with authentic parallel data helps to train a high quality MT system (Bojar and Tamchyna, 2011; Sennrich et al., 2016b; Poncelas et al., 2018; Popel, 2018). The usage of monolingual data in the target language provides the NMT system with more evidence on which words are more common and which are not (Koehn, 2017). We com-

[1] http://www.cfilt.iitb.ac.in/iitb_parallel/

Setting	Direction	Use synthetic	Use genuine	Transfer Learning	Avg (8 Last Models)
S1	EN-HI	✓	✗	✗	✓
S2	EN-HI	✓	✗	1M steps of EN-CS	✓
S3	EN-HI	✓	✓	1M steps of EN-CS	✓
S4	EN-HI	✗	✓	1M steps of EN-CS	✓
S5	EN-HI	✗	✓	1M steps of EN-CS	✓
S6	HI-EN	✗	✓	1M steps of CS-EN	✓

Table 2: Main differences between model settings.

bine transfer learning with back translation. We first train a CS-EN system, continue its training with HI-EN and then apply it to Hindi monolingual data to obtain a synthetic EN-HI corpus. The statistics of all the datasets are shown in Table 1.

4 Experiments

This section describes our experiments conducted for the translation task.

4.1 Tokenization and Vocabulary

We have used shared vocabulary of subword units, word pieces (Johnson et al., 2017), across both language pairs, where the word pieces handle tokenization automatically.

Our approach requires a shared vocabulary across the parent model (English to Czech) and the child model (English to Hindi). Our generated vocabulary contains 32k sub-word types.

4.2 NMT Model Description

In our approach, we train the parent language pair until the BLEU scores on the development set seem more or less stable and switch the training corpus to the child language pair without any hyper-parameter change.

We use the Transformer model as implemented in Tensor2Tensor (Vaswani et al., 2018) version 1.4.2. We have used the "Big Single GPU" configuration for our experiments. To fit the model to our GPUs (NVIDIA GeForce GTX 1080 Ti with 11 GB RAM), we set the batch size to 2300 and limit sentence length to 100 wordpieces. We use Noam learning rate decay[2] (Vaswani et al., 2017; Popel and Bojar, 2018) with the starting learning rate of 0.2 and 32000 warm up steps. In our experiments, we find that it is undesirable to reset the learning rate when switching to the child language pair as it leads to the loss of the performance gained in the parent model. Decoding uses the beam size of 8 and length normalization penalty is set to 1. The parent model was trained for 1M steps (approximately 6 days), the child models were trained for approximately 500k steps, which was sufficient for models to converge to the best performance. We selected the model with the best performance on the development test for the final evaluation on the test set. We also use checkpoint averaging which we confirmed to be effective for Transformer model (Popel and Bojar, 2018). We average the last 8 models.

4.3 Model Setups

We have used 6 settings for our English-to-Hindi and Hindi-to-English Translation Task as shown in Table 2 and described as follows:

1. *S1: TransBig (Back Translation, Averaging)* Transformer big, only back translation EN-HI. We have not used any parallel data for EN-HI, only the back translated EN-HI data, beam=8; alpha=0.8; averaging of last 8 models; stopped after 1300k steps.

   Here, we have applied the output correction by identifying the source language (English) texts in the S1 model's output and replacing them with the corresponding target language (Hindi) output generated from the model S2. This ensures that less English language text appears in the S1 model output. English segments which still remained untranslated in the output were substituted by outputs of the model S3.

2. *S2: TransBig (1M EN-CS Transfer Learning, Averaging)*

[2] `https://nvidia.github.io/OpenSeq2Seq/html/api-docs/optimizers.html`

Corpus	Task	Setting	BLEU
IITB	EN-HI	S1: Back Translation (EN-HI)	**20.28**
IITB	EN-HI	S2: 1M EN-CS Transfer Learning, Back Translation (EN-HI)	20.07
IITB	EN-HI	S3: 1M EN-CS Transfer Learning, Back Translation (EN-HI), Genuine (EN-HI)	17.63
IITB	EN-HI	S4: 1M EN-CS Transfer Learning, Genuine (EN-HI)	16.49
IITB	EN-HI	S5: 1M EN-CS Transfer Learning, Genuine (EN-HI)	14.20
IITB	HI-EN	S6: 1M CS-EN Transfer Learning, Genuine (HI-EN)	17.80

Table 3: WAT 2018 Official Automatic Evaluation Results of our Models. All setups use "Transformer-Big" and checkpoint averaging.

Transformer big, transfer learning from EN-CS 1M steps. We have not used any parallel data for EN-HI, we only used the back translated EN-HI data, beam=8; alpha=0.8; averaging of last 8 models; stopped after 700k steps.

We also applied a similar output correction as in S1. We resorted to outputs of the model S1 or eventually S3 if English was produced instead of Hindi.

3. *S3: TransBig (1M EN-CS Transfer Learning, Back Translation (EN-HI Back + EN-HI Genuine, Averaging)*
Transformer big, transfer learning from EN-CS 1M steps, followed by only back translation EN-HI for 300k steps, followed by genuine EN-HI for 500k steps, beam=8; alpha=0.8; averaging of last 8 models.

4. *S4: TransBig (1M EN-CS Transfer Learning, Averaging)*
Transformer big, transfer learning from EN-CS 1M steps, only genuine EN-HI, beam=8; alpha=0.8; averaging of last 8 models; stopped after 230k steps.

5. *S5: TransBig (1M EN-CS Transfer Learning, Averaging)*
Baseline, transformer big only EN-HI, beam=8, alpha=0.8, averaging 8 steps; stopped after 330k steps.

6. *S6: TransBig (1M CS-EN Transfer Learning, Averaging)*
Transformer big, transfer learning from CS-EN 1M steps, only genuine HI-EN, beam=8; alpha=0.8; averaging of last 8 models; stopped after 230k steps. This model used primarily in back translation but we also submitted it to the HI-EN task.

5 Official Results

This section shows the official results of our models as published by WAT 2018 using automatic and manual evaluation. Further details on the evaluation can be found in Nakazawa et al. (2018) and all scores are available on the WAT 2018 website.[3]

We report the official automatic evaluation results of all our models for the test dataset here in Table 3. We see that the model S1 performed best in automatic evaluation. We observed similarly high BLEU scores on the development set but a small manual validation revealed that the translation quality is actually better in model S2 and S3, esp. due to Hindi grammar and word selection. Figure 1 provides an illustration. The output by S1 is a little shorter, so it risks fewer incorrect n-grams in BLEU evaluation (and the brevity penalty still does not strike too hard).

Based on this small manual analysis, we decided to submit models S2 and S3 and not model S1 for manual evaluation. (Participants could submit up to two models for manual evaluation.) The WAT2018 official manual scores for our systems and a competitor are shown in Table 4. This larger evaluation confirms our observation that BLEU does not correlate well with human judgment in this setting. We see that the model S3 outperformed S2 and also the competing system from another team by a large margin. In sum, it was S3, our third setup in terms of BLEU, that topped among all submission for EN-HI task in WAT2018.

[3] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/

English Input:
Politicians are loath to raise the tax even one penny when gas prices are high.

S1 Translated Output:
जब गैस की कीमतें ऊंची होती हैं तो राजनीतज्‌ञ कर एक पैसा भी बढ़ा देते हैं .
Gloss: When gas prices are high then politician tax one penny high

S2 Translated Output:
गैस की कीमतें ज्यादा होने पर एक पैसा भी टैक्स बढ़ाने के लिए राजनीतज्‌ञ लोन ले रहे हैं ।
Gloss: In case of gas prices are high politicians take loan even to increase one penny of tax

S3 Translated Output:
राजनीतज्‌ञों को गैस की कीमतें ऊंची होने पर भी एक पैसे का कर बढ़ाने की घृणा है ।
Gloss: Politicians are hate to increase one penny of tax even though gas prices are high

Figure 1: Sample Hindi Output Generated by the Settings S1, S2, and S3.

Team	Task	System	BLEU	Human	Note
CUNI	EN-HI	S3	17.63	**77.00**	
competitor	EN-HI	ConvS2S	19.69	69.50	Used external data
CUNI	EN-HI	S2	**20.07**	60.00	
competitor	EN-HI	ConvS2S	16.77	50.50	
competitor	HI-EN	ConvS2S	**20.63**	**72.25**	Used external data
CUNI	HI-EN	S6	17.80	67.25	

Table 4: WAT2018 Official Automatic and Manual Evaluation Results for IITB corpora.

6 Conclusion and Future Plans

In this system description paper, we presented our English-Hindi NMT system. We have highlighted the benefits of synthetic data and transfer learning. Our model that used all our components (synthetic data, genuine data and transfer learning from an unrelated English-Czech dataset) performed best in the official manual evaluation. We observed a clear mismatch of BLEU and manual evaluation.

As the next step, we plan to investigate corpus filtering, and iterative augmentation for performance improvement. Further exploration of the poor BLEU performance in this setting is also highly desirable.

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

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

Ondřej Bojar and Aleš Tamchyna. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT@EMNLP 2011, Edinburgh, Scotland, UK, July 30-31, 2011*, pages 330–336, 2011. URL `https://aclanthology.info/papers/W11-2138/w11-2138`.

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

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, 2017.

Chenhui Chu and Rui Wang. A Survey of Domain Adaptation for Neural Machine Translation. *arXiv preprint arXiv:1806.00258*, 2018.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, 2017.

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. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. URL `http://aclweb.org/anthology/Q17-1024`.

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

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

Tom Kocmi, Roman Sudarikov, and Ondřej Bojar. CUNI Submissions in WMT18. In *Proceedings of the 3rd Conference on Machine Translation (WMT)*, Brussels, Belgium, November 2018a.

Tom Kocmi, Dušan Variš, and Ondřej Bojar. CUNI Basque-to-English Submission in IWSLT18. *IWSLT. Bruges, Belgium*, 2018b.

Philipp Koehn. Neural machine translation. *arXiv preprint arXiv:1709.07809*, 2017.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, 2017.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. The IIT Bombay English-Hindi Parallel Corpus. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9.

Surafel Melaku Lakew, Mattia Antonino Di Gangi, and Marcello Federico. Multilingual Neural Machine Translation for Low Resource Languages. In *CLiC-it*, 2017.

Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Win Pa Pa, Isao Goto, Hideya Mino, Katsuhito Sudoh, and Sadao Kurohashi. Overview of the 5th Workshop on Asian Translation. In *Proceedings of the 5th Workshop on Asian Translation (WAT2018)*, Hong Kong, China, December 2018.

Toan Q. Nguyen and David Chiang. 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, 2017. URL `http://aclweb.org/anthology/I17-2050`.

Shantipriya Parida and Ondřej Bojar. Translating short segments with nmt: A case study in english-to-hindi. In *21st Annual Conference of the European Association for Machine Translation*, page 229, 2018.

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

Martin Popel. Cuni transformer neural mt system for wmt18. In *Proceedings of the Third Conference on Machine Translation*, pages 486–491, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W18-64051`.

Martin Popel and Ondřej Bojar. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70, 2018.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, August 2016a. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P16-1162`.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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)*, volume 1, pages 86–96, 2016b.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. Byte pair encoding: A text compression scheme that accelerates pattern matching. Technical report, Technical Report DOI-TR-161, Department of Informatics, Kyushu University, 1999.

Martin Thoma. The wili benchmark dataset for written language identification. *arXiv preprint arXiv:1801.07779*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 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. Association for Machine Translation in the Americas, 2018. URL `http://aclweb.org/anthology/W18-1819`.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1163. URL `http://www.aclweb.org/anthology/D16-1163`.