CUNI Submission for Low-Resource Languages in WMT News 2019

Tom Kocmi **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

This paper describes the CUNI submission to the WMT 2019 News Translation Shared Task for the low-resource languages: Gujarati-English and Kazakh-English. We participated in both language pairs in both translation directions. Our system combines transfer learning from different high-resource language pair followed by training on backtranslated monolingual data. Thanks to the simultaneous training in both directions, we can iterate the backtranslation process. We are using the Transformer model in a constrained submission.

1 Introduction

Recently, the rapid development of Neural Machine Translations (NMT) systems led to the claims, that human parity has been reached (Hassan et al., 2018) on a high-resource language pair Chinese-English. However, NMT systems tend to be very data hungry as Koehn and Knowles (2017) showed the NMT lacks behind phrase based approaches in the low-resource scenarios. This lead to the rise of attention in the low-resource NMT in recent years, where the goal is to improve the performance of a language pair that have only a limited available parallel data.

In this paper, we describe our approach to low-resource NMT. We use standard Transformer-big model (Vaswani et al., 2017) and apply two techniques to improve the performance on the low-resource language, namely transfer learning (Kocmi and Bojar, 2018) and iterative backtranslation (Hoang et al., 2018).

A model trained solely on the authentic parallel data of the low-resource NMT model has poor performance, thus using it directly for the backtranslation of monolingual data lead to poor translation. Hence the transfer learning is as a great tool to first improve the performance of the NMT system later used for backtranslating the monolingual data.

The structure of this paper is organized as follows. First, we describe the transfer learning and backtranslation, followed by a description of used datasets and the NMT model architecture. Next, we present our experiments, final submissions, and followup analysis of synthetic training data usage. The paper is concluded in Section 5.

2 Background

In this chapter, we first describe the technique of transfer learning and iterative backtranslation, followed by our training procedure that combines both approaches.

2.1 Transfer learning

Kocmi and Bojar (2018) presented a trivial method of transfer learning that uses a high-resource language pair to train the parent model. After the convergence, the parent training data are replaced with the training data of the low-resource language pair, and the training continues as if the replacement would not happen. The training continues without changing any parameters nor resetting moments or learning rate.

This technique of fine-tuning the model parameters is often used in a domain adaptation scenario on the same language pair. However, when using for different language pairs, there emerges a problem with vocabulary mismatch. Kocmi and Bojar (2018) overcome this problem by preparing the shared vocabulary for all languages in both language pairs in advance. Their approach is to prepare mixed vocabulary from training corpora of both languages and generate wordpiece vocabulary (Vaswani et al., 2017) from it.

We use the *balanced vocabulary* approach, that combines an equal amount of parallel data from both training corpora, low-resource as well as the same amount from high-resource language pair. Hence the low-resource language subwords are

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 234–240
Florence, Italy, August 1-2, 2019.

Corpora	Language pair	Sentence pairs	Words 1st lang.	Words in English
Commoncrawl	Russian-English	878k	17.4M	18.8M
News Commentary	Russian-English	235k	5.0M	5.4M
UN corpus	Russian-English	11.4M	273.2M	294.4M
Yandex	Russian-English	1000k	18.7M	21.3M
CzEng 1.7	Czech-English	57.4M	546.2M	621.9M
Crawl	Kazakh-English	97.7k	1.0M	1.3M
News commentary	Kazakh-English	9.6k	174.1k	213.2k
Wiki titles	Kazakh-English	112.7k	174.9k	204.5k
Bible	Gujarati-English	7.8k	198.6k	177.1k
Dictionary	Gujarati-English	19.3k	19.3k	28.8k
Govincrawl	Gujarati-English	10.7k	121.2k	150.6k
Software	Gujarati-English	107.6k	691.5k	681.3k
Wiki texts	Gujarati-English	18.0k	317.9k	320.4k
Wiki titles	Gujarati-English	9.2k	16.6k	17.6k

Table 1: The parallel training corpora used to train our models with counts of the total number of sentences as well as the number of words (segmented on space). More details on the individual corpora can be obtained at http://statmt.org/wmt19/.

represented in the vocabulary in the roughly same amount as the high-resource language pair.

As Kocmi and Bojar (2018), showed the language pair does not have to be linguistically related, and the most important criteria is the amount of parent parallel data. For this reason, we have selected Czech-English as a parent language pair for Gujarati-English and Russian-English as a parent for the Kazakh-English. The Russian was selected due to the use of Cyrillic and being a high-resource language pair. All language pairs share English. We prepare Gujarati-English and Kazakh-English systems separately from each other.

2.2 Backtranslation

The amount of available monolingual data typically exceeds the amount of available parallel data. The standard technique of using monolingual data in NMT is called backtranslation (Sennrich et al., 2016). It uses a second model trained in the reverse direction to translate monolingual data to the source language of the first model.

Backtranslated data are aligned with their monolingual sentences to create synthetic parallel corpora. The standard practice is to mix the authentic parallel corpora to the synthetic. Although it is not the only approach. (Popel, 2018) proposed a scenario of alternating the training between using only synthetic and only authentic corpora instead of mixing them.

This new corpus is used to train the first model by using backtranslated data as the source and the monolingual as the target side of the model.

Hoang et al. (2018) showed that backtranslation can be iterated and with the second round of backtranslation, we improve the performance of both models. However, the third round of backtranslation does yield better results.

The performance of the backtranslation model is essential. Especially in the low-resource scenario, the baseline models trained only on the authentic parallel data have a poor score (2.0 BLEU for English→Gujarati) generate very low quality backtranslated data. We have improved the baseline with the transfer learning to improve performance and generate the synthetic data of better quality.

2.3 Training procedure

We are training two models in parallel, one for each translation direction. Our training procedure is as follows. We train four parent models on the high-resource language pair until convergence: two models, one for each direction, for both directions. We stop training the models if there was no improvement bigger than 0.1 BLEU in the last 20% of the training time.

At this point, we run a hyperparameter search on the Gujarati→English and update the parameters for all following steps of all language pairs.

Afterward, we apply transfer learning on the authentic dataset of the corresponding low-resource language pair. We preserve the English side, thus Czech→English is a parent to Gujarati→English

Corpora	Lang.	Sent.	Words
News crawl 2018	EN	15.4M	344.3M
Common Crawl	KK	12.5M	189.2M
News commentary	KK	13.0k	218.7k
News crawl	Kk	772.9k	10.3M
Common Crawl	GU	3.7M	67.3M
Newscrawl	GU	244.9k	3.3M
Emille	GU	273.2k	11.4M

Table 2: Statistics of all monolingual data used for the backtranslation. It shows the number of sentences in each corpus and the number of words segmented on space. We mixed together all corpora for each language separately.

and English→Czech to English→Gujarati, likewise for the Russian-Kazakh.

After transfer learning, we select one of the translation directions to translate monolingual data. As a starting system for the backtranslation process, we have selected the English→Gujarati and Kazakh→English. This decision is motivated by choosing the better performing model in Kazakh-English language pair, and since the Gujarati-English have a similar score for both directions, we decided to select a model with English target side in contrast to Kazakh-English.

Following the backtranslation, we create synthetic data by mixing them with authentic parallel data and using to improve the performance of the second system. We continue repeating this process: Use the better system to backtranslate the data, and use this data in order to build an even better system in reverse direction.

We make two rounds of backtranslation for both directions on Gujarati-English and only one round of backtranslation on Kazakh-English due to the time consumption of the NMT translation process.

At last, we take the model with the highest BLEU score on the devset and average it with seven previous checkpoints to create final model.

3 Datasets and Model

In this section, we describe the datasets used to train our final models. All our models were trained only on the data allowed for the WMT 2019 News shared task. Hence our submission is constrained.

All used training data are presented in Table 1. We used all available parallel corpora allowed and accessible by WMT 2019 except for the Czech-English language pair, where we used only the CzEng 1.7. We have not clean any of the parallel corpora except deduplication and removing pairs with the same source and target translations in Wiki Titles dataset.

We used official WMT testsets from previous years as a development set. The year 2013 for Czech-English and Russian-English. For the Gujarati-English, we used the official 2019 development set. Lastly, for the Kazakh-English, the organizers do not provide any development set. Therefore we separated the first 2000 sentence pairs from the News Commentary training set and used as our development set.

The monolingual data used for the backtranslation are shown in Table 2. We use all available monolingual data for Gujarati and Kazakh. For the English, we did not use all available English monolingual data due to the backtranslation process being time-consuming, therefore we use only the 2018 News Crawl.

The available monolingual corpora are usually of high quality. However, we noticed that the Common Crawl contains many sentences in a different language and also long paragraphs, that are not useful for sentence level translation.

Therefore, we used language identification tool by Lui and Baldwin (2012) on the Common Crawl corpus and dropped all sentences automatically annotated as a different language than Gujarati or Kazakh respectively. Followed by splitting the remaining sentences that are longer than 100 words on all full stops, which led to an increase of sentences.

3.1 Model

The Transformer model seems superior to other NMT approaches as documented by several language pairs in the manual evaluation of WMT18 (Bojar et al., 2018).

We are using version 1.11 of sequence-to-sequence implementation of Transformer called tensor2tensor[1]. We are using the Transformer "big single GPU" configuration as described in (Vaswani et al., 2017), model which translates through an encoder-decoder 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 layers.

[1] https://github.com/tensorflow/tensor2tensor

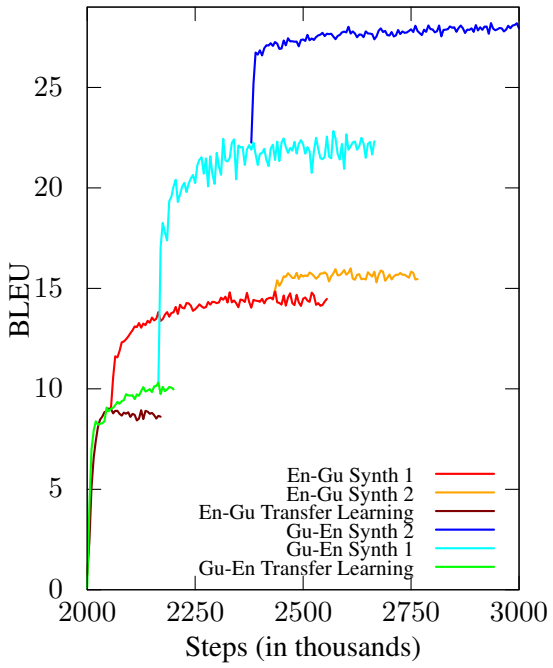

Figure 1: Learning curves for both directions of Gujarati-English models. The BLEU score is uncased and computed on the development set.

Popel and Bojar (2018) documented best practices to improve the performance of the model. Based on their observation, we are using as an optimizer Adafactor with inverse square root decay. Based on our previous experiments (Kocmi et al., 2018) we set the maximum number of subwords in a sentence to 100, which drops less than 0.1 percent of training sentences. However, it allows increasing the maximum size of the batch to 4500 for our GPU. The experiments are trained on a single GPU NVidia GeForce 1080 Ti.

4 Experiments

In this section, we describe our experiments starting with hyperparameter search, our training procedure, and supporting experiments.

All reported results are calculated over the testset of WMT 2019 and evaluated with case sensitive SacreBLEU (Post, 2018)[2] if not specified otherwise.

4.1 Hyperparameter search

Before the first step of transfer learning, we have done a hyperparameter search on Gujarati→English over the set of parameters that are not fixed from the parent (like dimensions of matrices or structure of layers). We examined the following hyperparameters: learning rate, dropout, layer prepostprocess dropout, label smoothing, and attention dropout.

The performance before hyperparameter search was 9.8 BLEU[3] for Gujarati→English, this score was improved to 11.0 BLEU. Based on the hyperparameter search we set the layer prepostprocess dropout and label smoothing both to 0.2 in the setup of Transformer-big.

These improvements show that transfer learning is not strictly associated with parent setup and that some parameters are possible to change. Although it must be noted, that we experimented only with a small subset of all hyperparameters and it is possible that other parameters could also be changed without damaging the parent model.

In this paper, we are using these parameters for all experiments (except for the parent models). Although applying hyperparameter search on each model separately or even between before each dataset switch is an interesting question, it is over the scope of this paper.

4.2 Problems with backtranslation

The synthetic data have a quality similar with the model by which they were produced. Since the low-resource scenario has an overall low quality, we observed, that the synthetic data contain many relics:

- Repeated sequence of words: The State Department has made no reference in statements, statements, statements, statements ...
- Sentences in Czech or Russian, most probably due to the parent model.
- Source sentences generated untranslated.

To avoid these problems, we cleaned all synthetic data in the following way. We had dropped all sentences, that contained any repetitive sequence of words. Then we checked the sentences by language identification tool (Lui and Baldwin, 2012) and dropped all sentences automatically annotated as a wrong language. The second step also filtered out some remaining gibberish translations.

We have not used beam search during backtranslation of monolingual data in order to speed up the translation process roughly 20 times compared to the beam search of 8.

[2]The SacreBLEU signature is BLEU + case.mixed + numrefs.1 + smooth.exp + tok.13a + version.1.2.12.

[3]This score is computed over devset with averaging of 8 latest models distanced one and half hour of training time.

Training dataset	EN→GU	GU→EN	EN→KK	KK→EN
Authentic (baseline)	2.0	1.8	0.5	4.2
Parent dataset	0.7	0.1	0.7	0.6
Authentic (transfer learning)	① 9.1	9.2	6.2	① 14.4
Synth generated by model ①	-	② 14.2	② 8.3	-
Synth generated by model ②	③ 13.4	-	-	17.3
Synth generated by model ③	-	④ 16.2	-	-
Synth generated by model ④	13.7	-	-	-
Averaging + beam 8	14.3	17.4	8.7	18.5

Table 3: Testset BLEU scores of our setup. Except for the baseline, each column shows improvements obtained after fine-tuning a single model on different datasets beginning with the score on a trained parent model.

4.3 Final models

Following the training procedure describe in Section 2.3, we trained the parent models for two million steps. One exception from the described approach is that we used a subset of 2M monolingual English data for the first round of backtranslation by the English→Gujarati model to cut down on the total consumed time.

Figure 1 shows the progress of training Gujarati-English models in both directions. The learning curves start at two millionth step as a visualization of the parent model training. We can notice that after each change of parallel data, there is a substantial increment of the performance. The learning curve is visualized on the development data, exact numbers for the testsets are in Table 3.

The baseline model in Table 3 is trained on the authentic data only, and it seems that the amount of parallel data is not sufficient to train the NMT model for the investigated language pairs. The rest of the rows shows incremental improvements of the models based on an undertaken step. The last step of model averaging takes the best performing model and averages it with the previous seven checkpoints that are distanced on average one and half hour of training time between each other.

We see that the transfer learning can be combined with iterated backtranslation on a low-resource language to obtain an improvement of 12.3 BLEU compared to the baseline in Gujarati→English and 15.6 in English→Gujarati.

For the final submission, we have selected models at following steps: step 2.99M for English→Gujarati, step 3.03M for Gujarati→English, step 2.48M for English→Kazakh and step 2.47M for Kazakh→English

4.4 Ratio of parallel data

Poncelas et al. (2018) showed that the balance between the synthetic and authentic data matters, and there should always be a part of authentic parallel data. We started our experiments with this intuition. However, the low-resource scenario complicates the setup since the amount of authentic data is several times smaller than synthetic. In order to balance the authentic and synthetic parallel data, we duplicated the authentic data several times.

We notice that the performance did not change from the setup that is using only synthetic data. Thus we prepare an experiment, where we do a second round of backtranslation on Gujarati→English with a various ratio of authentic and synthetic parallel data. For this experiment, we duplicated the full authentic parallel corpora of 173k sentences into a subsampled synthetic parallel corpus used in the second round of backtranslation. We have randomly selected 3.6M sentences from the synthetic corpora. The number of sentences is equal to 20x size of synthetic corpora. Therefore, we can present the ratio between authentic and synthetic corpora in percentage. The ratio in the legend of Figure 2 represent the actual ratio in the final corpus and not how much times the corpus has been duplicated. The synthetic is never duplicated, we only duplicate the authentic corpora. For example, the ratio "authentic:synthetic 1:2" means that the authentic has been multiplied ten times because the synthetic is twenty times bigger than the authentic corpora.

In Figure 2, we can see the difference between the amount of synthetic and authentic data. It seems that using only synthetic data generates the best performance, and whenever we increase the authentic part, the performance slowly decreases, contrary to the Poncelas et al. (2018). It could be

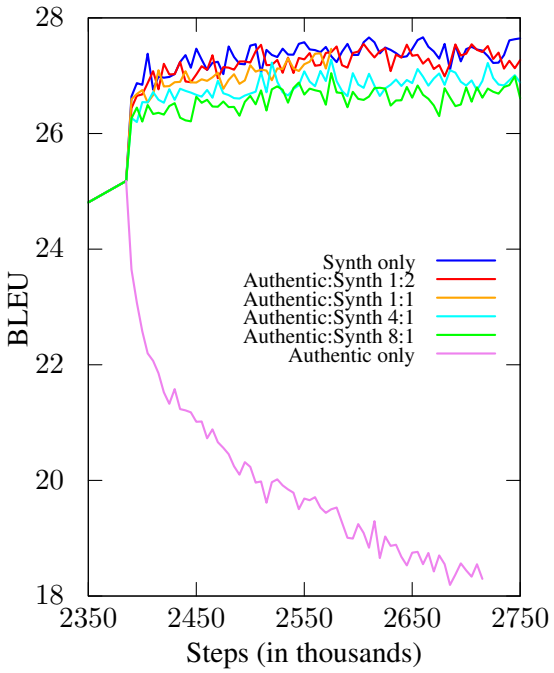

Figure 2: Comparison of different ratio of authentic and synthetic data.

due to the noise in the data, which implies that synthetic data are cleaner and more suitable for training the model.

4.5 Synthetic from Scratch

In the previous section, we have shown that during the iterative backtranslation of low-resource languages, the authentic data hurt the performance. In this section, we use the various ratios of training data and train the model from scratch without transfer learning or other backtranslation. Notably, all the parameters, as well as the wordpiece vocabulary, are the same.

Table 4 present the result of using synthetic data directly without any adaptation. It shows that having more authentic data hurt the low-resource languages. However, the most surprising fact is that training from scratch leads to significantly better model than the model trained by transfer learning and two rounds of the backtranslation by 0.7 (cased) BLEU. Unfortunately, we proposed this experiment after the submission. Therefore our final system has worse performance.

We believe it could be a result of unconscious overfitting to the development set because the performance on the development set is higher for our final model 26.9 BLEU compared to the performance of 25.8 BLEU for the synthetic only training. It could have been because we used development set three times during the training of the final model: first to select the best model from the transfer learning, then when selecting the best performing model in the first round of backtranslation and then third times during the second round of backtranslation. On the other hand, training on synthetic data from scratch used the development set only once for selection of the best performing model to evaluate.

Training dataset	cased	uncased
Authentic (baseline)	1.8	2.2
Synthetic only	**16.9**	**18.7**
Auth:Synth 1:1	16.8	18.4
Auth:Synth 2:1	16.3	17.8
Auth:Synth 4:1	15.2	16.8
Final model	16.2	17.9

Table 4: BLEU scores for training English→Gujarati from scratch on synthetic data from the second round of backtranslation. Neither of models uses the averaging or beam search. Thus the final model is our submitted model before averaging and beam search (the model ③). The scores are equal to those from http://matrix.statmt.org.

Another possible explanation is that the final model is already overspecialized on the data from the first round of backtranslation, that it is not able to adapt to the improved second synthetic data.

5 Conclusion

We participated in four translation directions on a low-resource language pairs in the WMT 2019 News translation Shared Task. We combined transfer learning with the iterated backtranslation and obtained significant improvements.

We showed that mixing authentic data and backtranslated data in a low-resource scenario does not affect the performance of the model: synthetic data is far more critical. This is a different result from what Poncelas et al. (2018) observed on higher-resource language pairs.

Lastly, in some scenarios, it is better to train the model on backtranslated data from scratch instead of fine-tuning the previous model.

In the future work, we want to investigate, why the training from scratch on backtranslated has led to better results. One of the reviewers suggested keep mixing the Czech→English corpus even during later stages of training as an additional source of parallel data, which we would like to compare.

Acknowledgments

This study was supported in parts by the grants SVV 260 453 of the Charles University, 18-24210S of the Czech Science Foundation and 825303 (Bergamot) of the European Union. 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 (LM2015071).

References

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*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.

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): Research Papers*, Brussels, Belgium.

Tom Kocmi, Roman Sudarikov, and Ondřej Bojar. 2018. Cuni submissions in wmt18. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 435–441, Belgium, Brussels. 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.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. 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. *arXiv preprint arXiv:1804.06189*.

Martin Popel. 2018. Machine translation using syntactic analysis. *Univerzita Karlova*.

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*.

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, 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.