CUNI Submission for the Inuktitut Language in WMT News 2020

Tom Kocmi

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

Abstract

This paper describes CUNI submission to the WMT 2020 News Translation Shared Task for the low-resource scenario Inuktitut–English in both translation directions. Our system combines transfer learning from a Czech–English high-resource language pair and backtranslation. We notice surprising behaviour when using synthetic data, which can be possibly attributed to a narrow domain of training and test data. We are using the Transformer model in a constrained submission.

1 Introduction

The rapid development of Neural Machine Translations (NMT) systems helped NMT in approaching human translation quality for high-resource language pairs like Chinese–English (Hassan et al., 2018) or English–Czech (Popel et al., 2020). This is not true for a low-resource scenario, where the lack of a large quantity of parallel data limits the performance of an NMT system. Thus, in recent years the research focused on low-resource NMT become important.

In this paper, we describe our approach to low-resource NMT of Inuktitut–English. We use the 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 backtranslation (Sennrich et al., 2016). We used a similar approach in WMT19 for Gujarati and Kazakh machine translation (Kocmi and Bojar, 2019).

Training low-resource model solely on authentic parallel data results in poor performance, and that results in low quality of generated backtranslation of monolingual data as well. Hence transfer learning is as an excellent method to first improve the performance of the NMT system that is later used for backtranslation of monolingual data.

2 Background

In this section, we describe the technique of transfer learning, backtranslation, and models that we used for training Inuktitut–English models.

2.1 Transfer Learning

Kocmi and Bojar (2018) presented a method of transfer learning that uses a high-resource language pair to train the "parent" model. After the training convergence, the parent training data are replaced with the training data of the low-resource language pair ("child"). Then the technique of fine-tuning continues without changing any parameters, resetting moments, nor changing the learning rate.

This technique has one shortcoming, and that is the problem with vocabulary mismatch. Kocmi and Bojar (2018) overcome this problem by preparing a shared vocabulary for all languages in both parent and child language pairs in advance. Their approach is to prepare a mixed subword vocabulary from the concatenation of training corpora for both languages.

We use their *balanced vocabulary* approach that combines an equal amount of parallel data from both training corpora, under-sampling the high-resource language pair as needed. Hence the low-resource language subwords are represented in the vocabulary with roughly the same prominence as the high-resource language pair ones.

Kocmi (2019) showed that parent and child language pairs do not have to be linguistically related, and more crucial criterion is the amount of parent parallel data. For this reason, we have selected Czech–English as a parent language pair, because it is one of the most resource-full languages allowed for WMT 2020.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 171–174
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

2.2 Backtranslation

The amount of available monolingual data typically exceeds the amount of available parallel data. One of the techniques for using monolingual data in NMT is called backtranslation (Sennrich et al., 2016). It uses a model trained in the reverse direction to translate monolingual data to the source language of the first model. Backtranslated sentences are then aligned with their monolingual sentences to create synthetic parallel corpora. The standard practice is to mix the authentic parallel corpora with the synthetic ones, although it is not the only possible approach. Popel (2018) obtained better results by repeatedly alternating between the training on the authentic and the synthetic portion of the parallel data instead of mixing them. We got inspired with this approach, and after training on synthetic data, we add a step to fine-tune again with authentic parallel data.

The performance of the backtranslation model is essential. Especially in the low-resource scenario, where the baseline models trained only on the authentic parallel data have a poor score, and they generate very low quality backtranslated data. Therefore, we first improve the performance of baseline with the transfer learning and generate the synthetic data of better quality.

Synthetic data can be noisy; therefore, we remove synthetic sentence pairs containing repetitive patterns, which is often a case of bad translation. We also remove sentences that contain Latin script in Inuktitut translations as Inuktitut has its script. This filtration reduced the number of synthetic sentences from 51.7M to 45.5M sentences.

2.3 Datasets and Model

All our models are trained only on the data allowed for the WMT 2020 News shared task. We use all available parallel data for Inuktitut–English prepared by Joanis et al. (2020). We use Czech–English corpus CzEng20 created by Kocmi et al. (2020) as a parent language pair. This corpus contains 61M authentic parallel sentences and also two sets of synthetic parallel sentences of similar size. For training parent model, we used all authentic parallel data plus one set of synthetic data with authentic English part. We ignored synthetic data with authentic Czech. The reason is that we assume the child model transfers knowledge mainly for the English language that is shared between both parent and child language pair. Additionally, we use monolingual English sentences from News Crawl 2018 and 2019 for backtranslation step. Results in section 3 are computed based on official WMT20 testset for Inuktitut–English (Joanis et al., 2020). All used training data are presented in table 1. We remove empty lines from Inuktitut–English training set.

Lang. pair	Sent. pairs	Words (CS/IU)	Words (EN)
CS–EN	137.2M	1913M	2176M
IU–EN	1.3M	8M	17M
Mono EN	51.7M	-	1756M

Table 1: More details on the training sizes of training corpora. Columns with words show number of words separated by space. All data are from http://statmt.org/wmt20/.

As for the model, we use 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. We use the version 1.11 of sequence-to-sequence implementation of Transformer called tensor2tensor.[1]

Popel and Bojar (2018) documented best practices to improve the performance of the model. Based on their observation, we use the Adafactor optimizer with inverse square root decay and 16000 warmup steps. Based on our previous experiments (Kocmi et al., 2018), we set the maximum number of subwords in a sentence to 100. The benefit is that the batch size can be increased to 4500 for our GPUs. The experiments are trained on a single GPU NVidia GeForce 1080 Ti or Quadro P5000.

3 Results

All reported results are calculated on the testset of WMT 2020 and evaluated with case sensitive SacreBLEU (Post, 2018).[2] The evaluation is done on unmodified outputs of our system. For final WMT20 submission, we have automatically corrected quotes to match the source. This step is not used for results in table 2.

The baseline models in table 2 are trained on the authentic data only. We have not focused on the backtranslation step for EN→IU as there are only 165k monolingual Inuktitut sentences available.

In IU→EN "Transfer from CS–EN" we get an

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

[2] The SacreBLEU signature is BLEU + case.mixed + numrefs.1 + smooth.exp + tok.13a + version.1.4.6

Training dataset	IU→EN	EN→IU
Authentic (baseline)	20.10	9.52
Transfer from CS–EN	22.98	10.41
Synthetic + auth	20.91	-
Authentic only	25.38	-

Table 2: Test set BLEU scores of our setup. Except for the baseline, each column shows improvements obtained after fine-tuning a model one line up on different datasets.

improvement of almost 3 BLEU. For a model, where we used a mix of "synthetic and authentic" data that is generated by our EN→IU model, we can notice a performance dropped from 22.98 to 20.91. However, following with fine-tuning this model again with authentic data, we get an increase in performance to 25.38. This is unexpected behaviour. Our understanding is that it could be attributed to a narrow domain of train and testset containing mainly speech transcripts. At the same time, the synthetic data are generated from English news articles, which is a more general domain. Therefore, while the model is trained on the general domain, it loses score on a domain-specific testset. This could be tested if we could obtain a testset on a different domain than speech transcription; however, we do not have such testset available.

During training on synthetic data, the model learns a general domain and loses performance on domain-specific testset to 20.91 BLEU, however after fine-tuning again on authentic domain-specific data only it reaches the highest performance of 25.38 BLEU.

4 Conclusion

We participated in the low-resource Inuktitut–English in the WMT 2020 News Translation Shared Task. We combined transfer learning with the back-translation and obtained significant improvements.

Surprisingly, we found out that although training the model on backtranslated data decreases the performance of the system in terms of BLEU score; it is still helpful when continued with fine-tuning on authentic data. We believe this is mainly because the Inuktitut–English training and test data are from a narrow domain of legal texts.

Acknowledgments

This study was supported in parts by the grants 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

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.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The nunavut hansard inuktitut–english parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2562–2572.

Tom Kocmi. 2019. *Exploring Benefits of Transfer Learning in Neural Machine Translation*. Ph.D. thesis, Charles University.

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 and Ondřej Bojar. 2019. CUNI Submission for Low-Resource Languages in WMT News 2019. In *Proceedings of the Fourth Conference on Machine Translation, Volume 2: Shared Task Papers*, Florence, Italy. Association for Computational Linguistics.

Tom Kocmi, Martin Popel, and Ondrej Bojar. 2020. Announcing czeng 2.0 parallel corpus with over 2 gigawords. *arXiv preprint arXiv:2007.03006*.

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.

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

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

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. 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, 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.