CUNI Systems for the Unsupervised and Very Low Resource Translation Task in WMT20

Ivana Kvapilíková **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 presents a description of CUNI systems submitted to the WMT20 task on unsupervised and very low-resource supervised machine translation between German and Upper Sorbian. We experimented with training on synthetic data and pre-training on a related language pair. In the fully unsupervised scenario, we achieved 25.5 and 23.7 BLEU translating from and into Upper Sorbian, respectively. Our low-resource systems relied on transfer learning from German–Czech parallel data and achieved 57.4 BLEU and 56.1 BLEU, which is an improvement of 10 BLEU points over the baseline trained only on the available small German–Upper Sorbian parallel corpus.

1 Introduction

An extensive area of the machine translation (MT) research focuses on training translation systems without large parallel data resources (Artetxe et al., 2018b,a, 2019; Lample et al., 2018a,b). The WMT20 translation competition presents a separate task on unsupervised and very low-resource supervised MT.

The organizers prepared a shared task to explore machine translation on a real-life example of a low-resource language pair of German (de) and Upper Sorbian (hsb). There are around 60k authentic parallel sentences available for this language pair which is not sufficient to train a high-quality MT system in a standard supervised way, and calls for unsupervised pre-training (Conneau and Lample, 2019), data augmentation by synthetically produced sentences (Sennrich et al., 2016a) or transfer learning from different language pairs (Zoph et al., 2016a; Kocmi and Bojar, 2018).

The WMT20 shared task is divided into two tracks. In the unsupervised track, the participants are only allowed to use monolingual German and Upper Sorbian corpora to train their systems; the low-resource track permits the usage of auxiliary parallel corpora in other languages as well as a small parallel corpus in German–Upper Sorbian.

We participate in both tracks in both translation directions. Section 2 describes our participation in the unsupervised track and section 3 describes our systems from the low-resource track. Section 4 introduces transfer learning via Czech (cs) into our low-resource system. We conclude the paper in section 5.

2 Unsupervised MT

Unsupervised machine translation is the task of learning to translate without any parallel data resources at training time. Both neural and phrase-based systems were proposed to solve the task (Lample et al., 2018b). In this work, we train several neural systems and compare the effects of different training approaches.

2.1 Methodology

The key concepts of unsupervised NMT include a shared encoder, shared vocabulary and model initialization (pre-training). The training relies only on monolingual corpora and switches between de-noising, where the model learns to reconstruct corrupted sentences, and online back-translation, where the model first translates a batch of sentences and immediately trains itself on the generated sentence pairs, using the standard cross-entropy MT objective (Artetxe et al., 2018b; Lample et al., 2018a).

We use a 6-layer Transformer architecture for our unsupervised NMT models following the approach by Conneau and Lample (2019). Both the encoder and the decoder are shared across languages.

We first pre-train the encoder and the decoder separately on the task of cross-lingual masked language modelling (XLM) using monolingual

Proceedings of the 5th Conference on Machine Translation (WMT), pages 1123–1128

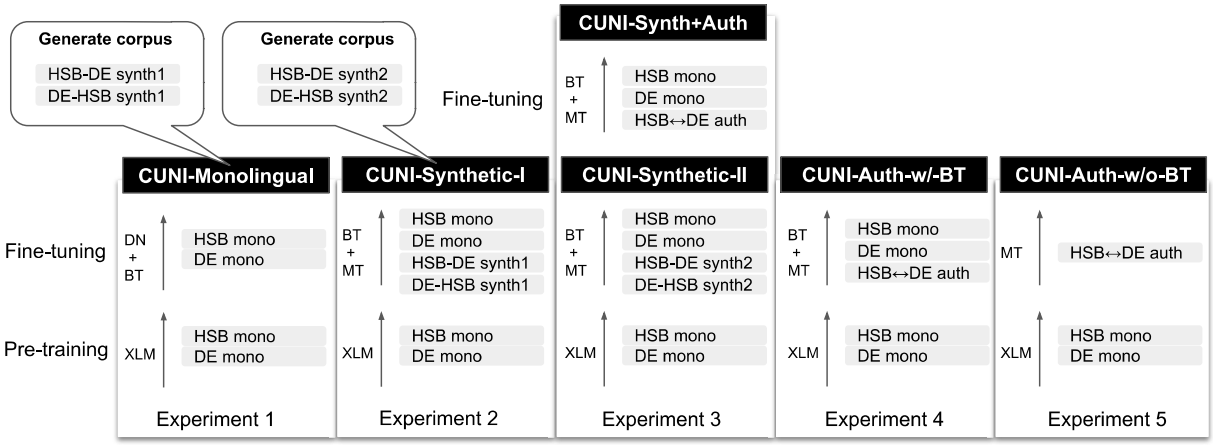

Figure 1: **An overview of selected CUNI systems.** Corpora are illustrated in gray boxes, system names in black boxes. Systems are trained with indicated training objectives: cross-lingual masked language modeling (XLM), denoising (DN), online back-translation (BT), and standard machine translation objective (MT). Monolingual training sets *DE mono* and *HSB mono* were available for both WMT20 task tracks, the parallel training set *HSB↔DE auth* was only allowed in the low-resource supervised track.

data only (Conneau and Lample, 2019). Subsequently, the initialized MT system (*CUNI-Monolingual*) is trained using de-noising and online back-translation. We then use this system to translate our entire monolingual corpus and train a new system (*CUNI-Synthetic-I*) from scratch on the two newly generated synthetic parallel corpora *DE-HSB synth1* and *HSB-DE synth1*. Finally, we use the new system to generate *DE-HSB synth2* and *HSB-DE synth2*, and repeat the training to evaluate the effect of another back-translation round (*CUNI-Synthetic-II*).

All unsupervised systems are trained using the same BPE subword vocabulary (Sennrich et al., 2016b) with 61k items generated using `fastBPE`.[1] An overview of the systems and their training stages is given in fig. 1.

2.2 Data

Our `de` training data comes from News Crawl; the `hsb` data was provided for WMT20 by the Sorbian Institute and the Witaj Sprachzentrum.[2] Most of the `hsb` data was of high quality but we fed the web-scraped corpus (*web_monolingual.hsb*) through a language identification tool `fastText`[3] to identify proper `hsb` sentences. All `de` data was also cleaned using this tool.

The final monolingual training corpora have 22.5M sentences (*DE mono*) and 0.6M sentences (*HSB mono*). Synthetic parallel corpora are generated from the monolingual data sets by coupling the sentences with their translation counterparts as described in section 2.1.

The parallel development (dev) and testing (dev test) data sets of 2k sentence pairs provided by WMT20 organizers are used for parameter tuning and model selection. The final evaluation is run on the blind test set *newstest2020*.

2.3 Results

The resulting scores measured on the blind *newstest2020* are listed in table 1 and table 2. The translation quality metrics BLEU (Papineni et al., 2002), TER (Snover et al., 2006), BEER (Stanojević and Sima'an, 2014) and CharacTER (Wang et al., 2016) provide consistent results. The best quality is reached when using synthetic corpora from the second back-translation iteration, although the second round adds only a slight improvement. A similar observation is made by Hoang et al. (2018) who show that the second round of back-translation does not enhance the system performance as much as the first round. Additionally, the third round does not produce any significant gains.

When training on synthetic parallel corpora, it is still beneficial to perform back-translation on-the-fly (Artetxe et al., 2018b) whereby new training instances of increasing quality are generated in every training step. This method adds 1 - 2 BLEU points to the final score as compared to training

[1] https://github.com/glample/fastBPE
[2] http://www.statmt.org/wmt20/unsup_and_very_low_res/
[3] https://github.com/facebookresearch/fastText/

		newstest2020					*dev test set*
	System Name	**BLEU**	**BLEU-cased**	**TER**	**BEER 2.0**	**CharacTER**	**BLEU**
a	**CUNI-Monolingual**	**23.7**	**23.4**	0.606	**0.530**	0.559	23.4
	CUNI-Synthetic-I	23.4	23.2	0.617	0.531	0.575	22.2
	CUNI-Synthetic-II*	**23.7**	**23.4**	**0.618**	**0.530**	**0.563**	**23.7**
b	**CUNI-Supervised-Baseline**	43.7	43.2	0.439	0.670	0.382	38.7
	CUNI-Auth-w\o-BT	51.6	51.2	0.362	0.710	0.332	48.3
	CUNI-Auth-w\-BT	**54.3**	**53.9**	**0.337**	**0.726**	**0.310**	**52.1**
	CUNI-Synth+Auth*	53.8	53.4	0.343	0.721	0.315	50.5

Table 1: Translation quality of the unsupervised (a) and low-resource supervised (b) `hsb` → `de` systems on *newstest2020* and the unofficial test set. The asterisk * indicates systems submitted into WMT20.

		newstest2020					*dev test set*
	System Name	**BLEU**	**BLEU-cased**	**TER**	**BEER 2.0**	**CharacTER**	**BLEU**
a	**CUNI-Monolingual**	21.7	21.2	0.670	0.497	0.557	20.4
	CUNI-Synthetic-I	24.9	24.5	0.599	0.535	0.521	25.1
	CUNI-Synthetic-II*	**25.5**	**25.0**	**0.592**	**0.540**	**0.516**	**25.3**
b	**CUNI-Supervised-Baseline**	40.8	40.3	0.452	0.655	0.373	38.3
	CUNI-Auth-w\o-BT	47.5	47.1	0.390	0.689	0.336	47.1
	CUNI-Auth-w\-BT	**52.3**	**51.8**	**0.350**	**0.718**	**0.301**	**52.4**
	CUNI-Synth+Auth*	50.6	50.1	0.368	0.703	0.326	50.4

Table 2: Translation quality of the unsupervised (a) and low-resource supervised (b) `de` → `hsb` systems on *newstest2020* and the unofficial test set. The asterisk * indicates systems submitted into WMT20.

only on sentence pairs from the two synthetic corpora so we use it in all our unsupervised systems.

We used the `XLM`[4] toolkit for running the experiments. Language model pre-training took 4 days on 4 GPUs[5]. The translation models were trained on 1 GPU[6] with 8-step gradient accumulation to reach an effective batch size of 8 × 3400 tokens. We used the Adam (Kingma and Ba, 2015) optimizer with inverse square root decay ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $lr = 0.0001$) and greedy decoding.

3 Very Low-Resource Supervised MT

3.1 Methodology

Our systems introduced in this section have the same model architecture as described in section 2, but now we allow the usage of authentic parallel data. We pre-train a bilingual XLM model and fine-tune with either only authentic parallel data (*CUNI-Auth-w\o-BT*) or both parallel and monolingual data, using a combination of standard MT training and online back-translation (*CUNI-Auth-w\-BT*). Finally, we utilize the trained model *CUNI-Synthetic-II* from section 2 and fine-tune it on the authentic parallel corpus, again using standard supervised training as well as online back-translation (*CUNI-Synth+Authentic)*.

All systems are trained with the same BPE subword vocabulary of 61k items.

3.2 Data

In addition to the data described in section 2.2, we used the authentic parallel corpus of 60k sentence pairs provided by Witaj Sprachzentrum mostly from the legal and general domain.

3.3 Results

The resulting scores are listed in the second part of table 1 and table 2. We compare system performance against a supervised baseline which is a vanilla NMT model trained only on the small parallel train set of 60k sentences, without any pre-training or data augmentation.

Our best system gains 11.5 BLEU over this baseline, utilizing the larger monolingual corpora for XLM pre-training and online back-translation. Fine-tuning one of the trained unsupervised systems on parallel data leads to a lower gain of ∼10 BLEU points over the baseline.

The translation models were trained on 1 GPU[7] with 8-step gradient accumulation to reach an effective batch size of 8 × 1600 tokens. Other training details are equivalent to section 2.1.

[4] https://github.com/facebookresearch/XLM

[5] GeForce GTX 1080, 11GB of RAM

[6] Quadro P5000, 16GB of RAM

[7] GeForce GTX 1080 Ti, 11GB of RAM

System Name	BLEU	BLEU-cased	TER	BEER 2.0	CharacTER
Helsinki-NLP	60.0	59.6	0.286	0.761	0.267
NRC-CNRC	59.2	58.9	0.290	0.759	0.268
SJTU-NICT	58.9	58.5	0.296	0.754	0.274
CUNI-Transfer	57.4	56.9	0.307	0.746	0.285
Bilingual only	47.8	47.4	0.394	0.695	0.356

Table 3: Translation quality of `hsb` → `de` systems on newstest2020.

System Name	BLEU	BLEU-cased	TER	BEER 2.0	CharacTER
SJTU-NICT	61.1	60.7	0.283	0.759	0.250
Helsinki-NLP	58.4	57.9	0.297	0.755	0.255
NRC-CNRC	57.7	57.3	0.300	0.754	0.255
CUNI-Transfer	56.1	55.5	0.315	0.743	0.265
Bilingual only	46.8	46.4	0.389	0.692	0.335

Table 4: Translation quality of `de` → `hsb` systems on newstest2020.

4 Very Low-Resource Supervised MT with Transfer Learning

One of the main approaches to improving performance under low-resource conditions is transferring knowledge from different high-resource language pairs (Zoph et al., 2016b; Kocmi and Bojar, 2018). This section describes the unmodified strategy for transfer learning as presented by Kocmi and Bojar (2018), using German–Czech as the parent language pair. Since we do not modify the approach nor tune hyperparameters of the NMT model, we consider our system a transfer learning baseline for low-resource supervised machine translation.

4.1 Methodology

Kocmi and Bojar (2018) proposed an approach to fine-tune a low-resource language pair (called "child") from a pre-trained high-resource language pair (called "parent") model. The method has only one restriction and that is a shared subword vocabulary generated from the corpora of both the child and the parent. The training procedure is as follows: first train an NMT model on the parent parallel corpus until it converges, then replace the training data with the child corpus.

We use the Tensor2Tensor framework (Vaswani et al., 2018) for our transfer learning baseline and model parameters "Transformer-big" as described in (Vaswani et al., 2018). Our shared vocabulary has 32k wordpiece tokens. We use the Adafactor (Shazeer and Stern, 2018) optimizer and a reverse square root decay with 16 000 warm-up steps. For the inference, we use beam search of size 8 and alpha 0.8.

4.2 Data

In addition to the data described in section 3.2, we used the `cs-de` parallel corpora available at the OPUS[8] website: OpenSubtitles, MultiParaCrawl, Europarl, EUBookshop, DGT, EMEA and JRC. The `cs-de` corpus has 21.4M sentence pairs after cleaning with the `fastText` language identification tool.

4.3 Results

We compare the results of our transfer learning baseline called *CUNI-Transfer* with three top performing systems of WMT20. These systems use state-of-the-art techniques such as BPE-dropout, ensembling of models, cross-lingual language modelling, filtering of training data and hyperparameter tuning. Additionally, we also include results for a system we trained without any modifications solely on bilingual parallel data (*Bilingual only*).[9]

The results in table 4 show that training solely on German–Upper Sorbian parallel data leads to a performance of 47.8 BLEU for `de`→`hsb` and 46.7 BLEU for `hsb`→`de`. When using transfer learning with a Czech–German parent, the performance increases by roughly 10 BLEU points to 57.4 and 56.1 BLEU. As demonstrated by the winning system, the performance can be further boosted using additional techniques and approaches to 60.0 and 61.1 BLEU. This shows that transfer learning plays an important role in the low-resource scenario.

[8] `http://opus.nlpl.eu/`

[9] The model *Bilingual only* is trained on the same data as *CUNI-Supervised-Baseline* but uses a different architecture and decoding parameters.

5 Conclusion

We participated in the unsupervised and low-resource supervised translation task of WMT20.

In the fully unsupervised scenario, the best scores of 25.5 (hsb→de) and 23.7 (de→hsb) were achieved using cross-lingual language model pre-training (XLM) and training on synthetic data produced by NMT models from earlier two iterations. We submitted this system under the name *CUNI-Synthetic-II*.

In the low-resource supervised scenario, the best scores of 57.4 (hsb→de) and 56.1 (de→hsb) were achieved by pre-training on a large German–Czech parallel corpus and fine-tuning on the available German–Upper Sorbian parallel corpus. We submitted this system under the name *CUNI-Transfer*.

We showed that transfer learning plays an important role in the low-resource scenario, bringing an improvement of ∼10 BLEU points over a vanilla supervised MT model trained on the small parallel data only. Additional techniques used by other competing teams yield further improvements of up to 4 BLEU over our transfer learning baseline.

Acknowledgments

This study was supported in parts by the grants 19-26934X and 18-24210S of the Czech Science Foundation, SVV 260 575 and GAUK 1050119 of the Charles University. 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 (LM2018101).

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on EMNLP*, Brussels. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the ACL*, Florence. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

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*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

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*, Brussels. Association for Computational Linguistics.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations (ICLR 2018)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on EMNLP*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the ACL*, Philadelphia. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the ACL*, Berlin. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th of the AMTA*, Cambridge. Association for Machine Translation in the Americas.

Miloš Stanojević and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on EMNLP*, Doha, Qatar. 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 AMTA (Volume 1: Research Papers)*, Boston, MA. Association for Machine Translation in the Americas.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, Berlin, Germany. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016a. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on EMNLP*, Austin, Texas. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016b. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on EMNLP*, Austin, Texas. Association for Computational Linguistics.