The LMU Munich Systems for the WMT21 Unsupervised and Very Low-Resource Translation Task

Jindřich Libovický and **Alexander Fraser**
Center for Information and Language Processing
LMU Munich
{libovicky,fraser}@cis.lmu.de

Abstract

We present our submissions to the WMT21 shared task in Unsupervised and Very Low-Resource machine translation between German and Upper Sorbian, German and Lower Sorbian, and Russian and Chuvash. Our low-resource systems (German↔Upper Sorbian, Russian↔Chuvash) are pre-trained on high-resource pairs of related languages. We fine-tune those systems using the available authentic parallel data and improve by iterated back-translation. The unsupervised German↔Lower Sorbian system is initialized by the best Upper Sorbian system and improved by iterated back-translation using monolingual data only.

1 Introduction

In this paper, we describe systems for translation between German (*de*) and Upper Sorbian (*hsb*), German (*de*) and Lower Sorbian (*dsb*), and Russian (*ru*) and Chuvash (*cv*) developed at LMU Munich for the WMT21 shared task on unsupervised and very low resource machine translation (MT).

Upper Sorbian is a minority language spoken by around 30,000 people in today's German federal state of Saxony, Lower Sorbian has around 7,000 speakers and is spoken in the German federal state of Brandenburg. With such a small number of speakers, machine translation and automatic processing of Sorbian language is an inherently low-resource problem without any chance that the resources available for Sorbian would ever approach the size of resources for languages spoken by millions of people. On the other hand, being Western Slavic languages related to Czech and Polish, it is possible to take advantage of relatively rich resources collected for these two languages.

Unlike our last year's submission for Upper Sorbian (Libovický et al., 2020), we decided not to use synthetic data from unsupervised translation between Czech and Upper Sorbian and only did iterative back-translation. Despite having more authentic parallel data than last year, our system reaches approximately the same translation quality. Our Upper Sorbian systems ranked third out of six systems in the official ranking.

We leverage the relatedness between the Sorbian languages and use the Upper Sorbian system as a starting point for iterative back-translation using monolingual data only. Our Lower Sorbian Systems ranked second (de→dsb) and third (dbs→de) out of four teams in the official ranking.

Chuvash is a minority language spoken in the Volga region in the southwest of Russia. Although it uses the Cyrillic script, it is not related to eastern Slavic languages, but it is a Turkic language, relatively isolated in the Turkic language family. As a language with the highest number of speakers in this shared task, it also has the highest amount of available parallel data. We adopt a similar approach as for German-Upper Sorbian translation and pre-train our models on the related Kazakh language. In addition, we experiment with character-level models in the hope that they will be particularly effective for agglutinative morphology.

2 Experimental Setup

Most of our experimental setup is shared across all the language pairs. All our models use the Transformer architecture (Vaswani et al., 2017) as implemented in FairSeq (Ott et al., 2019).

All data is segmented using BPE (Sennrich et al., 2016b) with 16k merge operations as implemented in YouTokenToMe[1] without previous explicit tokenization. The merges are computed using a concatenation of all training data: German, Czech, Upper and Lower Sorbian in the first set of experiments, Russian, Kazakh, and Chuvash in the second set of experiments.

For the supervised task, we first pre-train mod-

[1] https://github.com/VKCOM/YouTokenToMe

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 989–994

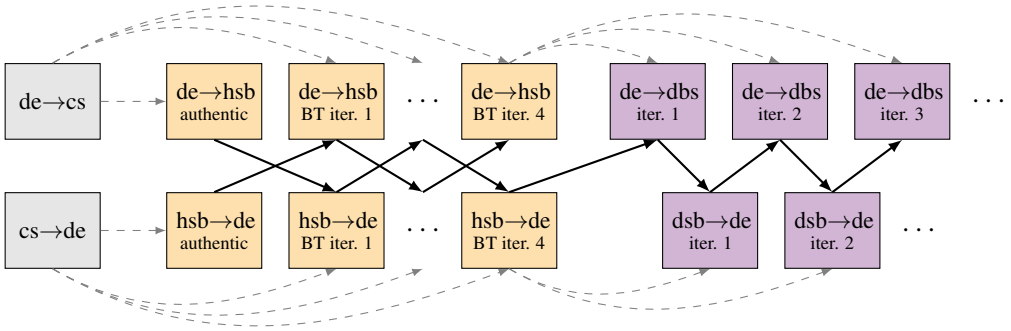

Figure 1: A diagram of the training procedure of the German↔Upper/Lower Sorbian systems. Gray dashed arrows (- -▸) denote model initialization, solid black arrows (⟶) denote syntetic data generation by back-translation.

els on high-resource related languages: Russian-Kazakh for Chuvash and German-Czech for Upper Sorbian. We first train Transformer Base models on authentic data. These systems are used to generate back-translation (Sennrich et al., 2016a) of monolingual data. Using tagged back-translation (Caswell et al., 2019), we trained Transformer Big models for German↔Czech and Russian↔Kazakh translation. All back-translation steps use sampling and simple length-based filtering as proposed by Edunov et al. (2018)[2]. We upsample the authentic parallel data to match the size of the synthetic data.

We keep most default hyperparameters from the predefined architectures in FairSeq (`transformer` for the Base model, `transformer_wmt_en_de_big_t2t` for the Big model. The batch size is 6k tokens for the Base models, 2k tokens for Big models on a single GPU, Because we always start with high-resource training, we keep the dropout on the standard value of 0.1.

We use these models to initialize the weights (Nguyen and Chiang, 2017; Kocmi and Bojar, 2018) of the supervised low-resource models without restarting the optimizer. Because the learning rate is already low at that stage of training, we do not need to change the dropout to prevent overfitting. First, we train the supervised models using the authentic parallel data only, then we continue with iterated back-translation. The best Upper Sorbian-to-German model is used to translate Lower Sorbian monolingual data into German. In the next steps, we continue with a standard iterative back-translation procedure for unsupervised neural machine translation (Artetxe et al., 2018; Lample et al., 2018).

[2]We re-used the published code `https://github.com/pytorch/fairseq/tree/master/examples/backtranslation`.

Our final submission is an ensemble (with the vote strategy) of the best-scoring systems in the process of iterated back-translation. Language-pair-specific descriptions and results are discussed in the following sections.

We evaluate our systems using the BLEU Score (Papineni et al., 2002), chrF score (Popović, 2015) as implemented in SacreBLEU (Post, 2018).[3] Further, we evaluate the models using BERTScore (Zhang et al., 2020)[4] with XLM-RoBERTa Large (Conneau et al., 2020) as an underlying model for German and Russian and mBERT (Devlin et al., 2019) for Chuvash. Similar to the official task evaluation, we also report for each system the number of significantly worse systems in each metric at the significance level 0.95 with bootstrap resampling (Koehn, 2004) with 1k samples. For each metric, each system receives one point for each system it significantly outperforms in the metric at the significance level of 0.95.

3 German ↔ Upper Sorbian

Pre-training. For training the German↔Czech systems, we followed the same setup as in our last year's submission (Libovický et al., 2020). We used all parallel datasets from the Opus project (Tiedemann, 2012), which was 15.4M sentences after filtering by length and language identity. We trained a Transformer Base model on this data and used this model to generate back-translation. We used 20M Czech and 20M German sentences from the WMT News Crawl. We mix the back-translated and authentic parallel data one-to-one and train Transformer Big models on it.

[3]BLEU score signature `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0` chrF score signature `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.0.0`

[4]`https://github.com/Tiiiger/bert_score`

	hsb → de				de → hsb		
	BLEU	chrF	BERTScore	Points	BLEU	chrF	Points
Authentic data only	53.4 [0]	.763 [0]	.933 [0]	[0]	54.9 [0]	.769 [0]	[0]
BT iter 1	55.2 [0]	.773 [0]	.936 [1]	[1]	56.4 [0]	.778 [0]	[0]
BT iter 2	55.8 [1]	.777 [1]	.937 [2]	[4]	56.5 [0]	.778 [0]	[0]
BT iter 3	55.8 [1]	.777 [1]	.937 [3]	[5]	56.2 [0]	.778 [0]	[0]
BT iter 4	56.1 [1]	.779 [1]	.938 [5]	[7]	56.0 [0]	.776 [0]	[0]
Ensemble	56.2 [1]	.779 [1]	.938 [4]	[6]	56.4 [0]	.779 [0]	[0]

Table 1: Quantitative results of the German↔Upper Sorbian translation systems on the development test data.

Sorbian data. We used all Upper Sorbian data provided for the shared task, i.e., 148k parallel sentence pairs (this is 88k sentence pairs more than last year), we did not apply any filtering on the parallel dataset. The development validation and the development test set of 2k sentences were the same as the last year.

Back-translation. We used 15M German sentences from the WMT News Crawl and all available monolingual Upper Sorbian data, 696k sentences, for back-translation. We applied the same rule-based statistical fixing of hyphenation-related OCR errors as the last year (Libovický et al., 2020, § 3.1). To better leverage the limited amount of monolingual data, we sample the Upper Sorbian translations 5×. We iterated the back-translation 4 times, always initializing the model with the Czech-German models (see Figure 1).

Results. The results are presented in Table 1. In the translation direction into German, the translation quality gradually increased between the back-translation steps. In the opposite direction, the translation quality oscillated. We attribute this to a larger amount of authentic German sentences. Ensembling only has a negligible effect. Note also that for translation into Sorbian, no differences between the models are statistically significant. In the opposite direction, the BLEU and the chrF score only separate the systems into two clusters, whereas the differences among BERTScores are always significant in the bootstrap testing, even though the absolute score differences are smaller. The best system for translation into German is a single from the last iteration of back-translation despite scoring slightly worse in the BLEU score.

4 German ↔ Lower Sorbian

Data. Because this is a purely unsupervised task, we did not use any Lower Sorbian parallel data.

		BLEU	chrF	BERTScore
dsb→de	Single	33.7	.606	.873
	Ensemble	33.8	.602	.874
de→dsb	Single	30.1	.587	—
	Ensemble	30.1	.588	—

Table 2: Automatic scores for the best German↔Lower Sorbian Systems.

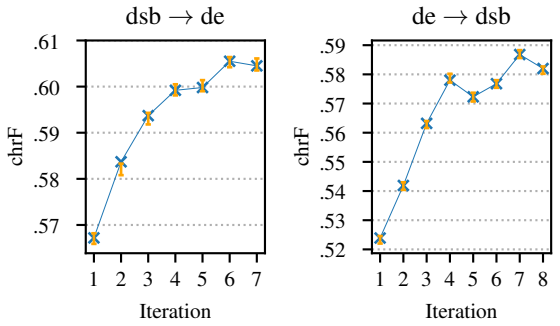

Figure 2: chrF scores during iterative back-translation for unsupervised German↔Lower Sorbian translation. The orange vertical lines denote 95%-confidence intervals using bootstrap resampling.

We used the same German monolingual data as we used for back-translation for Upper Sorbian. We use all the Lower Sorbian monolingual data, 145k sentences, provided by the organizers.

Iterative back-translation. Similarly to Upper Sorbian, we sample the back-translation of Lower Sorbian 10× for higher diversity in the training data.

Results. The final results are tabulated in Table 2. Figure 2 shows the translation quality in terms of chrF score during back-translation iterations. Similar to Upper Sorbian, the direction into German that uses larger monolingual data tends to improve more smoothly than the opposite direction. Also, the ensembling of the three best-scoring systems only has a negligible effect. The single system and

	cv → ru				ru → cv			
	BLEU	chrF	BERTScore	Points	BLEU	chrF	BERTScore	Points
Authentic data only	20.5 [2]	.451 [2]	.847 [3]	[7]	18.4 [0]	.486 [2]	.854 [3]	[5]
BT iteration 1	19.1 [0]	.443 [2]	.846 [2]	[4]	18.6 [0]	.487 [2]	.854 [4]	[6]
BT iteration 2	20.3 [2]	.450 [2]	.848 [4]	[8]	18.5 [0]	.487 [2]	.854 [2]	[4]
Ensemble of the two above	20.0 [2]	.450 [2]	.848 [4]	[8]	18.8 [1]	.489 [2]	.855 [5]	[8]
BT iteration 1 to char	18.0 [0]	.423 [0]	.843 [1]	[1]	16.9 [0]	.457 [0]	.850 [0]	[0]
BT iteration 2 to char	17.4 [0]	.420 [0]	.841 [0]	[0]	17.1 [0]	.463 [0]	.851 [1]	[1]
Ensemble of the two above	20.0 [2]	.450 [2]	.848 [4]	[8]	18.9 [1]	.490 [2]	.855 [5]	[8]

Table 3: Quantitative results of the Russian↔Chuvash translation systems on the development test data.

the ensemble do not significantly differ in any of the metrics.

5 Russian ↔ Chuvash

Pre-training. Similar to Upper Sorbian systems, we pre-train the systems on high-resource related language pair, Kazakh-Russian. We used the crawled Kazakh-Russian corpus of 5M sentence pairs published for WMT19 (Barrault et al., 2019) to train a Transformer Base model. We used these models to back-translation 3M Kazakh and 3M Russian sentences from the WMT News Crawl from the most recent years.

Chuvash data. We used all parallel data provided by the organizers, 717k sentence pairs, without any filtering. For back-translation, we used all 2.8M monolingual Chuvash sentences provided for the competition. For Russian, we used 18M monolingual sentences from the WMT News Crawl.

Back-translation. We ran two iterations of back-translation. We sample from the model during back-translation. We sampled 4 different translations for each Chuvash sentence to increase the training data diversity. We mix the authentic and synthetic parallel training data in the one-to-one ratio. All models are initialized by the Russian↔Kazakh models.

Character models. We further experiment with finetuning the system to the character level. Libovický and Fraser (2020) managed to train a character-level system for another Turkic language, English-to-Turkish translation. Here, we test if this is a property of Turkic languages or an artifact of the dataset English-Turkish dataset. We follow Libovický and Fraser (2020) and finetune the subword model to the character level.

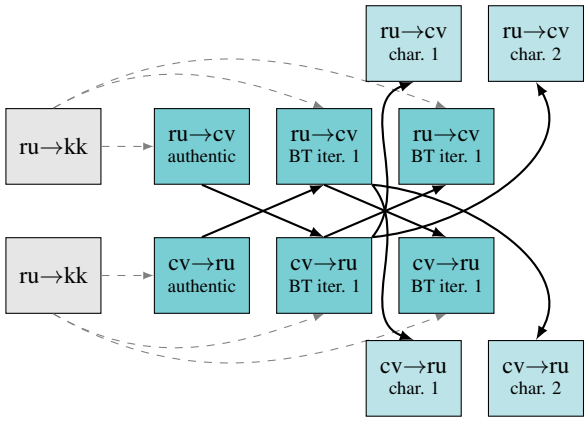

Figure 3: A diagram of the training procedure of the Russian↔Chuvash. Gray dashed arrows (- - ▸) denote model initialization, solid black arrows (⟶) denote syntetic data generation by back-translation.

Results. The results are presented in Table 3. Compared to other language pairs, back-translation had a surprisingly small effect on the translation quality. We suspect this result might be due to errors in data processing or signalize a need for a better data filtering technique. Model ensembling has no effect here. The character-level systems are on average 2 BLEU points worse than their subword counterparts, which is consistent with the results of character-level models on high-resource languages (Libovický and Fraser, 2020). Surprisingly, the character-level models seem to have much larger gains from model ensembling than the subword-based models. In fact, the ensemble of the character-level models is statistically indistinguishable from the best subword-based models.

6 Conclusions

We presented our systems for low-resourced translation between German and Upper Sorbian, unsu-

pervised translation between German and Lower Sorbian, and translation between Chuvash and Russian.

Our systems used standard state-of-the-art techniques for low-resource and unsupervised machine translation but did not exhaust all available methods. Better results could be achieved using more monolingual data and by more careful filtering of the synthetic parallel data.

Acknowledgments

This work was also supported by the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement #640550) and by the DFG (grant FR 2829/4-1).

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

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*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jindřich Libovický and Alexander Fraser. 2020. Towards reasonably-sized character-level transformer NMT by finetuning subword systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2572–2579, Online. Association for Computational Linguistics.

Jindřich Libovický, Viktor Hangya, Helmut Schmid, and Alexander Fraser. 2020. The LMU Munich system for the WMT20 very low resource supervised MT task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1104–1111, Online. Association for Computational Linguistics.

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, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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, Brussels, Belgium. 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 Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, 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 Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Training hyper-parameters

We use the following command line options for `fairseq-train` command in all experiments. For the Transformer Base models, we use the pre-defined `transformer` architecture, for Transformer Big, we use `transformer_wmt_en_de_big_t2t`. The batch size is 6000 tokens for the Base models and 2000 tokens for the Big models.

```
fairseq-train \
  $DATA \
  --arch $ARCHITECTURE \
  --share-all-embeddings \
  --label-smoothing 0.1 \
  --criterion \
     label_smoothed_cross_entropy \
  --optimizer adam \
  --adam-betas '(0.9, 0.998)' \
  --clip-norm 5.0 \
  --lr 5e-4 \
  --lr-scheduler inverse_sqrt \
  --warmup-updates 16000 \
  --max-tokens $TOKENS
```