Findings of the WMT 2022 Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT

Marion Weller-Di Marco and Alexander Fraser
Center for Information and Language Processing
LMU Munich
{dimarco,fraser}@cis.uni-muenchen.de

Abstract

We present the findings of the WMT2022 Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT with experiments on the language pairs German to/from Upper Sorbian, German to/from Lower Sorbian and Lower Sorbian to/from Upper Sorbian. Upper and Lower Sorbian are minority languages spoken in the Eastern parts of Germany. There are active language communities working on the preservation of the languages who also made the data used in this Shared Task available.

In total, four teams participated on this Shared Task, with submissions from three teams for the unsupervised sub task, and submissions from all four teams for the supervised sub task. In this overview paper, we present and discuss the results.

1 Introduction

For a large majority of the world's languages, only limited resources are available to train and provide NLP tools. The need for parallel data in a (supervised) translation scenario aggravates this problem further. The Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT aim at promoting the research on translating low and very low resourced languages.

Following the Shared Tasks in the two previous years (Libovický and Fraser, 2021; Fraser, 2020), we continue to cooperate with the Sorbian community, namely the Sorbian Institute[1] and the Witaj Sprachzentrum (Witaj Language Center)[2] for this year's Shared Task. We offer all translation directions between the languages Upper Sorbian, Lower Sorbian and German, for both supervised and unsupervised translation.

Upper and Lower Sorbian are minority languages spoken in the eastern part of Germany in the federal states of Saxony and Brandenburg. With only 30k and 7k native speakers, there are only few resources available. However, as western Slavic languages, Upper and Lower Sorbian are closely related to Polish and Czech and can thus make use of the comparatively large data sets available for those languages.

In this year, four teams participated in the Shared Task, resulting in three to four submissions for each language pair for both the supervised and unsupervised variants.

2 Tasks and Evaluation

In contrast to the previous Shared Tasks, all language combinations between Upper Sorbian, Lower Sorbian and German are considered, resulting in the six following translation pairs:

- Upper Sorbian ↔ German
- Lower Sorbian ↔ German
- Upper Sorbian ↔ Lower Sorbian

Factoring in the variants supervised and unsupervised translation for each language pair, there is a total of 12 translation pairs.

For the evaluation, we follow the strategy employed in the previous Shared Task and use BLEU scores (Papineni et al., 2002) and chrF scores (Popović, 2015) as implemented in sacreBLEU (Post, 2018).[3] Furthermore, we evaluate the submissions using BERTScore (Zhang et al., 2020)[4] with XLM-RoBERTa Large (Conneau et al., 2020) as an underlying model for translations into German[5].

[1] https://www.serbski-institut.de/en/Institute/
[2] https://www.witaj-sprachzentrum.de/

[3] BLEU score signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.0`
chrf2 score signature: `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.0`

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

[5] BERTScore signatures: `xlm-roberta-large_L17_no-idf_version=0.3.11(hug_trans=4.22.2)_fast-tokenizer` and `xlm-roberta-large_L17_idf_version=0.3.11(hug_trans=4.22.2)_fast-tokenizer`

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 801–805
Abu Dhabi, December 7–8, 2022.

HSB ↔ DE	449.057	parallel sentences
DSB ↔ DE	40.193	parallel sentences
DSB ↔ HSB	62.564	parallel sentences

DSB	220.419	monolingual sentences
HSB	1.132.850	monolingual sentences

Table 1: Training data per language pair. The data sets have been made available by the Sorbian Institute (monolingual data) and The Witaj Sprachzentrum (both parallel and monolingual data).

We decided against using COMET Scores (Rei et al., 2020). This metric considers both the source language and the target language, but because it relies on XLM-R models, it does not support the Sorbian languages.

3 Data

To allow for a direct comparison between the different submissions, we only allowed training based on the resources released for the task, as well as resources for related languages (German, Czech and Polish data from the WMT news tasks[6] or available in the OPUS project[7]). In particular, the use of large models pre-trained on large data sets was not allowed. Table 1 gives an overview of the parallel and monolingual training data for the Sorbian languages.

For the unsupervised translation sub-task, we restricted the the data set as follows: all released Upper/Lower Sorbian data could be used, with the exception of the parallel Upper Sorbian ↔ Lower Sorbian corpus. Furthermore, the German side of the parallel German ↔ Upper Sorbian and German ↔ Lower Sorbian training corpora was excluded. This setup allowed us to make maximum use of the low-resourced languages without providing parallel data.

4 Submitted Systems

Four teams participated in the supervised sub-task[8], and three teams participated in the unsupervised sub-task. We present a brief system description of each team's submission, with an overview of the results listed in tables 2 to 7. Table 8 gives a brief overview of some relevant details; for more detailed information, we refer the reader to the respective system description papers.

AIC (Shapiro et al., 2022) For the unsupervised system, they trained an unsupervised phrase-based statistical machine translation (UPBSMT) system on each pair independently. They trained a De-Slavic mBART model from Scratch (Random initialization) on the following languages: Polish (pl), Czech (cs), German (de), Upper Sorbian (hsb), and Lower Sorbian (dsb). They then fine-tuned their mBART on the synthetic parallel data generated by the UPBSMT model along with authentic parallel data (de ↔ pl, de ↔ cs). They further fine-tuned their unsupervised system on authentic parallel data (hsb ↔ dsb, de ↔ dsb, de ↔ hsb) to submit the supervised low-resource system.

MUNI NLP (Signoroni and Rychlý, 2022) This team submitted supervised NMT systems for the Lower Sorbian-German and Lower Sorbian-Upper Sorbian language pairs, in both translation directions. They employed a new subword tokenization algorithm, High Frequency Tokenizer (HFT), to obtain more meaningful subword vocabularies. They tested this against BPE in the first round of experiments where they trained two different models on the data tokenized with each tokenizer, so four systems in total: two standard Transformers and two Transformers with hyperparameters optimized for the dataset size. They then followed the Data Diversification procedure (Nguyen et al., 2020) generating and collating authentic and synthetic data alternatively from each previous system and the original parallel data to create an augmented dataset. Then, they trained a Transformer model on these new data, tokenized with HFT, to obtain the final system. Thus, the approach is based only on the original parallel corpus.

Huawei TSC (Li et al., 2022) Huawei Translation Services Center participated in all 6 supervised tracks. Their systems are build on deep Transformer models with a large filter size. First, they selected a base multilingual model with German-Czech (DE-CS) and German-Polish (DE-PL) parallel data for all of the 6 tracks. They then utilized regularized dropout (R-Drop), back translation, fine-tuning and ensemble multilingual models to improve on the best individual system performance. For the unsupervised task submission, they applied their pre-trained multilingual system with zero-shot.

[6] https://www.statmt.org/wmt22/translation-task.html

[7] https://opus.nlpl.eu/

[8] There were submissions by a fifth team in for the supervised task. We do not have system descriptions for this team's submissions, and thus listed their results separately in table 9.

	DE-DSB	
System	BLEU	chrF2
HuaweiTSC	73.9	87.1
MUNI-NLP	50.5	74.1
AIC	48.2	73.0
PICT-NLP	20.8	44.1

	DSB-DE			
System	BLEU	chrF2	BERT_F	BERT_{F_IDF}
HuaweiTSC	62.5	80.9	0.9792	0.9764
MUNI-NLP	49.5	73.0	0.9664	0.9613
AIC	39.4	66.2	0.9542	0.9463
PICT-NLP	25.4	51.3	0.9246	0.9125

Table 2: Results for supervised DE-DSB and DSB-DE translation.

	DE-HSB	
System	BLEU	chrF2
HuaweiTSC	70.7	85.5
AIC	51.0	73.7
PICT-NLP	25.7	49.1

	HSB-DE			
System	BLEU	chrF2	BERT_F	BERT_{F_IDF}
HuaweiTSC	71.9	85.3	0.9843	0.9825
AIC	47.5	71.4	0.9637	0.9574
PICT-NLP	29.7	53.8	0.9317	0.9207

Table 3: Results for supervised DE-HSB and HSB-DE translation.

	DSB-HSB	
System	BLEU	chrF2
HuaweiTSC	86.8	94.0
MUNI-NLP	72.2	87.5
AIC	65.8	83.9
PICT-NLP	49.1	65.5

	HSB-DSB	
System	BLEU	chrF2
HuaweiTSC	88.0	94.4
MUNI-NLP	72.3	87.5
AIC	66.6	84.3
PICT-NLP	50.7	66.9

Table 4: Results for supervised DSB-HSB and HSB-DSB translation.

	DE-DSB	
System	BLEU	chrF2
HuaweiTSC	9.0	32.6
AIC	1.2	22.1
PICT-NLP	0.2	8.1

	DSB-DE			
System	BLEU	chrF2	BERT_F	BERT_{F_IDF}
HuaweiTSC	11.5	33.9	0.9141	0.8970
AIC	4.0	26.9	0.8567	0.8434
PICT-NLP	0.0	5.0	0.7822	0.7693

Table 5: Results for unsupervised DE-DSB and DSB-DE translation.

	DE-HSB	
System	BLEU	chrF2
AIC	17.9	48.5
HuaweiTSC	10.4	33.4
PICT-NLP	0.5	14.3

	HSB-DE			
System	BLEU	chrF2	BERT_F	BERT_{F_IDF}
AIC	18.0	46.9	0.9046	0.8937
HuaweiTSC	13.5	35.8	0.9162	0.8996
PICT-NLP	0.3	13.6	0.8306	0.8194

Table 6: Results for unsupervised DE-HSB and HSB-DE translation.

	DSB-HSB	
System	BLEU	chrF2
AIC	44.2	72.9
HuaweiTSC	–	–
PICT-NLP	10.4	48.6

	HSB-DSB	
System	BLEU	chrF2
AIC	35.9	67.4
PICT-NLP	9.3	44.2
HuaweiTSC	2.4	16.1

Table 7: Results for unsupervised DSB-HSB and HSB-DSB translation.

team	data (in addition to the provided de/hsb/dsb corpora)	synthetic data/ back translation	segmentation (vocab. size)	toolkits
AIC	DE (431.4M), CS (111.1M) PL (13.4M), PL-DE (12.4M) CS-DE (12.4M)	synthetic data through UPBSMT	SentencePiece (32k)	Fairseq
HUAWEI	DE-CS (55.9M), DE-PL (66.5M), DE (20M)	back-translation with sampling (Graça et al., 2019)	SentencePiece (40k)	Fairseq Marian
MUNI	–	Data diversification (Nguyen et al., 2020)	High Frequency Tokenizer (4k)	Fairseq
PICT	DE (53.3k)	–	BPE	Fairseq Facebook's XLM

Table 8: Overview of methods and data.

PICT NLP (Vyawahare et al., 2022) They implemented the XLM's Masked Language Model (MLM) for unsupervised learning. They trained it only using the monolingual data provided by the organizers and the OPUS project. Finally, they also applied XLM preprocessing to the data before training.

For supervised learning, they trained language models such as LSTM and attention based transformer models with the help of the Fairseq library. They trained it using monolingual data provided by the organizers. They applied the inbuilt tokenization provided by Fairseq on the data.

5 System Results

Tables 2 to 7 list the results of the submitted systems in terms of BLEU and chrF2 for all systems, and additionally BERT scores for those translating into German. For the BERT scores, we list both BERT_F and BERT_F with idf weighting to give less weight to commonly occuring words. The ordering of the systems is consistent across all metrics.

The supervised systems obtain higher results than the unsupervised systems. The language pair DSB $\leftrightarrow$ HSB obtained comparatively high scores for both supervised and unsupervised translation which is very probably due to the high similarity between the two languages.

Overall, we see no winner across all tasks: HuaweiTSC has the best scores across all supervised translation tasks, followed by MUNI-NLP for the DE-to/from-DSB and DSB-to/from-HSB translations. These two language pairs only have comparably small parallel data sets which are, notably, the sole basis of MUNI-NLP's submissions.

For the unsupervised translation (where MUNI-NLP did not participate), AIC has the strongest results with the exception of DSB-to/from-DE, where HuaweiTSC is leading.

6 Conclusion

In the WMT 2022 Shared Task on Unsupervised and Very Low Resource MT, we provided the participants with resources for all possible translation directions for the three languages Upper Sorbian, Lower Sorbian and German, of which Upper Sorbian $\leftrightarrow$ Lower Sorbian is a new language pair in comparison to last year's shared task.

The participating teams submitted strong systems relying on a wide range of methods. Using modeling techniques such as pre-training on parallel data of related languages is important, as is the creation of synthetic data for which we saw the application of different methods. However we also saw that careful modeling on a small data set only can lead to good results.

We hope that this Shared Task will continue to increase the interest in research on methods for under-resourced languages, both for supervised and unsupervised approaches.

Acknowledgements

This work was supported by the DFG (grant FR 2829/4-1). We would also like to thank Marko Měškank and Olaf Langner (Witaj Sprachzentrum) and Hauke Bartels and Marcin Szczepański (Sorbian Institute).

References

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.

Alexander Fraser. 2020. Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Online. Association for Computational Linguistics.

Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. 2019. Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 45–52, Florence, Italy. Association for Computational Linguistics.

Shaojun Li, Yuanchang Luo, Daimeng Wei, Zongyao Li, Hengchao Shang, Xiaoyu Chen, Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Yuhao Xie, Lizhi Lei, Hao Yang, and Ying Qin. 2022. HW-TSC Systems for WMT22 Very Low Resource Supervised MT Task. In *Proceedings of the Seventh Conference on Machine Translation*. Association for Computational Linguistics.

Jindřich Libovický and Alexander Fraser. 2021. Findings of the WMT 2021 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 726–732, Online. Association for Computational Linguistics.

Xuan-Phi Nguyen, Shafiq R. Joty, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. In *NeurIPS*.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ahmad Shapiro, Mahmoud Tarek Salama, Omar Khaled Abdelhakim, Mohamed Essam Fayed, Ayman Khalafallah, and Noha Adly. 2022. The AIC System for the WMT 2022 Unsupervised MT and Very Low Resource Supervised MT Task. In *Proceedings of the Seventh Conference on Machine Translation*. Association for Computational Linguistics.

Edoardo Signoroni and Pavel Rychlý. 2022. MUNI-NLP Systems for Lower Sorbian-German and Lower Sorbian-Upper Sorbian Machine Translation @ WMT22. In *Proceedings of the Seventh Conference on Machine Translation*. Association for Computational Linguistics.

Aditya Vyawahare, Rahul Tangsali, Aditya Mandke, Onkar Litake, and Dipali Kadam. 2022. PICT-NLP@WMT22-EMNLP2022: Unsupervised and Very-Low Resource Supervised Translation on German and Sorbian Variant Languages. In *Proceedings of the Seventh Conference on Machine Translation*. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A Further Results

Table 9 lists the results of another submission that did not provide further details.

	BLEU	chrF2	BERT_F	BERT_{F_IDF}
DE-DSB	58.2	79.5	–	–
DSB-DE	61.5	80.4	0.9784	0.9755
DE-HSB	67.3	83.9	–	–
HSB-DE	71.2	85.1	0.9840	0.9821
DSB-HSB	72.8	87.7	–	–
HSB-DSB	72.2	87.6	–	–

Table 9: Results for supervised translation of a team that we were not able to contact.