# Findings of the WMT 2021 Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT

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

## Abstract

We present the findings of the WMT2021 Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT. Within the task, the community studied very low resource translation between German and Upper Sorbian, unsupervised translation between German and Lower Sorbian and low resource translation between Russian and Chuvash, all minority languages with active language communities working on preserving the languages, who are partners in the evaluation. Thanks to this, we were able to obtain most digital data available for these languages and offer them to the task participants. In total, six teams participated in the shared task. The paper discusses the background, presents the tasks and results, and discusses best practices for the future.

## 1 Introduction

For some languages, machine translation (MT) reached such a high quality that allows a discussion of whether and under what circumstance human parity might have been reached (Popel et al., 2020; Läubli et al., 2020). This is the case, however, for only a small minority of the world's language. For most of the 7k languages spoken in the world only very limited resources exist. The goal of the WMT Shared Task on Unsupervised and Very Low Resource MT is to promote research on methods for MT that alleviate such data sparsity in a real-world setup.

A task on unsupervised MT was already held at WMT in 2018 (Bojar et al., 2018) and 2019 (Barrault et al., 2019), where the lack of parallel data was simulated on high-resource language pairs: English–German in 2018 and German–Czech in 2019.

Starting from last year, we cooperate with local communities working on preserving their languages. In cooperation with the Sorbian Institute[1] and the Witaj Sprachzentrum[2], we offered a shared task in translation between German and Upper Sorbian in low-resource and unsupervised tracks (Fraser, 2020). For this year, we kept the low-resource track for Upper Sorbian and added unsupervised translation between German and Lower Sorbian. Upper and Lower Sorbian are minority languages spoken in the east part of Germany in the federal states of Saxony and Brandenburg. Having only 30k and 7k native speakers, processing of the languages is an inherently low-resource problem, without any chance that the size of available resources would ever get close to the size of resources available for languages with millions of speakers. On the other hand, being western Slavic languages, the Sorbian languages can take advantage of existing resources for Czech and Polish.

Additionally, in cooperation with the Chuvash Language Laboratory[3], we added another low-resource task, translation between Russian and Chuvash. Chuvash is a minority Turkic language spoken by approximately one million people in the Volga region in the southwest of Russia. There is a larger amount of training data available for Chuvash, but the language is rather isolated in the Turkic language family, so unlike Sorbian, it cannot benefit that much from the existence of closely related languages.

Five teams participated in the German-Upper Sorbian task, six teams in the German-Lower Sorbian task, and two teams in the Russian-Chuvash task.

## 2 Tasks and Evaluation

This year, there were three tasks for very low resource and unsupervised translation were:

[1] https://www.serbski-institut.de
[2] https://www.witaj-sprachzentrum.de/
[3] https://en.corpus.chv.su/content/about.html

*Proceedings of the Sixth Conference on Machine Translation (WMT)*, pages 726–732
November 10–11, 2021. ©2021 Association for Computational Linguistics

- Very Low Resource Supervised Machine Translation: *German ↔ Upper Sorbian.*
- Unsupervised Machine Translation: *German ↔ Lower Sorbian.*
- Low Resource Supervised Machine Translation: *Russian ↔ Chuvash.*

To make the submissions better comparable with each other, we only allowed using resources released for the task (see Section 3) and resources for related languages commonly used in other WMT tasks. The use of large models pre-trained on large datasets was not allowed. By this decision, we wanted to motivate the participants to find better use of limited language resources.

**German↔Upper Sorbian.** There is only a very limited amount of parallel data between Upper Sorbian and German. However, because Upper Sorbian is closely related to Czech and Polish, we encouraged the use of all German, Czech and Polish data released for WMT. Other parallel data released from the WMT News Task were also allowed, but the participants were recommended not to use them. Unlike last year, there was no unsupervised task for Upper Sorbian.

**German↔Lower Sorbian.** For this task, no parallel training data were available, as the only available Lower Sorbian data were monolingual. Lower Sorbian is closely related to other Western Slavic languages, so the same related language data as for the Upper Sorbian task was allowed.

**Russian↔Chuvash.** The Chuvash language is not that critically low-resource as the Sorbian languages, but it is still affected by being a minority language. The participants were provided with parallel and monolingual data that we released for the task. Additional data that might be used: Chuvash-Russian part of the JW300 corpus (Agić and Vulić, 2019). In addition, the participants were encouraged to use the Kazakh–Russian corpus and monolingual Kazakh data from WMT19 (Barrault et al., 2019) and monolingual Russian data made available for the WMT News tasks.

**Evaluation.** Following the recent literature on MT evaluation (Mathur et al., 2020; Marie et al., 2021; Kocmi et al., 2021), we evaluate the systems using multiple evaluation measures, both string-based and model-based, and perform statistical testing to decide the ranking of the systems. In particular, we use the BLEU Score (Papineni et al., 2002), chrF score (Popović, 2015) as implemented in SacreBLEU (Post, 2018).[5] Further, we evaluate the models using BERTScore (Zhang et al., 2020)[6] with XLM-RoBERTa Large (Conneau et al., 2020) as an underlying model for German and Russian

| Dataset | # lines | # chars. |
|---|---|---|
| *German↔Upper Sorbian* | | |
| WMT20 parallel data | 60k | 11M |
| Parallel data provided by the Witaj Sprachzentrum, collected for the development of its own tranlator SoTra[4]. | | |
| Additional parallel data | 87k | 17M |
| Additional parallel Witaj Sprachzentrum collected since the last year. | | |
| Sorbian Institute mono | 340k | 39M |
| Upper Sorbian monolingual data provided by the Sorbian Institute. This contains a high quality corpus and some medium quality data which were mixed together. | | |
| Witaj mono | 222k | 19M |
| Upper Sorbian monolingual data provided by the Witaj Sprachzentrum (high quality). | | |
| Web monolingual | 134k | 12M |
| Upper Sorbian monolingual data scraped from the web by CIS, LMU. This should be used with caution, it is probably noisy, it might erroneously contain some data from related languages. | | |
| *German↔Lower Sorbian* | | |
| Sorbian Institute mono | 145k | 14M |
| The sentences come from the Lower Sorbian reference corpus and were provided by the Sorbian Institute. | | |
| *Russian↔Chuvash* | | |
| Parallel corpus | 714k | 181M |
| A parallel corpus being collected by the Chuvash Language Laboratory since 2016 with the goal of promoting automatic processing of Chuvash. | | |
| Bilingual dictionary | 74k | 182k |
| Monolingual Chuvash | 5.6M | 749M |
| The dataset contains monolingual sentences from various publicly available sources including Wikipedia, web crawl and fiction. | | |

Table 1: Overview of the data made available for the shared task.

[4] `https://sotra.app/`

[5] 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`

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

| Team | Architecture | Pretraining | Pretraining data | German / Russian mono. | BT iter. | BT filtering | Data tricks | Segmentation | Ensembling | Tookit |
|---|---|---|---|---|---|---|---|---|---|---|
| *German ↔ Upper Sorbian* | | | | | | | | | | |
| NoahNMT | Big | de-cs | 15M | 100M | 5 | None | | BPE | Yes | Inhouse |
| NRC-CNRC | Base | de-cs | 16.5M | 5M | 2 | Moore and Lewis (2010) | Tagged BT, BPE Dropout, Lang. tags | BPE | Yes | Sockeye |
| IICT-Yverdon | Base | de-cs | 3M | 1M | 1 | Length | | SP | Yes | OpenNMT |
| CFILT | Base | mono de, hsb | 2×.7M | .7M | 60 | None | BPE Dropout | BPE | No | MASS |
| LMU Munich | Big | de-cs | 25M | 15M | 4 | Length | Tagged BT | BPE | Yes | Fairseq |
| *German ↔ Lower Sorbian* | | | | | | | | | | |
| NRC-CNRC | Base | de-cs, de-hsb | 16.5M | 147k to 5.2M | 2 | Moore and Lewis (2010) | BPE Dropout | BPE | Yes | Sockeye |
| IICT-Yverdon | Base | de-hsb | 150k | 1M | 1 | Length | | SP | Yes | OpenNMT |
| CFILT | Base | de-hsb | 3×.7M | .7M | 60 | None | BPE Dropout | BPE | No | MASS |
| CL_RUG | XLM | de-cs, de-hsb | 8.5M +.8M | 10.6M | 2 | None | | BPE | No | MASS |
| LMU Munich | Big | de-cs, de-hsb | 45M | 15M | 8 | Length | | BPE | Yes | Fairseq |
| *Russian ↔ Chuvash* | | | | | | | | | | |
| NoahNMT | Big | en-ru | 17M | 110M | 3 | None | Domain adap. | BPE | Yes | Inhouse |
| LMU Munich | Big | ru-kk | 11M | 18M | 2 | Length | Tagged BT | BPE | Yes | Fairseq |

Table 2: Overview of the method used by the task participants. SP stands for SentencePiece, BT for backtranslation.

and mBERT (Devlin et al., 2019) for Chuvash. We conduct the significance test using bootstrap resampling (Koehn, 2004) at a significance level of 0.95.

The final ranking is determined by the number of points each system gets. The systems get one point for each system that is significantly worse in each of the metrics. This means that if a system is significantly better than 1 system in the BLEU score, 2 systems in the chrF score, and 3 systems in the BERTScore, it gets 6 points in total.

## 3 Data

**Upper Sorbian.** The data for this task was provided by the Sorbian Institute (monolingual data) and The Witaj Sprachzentrum (Witaj Language Center) (both parallel and monolingual data).

The development and test data for Upper Sorbian are the same as the last year. There was a different blind test set than the last year.

**Lower Sorbian.** As far as we know, there is no parallel data for Lower Sorbian except for the development and test data provided for this task.

**Chuvash.** The validation data are sampled from the training set. The development test data and blind test data were also sampled from the parallel corpus and manually filtered by a native speaker.

In addition to the described data, the use of other parallel and monolingual data available for WMT News Tasks was allowed (see Section 2).

## 4 Submitted systems

Six teams participated in the shared task, five teams in Upper Sorbian-German, slightly different five in Lower Sorbian-German, and two in the Russian-Chuvash direction. An overview of the systems is in Table 2, a brief description of the systems follows. For detailed information, we refer the reader to the respective system description papers.

**NoahNMT (Zhang et al., 2021b).** NoahNMT submitted their systems into the supervised tasks. The NoahNMT submission is a standard Transformer model equipped with our recent technique of dual transfer (Zhang et al., 2021a). Compared to other systems, these submissions used a significantly larger amount of monolingual data.

**NRC-CNRC (Knowles and Larkin, 2021).** The Upper Sorbian-German system is an ensemble of eight systems with 25k BPE vocabulary, incorporating transfer learning (from cs–de) with continued training, monolingual data filtering, backtranslation (Sennrich et al., 2016), BPE-dropout (Provilkov et al., 2020), and multilingual models.

Upper Sorbian → German

| Team | BLEU | | chrF | | BERTScore | | Points |
|---|---|---|---|---|---|---|---|
| NRC-CNRC | 67.3 | 3 | 83.6 | 3 | .981 | 4 | 10 |
| NoahNMT | 67.7 | 3 | 83.4 | 3 | .981 | 3 | 9 |
| LMU | 64.3 | 2 | 81.9 | 2 | .979 | 2 | 6 |
| IICT-Yverdon | 61.4 | 0 | 80.2 | 0 | .976 | 1 | 1 |
| CFILT | 60.1 | 0 | 79.2 | 0 | .975 | 0 | 0 |

German → Upper Sorbian

| Team | BLEU | | chrF | | BERTScore | Points |
|---|---|---|---|---|---|---|
| NRC-CNRC | 66.3 | 3 | 83.7 | 3 | — | 6 |
| NoahNMT | 65.9 | 3 | 83.3 | 3 | — | 6 |
| LMU | 63.3 | 1 | 81.9 | 2 | — | 3 |
| CFILT | 60.2 | 0 | 79.6 | 0 | — | 0 |
| IICT-Yverdon | 61.6 | 0 | 80.6 | 0 | — | 0 |

Lower Sorbian → German

| Team | BLEU | | chrF | | BERTScore | | Points |
|---|---|---|---|---|---|---|---|
| NRC-CNRC | 33.5 | 1 | 63.8 | 1 | .953 | 2 | 4 |
| CL_RUG | 32.4 | 1 | 62.2 | 1 | .953 | 2 | 4 |
| LMU | 33.3 | 1 | 62.0 | 1 | .952 | 1 | 3 |
| CFILT | 5.9 | 0 | 31.6 | 0 | .884 | 0 | 0 |

German → Lower Sorbian

| Team | BLEU | | chrF | | BERTScore | Points |
|---|---|---|---|---|---|---|
| NRC-CNRC | 29.9 | 3 | 59.9 | 3 | — | 6 |
| LMU | 27.5 | 3 | 57.9 | 3 | — | 6 |
| CL_RUG | 24.1 | 2 | 54.2 | 2 | — | 4 |
| IICT-Yverdon | 8.0 | 0 | 32.1 | 1 | — | 1 |
| CFILT | 6.4 | 0 | 29.0 | 0 | — | 0 |

Chuvash → Russian

| Team | BLEU | | chrF | | BERTScore | | Points |
|---|---|---|---|---|---|---|---|
| NoahNMT | 23.4 | 0 | 47.6 | 0 | .944 | 1 | 1 |
| LMU | 22.0 | 0 | 46.3 | 0 | .942 | 0 | 0 |

Russian → Chuvash

| Team | BLEU | | chrF | | BERTScore | | Points |
|---|---|---|---|---|---|---|---|
| NoahNMT | 22.1 | 0 | 51.3 | 0 | .857 | 1 | 1 |
| LMU | 20.9 | 0 | 50.1 | 0 | .856 | 0 | 0 |

Table 3: The main results of the task. Points awarded in the particular metrics are in gray.

In the opposite direction, the submission is an ensemble of 7 systems. The Lower Sorbian-German and German-Lower Sorbian systems are ensembles of 2 and 4 systems, respectively, with 20k BPE vocabulary, incorporating transfer learning from hsb–de and de-hsb systems along with iterative backtranslation.

**IICT-Yverdon (Atrio et al., 2021).** The system used the Transformer architecture with back-translation of large German corpora and parent-language initialization using Czech-German data. The final submission is an ensemble of different models with some changes in their training setups to maximize the diversity among the models.

**CFILT.** The submitted systems cover four language pairs: German↔Upper Sorbian German↔Lower Sorbian. For de↔hsb, the system pre-trained using the MASS objective (Song et al., 2019) and finetuned using iterative back-translation. Final finetuning is performed using the provided parallel data for translation objective. For de↔dsb, no parallel data is provided in the task. The final de↔hsb model is used for initialization of the de↔dsb model, which is further trained using iterative back-translation, using the same vocabulary as used in the de↔hsb model.

**CL_RUG (Edman et al., 2021).** CL_RUG's submission uses the MASS model, focusing pre-training on 2 languages at a time, from least to most related to Lower Sorbian. The largest improvement comes from a novel method for initializing the Lower Sorbian word embeddings from Upper Sorbian, using a bilingual dictionary created in an unsupervised fashion.

**LMU Munich (Libovický and Fraser, 2021).** The LMU submissions for all tasks are Transformer models first pre-trained on related languages and then finetuned on the low-resource languages. For the Sorbian languages, the systems are pre-trained on German–Czech translation. The system is fine-tuned using the authentic German–Upper Sorbian data, which is the starting point for four iterations of tagged back-translation. The unsupervised German–Lower Sorbian translation is trained by iterative backtranslation using the monolingual data only. The Upper Sorbian–German system is used to generate the first translation of Lower Sorbian. The Russian–Chuvash systems were pretrained on Russian–Kazakh translation and finetuned using the provided parallel data.

## 5 Results

The results are presented in Table 3. The most successful teams were NRC-CNRC, which was the best or on par with the best systems in all Sorbian tasks, and NoahNMT which were the best in the Chuvash tasks, on par with the best systems in German-Upper Sorbian translation and the second in the Upper Sorbian-German direction.

In German-Upper Sorbian translation, the best two system, NRC-CNRC and NoahNMT reach very similar results although they use significantly different sizes of monolingual data for backtranslation. NRC-CNRC manage to compensate for the smaller data size by accumulating minor tricks including monolingual data selection (Moore and Lewis, 2010), tagged backtranslation (Caswell et al., 2019), BPE dropout (Provilkov et al., 2020), and language tags in multilingual training. LMU, which used data of a similar size to NRC-CNRC but did not use most of the further tricks, ranked below these two.

In Upper Sorbian-German translation, all teams used German-Czech parallel data for pre-training, except for CFILT who only used monolingual data for pre-training and scored 0 points in both directions.

In the unsupervised German-Lower Sorbian task, CL_RUG ranked on par with NRC-CNRC in translation into German (despite not using ensembling), but at third place in the opposite direction. This suggests that CL_RUG's innovative vocabulary transfer method works better on the encoder side than on the decoder side.

In the Russian-Chuvash translation, NoahNMT outperformed LMU Munich by using larger datasets and a more advanced transfer learning technique.

## 6 Conclusions

In WMT 2021 shard task on Unsupervised and Very Low Resource MT, we created realistic benchmarks for low-resource minority language which reflect the needs of the language communities trying to preserve their languages. In the task, we provided the participants with comprehensive resource for translation between German and Upper and Lower Sorbian and for translation between Russian and Chuvash. We hope that this will increase the interest of the community in these languages.

The six teams that participated in the task used state-of-the-art MT techniques to develop high quality systems. The main technical takeaway from the results are that pre-training on parallel data in related language is important and that carefully applying known tricks can to a large extent compensate for using smaller datasets.

## 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). Thanks very much to coorganizers: Hauke Bartels – Sorbian Institute, Olaf Langner – Witaj Sprachzentrum, Marcin Szczepanski – Sorbian Institute, Alexander Antonov – Chuvash Language Laboratory. Many thanks also to Tom Kocmi and Christian Federmann from Microsoft for providing an instance of Ocelot for the system submissions.

## References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Àlex R. Atrio, Gabriel Luthier, Axel Fahy, Giorgos Vernikos, Andrei Popescu-Belis, and Ljiljana Dolamic. 2021. The iict-yverdon system for the wmt 2021 unsupervised mt and very low resource supervised mt task. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.

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.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. 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.

Lukas Edman, Ahmet Üstün, Antonio Toral, and Gertjan van Noord. 2021. Unsupervised translation of german–lower sorbian: Exploring training and novel transfer methods on a low-resource language. In *Proceedings of the Sixth Conference on Machine Translation*. 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.

Rebecca Knowles and Samuel Larkin. 2021. NRC–CNRC systems for upper sorbian-german and lower sorbian-german machine translation 2021. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. *CoRR*, abs/2107.10821.

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.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human-machine parity in language translation. *J. Artif. Intell. Res.*, 67:653–672.

Jindřich Libovický and Alexander Fraser. 2021. The lmu munich systems for the wmt21 unsupervised and very low-resource translation task. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. 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.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15.

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.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. 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.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Meng Zhang, Liangyou Li, and Qun Liu. 2021a. Two parents, one child: Dual transfer for low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2726–2738, Online. Association for Computational Linguistics.

Meng Zhang, Minghao Wu, Pengfei Li, Liangyou Li, and Qun Liu. 2021b. Noahnmt at wmt 2021: Dual transfer for very low resource supervised machine translation. In *Proceedings of the Sixth 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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.