Findings of the WMT 2020 Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT

Alexander Fraser

Center for Information and Language Processing, LMU Munich, Germany
fraser@cis.lmu.de

Abstract

We describe the WMT 2020 Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT. In both tasks, the community studied German↔Upper Sorbian MT, which is a very realistic machine translation scenario (unlike the simulated scenarios used in particular in much of the unsupervised MT work in the past). We were able to obtain most of the digital data available for Upper Sorbian, a minority language of Germany, which was the original motivation for the Unsupervised MT shared task. As we were defining the task, we also obtained a small amount of parallel data (about 60000 parallel sentences), allowing us to offer a Very Low Supervised MT task as well. Six primary systems participated in the unsupervised shared task, two of these systems used additional data beyond the data released by the organizers. Ten primary systems participated in the very low resource supervised task. The paper discusses the background, presents the tasks and results, and discusses best practices for the future.

1 Introduction

There is no machine translation available for most of the approximately 7000 languages spoken on the planet Earth. This is because very limited or no parallel corpora are available. Research on unsupervised and very low resource machine translation is important for alleviating this problem. Unsupervised machine translation requires only monolingual data, while very low resource supervised machine translation uses very limited parallel data.

At WMT 2018 and WMT 2019, the first shared task (Bojar et al., 2018) and second shared task (Barrault et al., 2019) on Unsupervised Machine Translation (UMT), were held as part of the news translation track. In both 2018 and 2019 the scenarios simulated low resource setups using medium or high resource languages. In 2018, the language pairs were Turkish-English, Estonian-English and German-English. In 2019, we tested unsupervised systems for German to Czech unsupervised translation (where no German/Czech parallel data was allowed).

In the 2020 shared task we proposed a third edition on UMT, which aimed at a more realistic scenario, German to Upper Sorbian (and Upper Sorbian to German) translation. Upper Sorbian is a minority language of Germany that is in the Slavic language family (e.g., related to Lower Sorbian, Czech and Polish), and we provide here most of the digital data that is available for Upper Sorbian, as far as we know. The amount of monolingual data available for Upper Sorbian is quite small (see below), making unsupervised machine translation very challenging.

While working on the UMT task we were able to obtain a very small amount of parallel data (about 60000 parallel sentences) for this language pair, which allowed us to additionally offer the very low resource supervised translation task.

The exact tasks studied in the shared tasks were:

- Unsupervised Machine Translation: German to Upper Sorbian. Upper Sorbian to German.
- Very Low Resource Supervised Machine Translation: German to Upper Sorbian. Upper Sorbian to German.

The data used can be downloaded from the shared task webpage.[1]

This paper will give some background for the task, discuss the data made available (with a particular focus on considerations for benchmarks for future UMT and very low resource MT research), discuss how the tasks went with a presentation of the results, and then conclude.

[1] http://www.statmt.org/wmt20/unsup_and_very_low_res/

Proceedings of the 5th Conference on Machine Translation (WMT), pages 765–771
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

2 Background

The 2020 shared tasks focused on the very low resource language Upper Sorbian (a Slavic minority language spoken in the Eastern part of Germany). The Sorbian language (referring to Upper Sorbian and Lower Sorbian) has a special protected status in the "Grundgesetz" of Germany (similar to a constitution).

Working with the Sorbian Institute[2] we initially prepared an unsupervised MT task. We expect this task to become a standard benchmark for unsupervised MT development. This task particularly relies on having high quality Upper Sorbian monolingual data, which we obtained from the Sorbian Institute and the Witaj Sprachzentrum. We also offered data which LMU Munich obtained through web crawling. Finally, the Sorbian Institute has also provided medium quality data which has not been quality checked.

The Witaj Sprachzentrum (Witaj Language Center[3]) then provided a small training corpus of German/Upper Sorbian parallel data, which was used in the Very Low Resource Supervised Machine Translation task. We expect this task to become a standard task for very low resource scenarios, and note that the results are important as they will directly inform efforts to create state-of-the-art machine translation systems for use by the Upper Sorbian community.

The Witaj Sprachzentrum provided development and "development test" (not blind test) sets for German to Upper Sorbian and Upper Sorbian to German translation. The development set is used to tune parameters, while the devtest set is used to measure progress using automatic metrics. The Witaj Sprachzentrum also provided the blind test sets, which were released just in time for the evaluation.

We note that CIS (LMU Munich), the Sorbian Institute and the Witaj Sprachzentrum hope to organize a task for Unsupervised Lower Sorbian translation in the near future.

LMU Munich would be interested in adding new typologically diverse languages as unsupervised and/or very low resource tasks in future WMT Shared Tasks.

3 Basic Tasks and Evaluation

We provide some basic details about the tasks and evaluation here.

Unsupervised MT: Unsupervised translation from German to Upper Sorbian, and unsupervised translation from Upper Sorbian to German. We allow the use of all German data released for WMT, except that the German side of the small parallel German/Upper Sorbian training corpus may not be used. All Upper Sorbian data we release may be used. No other language data may be used.

Very Low Resource Supervised MT: Supervised translation from German to Upper Sorbian, and supervised translation from Upper Sorbian to German. We allow the use of all German and Upper Sorbian data released for WMT, including the small parallel German/Upper Sorbian training corpus. Other WMT data for other languages may also be used. Upper Sorbian is a Slavic language which has strong similarities to Czech, so the German/Czech data we discuss below may be of particular interest for multilingual systems.

We used automatic metrics for the evaluation of this task. We believe that manual evaluation may not be so necessary for unsupervised MT and very low resource MT development, because automatic metrics worked well at relatively low translation quality levels in the past. Translation to Upper Sorbian would have been fairly difficult to evaluate in a human evaluation, as we do not have easy access to a large number of native speakers.

4 Data

We describe the data released for the two scenarios, Unsupervised and Very Low Resource Supervised.

The Unsupervised MT scenario allowed the use of all German data released for WMT, except that the German side of the small parallel German/Upper Sorbian training corpus could not be used. All Upper Sorbian data available on the web page was allowed. No other language data could be used (no parallel, no other monolingual data sets for any language except those explicitly listed as usable for Unsupervised).

The Very Low Resource Supervised MT scenario allowed the use of all German and Upper Sorbian data released for WMT, including the 60000 sentence parallel German/Upper Sorbian training corpus. Other WMT 2020 data for other languages were also allowed. Upper Sorbian is a Slavic language which is related to Czech, so the Ger-

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

[3] https://www.witaj-sprachzentrum.de/

man/Czech parallel data we briefly describe next were of particular interest for building multilingual systems. We would also like to express our thanks to the Opus project for the German/Czech parallel data.

In addition to training data, we also provided **development data**. Specifically, we provided development and test sets for German to Upper Sorbian and Upper Sorbian to German. These were usable for both Unsupervised and Very Low Resource. The dev set was intended for use for parameter tuning (and the participants were asked to not use it as a parallel training corpus). The test set was intended for system evaluation during development (likewise, the participants were asked to not use it as a parallel training corpus). We would like to thank Jindřich Libovický for creating the data splits and for work on the training data.

The specific data sets are presented in Table 1.

The translation output was submitted as real case, detokenized, and in SGML format. We used the Matrix for submission, thanks to Barry Haddow for assistance. The standard script wrap-xml.perl was used to create the sgm files for upload.

5 Results

In this section we discuss the results for the Unsupervised track, a non-constraint system that does not use parallel German/Sorbian data but does use German-English parallel data (which does not meet the constraints for the Unsupervised Task), and the Very Low Resource task.

We note that the absolute BLEU scores are surprisingly high, this may be due to unusually homogeneous train and test sets.

5.1 Unsupervised MT

There were four submissions to the Unsupervised MT which used only the allowed data for the track, see Table 2. The citations are listed in the table. Highlights of systems here included the use of transfer learning to obtain better initialization for the lower resource language and refinements of BPE, see the papers (and the analysis below) for further details.

5.2 Unsupervised MT with Multilingual Transfer Learning

Two primary submissions to Unsupervised MT were not restricted to the allowed data for the track, see Table 3. One system used data mined from German and Upper Sorbian Wikipedia, see (Dutta et al., 2020) for more details. The creators of the other system (Li et al., 2020) used parallel data for English to German to initialize an unsupervised system, which is a reasonable scenario to consider. This was highly effective. We suggest that this result be used as an initial benchmark for multilingual transfer-based unsupervised systems. We consider these results to be separate from the simpler unsupervised benchmark that was previously proposed. See the subsequent analysis and discussion for more details.

5.3 Very Low Resource MT

A Moses baseline from the Witaj Sprachzentrum using only the data from the shared task web page scored 46.36 for DE-HSB (BLEU-cased, 11b) and 47.70 for HSB-DE (BLEU-cased, 11b). The results for the shared task systems are presented in Table 4. Note that the last system did not submit a shared task paper.

6 Analysis

Overall we proposed two new benchmark tasks for low resource MT. The Unsupervised MT track can be used as a benchmark for testing new unsupervised MT approaches. The Very Low Resource track can be used as a benchmark for testing MT systems when one language has very little parallel data. Additionally, one system which was conceptually unsupervised (in that it used no parallel German-Upper Sorbian data), effectively defined a new benchmark, which we will call Unsupervised MT with Multilingual Transfer, by using additional English-German parallel data, see below.

6.1 Task Definitions

Unsupervised MT: The data released for the Unsupervised MT track should be considered to be a new realistic benchmark for testing unsupervised MT systems. Artetxe et al. (2020) (a best practices paper at ACL 2020) made a number of suggestions in terms of how to setup future Unsupervised MT research. The guidelines we set agree with their guidelines except with respect to one point. We disagree with Artetxe et al. in terms of whether development sets should be made available. The dev set is traditionally used for setting a small number of open parameters. Artetxe et al. argue that this data should not be made available, as it represent parallel data that is not really in spirit with

Monolingual Upper Sorbian Data	
sorbian_institute_monolingual.hsb.gz	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_monolingual.hsb.gz	Upper Sorbian monolingual data provided by the Witaj Sprachzentrum (high quality).
web_monolingual.hsb.gz	Upper Sorbian monolingual data scraped from the web by CIS, LMU (thanks to Alina Fastowski). This should be used with caution, it is probably noisy, it might erroneously contain some data from related languages.
Monolingual German Data	
	See the news translation task web page for allowed monolingual German data. All monolingual German sets allowed in this years News task were allowed in both scenarios.
Upper Sorbian side of parallel training corpus	
train.hsb-de.hsb.gz	This file was usable for both the Unsupervised and Very Low Resource Supervised scenarios. In the Unsupervised scenario, it was used as a small high quality monolingual corpus.
German side of parallel training corpus	
train.hsb-de.de.gz	This file was not usable for Unsupervised.
Dev and Test Sets	
devtest.tar.gz	The participants were requested to please use dev to tune system parameters, and test to measure progress. These files were allowed in both tracks.
German/Czech Parallel Data	
	This data was not allowed for Unsupervised. For Very Low Resource MT we allowed all German/Czech parallel corpora obtainable from the Opus project. The de-cs corpora we particularly recommended using are: Europarl v8 and JW300 v1. These two corpora may be somewhat similar to the DE-HSB parallel training and test data, but this is far from certain.

Table 1: Data sets for both tasks.

System Name	DE-HSB	HSB-DE	citation
LMU Munich-NMT	35.0	31.6	(Chronopoulou et al., 2020)
CUNI-Synthetic	25.0	23.4	(Kvapilíková et al., 2020)
NITS-CNLP	15.4		(Singh et al., 2020)
rug_hsbde_unsup_sel		20.1	(Edman et al., 2020)

Table 2: Four primary submissions to the Unsupervised MT Task using only the allowed data for the track, sorted by DE-HSB BLEU score.

System Name	DE-HSB	HSB-DE	citation
SJTU-NICT	40.3	32.8	(Li et al., 2020)
UdS-DFKI	10.3	9.0	(Dutta et al., 2020)

Table 3: Two primary submissions to the Unsupervised MT Task using additional English-German parallel data (row 1) and German and Upper Sorbian monolingual data from Wikipedia (row 2).

the unsupervised MT paradigm. In practice one could create a "very very low resource" system by simply training a supervised system on this data. We however argue that system developers need to have access to standard dev and devtest sets in order to measure progress. Having access to a dev set simulates having access to, e.g., a native speaker of the low resource language, who is able to look at outputs during MT development and judge the changes in quality obtained through simple hyperparameter changes. In our view the use of a devtest set to measure BLEU (which is completely standard in all unsupervised MT research) is similar to this, and of course also represents a small amount of available parallel data. A further argument for our view is that having a designated dev set for parameter tuning strongly helps with replicability of results. However, it is important in our view that this data not be trained on (in terms of using it as a parallel corpus to training a supervised system). The restriction of the rest of the data to be monolingual and carefully controlled makes sense for the straightforward testing of bilingual unsupervised systems, but we describe an interesting different multilingual scenario next.

Unsupervised MT with Multilingual Transfer: SJTU-NICT submitted a system to Unsupervised MT which looked at the reasonable scenario of assuming that the high resource language (German for German-Upper Sorbian) has parallel corpora with another language. In their system they used an English-German parallel corpus to initialize their German-Upper Sorbian Unsupervised MT system. This makes their system conceptually a multilingual system. Another obvious choice would be to try German-Czech and/or German paired with another Slavic language (and in fact, in the Very Low Resource Track, this was a common strategy). There were no competing systems for this unplanned benchmark (which was essentially created by SJTU-NICT's submission), but we suggest that this result is also interesting for comparison in future research and intend to offer a new track for this scenario in the shared task next year.

Very Low Resource MT: Please see the shared task system descriptions to see the wide variation in the exact data used for the Very Low Resource task.

There was obviously a high level of interest overall in this task, and we already have plans to offer more tasks of this kind as well. This task is very interesting as even if an unsupervised MT system is actually deployed for a particular language pair, the users of the system can post-edit the output of that system and will soon therefore be interested in training higher quality supervised systems using the Very Low Resource scenario. The parallel data used in this track was in fact created manually by the Witaj Sprachzentrum in order to be used by the Upper Sorbian community in Germany (to train the baseline Moses system we mentioned in the results section, and for future neural machine translation work which will be informed by the results produced by the research community). As the result of the existence of this pipeline more data may become available in the future.

6.2 What Worked Well

We highlight four interesting trends in terms of the results. Two of the trends we observe here involve transfer learning (i.e., pretraining). For all three

System Name	DE-HSB	HSB-DE	citation
SJTU-NICT	60.7	58.5	(Li et al., 2020)
Helsinki-NLP	57.9	59.6	(Scherrer et al., 2020)
NRC-CNRC	57.3	58.9	(Knowles et al., 2020)
LMU-supervised-ensemble	56.5	57.6	(Libovický et al., 2020)
CUNI-Transfer	55.5	56.9	(Kvapilíková et al., 2020)
Brown-NLP-b	46.2	45.7	(Berckmann and Hiziroglu, 2020)
IITBHU-NLPRL-DE-HSB	45.9	47.9	(Baruah et al., 2020)
Adobe-AMPS	45.2	47.6	(Singh, 2020)
UdS-DFKI	40.9		(Dutta et al., 2020)
HierarchicalTransformer	38.2	40.1	

Table 4: Ten primary systems submitted to the Very Low Resource Task, sorted by DE-HSB BLEU score.

benchmarks (Unsupervised, Unsupervised with Multilingual Transfer, and Very Low Resource) transfer learning is a critical component of many or all systems. The trends we will discuss are transfer learning using monolingual corpora and transfer learning using bilingual corpora. For Unsupervised, transfer learning using monolingual corpora was an important aspect of successful systems. Please see the system description papers for more detail about how the monolingual data was used to initialize successful Unsupervised systems. For Unsupervised with Multilingual Transfer, the use of English-German bilingual corpora to initialize the Unsupervised system seems to have been highly effective, with a particularly strong result for the DE-HSB translation direction, perhaps due to a very strong starting point for the German encoder. For Very Low Resource, heavy usage of both types of transfer were made as well, with one particular focus being on efforts to leverage the similarity of Czech and Upper Sorbian using Czech/German parallel corpora.

The third trend was that another area of study for many systems was word segmentation (typically with a variant of BPE) and/or the use of morphological information, sometimes learned in translation at BPE or character level, sometimes monolingually. Finally the fourth trend was that there was also a significant focus on trying to make Czech more like Upper Sorbian using a variety of techniques, and/or sampling German data like Upper Sorbian data.

7 Conclusion

The WMT 2020 Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT have created three interesting new benchmarks for future research. In addition to the initially proposed benchmark on straight-forward bilingual Unsupervised MT, we suggest the use of the Unsupervised Multilingual Transfer result as an additional new benchmark. The very low resource MT benchmark also generated strong participation from the research community.

We hope that additionally we have played a role in raising the interest of the NLP community in Upper Sorbian language processing, and that this interest will extend to Lower Sorbian language processing and in fact extend to working on other understudied languages as well.

We plan to organize similar tasks for the three benchmarks next year. Options include, e.g., more parallel data for Upper Sorbian, an Unsupervised Task for Lower Sorbian, and more ambitiously the inclusion of new typologically diverse languages. Please don't hesitate to contact us if you are interested, particularly if you have access to appropriate data.

8 Acknowledgments

This work has received funding from the European Research Council (ERC) under grant agreement No. 640550. Thanks very much to coorganizers: Hauke Bartels - Sorbian Institute, Olaf Langner - Witaj Sprachzentrum, Marcin Szczepanski - Sorbian Institute.

References

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *the 58th Annual Meeting of the Association for Computational Linguistics*.

Loïc Barrault, Ondrej 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 *the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*.

Rupjyoti Baruah, Rajesh Kumar Mundotiya, Amit Kumar, and Anil Kumar Singh. 2020. NLPRL system for very low resource supervised machine translation. In *the Fifth Conference on Machine Translation*.

Tucker Berckmann and Berkan Hiziroglu. 2020. Low resource translation as language modeling. In *the Fifth Conference on Machine Translation*.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *the Third Conference on Machine Translation, Volume 2: Shared Task Papers*.

Alexandra Chronopoulou, Dario Stojanovski, Viktor Hangya, and Alexander Fraser. 2020. The LMU Munich system for the WMT 2020 unsupervised machine translation shared task. In *the Fifth Conference on Machine Translation*.

Sourav Dutta, Jesujoba O. Alabi, Saptarashmi Bandyopadhyay, Dana Ruiter, and Josef van Genabith. 2020. UdS-DFKI@WMT20: Unsupervised MT and Very Low Resource Supervised MT for German - Upper Sorbian. In *the Fifth Conference on Machine Translation*.

Lukas Edman, Antonio Toral, and Gertjan van Noord. 2020. Data selection for unsupervised translation of German–Upper Sorbian. In *the Fifth Conference on Machine Translation*.

Rebecca Knowles, Samuel Larkin, Darlene Stewart, and Patrick Littell. 2020. NRC systems for low resource German-Upper Sorbian machine translation 2020: Transfer learning with lexical modifications. In *the Fifth Conference on Machine Translation*.

Ivana Kvapilíková, Tom Kocmi, and Ondřej Bojar. 2020. CUNI systems for the unsupervised and very low resource translation task in WMT20. In *the Fifth Conference on Machine Translation*.

Zuchao Li, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2020. SJTU-NICT's supervised and unsupervised neural machine translation systems for the WMT20 news translation task. In *the Fifth Conference on Machine Translation*.

Jindřich Libovický, Viktor Hangya, Helmut Schmid, and Alexander Fraser. 2020. The LMU Munich system for the WMT20 very low resource supervised MT task. In *the Fifth Conference on Machine Translation*.

Yves Scherrer, Stig-Arne Grönroos, and Sami Virpioja. 2020. The University of Helsinki and Aalto University submissions to the WMT 2020 news and low-resource translation tasks. In *the Fifth Conference on Machine Translation*.

Keshaw Singh. 2020. Adobe AMPS's submission for very low resource supervised translation task at WMT20. In *the Fifth Conference on Machine Translation*.

Salam Michael Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2020. The NITS-CNLP system for the unsupervised machine translation task at WMT 2020. In *the Fifth Conference on Machine Translation*.