Data Selection for Unsupervised Translation of German–Upper Sorbian

Lukas Edman **Antonio Toral** **Gertjan van Noord**
Center for Language and Cognition
University of Groningen
{j.l.edman, a.toral.ruiz, g.j.m.van.noord}@rug.nl

Abstract

This paper describes the methods behind the systems submitted by the University of Groningen for the WMT 2020 Unsupervised Machine Translation task for German–Upper Sorbian. We investigate the usefulness of data selection in the unsupervised setting. We find that we can perform data selection using a pretrained model and show that the quality of a set of sentences or documents can have a great impact on the performance of the unsupervised neural machine translation (UNMT) system trained on it. Furthermore, we show that document-level data selection should be preferred for training the state-of-the-art UNMT model, the XLM model, when possible. Finally, we show that there is a trade-off between quality and quantity of the data used to train UNMT systems.

1 Introduction

Unsupervised Neural Machine Translation (UNMT) has recently become the dominant paradigm for unsupervised MT, with the advent of cross-lingual language model pretraining as used in the XLM model (Conneau and Lample, 2019). However, much of the existing research in UNMT assumes that the amount of data available for one language is roughly equivalent to the other. The WMT 2020 Unsupervised Machine Translation task is unique in that monolingual data is abundant for one language (German), with hundreds of millions of sentences available, and sparse for the other (Upper Sorbian), which only has around 750 thousand sentences available. With a wealth of data available on the German side, it is natural to ask: how can we best use this data? Viewing this under the lens of data selection, we break this broad question down into 3 concrete sub-questions, tailored for the unsupervised setting. They are as follows:

- How can we determine the quality of training data?
- What kinds of data selection are best for training an XLM model?
- Is quality or quantity more important when it comes to training data for UNMT?

Section 2 describes the general setup pertaining to every experiment, including datasets, data processing steps, model architecture, and training details. In Section 3, we detail our individual experiments and their corresponding results. Finally, in Section 4, we make our conclusions and discuss paths for future work.

2 Setup

For Upper Sorbian, we use the 3 monolingual datasets provided by the Sorbian Institute, the Witaj Sprachzentrum, and the web data from CIS, LMU. We also use the Upper Sorbian side of the parallel corpus from `train.hsb-de.hsb.gz`. For German, we use monolingual data from News Crawl and Common Crawl. For validation and testing, we use the data provided in `devtest.tar.gz`.

All data is tokenized and truecased using the Moses toolkit (Koehn et al., 2007). For BPE segmentation (Sennrich et al., 2016), we apply a joint segmentation for both languages. This is done by first taking a sample of the German data of the same length as the Upper Sorbian data (around 750 thousand sentences). The BPE codes are learned and applied using FastBPE.[1] After BPE is applied, we remove duplicate sentences while retaining the order of the corpora.[2]

We used the XLM model (Conneau and Lample, 2019) using the default parameters, with the excep-

[1] https://github.com/glample/fastBPE

[2] For document-level filtering, we do not remove duplicates.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 1099–1103

tion of allowing for sentences of max length 200 rather than 100.[3] The language model pretraining step includes only masked language modelling, and training is limited to 24 hours. The NMT step is also limited to 24 hours, with the additional stopping criterion of no improvement on the DE→HSB validation set for 10 epochs.[4]

3 Experiments

For all of our data selection experiments, we start by training an initial model. Our initial model is trained on 10 million German sentences and all of the available Upper Sorbian sentences. The 10 million German sentences include all of the data from years 2007 and 2010, and the remaining sentences are taken from 2014.[5] Our initial model achieves BLEU scores of 17.43 and 19.05 for DE→HSB and HSB→DE respectively.

3.1 Data Selection

We apply two forms of data selection: sentence-level and document-level. As we have an abundance of German data ($\mathcal{D}$) and limited Upper Sorbian data ($\mathcal{H}$), we are only concerned with data selection for German. To select from $\mathcal{D}$, we first must score our data in terms of its potential to improve the performance of our NMT model. Drawing inspiration from Moore and Lewis (2010), our scoring function is as follows:

$$Score(s) = \frac{LM_{\mathcal{H}\rightarrow\mathcal{D}'}(s) - LM_{\mathcal{D}}(s)}{|s|}$$

In this equation, s refers to any sentence in the German data, $|s|$ to its token length, $LM_{\mathcal{X}}(s)$ to the log probability of s using a language model trained on dataset $\mathcal{X}$, and $\mathcal{H} \rightarrow \mathcal{D}'$ to the dataset obtained by translating $\mathcal{H}$ into German using the initial system. A high scoring sentence is thus a sentence that has a high probability according to the Upper Sorbian language model compared to that of the German language model.[6]

[3]The max length increase was found to perform slightly better in early testing.

[4]Both steps are limited to 24 hours as there was little to no improvement observed beyond 24 hours in preliminary tests.

[5]We choose these years because we found that the frequencies of "20XX" in the Upper Sorbian data peak at 2005, 2010, and 2014, and 2007 is the earliest News Crawl data available.

[6]The intuition behind subtracting the score of the German language model is that without it a sentence may have a high score due to it containing frequent words in general (e.g. "the") rather than words that are particularly frequent in the Upper Sorbian dataset (e.g. "Sorbia").

Selection Type	DE→HSB	HSB→DE
Sentence - Low	5.21	5.91
Sentence - Random	**16.98**	**18.45**
Sentence - High	15.08	18.05
Document - Low	9.32	8.46
Document - Random	17.03	18.19
Document - High	**17.60**	**19.23**

Table 1: BLEU scores for XLM trained on data selected with the lowest and highest sentence and document-level scores, as well as randomly selected sentences and documents.

The language model we use is KenLM (Heafield et al., 2013). We use a trigram model, with all other parameters being the default values. Since we require a portion of the German dataset to train the model, we choose N sentences randomly, with N being equal to the number of sentences in $\mathcal{H}$.[7] These sentences are not included during the selection process.

For sentence-level selection, we simply order each sentence based on score and select the sentences with the highest scores. For document-level selection, we score each document by averaging its sentence-level scores, and select the documents with the highest scores.

To answer our first research question, we show that systems trained on the highest scoring sentences and documents perform significantly better than those trained on the lowest scoring sentences and documents. For this experiment, we start with 10 million sentences from News Crawl 2015, and score each sentence and document. We then train models on the 2 million lowest and highest scoring sentences, as well as the lowest and highest scoring documents which total 2 million sentences in length. The results are shown in Table 1.

The results show a drastic improvement from using the lowest quality sentences to the highest according to our scoring function. This applies both at the sentence and document level. However only document-level filtering outperforms random selection. We speculate that this is due to a potential lack of variety in the sentence-level filtering, as it may select sentences with substantial trigram overlap, due to their similarly high score. This would be less of an issue on the document-level, since there is a smaller likelihood for two documents to have a high degree of overlap. A potential solution

[7]The choice of N follows Moore and Lewis (2010).

to this lack of variety would be to select sentences sequentially, enforcing a word overlap constraint. This would limit the number of words a sentence could share with previously selected sentences.

3.2 Document-level versus sentence-level

We see from Table 1 that document-level selection outperforms sentence-level selection. This could be for 2 reasons: either the sentences selected are higher quality on average or the language model pretraining step for the XLM model benefits more from documents than sentences. To further explain the latter reason, the pretraining step for XLM uses streams of text which can contain multiple sentences, so sentences being in order should be beneficial for training the language model. To test this, we take the document-level selected sentences and shuffle their order and train a new model. With a shuffled dataset, we obtained far lower BLEU scores of 12.84 and 16.73 for DE→HSB and HSB→DE respectively. As these BLEU scores are lower than even the scores obtained via sentence-level selection, we can conclude that the XLM model greatly benefits from sentences being in order for pretraining. However, it does appear that sentence-level selection provides higher quality sentences individually.

3.3 Quality versus quantity

With both selection methods, we can choose a threshold to determine how many sentences we should use for training our model. We start by selecting roughly 93 million sentences from News Crawl 2007-2019.[8] We chose the first 10 million sentences from each year, apart from 2008 and 2009, which only contain roughly 6.5 million sentences each. The sentences are chosen at the document-level. From the 93 million sentences combined, we use document-level selection to choose various amounts of data, varying from 1 million to 20 million sentences, and train models on each. The results are shown in Table 2.

As we can see, selecting 5 million sentences results in the highest BLEU scores. As data is either added or removed, the performance drops by around 1-2 BLEU. Given the nature of attention-based neural models, it is somewhat surprising to see that using more data is not helpful and in fact potentially harmful. Whether this is a peculiarity of the German–Upper Sorbian data or not requires further investigation.

Sentences (M)	DE→HSB	HSB→DE
1	16.01	17.14
2	15.20	16.61
5	**17.18**	**19.32**
10	16.78	18.65
20	16.09	17.75

Table 2: BLEU scores of models trained on varying amounts of document-level selected data.

Sentences (M)	DE→HSB	HSB→DE
2	17.76	19.19
5	**18.04**	**19.57**

Table 3: BLEU scores of models trained using 5 million sentences from News Crawl and various amounts of sentences from Common Crawl.

3.4 Using Common Crawl data

As a portion of the Upper Sorbian data is crawled from the web, we also perform data selection on Common Crawl. Since document boundaries are not available for Common Crawl, we can only use sentence-level selection. [9] We tested using various amounts of data in addition to the 5 million News Crawl sentences and report results in Table 3.

As we can see the system with 5 million News Crawl sentences and 5 million Common Crawl sentences performed the best. While the improvements are marginal, this may be due to a similar phenomenon as in Table 2, where too much monolingual data is not beneficial.

3.5 Iterative data selection

Since we saw improvements from one round of data selection, it would stand to reason that using a more accurate model to translate the Upper Sorbian data to German would result in potentially better data selection. As such, we use our model trained on 5 million sentences selected from News Crawl to translate the Upper Sorbian data into German, and apply the same data selection process on the roughly 93 million sentences as before.

The results on the second iteration are markedly worse, with BLEU scores of 15.9 and 17.45, on DE→HSB and HSB→DE, respectively, compared

[8] We exclude years 2007, 10, and 14 as they are used for training our initial model and thus may affect the selection.

[9] Our finding that randomly selected sentences indeed perform better was done post-hoc, which is why we use sentences selected with the highest scores.

to the original scores of 17.18 and 19.32. We suspect that this is due to the same data being used for training the NMT system and for selection, despite the data being used to train the KenLM models being different.[10]

This highlights a major downside of data selection using our methods: data cannot be used both for training a selection model and for the selection itself. The most likely reason for this is that the model will give all sentences that appear in the original training set higher scores, and documents which include the same or similar sentences will be chosen over documents that are more unique, effectively leading to an overfitting problem. This then raises a question of trade-off: is it better to use worse quality data to train the initial model and to then select from better quality data, or vice versa? Our results seem to indicate the former, but further research is required to get a definitive answer.

3.6 Further Analysis

To further analyze the data selected by the model, we look at the frequencies of words that appear in the selected data. We compare our document-filtered data from Section 3.1 with the data from the Upper Sorbian side for 10 word roots in Table 4. These word roots are selected manually as the correctly translated root is easy to verify (with Wikipedia and Wiktionary), and the translations are also one-to-one (ignoring the suffixes). We also select roots with varying frequency within the Upper Sorbian dataset.

As we can see, the high-quality document-filtered data has higher relative frequencies for the first 7 out of 10 word roots, and the lower-quality data has higher frequencies for the last 3. As the words are in order of frequency within the Upper Sorbian dataset, this indicates that the higher quality filtered data better represents the topics found in the Upper Sorbian dataset. Roots such as Sorbia- and Bautzen (a city where Sorbian is spoken) appear far more often in the higher quality data, despite being relatively uncommon in the German dataset. The last 3 words are relatively rare in the Upper Sorbian data, so it makes sense that the higher quality filtered data would have fewer occurrences of these words. Although most of the examples are related to locations, we do see that Domowin- (the root for Domowina, a non-profit organization) and Catholic- appear to show the same trends.

We also looked at the relative frequencies of the years 2000-2025 across our various models to see the effect of our filtering methods in matching the Upper Sorbian data according to year. We expect that the filtered German data with the frequency distribution most closely matching the frequency distribution of the Upper Sorbian data will have the strongest NMT performance. We show the results in Figure 1.

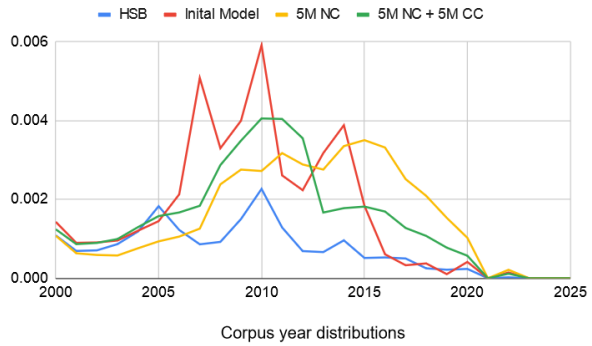

Figure 1: Relative frequencies of the years 2000-2025 within the various datasets. The frequencies are relative to the total number of sentences in that dataset.

Our initial model predictably has spikes in frequency at 2007, 2010, and 2014 as we manually chose data from these years to somewhat match the frequency of the Upper Sorbian data. Meanwhile, the 5 million document-level selected sentences from News Crawl seems to more closely match the frequencies in the Upper Sorbian data from 2000 to 2010, but has larger relative frequencies for years 2010 to 2020. We suspect that this is due to the limitation of the data available for selection, as earlier years have fewer sentences for the selection model to choose. Finally, the model using 5 million News Crawl and 5 million Common Crawl sentences has a frequency graph that most closely matches the graph of the Upper Sorbian data. The similarity of the Upper Sorbian graph to the other graphs seems to correlate with the resulting BLEU scores of the NMT model.

4 Conclusion

In the UNMT setting where one has access to a wealth of resources for one language, we investigated the feasibility of data selection. We attempt both document-level and sentence-level selection,

[10] We also saw similar performance drops when trying to include the data from years 2007, 10, and 14 in our original model trained on selected data, as these years were used to train the initial system used for selection.

Root			Count		Frequency %	
EN	DE	HSB	HSB	DE	Doc Low	Doc High
Sorbia-	Sorb-	Serbsk-	66105	187	0	97.3
German-	Deutsch-	Němsk-	17070	445203	18	21.8
Bautzen	Bautzen	Budyšin	11015	212	8.5	50.5
Lusatia-	Lausitz	Łužic-	10170	633	2.8	51.8
Domowin-	Domowin-	Domowin-	7835	32	0	100
Saxon-	Sachsen-	Saksk-	5163	10861	14.4	24.8
Catholic-	Kathol-	Katolsk-	4530	8515	12.7	28.9
Asia-	Asi-	Azij-	735	12175	31.4	11.3
Africa-	Afrik-	Afrik-	512	9967	23.2	15.7
Iran-	Iran-	Iran-	199	26714	53.9	4.6

Table 4: Frequencies of word roots within the Upper Sorbian (HSB), and relative frequencies of low-quality document-filtered (Doc Low) and high-quality document-filtered (Doc High) datasets. Relative frequency is based on the total frequency of each root within the 10 million sentences that the sets are selected from (i.e. the DE count column). Case is ignored when determining frequency.

finding that both methods are capable of distinguishing low quality data from high quality data, with quality in this case defined as the efficacy for training an XLM model. We found that while document-level selection chooses poorer sentences on average, the XLM model can leverage the inter-sentence information to achieve better results than when simply using the highest quality sentences. We also found that there appears to be a point where adding more monolingual data is not beneficial, but rather potentially harmful, indicating a need for data selection. Finally, we noted some potential drawbacks to using this form of data selection, particularly that data cannot be used for both initial training of the NMT model and subsequent selection. Future work could continue along many avenues, such as the effectiveness of data selection on other language pairs, or even on the Upper Sorbian side.

References

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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.