Bilingual Sentence Mining for Low-Resource Languages: a Case Study on Upper and Lower Sorbian

Shu Okabe[1,2] and **Alexander Fraser**[1,2,3]
[1]School of Computation, Information and Technology
Technische Universität München (TUM)
[2]Munich Center for Machine Learning
[3]Munich Data Science Institute
shu.okabe@tum.de, alexander.fraser@tum.de

Abstract

Parallel sentence mining is crucial for downstream tasks such as Machine Translation, especially for low-resource languages, where such resources are scarce. In this context, we apply a pipeline approach with contextual embeddings on two endangered Slavic languages spoken in Germany, Upper and Lower Sorbian, to evaluate mining quality. To this end, we compare off-the-shelf multilingual language models and word encoders pre-trained on Upper Sorbian to understand their impact on sentence mining. Moreover, to filter out irrelevant pairs, we experiment with a post-processing of mined sentences through an unsupervised word aligner based on word embeddings. We observe the usefulness of additional pre-training in Upper Sorbian, which leads to direct improvements when mining the same language but also its related language, Lower Sorbian.

1 Introduction

Machine Translation (MT) essentially relies on parallel corpora, which are widely available for 'winner' languages (Joshi et al., 2020). Yet, when it comes to lower-resourced languages, they become rarer, and such resources are more costly to obtain compared to monolingual corpora. This is why, to circumvent situations with too few or even no parallel sentences, parallel sentence mining is a task to find parallel sentences automatically in monolingual corpora. Research on parallel sentence mining is intertwined with MT since improving mining quality often leads to a better translation model.

The BUCC Shared Tasks (Zweigenbaum et al., 2017; Pierre Zweigenbaum and Rapp, 2018) notably focus on parallel sentence mining and acts as a benchmark. However, only four well-resourced language pairs are represented there. Hence, we try to fill this gap by evaluating sentence mining for low-resource languages.

In this work, we consider Upper Sorbian and Lower Sorbian, paired with German, which can be seen as a case study for low-resource sentence mining. We can effectively observe two data conditions (the former has more data than the latter) and also the impact of relatedness between the two languages.

We will try to answer the following questions: How well can we mine parallel sentences for a language with off-the-shelf word encoders? How useful is it to pre-train a model with the available monolingual data? How helpful is it to pre-train a model on a related language?

We consider two scenarios: (i) when computing resources are limited, we use already pre-trained models; (ii) otherwise, we fine-tune a language model on the available monolingual corpus in the low-resource language.

As such, we aim to foster further research on bilingual mining for low-resource languages and its challenges. We hope that this study provides important lessons useful even in a more data-restricted scenario.

To this end, we propose (a) two BUCC-style mining corpora, (b) a comparison of two state-of-the-art language models in mining Sorbian-German parallel sentences, (c) word encoders with different amounts of pre-training sentences in Upper Sorbian, and (d) an alignment post-processing to improve the mining quality. Thus, our work can serve as a benchmark for two low-resource languages in a realistic scenario. We release the corpora, the mining pipeline, and all related code material[1].

Section 2 will focus on the two languages and the creation of the corpora, while Section 3 compares the considered language models, the pre-training strategy and explains the mining method. Section 4 presents and analyses the mining results.

[1]At https://github.com/shuokabe/PaSeMiLL.

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 11–19

2 Languages and datasets

2.1 On Upper and Lower Sorbian

Upper Sorbian (ISO code: hsb) and Lower Sorbian (dsb) are two West Slavic languages and constitute the Sorbian branch. Both are spoken in Germany (Saxony for the former and Brandenburg for the latter) and are currently classified as endangered according to Ethnologue (Eberhard et al., 2024). There are state-level laws that notably guarantee the use and teaching of both languages. For instance, the Witaj Sprachzentrum (Witaj Language Center) offers language courses in certain kindergartens and schools.

The NLP community has lately focused on the two Sorbian languages in cooperation with them and the Sorbian Institute. They both provided data for the successive WMT Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT (Fraser, 2020; Libovický and Fraser, 2021; Weller-Di Marco and Fraser, 2022).

Hence, we focus on the Upper Sorbian-German and Lower Sorbian-German language pairs in this work. Previously, only Kvapilíková and Bojar (2023) focused on Upper Sorbian-German parallel sentence mining with a pre-training of XLM (Conneau and Lample, 2019), but the task has not been addressed on Lower Sorbian yet. It is the closest work, but their aim was to train a MT model, and their pre-trained encoder notably required 500K sentences in Upper Sorbian and German, which is already a large amount of available data and, hence, not a realistic scenario for most low-resource languages.

2.2 BUCC-style dataset creation

For our experiments, we apply the methodology of the BUCC 2017 Shared Task (Zweigenbaum et al., 2017) to Upper and Lower Sorbian by injecting known parallel sentences into their respective monolingual corpus.

This evaluation is an artificial approach, which can introduce some biases, such as parallel sentences that may stand out from the original monolingual sentences. However, this task is more difficult than the related sentence matching and gives a more realistic setting for bilingual mining.

We rely on the data provided for the above-mentioned WMT Shared Tasks and select its 2020 edition for Upper Sorbian and 2022 for Lower Sorbian for both monolingual and parallel sentences.

More precisely, for Upper Sorbian, we rely on the WMT 2020 Shared Task data and use the monolingual corpus provided by the Sorbian Institute (339,822 sentences). The monolingual German data comes from the Leipzig news corpora[2] (2020) (Goldhahn et al., 2012) and has 300K sentences. We chose to insert the development and development test data from the Shared Task (4,000 sentences) as parallel data.

For Lower Sorbian, we use the WMT 2022 Shared Task data for its monolingual corpus (66,408 sentences) and the parallel sentences from the development and development test datasets (1,353 sentences). The monolingual German data comes from the Leipzig news corpora of 2022 and contains 100K sentences.

Compared to the original BUCC methodology, presented in Zweigenbaum et al. (2017), we modified the following points. Instead of inserting a parallel sentence in a section of the monolingual corpus which deals with similar topics, we chose to shuffle all sentences. While we lose the context of each sentence, our mining pipeline does not take it into account. Besides, short sentences have been kept, while very long sentences of more than 40 words have been removed in the monolingual corpora, which explains the smaller datasets. Finally, we lower the possible amount of 'natural' parallel sentences (i.e., parallel sentences in the original monolingual corpora) by using the Leipzig news corpora, which is not directly related.

Table 1 presents the number of sentences in the Upper and Lower Sorbian datasets after inserting parallel sentences and shuffling. We also split the dataset into training and test subsets in a similar proportion of parallel sentences as in the German-English pair in the BUCC Shared Task.

	train	test
Upper Sorbian corpus	34,001	101,751
German corpus	32,915	98,747
of which parallel	1,000	3,000
Lower Sorbian corpus	22,303	44,616
German corpus	33,756	67,513
of which parallel	451	902

Table 1: Number of sentences in the Upper Sorbian (top) and Lower Sorbian (bottom) datasets.

[2] https://wortschatz.uni-leipzig.de/en/download/German.

2.3 Dataset difficulty

We verify whether the BUCC-style datasets created in Section 2.2 are suited to evaluate the mining task or not. If parallel sentences stand out from the other sentences which were originally in the unrelated monolingual corpus, the artificial dataset is deemed too easy.

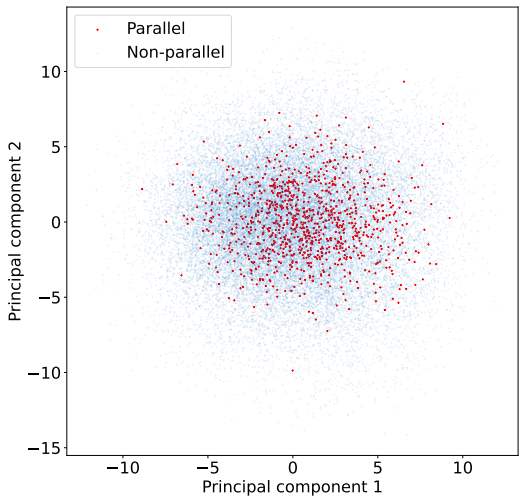

Figure 1: Distribution of embeddings of the sentences in the corpora according to the first two principal components for the created German dataset

We use the state-of-the-art sentence encoder LaBSE (Feng et al., 2022) to encode the well-resourced German dataset. We reduce the embedding dimension through a principal component analysis (PCA). Figure 1 displays each sentence embedding of the dataset according to the first two principal components. We can see that both parallel and non-parallel sentences are situated in similar regions with no clear cluster of sentences. Therefore, the task is not too easy.

3 Sentence mining methodology

3.1 Baseline models

We mainly study two multilingual pre-trained models to represent words. First, XLM-RoBERTa or XLM-R (Conneau et al., 2020) is a frequently used multilingual language model; we use its base version from the Transformers library. The other model is Glot500-m (Imani et al., 2023), which is an extension of XLM-R with additional pre-training for more than 500 low-resource languages.

It must be noted that XLM-R has seen German and two Slavic languages, namely Czech and Polish, during pre-training. Glot500-m has additionally been trained on 105K sentences in Upper Sorbian. From this perspective, both Sorbian languages are in a better situation than many other low-resource languages, which might not have as much available data or related well-resourced languages in the model pre-training.

3.2 Pre-training XLM-R in Upper Sorbian

Given that we have access to a monolingual corpus in Upper Sorbian, we also pre-train XLM-R on the available Upper Sorbian monolingual corpus. This model will also enable us to see how the additional pre-training in Upper Sorbian can indirectly help in the more closely related Lower Sorbian.

We replicate the pre-training strategy of Glot500-m (Imani et al., 2023). To gauge the amount of needed data to reach similar (or better) performance, we use different sizes and compositions of pre-training datasets.

In practice, we use the shuffled monolingual corpora presented in Section 2.2 for Upper Sorbian and German to pre-train XLM-R with a standard masked language modelling (MLM) objective. We obtain three models, named PT-HSB-3, PT-HSB-6, and PT-HSB-9, with different amounts of pre-training data: respectively, 30K, 60K, and 90K monolingual sentences in Upper Sorbian, coupled with and at least 30K sentences in German.

Providing bilingual cues Moreover, since Upper Sorbian is a Slavic language, we leverage additional data from the same language family. In our case, we choose to use parallel sentences in Czech and German, a better-resourced pair. Such a choice can be applied to other language pairs by considering neighbouring or related languages.

Hence, we carry out an additional pre-training on top of PT-HSB-9 with a MLM objective on a bilingual Europarl corpus in Czech and German from OPUS (Tiedemann, 2012), where we concatenate parallel sentences as one sentence for the model. We denote this model PT-HSB-9 + CS-DE. We restrict the training size to 220K sentences. The idea is twofold: give bilingual cues to the model, which is known to help the model, even when the language pair is different (Kvapilíková et al., 2020), and to indirectly improve the Upper Sorbian word representation thanks to Czech.

Experimental conditions To pre-train the models, we first relied on vocabulary extension, following the methodology of Imani et al. (2023). For

each pre-training setting, we extend the vocabulary to the used monolingual or bilingual corpora. Then, the pre-training itself uses the default parameters and approach given in the Transformers library.

All mining experiments have been carried out on 1 GPU (NVIDIA Tesla V100). The additional pre-training of XLM-R in Upper Sorbian or with the Czech-German corpus has been done on 1 to 4 of the same GPUs for 5 epochs. The longest pre-training is with the bilingual cues, due to a higher number of sentences and longer length; this took almost one week effectively. The other pre-trained models required a few days.

3.3 Mining and evaluation methods

The overall mining pipeline follows (Hangya and Fraser, 2019). First, we derive sentence representations by mean-pooling word embeddings with our encoders, which is a more effective approach than max-pooling (Kvapilíková et al., 2020). Then, we compute the similarity between a source (Sorbian) sentence and a target (German) sentence in the multilingual embedding space using the CSLS (Cross-Domain Similarity Local Scaling) score (Conneau et al., 2018)[3]. This metric is known to better deal with the hubness issue than the standard cosine similarity (Dinu et al., 2015). Formally, for two sentence vectors x and y, it is computed as in Equation (1):

$$\begin{aligned}\mathrm{CSLS}(x,y) &= 2\cos(x,y) \\ &- \sum_{z\in \mathrm{NN}_k(x)} \frac{\cos(x,z)}{k} - \sum_{z\in \mathrm{NN}_k(y)} \frac{\cos(y,z)}{k}, \end{aligned} \quad (1)$$

where $\mathrm{NN}_k(x)$ indicates the k-nearest neighbours of vector x. We choose $k = 20$.

Finally, we consider a source sentence and its most similar target sentence to be parallel according to a threshold that is chosen dynamically on the training dataset. Defined as in Equation (2) by Hangya et al. (2018), the threshold value depends on the mean and standard deviation (σ) from the found similarity values (S):

$$\text{threshold} = \text{mean}(S) + \lambda \times \sigma(S), \quad (2)$$

where λ is the tuneable hyper-parameter.

We evaluate the mining quality by computing the usual Precision, Recall, and F-score, following the BUCC Shared Task methodology. We also report the number of mined sentences (N_{sent}).

[3]This method is related to the margin-based methods presented by Artetxe and Schwenk (2019a); we observed comparable results on our dataset whether with CSLS or a ratio margin.

4 Experimental results

4.1 Mining results

embeddings	P (%)	R (%)	F (%)	N_{sent}
XLM-R	3.64	2.03	2.61	1,675
Glot500-m	32.82	20.63	25.34	1,886
PT-HSB-3	22.36	8.77	12.60	1,176
PT-HSB-6	34.54	17.23	22.99	1,497
PT-HSB-9	34.36	17.50	23.19	1,528
+ CS-DE	**36.96**	**26.30**	**30.73**	2,135

Table 2: Evaluation on the test dataset of the **Upper** Sorbian corpus.

Upper Sorbian Table 2 presents the quality of the mined parallel sentences with different word embeddings on the Upper Sorbian-German dataset. XLM-R's performance indicates that these embeddings are not suited for Upper Sorbian and cannot extend well to this language based on related pre-trained languages only. On the contrary, Glot500-m, which has seen a number of sentences in Upper Sorbian, has higher scores than XLM-R: the additional pre-training does indeed help to get a better word and, hence, sentence representation.

The bottom half of the table shows the performance of the different XLM-R models pre-trained on Upper Sorbian and German, as presented in Section 3.2. Not surprisingly, increasing amounts of Upper Sorbian data improve mining performance, reaching scores similar to Glot500-m, which was trained with roughly 100K Upper Sorbian sentences. Furthermore, using a bilingual cue from a related language pair (here Czech-German) enables us to go further, with an F-score of 31 for PT-HSB-9 + CS-DE. It is worth noting that this additional pre-training mainly helps with recall.

Lower Sorbian We use the same experimental methodology on the Lower Sorbian corpus and show the results in Table 3. XLM-R mines sentences of similar quality in both Sorbian languages: with no prior knowledge of the language, they equally struggle with F-scores of less than 3. Moreover, since Glot500-m has not seen any Lower Sorbian sentence, it also has a very low F-score compared to the Upper Sorbian case: pre-training in the language is indeed crucial, especially when mining

embeddings	P (%)	R (%)	F (%)	N_{sent}
XLM-R	5.88	1.88	2.85	289
Glot500-m	6.75	5.65	6.15	756
PT-HSB-3	5.99	5.21	5.57	785
PT-HSB-6	8.85	5.21	6.56	531
PT-HSB-9	10.06	6.87	8.17	616
+ CS-DE	**11.01**	**11.75**	**11.37**	963

Table 3: Evaluation on the test dataset of the **Lower** Sorbian corpus.

with averaged word embeddings. Here, the related languages help, with 3 points of F-score above the standard XLM-R, but only in a limited fashion.

Regarding pre-trained XLM-R models, they exhibit a comparable trend as for Upper Sorbian: more pre-training sentences improve the mining quality. The models see no Lower Sorbian during pre-training; the increase in performance is only due to the transfer between the related Slavic languages.

4.2 Precision-recall trade-off

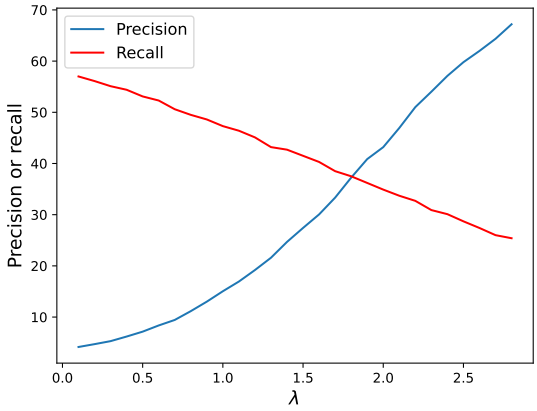

Figure 2: Evolution of the precision and recall for the best-performing PT-HSB-9 + CS-DE model on the *training* dataset in Upper Sorbian.

We defined the mining threshold by maximising the F-score on the training dataset in section 3.3. However, in real-life scenarios, the relevant criteria might differ depending on the use case. Another possible strategy would be to aim for a higher recall because, once mined, the precision can be increased through post-processing by filtering out wrong pairs.

Besides, as Figure 2 shows for the best model on Upper Sorbian, we notice that precision tends to rise faster than the decline in recall when increasing the threshold parameter λ. This means that voluntarily choosing a sub-optimal λ with a higher recall and post-processing could lead to higher F-scores.

Post-processing One approach is through manual annotation, which requires active involvement from the language community or speakers. This can be tedious, depending on the initial mining quality. A second method is to rely on unsupervised word aligners solely based on embeddings, such as SimAlign (Jalili Sabet et al., 2020). Given the lower amount of sentences, compared to the BUCC setting, for instance, this remains a reasonable option regarding the computing time and cost[4]. Moreover, since we focused on word encoders for Upper Sorbian in this work, embedding-based aligners can also benefit from the additional pre-training.

In this experiment, we select a very low threshold, $\lambda = 0.1$, to compromise between a scalable amount of sentences to align and a high enough recall. For all the kept pairs, we use SimAlign with our pre-trained embedding models[5]. Then, we compute a simple two-way alignment score by counting the found alignment links divided by the number of words in the sentence in both directions. Finally, using the dynamic threshold of Equation (2), we only consider sentence pairs above a threshold alignment score. This post-processing leads to a significant improvement when used on the best-performing PT-HSB-9 + CS-DE model in Upper Sorbian, for instance, with an F-score of 51.38 (to compare with 31, without alignment post-processing, in Table 2).

Since this approach relies on embeddings to correctly align words, it requires a decent modelling of the language. For instance, when we apply this post-processing method to Lower Sorbian (still with the PT-HSB-9 + CS-DE model), we only improve the F-score by 2 points, reaching 13.44.

4.3 Qualitative analysis

In Table 2, XLM-R obtained low metric scores on Upper Sorbian despite finding more than 1,000 sentences. This poor mining quality can be qualitatively seen in Figure 3, where the source Upper Sorbian and the found German sentences have nothing in common. Using the best model, in our case, PT-HSB-9 + CS-DE with alignment post-processing,

[4]In our experiments, the largest number of sentence pairs to align was 42,170, for the Upper-Sorbian test dataset, which took less than 10 minutes on one GPU.

[5]We use the 8th layer and align with the 'Argmax' strategy.

enables us to find the correct target German sentence.

HSB	Wón namjetuje moderěrowanu diskusiju wo tym.
XLM-R	Sie rechen das Laub der Laubbäume.
	They rake the leaves of the deciduous trees.
Best	Er schlägt eine moderierende Diskussion darüber an.
	He proposes a moderated discussion about this.

Figure 3: Example of mined sentences in Upper Sorbian. While XLM-R finds an unrelated sentence, PT-HSB-9 + CS-DE with alignment post-processing identifies the correct German sentence.

Figure 4 presents a pair of sentences wrongfully considered as parallel by the mining programme using PT-HSB-9 + CS-DE with alignment post-processing. One limitation of considering averaged word embeddings as sentence embedding is that nuances or details can get diluted in the final representation. A common issue is, hence, when two sentences have similar topics; even embedding-based word aligners will struggle in these cases. As such, the example sentences are incorrectly considered parallel because of a similar topic and structure. In the second half of the sentence, the dates and times do not correspond: Sunday, 15th of July ('njedźelu, 15. julija') at 17:00 ('17 hodź') in Upper Sorbian and Wednesday, 9th of December ('Mittwoch, den 9. Dezember') at 20:15 ('20.15 Uhr') in German. Nonetheless, this pair gets a high CSLC similarity score, and the computed align rate is 60%.

5 Related works

Parallel sentence mining has been extensively studied as an intermediate step geared towards Machine Translation, further stimulated by the BUCC Shared Tasks (Zweigenbaum et al., 2017; Pierre Zweigenbaum and Rapp, 2018). Previous works usually tackled parallel sentence mining with supervised bilingual and multilingual embeddings (e.g., Guo et al., 2018). When unsupervised, i.e. with no training parallel sentences, the embeddings stemmed from monolingual static embeddings such as fastText (Bojanowski et al., 2017) that were mapped to form bilingual or multilingual word embeddings (Hangya et al., 2018; Hangya and Fraser, 2018, 2019).

Then, static embeddings were replaced in the pipeline by multilingual contextual embeddings such as in (Kvapilíková et al., 2020; Kvapilíková and Bojar, 2023). The key point was to improve the bilingual (or multilingual) sentence representation, as proposed by Schwenk (2018).

Another reason to tackle sentence mining is to estimate the quality of embeddings; it is a simpler task computing-wise compared to the more resource-intensive machine translation. Similarly, an alternative method to assess the quality of multilingual word representations is sentence matching, where a parallel corpus is shuffled, and the true pairing must be found. It is more scalable to multiple languages due to the lower number of sentences to process.

An adjacent field of work worth mentioning here is on pre-trained multilingual embeddings, among which XLM-R (Conneau et al., 2020) and Glot500-m (Imani et al., 2023), that we consider here. The latter has notably been tested on the sentence-matching task to evaluate its word representation quality.

Finally, the task of parallel sentence mining itself is well-handled by multilingual sentence encoders, namely LASER (Artetxe and Schwenk, 2019b) and LaBSE (Feng et al., 2022), due to their massive training datasets and their specific objectives. More precisely, Costa-jussà et al. (2022) have actually striven to extend the initial LASER embeddings to more than 200 less-represented low-resource languages with LASER3 thanks to teacher-student distillation (Heffernan et al., 2022). By combining this approach with contrastive learning, Tan et al. (2023) get even further improvement on eight low-resource languages, with larger clean parallel data than Sorbian languages. These sentence embeddings are still unavailable for most low-resource languages, and extending them usually requires a significant amount of data or compute (if not both). The extension of our study to such embeddings goes beyond the scope of our current work but will be tackled in the future.

6 Conclusion

We studied the task of sentence mining, using averaged contextual word embeddings, by creating a benchmark for two Slavic low-resource languages: Upper and Lower Sorbian. We notably observed several advantages in carrying out additional pre-training on XLM-R for Upper Sorbian. The pre-trained model gets better word representations, which is reflected in a better mining capacity. Besides, the additional pre-training can improve the mining quality for related languages with even less data, in our case, for Lower Sorbian. Although

HSB	Kocorowy oratorij „Serbski kwas“ zaklinči po něhdźe dźesać lětach zaso, a to tutu njedźelu, 15. julija, w 17 hodź.
DE	Das große Finale von „Die Bachelorette“ läuft am Mittwoch, den 9. Dezember, um 20.15 Uhr bei RTL.

Figure 4: Example of a mining error in Upper Sorbian using PT-HSB-9 + CS-DE with alignment post-processing. Coloured parts respectively correspond to dates (in red) and times (in blue) in both languages but are not translations.

pre-training word encoders have a non-negligible computing cost, they open doors for other downstream tasks or parallel sentence post-processing with word aligners. Alternatively, if the language is already supported by Glot500-m, its word embeddings can be an off-the-shelf solution.

Our future work includes bringing the mining quality even higher by leveraging existing additional language resources (e.g., dictionaries). Besides, the natural downstream task would be machine translation, in a similar fashion to (Kvapilíková and Bojar, 2023) by using mined pseudo-parallel sentences during training.

More generally, we hope this work can foster further initiatives, namely real-life applications towards MT, for instance, together with language communities, in carrying out bilingual sentence mining for other low-resource languages. Our benchmark can also serve as a first place to evaluate upcoming tools before extending them to different languages. Indeed, it has yet to be confirmed whether our observations still hold true for other languages and language families.

Acknowledgments

We thank Viktor Hangya for his help and the anonymous reviewers for their comments. This work has received funding from the European Research Council (ERC) under grant agreement No. 101113091 - Data4ML, an ERC Proof of Concept Grant.

Limitations

We have focused on two low-resource languages, which might not be in the most challenging situation when it comes to pre-trained models: related Slavic languages such as Czech or Polish are commonly seen in the pre-training data, and both languages use the Latin alphabet. This is a favourable setting for an easier transfer between languages. The improvements we saw can hence be difficult to reproduce for languages with more different characteristics (grammar, morphology, language family, or script). Nonetheless, this work still represents an initial attempt towards parallel mining for low-resource languages, and we suggest that future researchers evaluate their tools initially on our benchmark.

Besides, since both Sorbian languages are close enough to two pre-training languages and German is also well-covered, some off-the-shelf *sentence* encoders, such as LASER or LaBSE, already have a high mining performance: with the latter model, the mining quality reaches a F-score of 73.17 on Upper Sorbian and 43.33 for Lower Sorbian. These results are tangential to our work, which focuses on improving *word* embeddings for Sorbian languages when mining sentences.

Finally, the task itself is only suitable for languages with a monolingual corpus large enough, which represents a subset of endangered languages; our work cannot handle left-behind or scraping-by languages (Joshi et al., 2020), where the essential challenge may indeed be to first create larger monolingual corpora in the first place (or to directly create parallel corpora so that sentence mining is not necessary).

References

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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.

Alexis Conneau and Guillaume Lample. 2019. *Cross-lingual language model pretraining*. Curran Associates Inc., Red Hook, NY, USA.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the Workshop Track at ICLR*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the World*, twenty-seventh edition. SIL International, Dallas, Texas. Online version.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. 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.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium. Association for Computational Linguistics.

Viktor Hangya, Fabienne Braune, Yuliya Kalasouskaya, and Alexander Fraser. 2018. Unsupervised parallel sentence extraction from comparable corpora. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 7–13, Brussels. International Conference on Spoken Language Translation.

Viktor Hangya and Alexander Fraser. 2018. An unsupervised system for parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 882–887, Belgium, Brussels. Association for Computational Linguistics.

Viktor Hangya and Alexander Fraser. 2019. Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234, Florence, Italy. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus

mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262, Online. Association for Computational Linguistics.

Ivana Kvapilíková and Ondřej Bojar. 2023. Boosting unsupervised machine translation with pseudo-parallel data. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 135–147, Macau SAR, China. Asia-Pacific Association for Machine Translation.

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.

Serge Sharoff Pierre Zweigenbaum and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Weiting Tan, Kevin Heffernan, Holger Schwenk, and Philipp Koehn. 2023. Multilingual representation distillation with contrastive learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1477–1490, Dubrovnik, Croatia. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Marion Weller-Di Marco and Alexander Fraser. 2022. Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.