Improving Low-Resource Languages in Pre-Trained Multilingual Language Models

Viktor Hangya[1,2], **Hossain Shaikh Saadi**[3] and **Alexander Fraser**[1,2]
[1]Center for Information and Language Processing, LMU Munich, Germany
[2]Munich Center for Machine Learning, Germany
[3]Technical University of Munich, Germany
{hangyav, fraser}@cis.lmu.de
shaikh.saadi@tum.de

Abstract

Pre-trained multilingual language models are the foundation of many NLP approaches, including cross-lingual transfer solutions. However, languages with small available monolingual corpora are often not well-supported by these models leading to poor performance. We propose an unsupervised approach to improve the cross-lingual representations of low-resource languages by bootstrapping word translation pairs from monolingual corpora and using them to improve language alignment in pre-trained language models. We perform experiments on nine languages, using contextual word retrieval and zero-shot named entity recognition to measure both intrinsic cross-lingual word representation quality and downstream task performance, showing improvements on both tasks. Our results show that it is possible to improve pre-trained multilingual language models by relying only on non-parallel resources.

1 Introduction

Pre-trained language models (LMs) have replaced static word embeddings, such as word2vec, fastText or GloVe (Mikolov et al., 2013a; Bojanowski et al., 2017; Jameel and Schockaert, 2016), due to their superior contextualized representations. Approaches such as mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020a) are trained on multiple languages simultaneously resulting in multilingual models which can be used for various cross-lingual transfer learning tasks (Conneau et al., 2018b; Schuster et al., 2019; Artetxe et al., 2020, inter alia).

However, multilingual LMs mainly focus on high resource languages, e.g., mBERT supports the top 104 languages based on Wikipedia sizes, while XLM-R supports the top 100 based on CommonCrawl data. Additionally, many of these languages are underrepresented leading to low model performance in both monolingual and cross-lingual setups. Due to the small data sizes of low-resource languages, subword tokenizers trained jointly on multiple languages tend to over-split the tokens of such languages and LMs are not able to learn good quality representations for them. Recent work have shown that pre-trained LMs can be improved on low-resource languages using vocabulary extension and model fine-tuning (Wang et al., 2020). The cross-lingual quality of LMs can further be improved by learning an additional alignment of language pairs (Aldarmaki and Diab, 2019; Wang et al., 2019) or fine-tuning the whole model (Cao et al., 2020). However these methods require cross-lingual data, often in the form of word aligned parallel sentences, to improve cross-linguality which is often missing for low-resource languages.

In this work we propose an unsupervised approach to improve the language support of low-resource languages without any parallel data. Relying on the initial cross-lingual quality of mBERT we mine word translation pairs from monolingual data of the source and target language pairs by leveraging contextualized cross-lingual word representations (CCWRs). More precisely, we build CCWRs for each token in the source and target corpora and look for the most similar token pairs by calculating their cosine similarity. Even though our approach does not rely on parallel corpora, we show that there are enough sentences with similar contexts (topics) containing at least one word translation pair that are detected by our mining approach. We then use the CCWRs of the mined word pairs to make their representations more similar, this way aligning source and target pairs. We use dedicated linear layers for both of the languages of the considered language pair to learn the alignment and we keep the LM's core frozen. An important contrast with previous work (Aldarmaki and Diab, 2019; Wang et al., 2019; Cao et al., 2020) is that we mine word pairs from sentences of similar contexts, while they were only able to extract them from parallel sentences, which are often not available for low resource languages.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11993–12006
December 7-11, 2022

Despite this, we show that CCWRs can be improved using our mined word pairs.

We conduct experiments on nine low-resource languages and test on two tasks: contextual cross-lingual word retrieval (Cao et al., 2020) and zero-shot named entity recognition (Rahimi et al., 2019). Our results show improved CCWRs for each of the languages and improved NER F_1 scores for eight out of the nine languages. Our analysis reveals our approach to be robust even when we mine noisy word pairs, because it benefits more from a larger quantity than a better quality of bootstrapped training set. Additionally, we experiment with vocabulary extension techniques (Wang et al., 2020) in case of languages using scripts other than Latin or Cyrillic, since the tokenizers of LMs tend to over-split the text of such languages. This step improves CCWR quality of the initial LMs which gives a further boost to our bootstrapping method, leading to the best performance on these languages. Finally, we run preliminary experiments on further four very low-resource languages that were not used for LM pre-training and show that the initial LM quality on such unseen languages is very low, which our approach can marginally improve. Our implementation is publicly available.[1]

2 Related Work

Pre-trained LMs (Devlin et al., 2019; Conneau et al., 2020a) provide the core of many NLP solutions. They are trained using accessible monolingual corpora of multiple languages resulting in multilingual LMs allowing them to be used in zero-shot transfer setups mitigating the issues of missing downstream task training data for many low-resource languages (Conneau et al., 2018b; Schuster et al., 2019; Artetxe et al., 2020). Although no parallel data is used, these models show remarkable cross-lingual quality, i.e., words with similar meaning in different languages are represented similarly by the models. Previous work investigated the reasons of this phenomenon. K et al. (2020) found that the structural similarity of languages is an important factor, while both Conneau et al. (2020b) and Artetxe et al. (2020) showed that a shared vocabulary is not necessary. Dufter and Schütze (2020) identified essential elements for multilinguality, such as shared special tokens or comparable training corpora.

On the other hand, multilingual LMs are less effective on low-resource languages. It was shown that a fixed sized model can only support up to a certain number of languages efficiently, while adding more languages deteriorates its performance (Conneau et al., 2020a). The small size of the available monolingual data for low-resource languages decreases model performance further (Wu and Dredze, 2020a; Lauscher et al., 2020). Additionally, data size imbalance of the used languages leads to an imbalanced subword vocabulary as well (Rust et al., 2021). To mitigate tokenization issues Wang et al. (2020) extend the vocabulary of pre-trained LMs with language specific tokens, while Pfeiffer et al. (2021) propose to learn language specific embeddings for low-resource languages. They show that with better subword vocabulary and model fine-tuning the model's performance can be improved on low-resource languages. We also rely on these techniques in our work.

To improve the cross-lingual quality of LMs various authors proposed steps on top of model pre-training. Based on embedding mapping approaches (Mikolov et al., 2013b; Conneau et al., 2018a) it was shown that representations of monolingual LMs can also be aligned (Schuster et al., 2019). In addition, mapping approaches can be applied to the representations of multilingual LMs as well for further improvements (Aldarmaki and Diab, 2019; Wang et al., 2019; Hämmerl et al., 2022). In contrast to the above approaches, Cao et al. (2020) proposes a word alignment based objective function that fine-tunes the whole model in order to build more similar token representations for the aligned word pairs, while Chi et al. (2021) introduced the denoising word alignment task and Wu and Dredze (2020b) relied on a contrastive alignment objective to encourage better cross-lingual performance. Others leverage cross-lingual training signals already in the model pre-training phase (Conneau and Lample, 2019; Hu et al., 2021). However, the above approaches require parallel data which is not available for many low-resource languages. In contrast, we show that LMs can be improved by using only monolingual corpora where source and target language sentences with similar contexts can be found.

3 Unsupervised Language Alignment

In order to improve the cross-lingual quality of pre-trained multilingual LMs we mine word translation pairs using only monolingual corpora of the source and target languages. We rely on the token representations given by LMs to look for the most similar

[1] https://cistern.cis.lmu.de/lowresCCWR

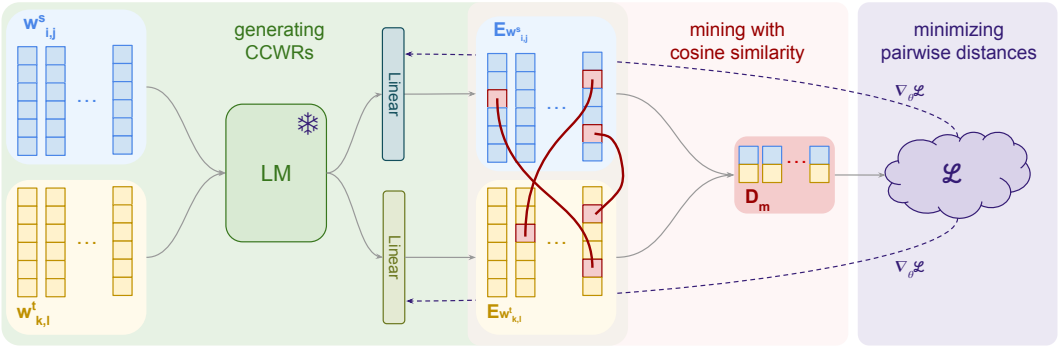

Figure 1: High-level overview of our proposed approach. First we generate CCWRs ($E_{w^{:}_{:,:}}$) of source ($w^s_{i,j}$) and target ($w^t_{k,l}$) language words in the input sentences followed by mining translation pairs (D_m) and making their representation more similar at the end. We keep the core LM parameters frozen (❄) and train only the linear layers on top of it. The process is repeated for multiple iterations.

source and target language word pairs in the mining step. Mined pairs are then used as training samples in the update step, where the training objective is to make their representations more similar. We iterate the mining and update steps until convergence. Even though our approach does not require parallel sentences, we show that it is possible to mine useful word pairs from sentences with similar contexts. We depict our approach in Figure 1 and detail the two main steps in the following.

3.1 Word Pair Mining

For the word translation pair extraction we assume that we have monolingual corpora for the source and target languages, D_s and D_t. We build contextualized cross-lingual word representations for each token in each sentence of both corpora by taking the corresponding output vectors of the used model's final layer. As explained in Section 3.2 we apply a linear layer on top of the used LM for alignment purposes. We take the output of this linear layer as CCWR instead of the average of multiple LM layers in contrast to previous work (Vulić et al., 2020) in order to directly benefit from model updates as our method progresses. If a word is split into multiple subwords by the LM's tokenizer we take the representation of the last subword token based on the findings of (Ács et al., 2021).[2] For each token $w^s_{i,j} \in D_s$ we look for the most similar token $w^t_{k,l} \in D_t$ by calculating:

$$w^t_{k,l} = \underset{w^t_{m,n} \in D_t}{\arg\max}\, cos(E_{w^s_{i,j}}, E_{w^t_{m,n}}) \quad (1)$$

where $w^x_{i,j}$ is the j^{th} token of the i^{th} sentence in D_x, $E(w^x_{i,j})$ is its CCWR and cos is the cosine similarity of two vectors. Finally, we filter out low quality word pairs by keeping only those pairs which have a similarity value larger than a given threshold value ($th = 0.2$). We mine word pairs in both language directions, i.e., we mine pairs for each $w^s_{i,j} \in D_s$ and in the reverse direction as well for each $w^t_{k,l} \in D_t$ this way boosting the number of training examples.

Due to the quadratic nature of the mining process, instead of using the full source and target corpora, we randomly sample $1K$ sentences in each iteration from D_s and D_t resulting in $\tilde{D}_s$ and $\tilde{D}_t$ respectively and use them for the word alignment instead of the full corpora. Calculating with 20 tokens per sentence on average means that there are $20K$ source tokens for which we look for the most similar pair in the $20K$ candidate target tokens in each iteration. Furthermore, we keep the mining process simple, fast and memory efficient, i.e., we use cosine similarity instead of CSLS (Conneau et al., 2018a) and do not rely on previously introduced word alignment techniques, such as *SimAlign* (Jalili Sabet et al., 2020). We present further analysis of the mining quality in Section 5.3 and experiments with up to $100K$ sample size instead of $1K$ per iteration and employ the efficient search method of *Faiss* (Johnson et al., 2019) in Appendix C.

3.2 Model Training

Motivated by mapping approaches (Mikolov et al., 2013b; Schuster et al., 2019) we use a dedicated linear layer on top of the pre-trained LM for both source and target languages in the given language pair to improve cross-lingual quality.[3] More precisely, we feed the output representations of the

[2] We ignore special tokens, such as `[CLS]` or `[SEP]`.

[3] We use a layer not only for the source language but for the target as well, since we found it to be beneficial.

LM's last hidden layer to the dedicated linear layer depending on the language of the input word, which has input and output sizes matching the LM's hidden size and has no bias term. Since we want to exploit the improved cross-linguality of the model in the mining step in each iteration, we initialize the dedicated linear layers with the identity matrix[4] and use the output of the linear layers as CCWRs for word pair mining. Given the mined word pairs we update the parameters of the dedicated linear layers while keeping the parameters of the LM frozen by minimizing the following loss function based on the work of Cao et al. (2020):

$$\mathcal{L} = \sum_{(w^s_{i,j}, w^t_{k,l}) \in D_m} \left\| E_{w^s_{i,j}} - E_{w^t_{k,l}} \right\|_2^2 \quad (2)$$

where $D_m \subseteq \tilde{D}_s \times \tilde{D}_t$ is the dataset containing the mined word pairs from step 1.

We note that we experimented with the proposed model of Cao et al. (2020), i.e., updating the full model using no linear layer on top of the original LM architecture but it resulted in significant degradation of the model performance when using our mined training samples. The main difference with the original method of Cao et al. (2020) is that they use parallel sentences where most of the words are aligned, while in our case most sentence pairs that are extracted by our method contain only one aligned word pair and the rest are unaligned, thus not used for training. Our conjecture is that due to the smaller information ratio per sentence pair and the smaller number of covered unique words in the training process (see Section 5.3) the full model update is more susceptible to over-fitting, thus the use of language specific linear layers is crucial for our approach. This shows that the simpler linear layer and the static LM is needed to prevent over-fitting and degrading the quality of monolingual language subspaces in case of the mined training data. We summarize our complete iterative method in Algorithm 1.

Algorithm 1 High level pseudo-code of our proposed method.

Require: D_s, D_t the full monolingual corpora of the source and target languages; Θ pre-trained model parameters; N number of update steps; th minimum word similarity threshold

for i *in* $0..N$ **do**

$\tilde{D}_s,\ \tilde{D}_t \leftarrow sample(D_s),\ sample(D_t)$

$D_m \leftarrow mining(\tilde{D}_s,\ \tilde{D}_t,\ \Theta,\ th)$

$\Theta \leftarrow update(D_m,\ \Theta)$

end for

4 Experimental Setup

4.1 Datasets and Model Parameters

We test our proposed model on a wide range of languages: Bengali, Basque, Macedonian, Malayalam, Afrikaans, Swahili, Kannada, Gujarati and Nepali. For each of the languages we used $1M$ randomly selected sentences from the full Wikipedia dumps as monolingual corpora (D_s and D_t) which we tokenized with the *IndicNLP* toolkit (Kakwani et al., 2020) in case of the Indian subcontinent languages[5] or with the *Moses* toolkit (Koehn et al., 2007) in case of the others.[6] We show language and dataset statistics in Table 1. In Appendix D we present simulated low-resource experiments indicating the effectiveness of our approach when only a small amount of monolingual sentences are available.

As the pre-trained LM we use *bert-base-multilingual-cased* (Devlin et al., 2019) which we refer to as mBERT in the following. To try to strengthen the baselines for underrepresented languages we perform vocabulary extension and model fine-tuning using the monolingual data of the target low-resource language. We follow the approach and suggested model parameters of Wang et al. (2020), i.e., we extend mBERT's original vocabulary with the most frequent $10K$ subword tokens of the low-resource language and run $100K$ fine-tuning steps on the full Wikipedia dumps using only the masked language modeling objective. We refer to the vocabulary extended models as eBERT.

For our approach we tune model parameters using the development set of the word retrieval task (see Section 4.2) on Nepali only and use the same parameters for the rest of the languages as well. The used parameters are the following: number of model update steps (N) $5K$ with batch size 2, gradient accumulation steps 6 (which means $60K$ extracted sentence pairs), $1K$ warm-up steps and learning rate 5×10^{-5}. We used the *Huggingface* library for the implementation of our techniques (Wolf et al., 2020). The runtime of our method ranges between 0.5 and 2 hours using a

[4] We add a uniform random noise with value up to 0.01 for better training performance.

[5] Bengali, Malayalam, Kannada, Gujarati and Nepali.

[6] The tokenization of monolingual corpora is only required for the word pair mining process.

	ISO	#Wiki	Para. corp.	#Para.	#NER
Bengali	bn	959	Bangla-NMT	300	10
Basque	eu	807	OpenSubtitles	805	10
Macedonian	mk	515	OpenSubtitles	3,399	10
Malayalam	ml	481	Samanantar	5,774	10
Afrikaans	af	369	OpenSubtitles + Bible	107	5
Swahili	sw	141	GoURMET	155	1
Kannada	kn	128	Samanantar	4,011	0.1
Gujarati	gu	115	Samanantar	3,017	0.1
Nepali	ne	102	Bible	68	0.1

Table 1: Language and dataset statistics. From left to right we indicate language ISO code, total number of Wikipedia articles in thousands (K), used parallel corpora, the number of parallel sentence pairs (K) and the number of WikiANN training sentences (K) for NER.

single GeForce GTX 1080 Ti.

4.2 Evaluation

Contextual cross-lingual word retrieval As an intrinsic evaluation of cross-lingual quality of LMs we perform the word retrieval task as defined by Cao et al. (2020). Given a word aligned parallel test corpus for the source and target languages the task is for each aligned token in the source corpus to retrieve its pair, i.e., the target word it is aligned to, given all tokens in the target corpus. Similarly as in our proposed mining process, we follow Equation 1 for word retrieval with the only exception that we use CSLS[7] instead of cosine as the vector similarity function, since its longer runtime and larger memory footprint is not an issue for the evaluation. Note that since a given word type is contained in multiple sentences on both the source and target language sides, thus it has multiple CCWRs, Equation 1 implicitly involves retrieving the parallel sentence pair of the source sentence and the aligned word pair of the source word just by measuring CSLS similarity of CCWRs. Accuracy is measured by the ratio of correctly retrieved aligned word pairs. We measure accuracy in both source to target and target to source directions and report their average.

In our experiments we consider English as the target language and the already mentioned low-resource languages as the source and use various parallel corpora including the corpora of Bangla-NMT (Hasan et al., 2020), OpenSubtitles (Lison and Tiedemann, 2016), the Samanantar (Ramesh et al., 2021) and GoURMET (Sánchez-Martínez et al., 2020) projects, as well as the Bible (Christodouloupoulos and Steedman, 2015). More details are shown in Table 1. We use the first 1024 sentence pairs as test. We reserved the next 1024 as development (recall that we only used the Nepali development data as discussed above). We used the rest only for training the supervised baseline models (Cao et al., 2020). We tokenized the datasets with the tools discussed in Section 4.1, performed word alignment using *fastAlign* (Dyer et al., 2013) and kept only the one-to-one pairs in the intersection, in order to obtain a high quality test set.

Named entity recognition To test the usefulness of our approach on downstream tasks as well we perform zero-shot cross-lingual NER, since it was shown to reflect the cross-lingual quality of LMs well (Wang et al., 2020). We use the multilingual WikiANN dataset which supports a large set of languages (Rahimi et al., 2019). We keep our sequence tagger simple so that the quality of token representations is the most influential factor in the final results. We only apply a dropout (probability 0.1) and a single linear layer as the classification head. Additionally, we freeze all model parameters except the final classifier layer during training. We train our models on English with batch size 32, learning rate 5×10^{-5}, warm-up steps $1K$, using early stopping on the development set of the target language. We report F_1 scores as our final results.

Compared systems We compare our approach to the off-the-shelf *mBERT* model (Devlin et al., 2019) and additionally to the vocabulary extended *eBERT* (Wang et al., 2020) model for languages not written using Latin or Cyrillic scripts. Additionally we evaluated the supervised model of Cao et al. (2020) which fine-tunes the whole model using the training portion of the parallel corpora (*mBERT_full_sup* and *eBERT_full_sup*). To test the effectiveness of using just a linear layer for language alignment instead of full model update in the supervised setup we run experiments with our proposed architecture but with parallel data as training instead of running the unsupervised mining step (*mBERT_linear_sup* and *eBERT_linear_sup*). Finally, we refer to our systems as *mBERT_linear_unsup* and *eBERT_linear_unsup*. As mentioned in Section 3.2 *mBERT_full_unsup* and *eBERT_full_unsup* (full model update with mined training data) did not converge, thus we omit it from our final results. Due to environmental considerations we report the results of a single run for each setup.

On top of pure BERT-based models in case of

[7]CSLS neighborhood size: 10.

	Bn	Eu	Mk	Ml	Af	Sw	Kn	Gu	Ne
mBERT	14.31	6.84	9.17	6.90	10.24	6.37	9.51	5.90	2.28
mBERT_linear_unsup	**24.12**	**10.55**	**12.83**	**12.04**	**13.82**	**13.49**	**13.76**	12.05	4.64
mBERT_full_sup	42.80	23.68	26.53	46.34	16.21	7.83	50.47	49.90	12.45
mBERT_linear_sup	30.98	13.46	18.17	17.53	18.50	19.06	22.53	17.42	8.81
eBERT	13.73	-	-	5.46	-	-	5.56	7.01	4.92
eBERT_linear_unsup	20.11	-	-	9.44	-	-	10.56	**12.25**	**7.16**
eBERT_full_sup	36.36	-	-	39.69	-	-	44.60	46.18	12.21
eBERT_linear_sup	26.48	-	-	12.48	-	-	13.64	15.49	11.10

Table 2: Accuracy (%) scores of the contextual cross-lingual word retrieval task on the low-resource languages. Languages are sorted based on their Wikipedia sizes in descending order. We do not test the vocabulary extended eBERT on languages with Latin or Cyrillic scripts. Our proposed method is listed as mBERT_linear_unsup and *eBERT_linear_unsup*. Best results ignoring the supervised methods are indicated with bold fonts.

the NER task, we experiment with aligning BERT representation using *VecMap* (Artetxe et al., 2018) similarly to Schuster et al. (2019) and Liu et al. (2019). More precisely, i) we build type-level representations (anchors) for each of the most frequent $50K$ words of a given language. We randomly sample 100 sentences[8] containing a given word (w), build its CCWRs (E_w) and take their dimension-wise average as the type-level representation. Then ii) we train an orthogonal alignment of the source and target language type-level embedding spaces with VecMap using iterative-refinement. We use two types of training signals: identical word pairs (*mBERT_vecmap_id* and *eBERT_vecmap_id*) which similarly to our approach does not need explicit cross-lingual resources; and word translation pairs from the MUSE project (Conneau et al., 2018a) as the supervised setup (*mBERT_vecmap_sup* and *eBERT_vecmap_sup*). Finally, iii) we initialize the weights of the linear layer[9] on top of BERT with the learned alignment before NER training to transfer it to the downstream task.

5 Results

5.1 Contextual Word Retrieval

We show accuracy results of the word retrieval task on the low-resource languages in Table 2. On a high level it can be seen that both baseline models, *mBERT* and *eBERT*, were improved by all of the used methods for each of the languages.

Additionally, LMs updated with our unsupervised mining method (*mBERT_linear_unsup* and *eBERT_linear_unsup*) show large improvements compared to the baselines although no parallel data was used. This shows that useful word translation pairs can be mined automatically relying on the initial cross-lingual quality of multilingual LMs and that parallel data is not necessary. We show mined examples in Section 5.3.

We built vocabulary extended *mBERT* models (*eBERT*) for languages that used a script other than Latin or Cyrillic. *eBERT* is effective on the lower resourced languages (Gu and Ne). Still our mining approach improves over *eBERT* achieving best scores on Gu and Ne while *mBERT_linear_unsup* achieves best scores on the others.

The supervised approaches, which can be considered as oracle systems since they assume the availability of good quality parallel data, achieve large improvements over our unsupervised approach. This is not surprising, since these approaches do not rely on the initial cross-lingual quality of the used LM which is low for the considered languages. Additionally, the aligned word pairs in the used parallel data cover a larger portion of the given language's vocabulary which means more information for the training process. In contrast, the mining process covers less unique word pairs (see Section 5.3). Among the two variations the fully fine-tuned model (*mBERT_full_sup* and *eBERT_full_sup*) achieves best performance, while the supervised model with linear layer (*mBERT_linear_sup* and *eBERT_linear_sup*) lies between our proposed unsupervised models and the supervised fully fine-tuned models. As mentioned before, the information density of parallel sentences is much higher than that of the mined pairs, since most of the words are aligned in case of the former, while in most of the cases only one word pair is contained per sentence pair in case of the latter. This shows that

[8] Schuster et al. (2019) and Liu et al. (2019) used 1000 sentences for type-level representations, however we found 100 to be similarly effective but much faster.

[9] The linear layer of the source language and use identity for the target.

	Bn	Eu	Mk	Ml	Af	Sw	Kn	Gu	Ne
mBERT	40.45	35.49	46.80	31.36	**50.28**	37.95	33.77	17.39	27.97
mBERT_vecmap_id	34.36	33.68	44.97	25.65	43.04	27.51	14.85	5.89	11.88
mBERT_linear_unsup	**43.57**	**40.91**	**49.94**	**34.04**	47.93	**38.15**	**33.87**	24.20	28.26
mBERT_vecmap_sup	29.72	35.73	42.56	-	42.61	-	-	-	-
mBERT_full_sup	59.07	49.24	59.39	47.32	57.64	31.46	49.78	50.53	31.75
mBERT_linear_sup	42.90	39.89	47.82	36.76	52.48	41.36	38.33	15.89	28.44
eBERT	36.50	-	-	29.68	-	-	30.30	29.02	38.13
eBERT_vecmap_id	36.29	-	-	29.52	-	-	30.94	29.02	40.40
eBERT_linear_unsup	34.47	-	-	32.64	-	-	30.33	**31.82**	**43.75**
eBERT_vecmap_sup	36.29	-	-	-	-	-	-	-	-
eBERT_full_sup	51.29	-	-	47.51	-	-	45.11	56.29	43.18
eBERT_linear_sup	38.25	-	-	34.85	-	-	36.02	34.58	39.69

Table 3: F_1 scores of zeros-shot cross-lingual named entity recognition. Models are trained on English and evaluated on the low-resource languages. Some supervised VecMap results are not presented due to lack of MUSE training dictionaries. Best results ignoring the supervised methods are indicated with bold fonts.

parallel data can be exploited better by updating the full LM, while the simpler linear layer and frozen LM parameters are needed to prevent over-fitting in case of the mined training data.

5.2 Named Entity Recognition

We show F_1 scores of our zero-shot cross-lingual experiments in Table 3. Similarly to the contextualized word retrieval task the best scores were achieved by fine-tuning the baseline models. Following the trend in the contextual word retrieval results, eBERT is effective in case of the lower resourced languages (Gu and Ne) and the supervised methods using a strong cross-lingual signal in the form of a word aligned parallel data achieve best results. However, the latter is only applicable if parallel data exists for the low-resource language. Our mining based approach improved on the baselines on all languages except Afrikaans.

Similarly to Cao et al. (2020), we found the mapping based approach (VecMap) to be ineffective. *mBERT_vecmap_id* which uses no explicit cross-lingual training signal to learn the alignment, only identical word pairs, achieves lower performance than *mBERT* especially in case of the lower-resource spectrum (Kn, Gu and Ne). In contrast, *eBERT_vecmap_id* is competitive with *eBERT* and even outperforms it on Kannada and Nepali. This shows that extending the vocabulary of *mBERT* is an important step to make VecMap based on identical pairs effective even on the higher-resource languages. On the other hand, our unsupervised mining approach achieves better scores than VecMap with the exceptions of Bengali and Kannada where *eBERT_vecmap_id* performs better than *eBERT_linear_unsup*, however *mBERT_linear_unsup* is the most effective in these cases. Finally, it can be seen that the supervised variations of the VecMap based approach (*mBERT_vecmap_sup* and *eBERT_vecmap_sup*) are not able to benefit further from the stronger cross-lingual training signal. This is in line with the findings of previous work which show identical pairs often to be competitive with dictionaries (Artetxe et al., 2017; Søgaard et al., 2018; Severini et al., 2022).

We present our preliminary results on unseen languages, i.e., languages which were not used for LM pre-training, in Appendix A.

5.3 Analysis

Mined word pairs In Table 4 we show English to Macedonian mined word pair examples. It can be seen that the sentence pairs selected by our method are indeed not parallel but their contexts are similar. In example 1 both sentences mention Eastern European armies, example 2 discusses controversial events, while example 3 mentions various English and German well-known individuals. The similar context of the sentences helps to build similar CCWRs for the selected word pairs. As we mentioned before, most of the mined sentence pairs contain only one aligned word pair as in example 1, however there are a few sentence pairs (about 25%) containing multiple word pairs, such as example 2. Finally, example 3 shows a mined word pair (identity – име (name)) which is not a correct translation, however the words as used have similar meanings. Such examples indicate that our approach is able to leverage word pairs with similar meanings as well.

1	SRC	Immediately thereafter, Ceaușescu presided over the CPEx (Political Executive Committee) meeting and assumed the leadership of the **army**.
	TRG	Веројатно преценето но несомнено византиската **војска** била значително поголема од бугарската.
	TR	*Probably overestimated but undoubtedly the Byzantine **army** was significantly larger than the Bulgarian.*
2	SRC	Although the probing effect **can**[1] **be**[2] controversial when it comes to explaining just why it happens, researchers attempt to explain through the sender behavioral adaptation.
	TRG	Понекогаш, разликата помеѓу умерено-планински и планински ја одредуваат самите организатори и самата поделба **може**[1] да **биде**[2] контроверзна.
	TR	*Sometimes, the difference between temperate and mountainous is determined by the organizers themselves and the division itself **can**[1] **be**[2] controversial.*
3	SRC	It is a consequence of the Philip Hall and Ernst Witt's eponymous **identity**.
	TRG	Албрехт Петар Кан е еден од најплодните и читани писатели на германски јазик, иако само мал број луѓе го познаваа неговото вистинско **име**.
	TR	*Albrecht Peter Kahn is one of the most prolific and well-read German-language writers, although only a handful of people knew his real **name**.*

Table 4: Mining examples of the English to Macedonian mining direction by *mBERT_linear_unsup*. Mined word pairs are bolded. Translations (TR) are provided using Google Translate.

	Bn	Eu	Mk	Ml	Af	Sw	Kn	Gu	Ne
bestBERT	40.45	35.49	46.80	31.36	50.28	37.95	33.77	29.02	38.13
intersection	+2.03	+5.38	+2.57	+2.16	**-1.45**	**+2.14**	-1.29	-1.57	+2.60
forward-backward	**+3.13**	**+5.42**	**+3.15**	**+2.68**	-2.35	+0.20	**+0.10**	**+2.80**	**+5.62**

Table 5: Relative F_1 change on the named entity recognition task using *intersection* (quality) or *forward-backward* (quantity) mining methods compared to the best performing baseline model on each language respectively (bestBERT).

As discussed in Section 4.1 the complete training process extracts $60K$ sentence pairs due to the fixed batch size, gradient accumulation number and model update steps. The numbers of unique mined word types however are relatively small. For English it varies between $11K$ and $15K$, while for the low-resource languages it varies between $8K$ and $14K$. Covered word types mainly involve frequent words of the given language's vocabulary. When a given language has less monolingual resources the number of words having good quality vector representation decreases as well due to the lower frequency of these words (making the pairing of the word difficult). However, the number of mined word types is still larger than the frequently used $5K$ pairs for mapping approaches (Conneau et al., 2018a). We show the exact number of unique words per language mined by our method in Table 9, Appendix B.

Quality vs. Quantity We compared two variations of our word pair mining method in Table 5. We call our method which was discussed previously *forward-backward*, since it mines word pairs in the source to target and target to source language directions using the method discussed in Section 3.1. We simply take the union of the output of the two directions as the final set of mined word pairs. In contrast, in order to increase the quality of word pairs at the expense of quantity we take mutually aligned word pairs in the two directions. We call this approach the *intersection* mining method.

The results in Table 5 shows that the larger but less precise set of mined word pairs resulted by the *forward-backward* method outperforms the smaller but more precise set of *intersection* in 7 out of 9 cases on the named entity recognition task. This shows that the alignment method is robust against incorrectly aligned pairs and can successfully leverage pairs that are not direct translations but are nevertheless similar in meaning. This also leads to a larger number of mined unique word types.

6 Conclusions

Low-resource languages are underrepresented in pre-trained multilingual LMs. In contrast to previous work using a parallel dataset to improve the language support of low-resource languages, we presented an unsupervised method to align language pairs by relying only on monolingual corpora. We showed that word translation pairs can be extracted from non-parallel sentence pairs by leveraging the cross-lingual contextualized representations of words which in turn can be used to align the vector spaces of languages. We tested our

approach on the intrinsic contextual word retrieval and the downstream named entity recognition tasks. Our results showed improved cross-lingual quality of the fine-tuned LMs. Our analysis revealed that the quantity of mined word pairs matters over their quality and that the vocabulary extension method is important for performance boost in case of the lowest resource languages. As future work we aim at leveraging easily accessible cross-lingual resources for better unseen language support.

Limitations

Our preliminary experiments on unseen languages in Appendix A show that our approach improves the baselines for these languages as well but the achieved performance is still low. The main reason for this is the initial low quality of mBERT and eBERT, thus the mining using poor CCWEs is ineffective. The supervised experiments using parallel sentences show that with a larger quantity and more precisely aligned word pairs further improvements can be achieved. However, since such resources are often unavailable for low-resource languages further methods relying on more easily accessible cross-lingual resources should be considered in future work.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback. The work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (No. 640550) and by the German Research Foundation (DFG; grant FR 2829/4-1).

References

Judit Ács, Ákos Kádár, and Andras Kornai. 2021. Subword pooling makes a difference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2284–2295, Online. Association for Computational Linguistics.

Hanan Aldarmaki and Mona Diab. 2019. Context-aware cross-lingual mapping. 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 3906–3911, Minneapolis, Minnesota. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

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.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual Alignment of Contextual Word Representations. In *International Conference on Learning Representations*.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment. 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 3418–3430, Online. Association for Computational Linguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. 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. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word Translation Without Parallel Data. In *Proceedings of the International Conference on Learning Representations*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, 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.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2022. Combining static and contextualised multilingual embeddings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2316–2329, Dublin, Ireland. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. 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.

Shoaib Jameel and Steven Schockaert. 2016. D-GloVe: A feasible least squares model for estimating word embedding densities. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1849–1860, Osaka, Japan. The COLING 2016 Organizing Committee.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In *International Conference on Learning Representations*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

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 Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Qianchu Liu, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2019. Investigating cross-lingual alignment

methods for contextualized embeddings with token-level evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 33–43, Hong Kong, China. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceeding of the 1st International Conference on Learning Representations*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting Similarities among Languages for Machine Translation. *CoRR*, abs/1309.4.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. 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 3118–3135, Online. Association for Computational Linguistics.

Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Mikel L. Forcada, Miquel Esplà-Gomis, Andrew Secker, Susie Coleman, and Julie Wall. 2020. An English-Swahili parallel corpus and its use for neural machine translation in the news domain. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 299–308, Lisboa, Portugal. European Association for Machine Translation.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Silvia Severini, Viktor Hangya, Masoud Jalili Sabet, Alexander Fraser, and Hinrich Schütze. 2022. Don't Forget Cheap Training Signals Before Building Unsupervised Bilingual Word Embeddings. In *Proceedings of the 15th Workshop on Building and Using Comparable Corpora*, pages 15–22.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727, Hong Kong, China. Association for Computational Linguistics.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020a. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020b. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

A Unseen languages

We ran initial experiments on unseen languages (Sindhi, Faroese, Upper Sorbian and Maori) that were not used for pre-training mBERT. Other than the mentioned parallel datasets in Table 1 we used Tanzil (Tiedemann, 2012) for Sindhi. There are no available parallel corpora aligned with English for Faroese and Upper Sorbian. Dataset statistics can be seen in Table 6. We follow the same data preprocessing (we use *IndicNLP* for Sindhi and *Moses* for the others) and use the same hyper-parameters as for the seen languages.

The results in Table 7 show that *mBERT* performs below 1% accuracy on the two unseen languages in terms of word retrieval which is improved by our mining approach. Vocabulary extension is an important step as *eBERT* performs better than *mBERT_linear_unsup* on Sindhi. The updated *eBERT_linear_unsup* model achieves further improvements compared to *eBERT*. However, these results are still low and further improvements are needed. The F_1 scores on the named entity recognition task shown in Table 8 are inconsistent. Our approach achieves improvements on Sindhi and Faroese but not on Upper Sorbian and Maori, although it achieves improvements on all unseen languages in terms of word retrieval accuracy (we do not have Faroese and Upper Sorbian retrieval results due to the lack of parallel data with English). Similarly as in case of the seen languages, the VecMap based approach performs lower than the baseline when no vocabulary extension is performed. In contrast, *eBERT_vecmap_id* outperforms *eBERT* but not *eBERT_linear_unsup*.

	ISO	#Wiki	Para. corp.	#Para.	#NER
Sindhi	sd	56	Tanzil	93	0.1
Faroese	fo	40	-	-	0.1
Upper Sorbian	hsb	35	-	-	0.1
Maori	mi	13	Bible	61	0.1

Table 6: Language and dataset statistics. From left to right we indicate language ISO code, total number of Wikipedia articles in thousands (K), used parallel corpora, the number of parallel sentence pairs (K) and the number of WikiANN training sentences (K) for NER.

	Sd	Mi
mBERT	0.09	0.17
mBERT_linear_unsup	0.16	**0.38**
mBERT_full_sup	6.75	12.68
mBERT_linear_sup	0.54	1.21
eBERT	0.66	-
eBERT_linear_unsup	**1.32**	-
eBERT_full_sup	6.07	-
eBERT_linear_sup	2.64	-

Table 7: Accuracy ($\%$) scores of the contextual cross-lingual word retrieval task on the low-resource unseen languages. Languages are sorted based on their Wikipedia sizes in descending order. We do not test the vocabulary extended eBERT on languages with Latin or Cyrillic scripts. Faroese and Upper Sorbian evaluation is omitted because we do not have access to Fo-En and Hsb-En parallel corpora. Our proposed method is listed as mBERT_linear_unsup and *eBERT_linear_unsup*. Best results ignoring the supervised methods are indicated with bold fonts.

The supervised methods using parallel corpora for training show that with good quality word pairs even unseen languages can be improved significantly, indicating that in case of low initial LM quality on a given language a stronger cross-lingual signal is needed for meaningful model improvement. On the other hand, parallel datasets are not available for many low-resource languages.

Finally, Table 9 shows that the number of mined unique words for the unseen languages are lower than the numbers for seen languages, ranging between $5K$ and $10K$. Due to smaller monolingual corpora word types are less frequent and there is stronger oversplitting by the subword tokenizer, leading to fewer words with high quality CCWRs.

B Mined Words

We show the number of unique word types that were mined by our approach for both seen and unseen languages in Table 9.

C Size of Mining Candidates

In our main experiments we use cosine similarity for word pair mining and randomly sample $1K$ sentences as the source and target datasets ($\tilde{D}_s$ and $\tilde{D}_t$) in each iteration. We experimented further with larger sampling sizes and a more efficient CCWR similarity method on top of cosine which, although quadratic in runtime and memory requirements, can be efficiently performed on GPUs using batching. The runtime of our cosine-based approach on

	Sd	Fo	Hsb	Mi
mBERT	8.57	50.97	**39.01**	**26.39**
mBERT_vecmap_id	3.24	28.13	23.26	3.56
mBERT_linear_unsup	7.94	**52.17**	31.87	11.69
mBERT_vecmap_sup	-	-	-	-
mBERT_full_sup	13.53	-	-	35.60
mBERT_linear_sup	4.95	-	-	32.93
eBERT	21.86	-	-	-
eBERT_vecmap_id	23.98	-	-	-
eBERT_linear_unsup	**25.71**	-	-	-
eBERT_vecmap_sup	-	-	-	-
eBERT_full_sup	26.40	-	-	-
eBERT_linear_sup	24.72	-	-	-

Table 8: F_1 scores of zeros-shot cross-lingual named entity recognition. Models are trained on English and evaluated on the low-resource languages. Supervised scores for Fo and Hsb are missing due to lack of parallel training data. Additionally, supervised VecMap results are missing due to lack of MUSE training dictionaries. Best results ignoring the supervised methods are indicated with bold fonts.

		En	Low-res.
Seen	Bn	15,257	11,915
	Eu	14,896	11,721
	Mk	11,295	8,514
	Ml	15,241	12,970
	Af	11,291	8,038
	Sw	14,717	10,321
	Kn	14,970	13,941
	Gu	15,123	11,196
	Ne	15,110	11,524
Unseen	Sd	15,249	9,946
	Fo	11,293	7,272
	Hsb	11,350	6,418
	Mi	14,858	5,447

Table 9: Number of English and low-resource language vocabulary entries mined by *mBERT_linear_unsup*.

Nepali ranges between only 0.5 ($|\tilde{D}_t| = 1K$) and 15 ($|\tilde{D}_t| = 100K$) hours. Since the main goal of using larger data samples is to increase the chance of including the translation of a given source word in the candidate set, we keep the size of $\tilde{D}_s$ at $1K$ (since we sample a new set in every iteration) and only vary the size of $\tilde{D}_t$. As the alternative search method we use *Faiss*, a library for efficient vector similarity calculation (Johnson et al., 2019). More precisely, we use an inverted file index with 100 lists for faster search and product quantizer with 8 bits encoding for memory efficiency.[10] The runtime of the Faiss based system ranges between 20 minutes and 1.5 hours.

Table 10 shows our findings on Nepali with *eBERT_linear_unsup*. It can be seen that as the size of candidates increases the performance slightly decreases in case of cosine and stays on the same level in case of Faiss. As expected, we found that the setups with larger target candidate sizes mine more pairs per iteration, thus our conjecture is that it is better to update the model with a few examples in the initial training steps which positively influences the mining quality in the later stages. Secondly, cosine similarity outperforms Faiss due to its exact search in exchange for computational efficiency.

[10]Faiss index factory code: `IVF100,PQ8`.

$\lvert\tilde{D}_t\rvert$	cosine	Faiss
1	7.16	6.50
5	7.15	6.45
10	7.10	6.49
50	6.99	6.46
100	6.97	6.51

Table 10: Comparing setups on Nepali contextual cross-lingual word retrieval (%) using *eBERT_linear_unsup* with different numbers of sampled target language sentences (in thousands) as candidates for mining using exact match with cosine similarity or runtime/memory optimized Faiss search.

$\lvert D_s\rvert$	Ne
5	7.19
10	7.27
50	7.21
100	7.14
500	7.18
1,000	7.16

Table 11: Comparing simulated low-resource setups on Nepali contextual cross-lingual word retrieval (%) using *eBERT_linear_unsup*. We test how our approach performs when only a small monolingual corpus (size given in thousands) is available for a given source language (Nepali).

D Size of Monolingual Data

We also run simulated low-resource experiments where we assume that the size of the overall monolingual corpus (D_s), from which we sample $1K$ sentences in each iteration ($\tilde{D}_s$), is limited. Table 11 shows the results on simulated Nepali contextual cross-lingual word retrieval. Recall that in our main experiments we used $1,000K$ sentences for both D_s and D_t. It can be seen that there is a negligible difference between the different dataset sizes, showing the robustness of our approach against lack of

data for mining. As discussed in Section 4.1 the complete training process extracts $60K$ sentence pairs due to the fixed batch size, gradient accumulation number and model update steps, thus even the lowest setup has enough possible sentence pairs to mine from ($5K \times 1,000K$).