Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation

Anonymous ACL submission

Abstract

The performance of multilingual pretrained models is highly dependent on the availability of monolingual or parallel text present in a target language. Thus, the majority of the world's languages cannot benefit from recent progress in NLP as they have no or limited textual data. To expand possibilities of using NLP technology in these under-represented languages, we systematically study strategies that relax the reliance on conventional language resources through the use of bilingual lexicons, an alternative resource with much better language coverage. We analyze different strategies to synthesize textual or labeled data using lexicons, and how this data can be combined with monolingual or parallel text when available. For 19 under-represented languages across 3 tasks, our methods lead to consistent improvements of up to 5 and 15 points with and without extra monolingual text respectively. Overall, our study highlights how NLP methods can be adapted to thousands more languages that are under-served by current technology.[1]

1 Introduction

Multilingual pretrained models (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020) have become an essential method for cross-lingual transfer on a variety of NLP tasks (Pires et al., 2019; Wu and Dredze, 2019). These models can be finetuned on annotated data of a down-stream task in a high-resource language, often English, and then the resulting model is applied to other languages. This paradigm is supposed to benefit under-represented languages that do not have annotated data. However, recent studies have found that the cross-lingual transfer performance of a language is highly contingent on the availability of monolingual data in the language during pretraining (Hu et al., 2020). Languages with more monolingual data tend to have better performance while languages not present during pretraining significantly lag behind.

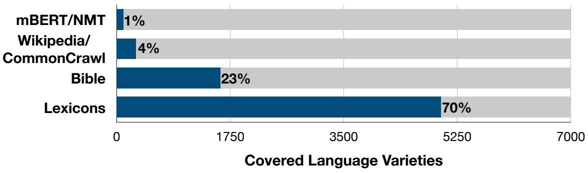

Figure 1: The percentage of the world's ≈7,000 languages covered by mBERT, monolingual data sources and lexicons.

Several works propose methods to adapt the pretrained multilingual models to low-resource languages, but these generally involve continued training using monolingual text from these languages (Wang et al., 2020; Chau et al., 2020; Pfeiffer et al., 2020, 2021). Therefore, the performance of these methods is still constrained by the amount of monolingual or parallel text available, making it difficult for languages with little or no textual data to benefit from the progress in pretrained models. Joshi et al. (2020) indeed argue that unsupervised pretraining makes the 'resource-poor poorer'.

Fig. 1 plots the language coverage of multilingual BERT (mBERT; Devlin et al., 2019), a widely used pre-trained model, and several commonly used textual data sources.[2] Among the 7,000 languages in the world, mBERT only covers about 1% of the languages while Wikipedia and CommonCrawl, the two most common resources used for pretraining and adaptation, only contain textual data from 4% of the languages (often in quite small quantities, partially because language IDs are difficult to obtain for low-resource languages (Caswell et al., 2020)). Ebrahimi and Kann (2021) show that continued pretraining of multilingual models on a small amount of Bible data can significantly improve the performance of uncovered languages. Although the Bible has much better language coverage of 23%, its relatively small data size and

[1] Code and data to reproduce experiments will be released.

[2] Statistics taken from Ebrahimi and Kann (2021) and `panlex.org`.

constrained domain limits its utility (see § 6)—and 70% of the world's languages do not even have this resource. The failure of technology to adapt to these situations raises grave concerns regarding the fairness of allocation of any benefit that may be conferred by NLP to speakers of these languages (Joshi et al., 2020; Blasi et al., 2021). On the other hand, linguists have been studying and documenting under-represented languages for years in a variety of formats (Gippert et al., 2006). Among these, bilingual lexicons or word lists are usually one of the first products of language documentation, and thus have much better coverage of the worlds' languages than easily accessible monolingual text, as shown in Fig. 1. There are also ongoing efforts to create these word lists for even more languages through methodologies such as "rapid word collection" (Boerger, 2017), which can create an extensive lexicon for a new language in a number of days. As Bird (2020) notes:

> After centuries of colonisation, missionary endeavours, and linguistic fieldwork, all languages have been identified and classified. There is always a wordlist. . . . In short, we do not need to "discover" the language ex nihilo (L1 acquisition) but to leverage the available resources (L2 acquisition).

However, there are few efforts on understanding the best strategy to utilize this valuable resource for adapting pretrained language models. Bilingual lexicons have been used to synthesize bilingual data for learning cross-lingual word embeddings (Gouws and Søgaard, 2015; Ruder et al., 2019) and task data for NER via word-to-word translation (Mayhew et al., 2017), but both approaches precede the adoption of pre-trained multilingual LMs. Khemchandani et al. (2021) use lexicons to synthesize monolingual data for adapting LMs, but their experimentation is limited to several Indian languages and no attempt was made to synthesize downstream task data.

In this paper, we conduct a systematic study of strategies to leverage this relatively under-studied resource of bilingual lexicons to adapt pretrained multilingual models to languages with little or no monolingual data. Utilizing lexicons from an open-source database, we create synthetic data for both continued pretraining and downstream task fine-tuning via word-to-word translation. Empirical results on 19 under-represented languages on 3 different tasks demonstrate that using synthetic data leads to significant improvements on all tasks (Fig. 2), and that the best strategy depends on the availability of monolingual data (§ 5, § 6). We further investigate methods that improve the quality of the synthetic data through a small amount of parallel data or by model distillation.

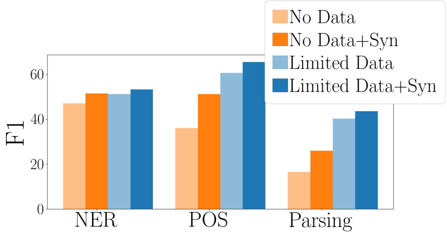

Figure 2: Results for baselines and adaptation using synthetic data for both resource settings across three NLP tasks.

2 Background

We focus on the cross-lingual transfer setting where the goal is to maximize performance on a downstream task in a target language T. Due to the frequent unavailability of labeled data in the target language, a pretrained multilingual model M is typically fine-tuned on labeled data in the downstream task $\mathcal{D}^S_{label} = \{(x^S_i, y^S_i)\}^N_{i=1}$ in a source language S where x^S_i is a textual input, y^S_i is the label, and N is the number of labeled examples. The fine-tuned model is then directly applied to task data $\mathcal{D}^T_{test} = \{x^T_i, y^T_i\}_i$ in language T at test time.[3] The performance on the target language T can often be improved by further adaptation of the pretrained model.

2.1 Adaptation with Text

There are two widely adopted paradigms for adapting pretrained models to a target language using monolingual or parallel text.

MLM Continued pretraining on monolingual text $\mathcal{D}^T_{mono} = \{x^T_i\}_i$ in the target language (Howard and Ruder, 2018; Gururangan et al., 2020) using a masked language model (MLM) objective has proven effective for adapting models to the target language (Pfeiffer et al., 2020). Notably, Ebrahimi and Kann (2021) show that using as little as several thousand sentences can significantly improve the model's performance on target languages not covered during pretraining.

[3] We additionally examine the few-shot setting where some task data $\mathcal{D}^T_{label}$ in T is available for fine-tuning in § 7.

Trans-Train For target languages with sufficient parallel text with the source language $\mathcal{D}^{ST}_{par} = \{(x^S_i, x^T_i)\}_i$, one can train a machine translation (MT) system that translates data from the source language into the target language. Using such an MT system, we can translate the labeled data in the source language $\mathcal{D}^S_{label}$ into target language data $\widehat{\mathcal{D}}^T_{label} = \{(\widehat{x}^T_i, y^S_i)\}^N_{i=1}$, and fine-tune the pretrained multilingual model on both the source and translated labeled data $\mathcal{D}^S_{label} \cup \widehat{\mathcal{D}}^T_{label}$. This method often brings significant gains to the target language, especially for languages with high-quality MT systems (Hu et al., 2020; Ruder et al., 2021).

2.2 Challenges with Low-resource Languages

Both methods above require $\mathcal{D}^T_{mono}$ or $\mathcal{D}^{ST}_{par}$ in target language T, so they cannot be directly extended to languages without this variety of data. Joshi et al. (2020) classified the around 7,000 languages of the world into six groups based on the availability of data in each language. The two groups posing the biggest challenges for NLP are:

"The Left-Behinds," languages with virtually no unlabeled data. We refer to this as the *No-Text* setting.

"The Scraping-Bys," languages with a small amount of monolingual data. We refer to this as the *Few-Text* setting.

These languages make up 85% of languages in the world, yet they do not benefit from the development of pretrained models and adaptation methods due to the lack of monolingual and parallel text. In this paper, we conduct a systematic study of strategies directly targeted at these languages.

3 Adapting to Under-represented Languages Using Lexicons

Since the main bottleneck of adapting to under-represented languages is the lack of text, we adopt a data augmentation framework (illustrated in Fig. 3) that leverages bilingual lexicons, which are available for a much larger number of languages.

3.1 Synthesizing Data Using Lexicons

Given a bilingual lexicon $\mathcal{D}^{ST}_{\text{lex}}$ between the source language S and a target language T, we create synthetic sentences $\widetilde{x}^T_i$ in T using sentences x^S_i in S via word-to-word translation, and use this synthetic data in the following adaptation methods.

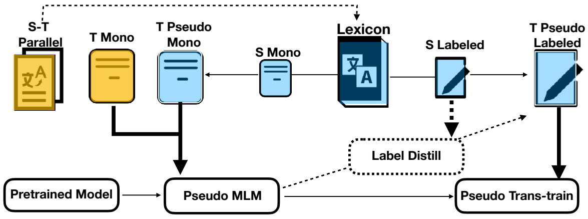

Figure 3: Pipelines for synthesizing data for both No-text and Few-text settings and utilizing extra data for the Few-Text setting. Solid lines indicate adaptation methods and dashed lines are synthetic data refinement methods.

Pseudo MLM Using monolingual text $\mathcal{D}^S_{mono} = \{x^S_i\}_i$, we generate pseudo monolingual text $\widetilde{\mathcal{D}}^T_{mono} = \{\widetilde{x}^T_i\}_i$ for T by replacing the words in x^S_i to its translation in T based on the lexicon $\mathcal{D}^{ST}_{\text{lex}}$. We then adapt the pretrained multilingual model on $\widetilde{\mathcal{D}}^T_{mono}$ using the MLM objective. For the Few-Text setting where some gold monolingual data $\mathcal{D}^T_{mono}$ is available, we can train the model jointly on the pseudo and the gold monolingual data $\widetilde{\mathcal{D}}^T_{mono} \cup \mathcal{D}^T_{mono}$.

Pseudo Trans-train Given the source labeled data $\mathcal{D}^S_{label} = \{(x^S_i, y^S_i)\}^N_{i=1}$, for each text example x^S_i we use $\mathcal{D}^{ST}_{\text{lex}}$ to replace the words in x^S_i with its corresponding translation in T, resulting in pseudo labeled data $\widetilde{\mathcal{D}}^T_{label} = \{(\widetilde{x}^T_i, y^S_i)\}^N_{i=1}$. We then fine-tune the model jointly on both pseudo and gold labeled data $\widetilde{\mathcal{D}}^T_{label} \cup \mathcal{D}^S_{label}$.

Since these methods only require bilingual lexicons, we can apply them to both No-Text and Few-Text settings. We can use either of the two methods or the combination of both to adapt the model.

Challenges with Pseudo Data Our synthetic data $\widetilde{\mathcal{D}}^T$ could be very different from the true data $\mathcal{D}^T$ because the lexicons do not cover all words in S or T, and we do not consider morphological or word order differences between T and S.[4] Nonetheless, we find that this approach yields significant improvements in practice (see Tab. 2). We also outline two strategies that aim to improve the quality of the synthetic data in the next section.

3.2 Refining the Synthetic Data

Label Distillation The pseudo labeled data $\widetilde{\mathcal{D}}^T_{label} = \{(\widetilde{x}^T_i, y^S_i)\}^N_{i=1}$ is noisy because the synthetic examples $\widetilde{x}^T_i$ could have a different label from the original label y^S_i (See Tab. 1). To alleviate

[4] In fact, we considered more sophisticated methods using morphological analyzers and inflectors, but even models with relatively broad coverage (Anastasopoulos and Neubig, 2019) did not cover many languages we used in experiments.

eng $x^S \in \mathcal{D}^S_{mono}$	Anarchism calls for the abolition of the state , which it holds to be undesirable , unnecessary , and harmful .
Pseudo Mono $\widetilde{x}^T \in \widetilde{\mathcal{D}}^T_{mono}$	Anarchism calls gal il abolition ta' il stat , lima hi holds gal tkun undesirable , bla bzonn , u harmful .
eng $x^S \in \mathcal{D}^S_{label}$	I suspect the streets of Baghdad will look as if a war is looming this week .
Pseudo Labeled $\widetilde{x}^T \in \widetilde{\mathcal{D}}^T_{label}$	jien iddubita il streets ta' Bagdad xewqa hares kif jekk a gwerra is looming dan ġimga .
Pseudo Labeled $y^S \in \widetilde{\mathcal{D}}^T_{label}$	PRON VERB DET NOUN ADP PROPN AUX VERB SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT
Label Distilled $\widetilde{y}^T \in \widetilde{\mathcal{D}}^T_{distill}$	PRON VERB DET NOUN ADP PROPN NOUN NOUN SCONJ SCONJ DET NOUN AUX VERB DET NOUN PUNCT

Table 1: Examples of pseudo monolingual data and pseudo labeled data for POS tagging for Maltese (mlt). Words in red have different labels between the source language and the label distilled data. This is because "xewqa" in Maltese is a noun meaning "desire,will", while the word "will" is not used as a noun in the original English sentence.

this issue, we propose to automatically "correct" the labels of pseudo data using a teacher model. Specifically, we fine-tune the pretrained multilingual model as a teacher model using only $\mathcal{D}^S_{label}$. We use this model to generate the new pseudo labeled data $\widetilde{\mathcal{D}}^T_{distill} = \{(\widetilde{x}^T_i, \widetilde{y}^T_i)\}^N_{i=1}$ by predicting labels $\widetilde{y}^T_i$ for the pseudo task examples $\widetilde{x}^T_i$. We then fine-tune the pretrained model on both the new pseudo labeled data and the source labeled data $\widetilde{\mathcal{D}}^T_{distill} \cup \mathcal{D}^S_{label}$.

Induced Lexicons with Parallel Data For the Few-Text setting, we can leverage the available parallel data $\mathcal{D}^{ST}_{par}$ to further improve the quality of the augmented data. Specifically, we use unsupervised word alignment to extract additional word pairs $\widetilde{\mathcal{D}}^{ST}_{lex}$ from the parallel data, and use the combined lexicon $\widetilde{\mathcal{D}}^{ST}_{lex} \cup \mathcal{D}^{ST}_{lex}$ to synthesize the pseudo data.

4 General Experimental Setting

In this section, we outline the tasks and data setting used by all experiments. We will then introduce the adaptation methods and results for the No-Text setting in § 5 and the Few-Text setting in § 6.

4.1 Tasks, Languages and Model

We evaluate on three different tasks with relatively good coverage of under-represented languages: named entity recognition (NER), part-of-speech (POS) tagging, and dependency parsing (DEP). We use two NER datasets: WikiAnn NER (Pan et al., 2017; Rahimi et al., 2019) and MasakhaNER (Adelani et al., 2021). We use the Universal Dependency 2.5 (Nivre et al., 2018) dataset for both the POS and DEP tasks.

We use English as the source language for all experiments. For each dataset, we use the English training data and select the checkpoint with the best performance on the English development set. For MasakhaNER, which does not have English training data, we follow Adelani et al. (2021) and use the CoNLL-2003 English NER training data.

We run each fine-tuning experiment with 3 random seeds and report the average performance. For NER and POS tagging, we follow the data processing and fine-tuning hyper-parameters in Hu et al. (2020). We use the Udify (Kondratyuk and Straka, 2019) codebase and configuration for parsing.

Languages For each task, we select languages that are not covered by the mBERT pretraining data. The list of languages we consider are in § A.1. Most selected languages fall under the Few-Text setting (Joshi et al., 2020). We employ the same set of languages to simulate the No-Text setting because very few languages in this category have suitable test data.

Model We use the multilingual BERT model (mBERT) because it has competitive performance on under-represented languages (Pfeiffer et al., 2020). We find that our mBERT performance on WikiNER and POS is generally comparable or exceeds the XLM-R large results in Ebrahimi and Kann (2021). We additionally verify our results also hold for XLM-R in § 7.

4.2 Adaptation Data

Lexicon We extract lexicons between English and each target language from the PanLex database.[5] The number of lexicon entries varies from about 0.5k to 30k, and most of the lexicons have around 5k entries. The lexicon statistics for each language can be found in § A.1.

Pseudo Monolingual Data English Wikipedia articles are used to synthesize monolingual data. We first tokenize the English articles using Stanza (Qi et al., 2020) and keep the first 200k sentences. To create pseudo monolingual data for a given target language, we replace each English word with its translation if the word exists in the bilingual lexicon. We randomly sample a target word if the English word has multiple possible

[5] https://panlex.org/snapshot/

translations because it is difficult to estimate translation probabilities due to lack of target text.

Pseudo Labeled Data Using the English training data for each task, we simply replace each English word in the labeled training data with its corresponding translation and retain its original label. For the sake of simplicity, we only use lexicon entries with a single word.

5 No-Text Setting

We analyze the results of the following adaptation methods for the setting where we do not have any monolingual data.

Pseudo MLM The mBERT model is trained on the pseudo monolingual data using the MLM objective. We train the model for 5k steps for the NER tasks and 10k steps for the POS tagging and Parsing tasks.

Pseudo Trans-train We fine-tune mBERT or the model adapted with Pseudo MLM for a downstream task on the concatenation of both the English labeled data and the pseudo labeled data.

Label Distillation We use the model adapted with Pseudo MLM as the teacher model to generate new labels for the pseudo labeled data, which we use jointly with the English labeled data to fine-tune the final model.

5.1 Results

The average performance of different adaptation methods averaged across all languages in each task can be found in Tab. 2.

Pseudo Trans-train is the best method for No-Text. Pseudo MLM and Pseudo Trans-train can both bring significant improvements over the mBERT baseline for all tasks. Pseudo Trans-train leads to the best aggregated result across all tasks, and it is also the best method or very close to the best method for each task. Adding Pseudo Trans-train on top of Pseudo MLM does not add much improvement. Label Distillation generally leads to better performance, but overall it is comparable to only using Pseudo Trans-train.

6 Few-Text Setting

We test same adaptation methods introduced in § 5 for the Few-Text setting where we have a small amount of gold data. First we introduce the additional data and adaptation methods for this setting.

6.1 Gold Data

Gold Monolingual Data We use the JHU Bible Corpus (McCarthy et al., 2020) as the monolingual data. Following the setup in Ebrahimi and Kann (2021), we use the verses from the New Testament, which contain 5000 to 8000 sentences for each target language.

Gold Parallel Data We can use the parallel data between English and the target languages from the Bible to extract additional word pairs. We use an existing unsupervised word alignment tool, eflomal (Östling and Tiedemann, 2016), to generate word alignments for each sentence in the parallel Bible data. To create high quality lexicon entries, we only keep the word pairs that are aligned more than once, resulting in about 2k extra word pairs for each language. We then augment the PanLex lexicons with the induced lexicon entries.

6.2 Adaptation Methods

Gold MLM The mBERT model is trained on the gold monolingual Bible data in the target language using the MLM objective. Following the setting in Ebrahimi and Kann (2021), we train for 40 epochs for the NER task, and 80 epochs for the POS and Parsing tasks.

Pseudo MLM We conduct MLM training on both the Bible monolingual data and the pseudo monolingual data in the target language. The Bible data is up-sampled to match the size of the pseudo monolingual data. We train the model for 5k steps for the NER task and 10k steps for the POS tagging and Parsing tasks.

6.3 Results

The average performance in each task for Few-Text can be found in Tab. 2.

Pseudo MLM is the competitive strategy for Few-Text. Unlike the No-Text setting, Pseudo Trans-train only marginally improves or even decreases the performance for three out of the four datasets we consider. On the other hand, Pseudo MLM, which uses both gold and pseudo monolingual data for MLM adaptation, consistently and significantly improves over Gold MLM for all tasks. Again, using Pseudo Trans-train on top of Pseudo MLM does not help and actually leads to relatively large performance loss for the syntactic tasks, such as POS tagging and Parsing.

	Method	Lexicon	WikiNER	Δ	MasakhaNER	Δ	POS	Δ	Parsing	Δ	Avg.	Δ
	mBERT	-	47.6	-	46.1	-	36.1	-	16.5	-	36.5	-
No-Text	Pseudo Trans-train	PanLex	49.8	2.2	54.4	8.3	**51.1**	15.0	25.9	9.4	**45.2**	8.7
	Pseudo MLM	PanLex	49.8	2.2	52.6	6.5	48.9	12.8	25.2	8.7	44.1	7.6
	Both	PanLex	48.5	0.9	**54.6**	8.5	48.7	12.6	25.9	9.4	44.4	7.9
	Both+Label Distillation	PanLex	**50.6**	2.1	53.5	-1.1	50.3	1.6	**26.0**	0.1	45.1	0.7
Few-Text	Gold MLM	-	49.5	-	53.6	-	60.6	-	40.2	-	50.9	-
	Pseudo Trans-train	PanLex	50.2	0.7	59.4	5.8	59.3	-1.3	37.0	-3.2	51.4	0.5
	Pseudo MLM	PanLex	50.7	1.2	57.4	3.8	65.4	4.8	**43.5**	3.3	54.2	3.3
		PanLex+Induced	52.2	1.5	58.5	0.9	64.7	-0.7	41.5	-2.0	54.2	0.0
	Both	PanLex	50.1	0.6	59.2	5.6	60.7	0.1	38.3	-1.9	52.0	1.1
		PanLex+Induced	52.6	2.5	**61.1**	1.9	59.5	-1.2	35.3	-3.0	52.0	0.0
	Both+Label Distillation	PanLex	51.7	1.6	58.4	-0.8	**66.2**	5.5	41.9	3.6	54.5	2.5
		PanLex+Induced	**53.2**	1.5	59.4	1.0	65.8	-0.4	40.7	-1.2	**54.7**	0.2

Table 2: Average F1 score for languages in each task. We record F1 of the LAS for Parsing. We compare three adaptation methods (Δ indicates gains over baselines): Pseudo Trans-train, Pseudo MLM, and Both. We also examine two data refinement methods: Label Distillation (Δ is gains over Both) and PanLex+Induced (Δ is gains over PanLex). **Bold** is the best result for each dataset, and underline indicates the best improvements among the three adaptation methods over the baselines.

Label Distillation brings significant improvements for the two syntactic tasks. Notably, it is the best performing method for POS tagging, but it still lags behind Pseudo MLM for Parsing. This is likely because Parsing is a much harder task than POS tagging to generate correct labels. The effect of Label Distillation on the NER task is less consistent—it improves over Pseudo Trans-train for WikiNER but not for MasakhaNER. This is because the named entity tags of the same words in different languages likely remain the same so that the pseudo task data probably has less noise for Label Distillation to have consistent benefits.

Adding Induced Lexicons We examine the effect of using the lexicons augmented by word pairs induced from the Bible parallel data. The results can be found in Tab. 2. Adding the induced lexicon significantly improves the NER performance, while it hurts the two syntactic tasks.

To understand what might have prevented the syntactic tasks from benefiting from the extra lexicon entries, we plot the distribution of the part-of-speech tags of the words in PanLex lexicons and the lexicons induced from the Bible in Fig. 4. PanLex lexicons have more nouns than the Bible lexicons while the Bible lexicons cover more verbs than PanLex. However, the higher verb coverage in induced lexicons actually leads to a larger prediction accuracy drop for verbs in the POS tagging task. We hypothesize that the pseudo monolingual data created using the induced lexicons would contain more target language verbs with the wrong word order, which could be more harmful for syntactic tasks than tasks that are less sensitive to word order such as NER.

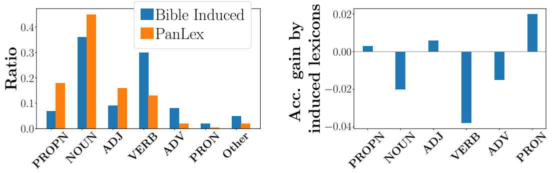

Figure 4: *left*: Ratio of words with different POS tags in each lexicon. right: POS accuracy gain of test words with different POS tags by using induced lexicons. The induced lexicons have more verbs but lead to worse performance on verbs.

Discrepancies between the two NER datasets While WikiNER, along with POS tagging and Parsing, benefit the most from Pseudo MLM for Few-Text, MasakhaNER achieves the best result with Pseudo Trans-train. One possible explanation is that MasakhaNER contains data from the news domain, while WikiNER is created from Wikipedia. The pseudo monolingual data used for MLM is created from English Wikipedia articles, which could benefit WikiNER much more than MasakhaNER. On the other hand, the English NER training data for MasakhaNER is from the news domain, which potentially makes Pseudo Trans-train a stronger method for adapting the model simultaneously to the target language and to the news domain. One advantage of Pseudo MLM is that the English monolingual data is much cheaper to acquire, while Pseudo Trans-train is constrained by the amount of labeled data for a task. We show in § A.5 that Pseudo MLM has more benefit for MasakhaNER when we use a subset of the NER training data.

	bam	glv	mlt	myv
Gold MLM (Ours)	59.7	64.1	58.5	70.6
Ebrahimi and Kann (2021)	60.5	59.7	59.6	66.6
+Pseudo Trans-train	57.4	63.2	69.1	63.8
+Pseudo MLM	68.5	67.5	**72.3**	73.8
+Both	60.3	64.5	69.3	65.9
+Both(Label Distillation)	**69.4**	**68.8**	72.1	**74.3**

Table 3: Results for POS tagging with XLM-R. Our methods follow similar trend as on mBERT and they lead to significant gains compared to prior work.

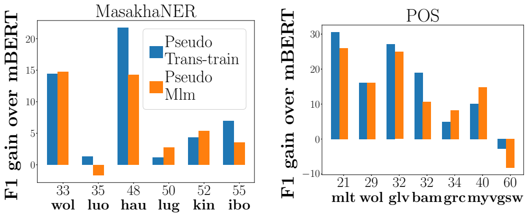

Figure 5: F1 gain over the baselines for languages with increasing baseline performance from left to right. Pseudo data tends to help more for languages with lower performance.

7 Analyses

Performance with XLM-R We mainly use mBERT because it has competitive performance for under-represented languages and it is more computationally efficient due to the smaller size. Here we verify our methods have the same trend when used on a different model XLM-R (Conneau et al., 2020). We focus on a subset of languages in the POS tagging task for the Few-Text setting and the results are in Tab. 3. We use the smaller XLM-R base for efficiency, and compare to the best result in prior work, which uses XLM-R large (Ebrahimi and Kann, 2021). Tab. 3 shows that our baseline is comparable or better than prior work. Similar to the conclusion in § 6, Pseudo MLM is the competitive strategy that brings significant improvements over prior work. While adding Pseudo Trans-train to Pseudo MLM does not help, using Label Distillation further improves the performance.

Effect of Baseline Performance Using pseudo data might be especially effective for languages with lower performance. We plot the improvement of different languages over the baseline in Fig. 5, where languages are arranged with increasing baseline performance from left to right. We mainly plot Pseudo MLM and Pseudo Trans-train for simplicity. Fig. 5 shows that for both resource settings, lower performing languages on the left tend to have more performance improvement by using pseudo data.

Using NMT Model to Synthesize Data One problem with the pseudo data synthesized using word-to-word translation is that it cannot capture the correct word order or syntactic structure in the target language. If we have a good NMT system that translates English into the target language, we might be able to get more natural pseudo monolingual data by translating the English sentences to the target language.

	WikiNER	MasakaNER	POS	Parsing
Lexicon	**45.0**	**56.0**	**63.7**	**40.7**
NMT	42.2	55.8	58.9	37.7

Table 4: F1 of using Pseudo MLM for Few-Text. Synthesizing data with NMT is consistently worse.

Since the target languages we consider are usually not supported by popular translation services, we train our own NMT system by fine-tuning an open sourced many-to-many NMT model on the Bible parallel data from English to the target language (details in § A.3). Instead of creating pseudo monolingual data using the lexicon, we can simply use the fine-tuned NMT model to translate English monolingual data into the target language.

The results of using NMT as opposed to lexicon for Pseudo MLM on all four tasks can be found in Tab. 4. Unfortunately, NMT is consistently worse than word-to-word translation using lexicons. We find that the translated monolingual data tend to have repeated words and phrases that are common in the Bible data, although the source sentence is from Wikipedia. This is because the NMT model overfits to the Bible data, and it fails to generate good translation for monolingual data from a different domain such as Wikipedia.

Comparison to Few-shot Learning Lauscher et al. (2020) found that using as few as 10 labeled examples in the target language can significantly outperform the zero-shot transfer baseline for languages included in mBERT. We focus on the zero-shot setting in this paper because the languages we consider have very limited data and it could be expensive or unrealistic to annotate data in every task for thousands of languages. Nonetheless, we experiment with k-shot learning to examine its performance on low-resource languages in the MasakhaNER task. Tab. 5 shows that using 10 labeled examples brings improvements over the mBERT baseline for a subset of the languages, and it is mostly worse than our best adapted model without using any labeled data. When we have access to 100 examples, few-shot learning begins

Method	hau	wol	lug	ibo	kin	luo
mBERT	48.7	33.9	50.9	55.2	52.4	35.3
Best Adapted	74.4	**60.3**	**61.6**	63.6	**63.8**	42.6
10-shot	44.5	49.1	52.7	56.2	51.2	46.2
100-shot	64.0	56.9	58.3	**65.5**	55.7	**51.6**
Best Adapt+100-shot	**76.1**	57.3	61.3	63.2	62.6	49.4

Table 5: Results on MasakhaNER for k-shot learning. We compare to the zero-shot mBERT baseline and our best adapted model.

to reach or exceed our zero-shot model. In general, few-shot learning seems to require more data to consistently perform well for under-represented languages while our adaptation methods bring consistent gains without any labeled data. Combining the best adapted model with few-shot learning leads to mixed results. More research is needed to understand the annotation cost and benefit of few-shot learning for low-resource languages.

8 Related Work

Several methods are proposed to adapt pretrained language models to a target language. Most of them rely on MLM training using monolingual data in the target languages (Wang et al., 2020; Chau et al., 2020; Muller et al., 2021; Pfeiffer et al., 2020; Ebrahimi and Kann, 2021), competitive NMT systems trained on parallel data (Hu et al., 2020; Ponti et al., 2021), or some amount of labeled data in the target languages (Lauscher et al., 2020). These methods cannot be easily extended to low-resource languages with no or limited amount of monolingual data, which account for more than 80% of the World's languages (Joshi et al., 2020).

Bilingual lexicons have been commonly used for learning cross-lingual word embeddings (Mikolov et al., 2013; Ruder et al., 2019). Among these, some work uses lexicons to synthesize pseudo bilingual (Gouws and Søgaard, 2015; Duong et al., 2016) or pseudo multilingual corpora (Ammar et al., 2016). Mayhew et al. (2017) propose to synthesize task data for NER using bilingual lexicons. More recently, Khemchandani et al. (2021) synthesize monolingual data in Indian languages for adapting pretrained language models via MLM. However, none of them provide systematic studies of methods that utilize lexicons and limited data resources for adapting pretrained language models to languages with no or limited text.

9 Conclusion and Discussion

We propose a pipeline that leverages bilingual lexicons, an under-studied resource with much better language coverage than conventional data, to adapt pretrained multilingual models to under-represented languages. Through comprehensive studies, we find that using synthetic data can significantly boost the performance of these languages while the best method depends on the data availability. Our results show that we can make concrete progress towards including under-represented languages into the development of NLP systems by utilizing alternative data sources.

Our work also has some limitations. Since we focus on different methods of using lexicons, we restrict experiments to languages in Latin script and only use English as the source language for simplicity. Future work could explore the effect of using different source languages and combining transliteration (Muller et al., 2021) or vocabulary extension (Pfeiffer et al., 2021) with lexicon-based data augmentation for languages in other scripts. We also did not test the data augmentation methods on higher-resourced languages as MLM fine-tuning and translate-train are already effective in that setting and our main goal is to support the languages with little textual data. Nonetheless, it would be interesting to examine whether our methods can deliver gains for high-resource languages, especially for test data in specialized domains.

We point to the following future directions: First, phrases instead of single word entries could be used to create pseudo data. Second, additional lexicons beyond PanLex could be leveraged.[6] Third, more effort could be spent on digitizing both existing monolingual data such as books (Gref, 2016) and lexicons into a format easily accessible by NLP practitioners. Although PanLex already covers over 5000 languages, some language varieties have only as little as 10 words in the database, while there exist many paper dictionaries that could be digitized through technologies such as OCR (Rijhwani et al., 2020).[7] Lexicon collection is also relatively fast, which could be a more cost effective strategy to significantly boost the performance of many languages without lexicons. Finally, the quality of synthetic data could be improved by incorporating morphology. However, we find that there is virtually no existing morphological analysis data or toolkits for the languages we consider. Future work could aim to improve the morphological analysis of these low-resource languages.

[6] We provide a list of resources in Appendix A.6.

[7] `https://panlex.org/source-list/` contains a list of undigitized dictionaries.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. Masakhaner: Named entity recognition for african languages. In *TACL*.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. 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)*, Hong Kong, China. Association for Computational Linguistics.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages. *arXiv preprint arXiv:2110.06733*.

Brenda Boerger. 2017. Rapid word collection, dictionary production, and community well-being.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *COLING*.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of EMNLP 2020*.

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 *ACL*.

Alexis Conneau and Guillaume Lample. 2019. Crosslingual language model pretraining. In *NeurIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. *arXiv preprint arXiv:1606.09403*.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. In *ACL*, Online. Association for Computational Linguistics.

Jost Gippert, Nikolaus Himmelmann, Ulrike Mosel, et al. 2006. *Essentials of language documentation*. Mouton de Gruyter Berlín.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390, Denver, Colorado. Association for Computational Linguistics.

Emily Kennedy Gref. 2016. *Publishing in North American Indigenous Languages*. Ph.D. thesis, University of London.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*, Online.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of ACL 2018*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *ICML*.

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 *ACL*, Online. Association for Computational Linguistics.

Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. 2021. Exploiting language relatedness for low web-resource language model adaptation: An Indic languages study. In *ACL*, Online. Association for Computational Linguistics.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *EMNLP*, Hong Kong, China.

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 *EMNLP*, Online. Association for Computational Linguistics.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *EMNLP*, Copenhagen, Denmark. Association for Computational Linguistics.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *LREC*, pages 2884–2892, Marseille, France. European Language Resources Association.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *NAACL*, Online.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, et al. 2018. Universal dependencies 2.2.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *ACL*, pages 1946–1958, Vancouver, Canada. ACL.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *EMNLP*, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs Everywhere: Adapting Multilingual Language Models to New Scripts. In *Proceedings of EMNLP 2021*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *ACL*, Florence, Italy.

Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulić, and Siva Reddy. 2021. Modelling latent translations for cross-lingual transfer. In *Arxiv*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *ACL*.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *ACL*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *EMNLP*, Online. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation. In *Proceedings of EMNLP 2021*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

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

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *EMNLP*.

Language	iso	Family	Task	Lex Count
Acehnese	ace	Austronesian	NER	0.5k
Bashkir	bak	Turkic	NER	3.4k
Crimean Turkish	crh	Turkic	NER	4.4k
Hakka Chinese	hak	Sino-Tibetan	NER	8.5k
Igbo	ibo	Niger-Congo	NER	3.6k
Ilokano	ilo	Austronesian	NER	4.0k
Kinyarwanda	kin	Niger-Congo	NER	4.7k
Eastern Mari	mhr	Uralic	NER	21.7k
Maltese	mlt	Afro-Asiatic	All	1.0k
Maori	mri	Austronesian	NER	13.8k
Hausa	hau	Niger-Congo	NER	5.6k
Wolof	wol	Niger-Congo	All	1.9k
Luganda	lug	Niger-Congo	NER	3.5k
Luo	luo		NER	0.7k
Bambara	bam	Mande	POS,Parsing	4.4k
Manx	glv	Indo-European	POS,Parsing	37.6k
Ancient Greek	grc	Indo-European	POS,Parsing	8.0k
Swiss German	gsw	Indo-European	POS,Parsing	2.5k
Erzya	myv	Uralic	POS,Parsing	7.4k

Table 6: Languages used for evaluation.

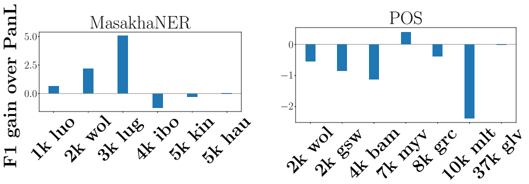

Figure 6: Improvements of using combined lexicons compared to PanLex lexicons for Pseudo MLM. Languages with fewer PanLex lexicons tend to benefit more from the combined lexicons.

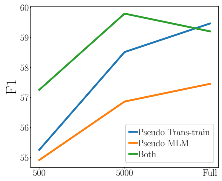

Figure 7: F1 on MasakhaNER with different amount of labeled data. Pseudo MLM becomes beneficial when the labeled training data is small.

A Appendix

A.1 Languages

The languages we used for experiments are listed in Tab. 6.

A.2 Experiment Details

For all experiments using MLM training for NER tasks, we train 5000 steps, or about equivalent to 40 epochs on Bible; for MLM training for POS tagging and Parsing, we train 10000 steps, or equivalent to 80 epochs on Bible. We use learning rate of $2e-5$, batch size of 32, and maximum sequence length of 128. We did not tune these hyperparameters because we mostly follow the ones provided in (Ebrahimi and Kann, 2021).

To finetune the model for a downstream task, we use learning rate of $2e-5$ and batch size of 32. We train all models for 10 epochs and pick the checkpoint with the best performance on the English development set.

A.3 NMT Models

We use the many-to-many NMT models provided in the fairseq repoo (Ott et al., 2019). We use the model with 175M parameters and finetune the NMT model for 50 epochs on the parallel data from the Bible.

We use beam size of 5 to generate translations.

A.4 Induced lexicons help languages with Fewer PanLex Entries

We plot the performance difference between using combined lexicons and PanLex for the Few-Text in Fig. 6. The languages are arranged from left to right based on increasing amount of PanLex entries. For MasakhaNER, the three languages with fewer entries in PanLex have much more significant gains by using the combined lexicon. While using the combined lexicons generally hurts POS tagging, the languages with fewer entries in PanLex tend to have less performance decrease.

A.5 Effect of Task Data Size

Our experiments in Tab. 2 show that MasakhaNER benefits more from Pseudo Trans-train, likely because the labeled data is closer to the domain of the test data. However, this result might not hold when the amount of labeled data is limited. One advantage of Pseudo MLM over Pseudo Trans-train is that it only requires English monolingual data to synthesize pseudo training data, while Pseudo Trans-train is constrained by the availability of labeled data. We subsample the amount of English NER training data for MasakhaNER and plot the average F1 score of Pseudo Trans-train, pseudo MLM and using both. Fig. 7 shows that the advantage of Pseudo Trans-train on MasakhaNER decreases as the number of labeled data decreases, and using both methods is more competitive when the task data is small.

A.6 List of Bilingual Lexicons

We provide a list of bilingual lexicons beyond PanLex:

- Swadesh lists in about 200 languages in Wikipedia[8]

[8] https://en.wiktionary.org/wiki/

- Words in 3156 language varietities in CLICS[9]
- Intercontinental Dictionary Series in about 300 languages[10]
- 40-item wordlists in 5,000+ languages in ASJP[11]
- Austronesian Basic Vocabulary Database in 1,700+ languages[12]
- Diachronic Atlas of Comparative Linguistics in 500 languages[13]

A.7 Lexicon Extraction

We use a simple python script to extract the lexicons from the PanLex database, and directly use them for synthesizing the pseudo data. We will open-source the script in our codebase.

Appendix:Swadesh_lists

[9] https://clics.clld.org/

[10] https://ids.clld.org/

[11] https://asjp.clld.org/

[12] https://abvd.shh.mpg.de/austronesian/

[13] https://diacl.ht.lu.se/