Code-Switch Pretraining for Improved Cross-Lingual Alignment in Low-Resource Languages

Ruan Visser[1*], Trienko Grobler[1], and Marcel Dunaiski[1]

Department of Computer Science, Stellenbosch University, Stellenbosch, South Africa
ruanvisser101@gmail.com, tlgrobler@sun.ac.za, marceldunaiski@sun.ac.za

Abstract. Cross-lingual language models enable the transfer of linguistic knowledge across languages, however, they often perform worse for low-resource or typologically distant languages. Prior work has explored alignment and adapter methods, but the use of code-switching remains limited and typically confined to fine-tuning with static word substitutions. In this work, we propose an approach that integrates code-switching directly into masked language model pretraining. Instead of applying word substitutions after pretraining, we introduce a multiview probabilistic translation strategy that samples candidate translations based on alignment likelihoods, applying substitutions only to unmasked tokens. This exposes the model to cross-lingual ambiguity and encourages more robust cross-lingual representations. Our results on a diverse set of eight language pairs show that this approach improves zero-shot cross-lingual natural language understanding performance across all languages relative to bilingual baselines. We further observe gains on downstream named entity recognition tasks in most languages when incorporating our code-switched pretraining approach.

1 Introduction

Cross-lingual understanding enables language models to transfer knowledge across languages, supporting scalable deployment in diverse linguistic contexts without relying on large pretraining datasets or expensive labeled downstream tasks [12,2,9]. However, target language performance on cross-lingual tasks still lags behind source language performance, particularly for low-resource languages and those that are typologically distant from source languages. This performance gap provides opportunities for more effective methods to improve cross-lingual alignment.

Many earlier cross-lingual solutions focus on creating language-agnostic models designed to generalize across diverse languages. For example, Artetxe et al. [2] introduced the LASER (Language-Agnostic SEntence Representations) model, which produces sentence embeddings that generalize across 93 languages without additional training. Similarly, Pheiffer et al. [18] proposed Multiple ADapters

* Corresponding author

for Cross-lingual transfer (MAD-X), a modular framework with language- and task-specific adapters, designed to enhance zero-shot cross-lingual transfer to low-resource target languages, and demonstrated significant improvements in this area.

More recently, the advantages of specialized models have become increasingly evident. For instance, LASER3 [13] introduced localized language family representations and combined supervised and self-supervised learning, and achieved better cross-lingual performance than its predecessors, especially for low-resource languages. Likewise, Parovic et al. [17] proposed BAD-X (Bilingual ADapters improve zero-shot Cross-lingual transfer) which focuses on bilingual transfer and outperforms more general approaches like MAD-X by specializing in specific language pairs. Both LASER3 and BAD-X highlight the effectiveness of specialized solutions in order to improve cross-lingual understanding.

The use of code-switching has also emerged as a promising technique to improve cross-lingual alignment. Code-switching refers to the practice of alternating between two or more languages within a corpus, sentence, or a single phrase, which reflects how multilingual speakers naturally mix languages. Qin et al. [20] enhanced the cross-lingual performance of multilingual BERT (mBERT) and XLM by fine-tuning them on code-switched data. Alternatively, Chaudhary et al. [6] introduced a modified Masked Language Modeling (MLM) objective, predicting translations of masked tokens based on the target language. However, these methods rely on static word substitutions and are typically applied to models after pretraining. This raises the question whether more dynamic probabilistic approaches can be beneficial to model learning if incorporated during pretraining itself. Lastly it should be noted that language-agnostic solutions [2,18,20,6], which apply translations from multiple languages and which do not tailor their models to specific language pairs, have been shown to be less effective [13,17].

To address these limitations, we evaluate the cross-lingual efficacy of specialized bilingual models that incorporate probabilistic multiview code-switching during pretraining. In this approach, tokens in the input sequences are dynamically altered by substituting words which are sampled from multiple plausible translations. Therefore, our approach ensures that the models see code-switched bilingual input with the aim to obtain more robust cross-lingual representations and to improve downstream task performances for low-resource and typologically distant languages.

2 Background

Cross-lingual alignment refers to the ability of language models to represent linguistic features similarly across languages. Researchers often assess this through translated word or sentence similarity or, more recently, via zero-shot cross-lingual transfer, where models trained on a source language are evaluated directly on target languages. Alternatively, Hammerl et al. [12] define cross-lingual alignment through two lenses, namely (1) the similarity of semantically equivalent representations across languages compared to dissimilar ones and (2) the

ability of task-specific prediction heads to generalize patterns across languages. While the first lens offers a useful theoretical framework, it is often more challenging to evaluate and may not directly translate into downstream performance gains.

2.1 Neural Models for Cross-Lingual Understanding

Massively multilingual models trained without any explicit alignment objectives have demonstrated strong cross-lingual capabilities. For example, models such as XLM-R [7], trained on over 100 languages, show significant cross-lingual transfer performance on benchmarks like XNLI [9].

Given the inherent multilingual capabilities of pretrained language models like XLM-R, researchers have explored various methods to further improve their cross-lingual understanding. For instance, Pheiffer et al. [18] implement an adapter framework, MAD-X, which incorporates language and task-specific adapters. Their adapters are trainable modular neural network layers that are inserted within transformer layers and then exchanged for zero-shot inference. They find that incorporating this adapter strategy in multilingual models leads to improved zero-shot cross-lingual performance on low-resource languages and languages not seen during initial pretraining. Parovic et al. [17] extend the work by Pheiffer et al. [18] to create BAD-X, a bilingual adapter framework that specializes in transfer between specific language pairs, rather than focusing on the more language-agnostic transfer as done in MAD-X. They sacrifice the modularity of exchangeable language adapters in favor of a bilingual adapter, which results in better performance compared to MAD-X across most languages tested.

In contrast to approaches that rely on existing massively multilingual models, more specialized methods have been developed to improve cross-lingual capabilities through targeted architectures and training strategies. For instance, Artetxe et al. [2] train LASER, a multilingual machine translation model, to create universal language agnostic sentence embeddings. They argue that their 5-layer BiLSTM model finetuned using exclusively English data can be used without any further training on 93 different languages. Subsequently, Heffernan et al. [13] created LASER3, a larger 12-layer Transformer model, that focuses on language family representations rather than being a language agnostic encoder model. Additionally, instead of using fully supervised parallel data as done by Artetxe et al. [2], they combine supervised parallel training with self-supervised monolingual training. They find that by using these modifications, LASER3 is able to significantly outperform LASER while increasing the language coverage to include many very low-resource African languages.

With improvements in contextual embeddings, many researchers focus on obtaining word pairs while taking contextual information into account to improve cross-lingual alignment of multilingual models. For instance, Wang et al. [24] demonstrate that aligning cross-lingual contextual word embeddings using learned linear transformations improves zero-shot dependency parsing compared to non-contextual methods. Their approach involves identifying word pairs

through unsupervised bidirectional word alignment on parallel corpora and obtaining contextualized word embeddings for these pairs. They then learn a linear transformation to align the embeddings of the target language with those of the source language. By applying these aligned embeddings to a dependency parser trained on the source language, they achieve significant improvements over static embeddings, with a 2.91% gain on the Universal Dependencies dataset [16].

Cao et al. [5] use the contextual word retrieval task to better understand mBERT's multilingual capabilities. In line with Hammerl et al. [12]'s first definition of cross-lingual alignment, they argue that models that have similar representations for word pairs in parallel sentences are contextually aligned. In this task, a model receives parallel corpora along with a source-language word within a sentence and is expected to output the corresponding word and the sentence containing it in the target corpus. Their evaluation shows that mBERT is partially aligned despite the absence of explicit alignment during pretraining. Furthermore, fine-tuning mBERT using the contextual word retrieval task leads to improved performance over previous rotation-based alignment methods, such as those proposed by Wang et al. [24], on zero-shot XNLI. Remarkably, they also find that their method, in some instances, performs on par with fully supervised translate-train models in zero-shot evaluation.

2.2 Code-Switching for Cross-Lingual Alignment

The use of code-switch data during training is an online approach that can be used to align embedding spaces of different languages. Qin et al. [20] employ code-switch finetuning to improve the zero-shot cross-lingual performance of mBERT and XLM. Specifically, they alter the finetuning data by inducing code-switching in the source sentence using Multilingual Unsupervised or Supervised word Embeddings (MUSE) words pairs from multiple languages. This fine-tuning approach leads to improved cross-lingual language understanding on various tasks compared to simply using massively multilingual models.

Alternatively, Yang et al. [26] propose code-switch pretraining for neural machine translation. They randomly replace sections of a source sentence with their translations, obtained through lexical induction using unsupervised word embeddings between source and target languages. This approach outperforms established models such as XLM [15] and can better handle code-switched inputs.

More related to this work, Chaudhary et al. [6] introduce DICT-MLM, a modified MLM objective which involves predicting translations of masked tokens given the target language. Unlike standard MLM, DICT-MLM incorporates language tokens with each word to specify the target language in which predictions should be made. Chaudhary et al. [6] argue that this objective enables better cross-lingual alignment and find that their model can outperform mBERT across a wide variety of languages. However, as noted by Ruder et al. [21], cross-lingual transfer typically relies on language-agnostic representations, and including language tokens during training could potentially degrade this property. As a result, methods that do not rely on explicit language markers may be preferable.

Addressing this concern, Tang et al. [22] created Align-MLM, a model that focuses on better aligning word embeddings between synonymous words from different languages without using language tokens. They introduce an additional loss term to reduce the distance between embeddings and find that this enables Align-MLM models to outperform or match both the MLM and DICT-MLM objectives.

2.3 Word Translations

Due to the lack of large parallel corpora for most languages, many approaches that aim to improve cross-lingual performance utilize word translations [5,24,6,20]. Techniques for identifying translation pairs in natural language text range from statistical approaches to newer unsupervised neural approaches.

IBM Model 1 [3], the first model to effectively leverage statistical learning for machine translation, is a statistical framework that can be trained on parallel corpora without the use of seed word translations, using the Expectation-Maximization (EM) algorithm. This model uses a translation table, a matrix that quantifies the probability that each target word is a translation of each source word, which is initialized uniformly, assuming all words of the target language are equally likely to be translations of each source word. The model then iteratively refines these probabilities by evaluating possible alignments across parallel sentences until convergence. Later IBM models, such as Models 2 through 5, introduced additional features like positional alignment, fertility modeling, and distortion parameters to improve translation accuracy [3,4].

More recently, Conneau et al. [8] leverage word embeddings trained on monolingual Wikipedia text to create state-of-the-art bilingual word pairs. Their approach involves several stages. First, they obtain the FastText word embeddings of two different languages in a shared semantic space. They then use adversarial training to obtain candidate translations, which are subsequently reduced to only include the most frequently used words and mutual word pairs. These translations are further improved by iteratively using the Procrustus solution, which finds the optimal orthogonal mapping between source and target embedding spaces using singular value decomposition (SVD). To address the hubness problem with nearest neighbor algorithms, where a single point is the nearest neighbor for many points, they use the Cross-Domain Similarity Local Scaling (CSLS) metric. CSLS ensures that the density of the embedding space is more uniform, thereby reducing the number of hubs and anti-hubs (points that are not the nearest neighbor of any other point). By following this approach along with the CSLS metric, Conneau et al. [8]'s unsupervised MUSE algorithm is able to outperform supervised methods on most bilingual lexicon induction tasks.

3 Methodology

Our primary objectives are to assess how code-switching during pretraining affects cross-lingual generalization and to compare the downstream performance

gains that can be obtained using probabilistic word translations instead of static ones. To this end, we developed encoder-based masked language models (MLMs) which we trained from scratch on various bilingual data mixtures, with code-switching incorporated into the pretraining process.

We expanded upon approaches such as DICT-MLM by incorporating code-switched tokens directly into the model's input. Unlike prior work, we do not explicitly train the model to predict code-switched tokens. Instead, source words are predicted in the presence of a code-mixed context window. This setup allows the model to implicitly learn relationships between source and target tokens, facilitating better cross-lingual alignment. This mechanism can also be interpreted as increasing the number of anchor points between languages in the training data. Anchor points can be seen as tokens that are consistent or shared across languages, such as numerals, named entities, and punctuations. These elements are believed to support cross-lingual alignment by providing reference points for the model to associate semantically similar contexts across languages [10,19,25].

In contrast to previous code-switching approaches which utilize many languages, we limit the number of languages to two. While this could make the model less language-agnostic, prior work, such as Parovic et al. [17] and Heffernan et al. [13], demonstrates that focusing on a smaller set of languages can enhance the cross-lingual understanding between these languages.

The following sections outline our translation strategies, model configurations, and evaluation procedures.

3.1 Translations

We train IBM translation models on 1 million sentence pairs sampled from the No Language Left Behind (NLLB) dataset, a collection of bitexts that were mined from the web using LASER3, and use these models to identify word translations. For IBM models, each source word has a number of candidate target translations paired with probabilities.

Given the IBM model's overconfidence in a single translation candidate, often assigning it the majority of the probability mass, we apply a temperature-controlled formula to more equally redistribute the probability mass among alternatives.

$$P(xw_i) = \frac{\left(\text{prob}_{xw_i}\right)^{1/T}}{\sum_j \left(\text{prob}_{xw_j}\right)^{1/T}}$$

In the above formula prob_{xw_i} refers to the IBM word translation probability while the T parameter refers to the temperature. We use a T value of 3, a relatively high temperature, which increases the probability of less probable translations being selected. To enable models to train on a range of different translations while reducing the number erroneous translations, we only sample translations with probabilities greater than 2%. Furthermore, to prevent the model from learning trivial translation mappings, we exclude stopwords and words shorter than four characters from translation.

We present these translations and probabilities in Table 1, which lists the most likely French translations for the English word *work* according to our trained IBM model, including *travail*, *travailler*, *travaille*, and *travaillent*. These variants offer different morphological perspectives for the same concept. We also provide examples for *excited* and *wounds* to illustrate how translation quality can vary, where some candidate translations are semantically relevant, while others (often function words) are unrelated or less informative. Frequent stopwords such as *a*, *le*, and *des* are excluded from the substitution table to avoid uninformative replacements.

Following Chaudhary et al. [6], we dynamically replace words with translations using a data loader that provides new replacements for each batch. Each instance has a 50% probability of containing replacements, and if selected, each word has a 3% probability of being replaced. We determined the optimal replacement rate by examining the performance 1%, 3%, 5% and 10% rates on the English-French languages pair. We then used this rate for all our language pairs. Additionally, we found that giving each instance a 50% probability of remaining unchanged led to better performance compared to applying code-switching to all instances.

We also evaluated the use MUSE [8], an unsupervised model that has been shown to produce higher-quality translations compared to supervised models such as the IBM model. MUSE does not require parallel corpora and can be trained solely on monolingual data. We pretrained a BERT model using a MUSE substitutions table for each language pair. Unlike the IBM word translation, MUSE translations are discrete and therefore each candidate word translation can be directly added to the substitution table. We therefore create substitution table where we map each token with all its possible MUSE translations. All languages in our experiments had available MUSE translation tables except for Xhosa and Swahili[1].

We observed that while the quality of MUSE word translations are on average better than those created by the IBM model, there are significantly fewer translation options and therefore less diversity. Similar to our IBM experiments, we also evaluated various replacement rates and found that a rate of 3% obtained the best downstream performance when trained on the English-French language pair.

In addition to the two code-switching strategies, we include a baseline model trained on the same bilingual corpora without code-switching, referred to as No CS (no code-switching).

3.2 Data Mixtures

We experiment with eight language pairs, each combining English as the high-resource language with the following target languages or simulated low-resource, obtained by down-sampling their available data: French, German, Portuguese,

[1] MUSE translation tables were obtained from the official repository: `https://github.com/facebookresearch/MUSE`

Table 1. Top translations with probabilities and normalized likelihoods (NL) for each word. We find that translations can be fairly noisy, with some translations having little to no relation to the source word.

Translation	Probability	NL	Eng Translation
		work	
travail	0.3689	0.717	work
travailler	0.2252	0.608	to work
travaille	0.0544	0.379	work
travaillent	0.0405	0.343	work
de	0.0312	0.314	of
pour	0.0244	0.289	for
et	0.0164	0.254	and
les	0.0154	0.248	the
à	0.0150	0.246	has
le	0.0136	0.238	the
des	0.0133	0.236	of the
travaillons	0.0112	0.223	lets work
fonctionnent	0.0098	0.214	work
		excited	
excité	0.1293	0.5057	excited
excités	0.0950	0.4563	excited
très	0.0748	0.4214	very
enthousiastes	0.0701	0.4124	enthusiastic
enthousiaste	0.0647	0.4015	enthusiastic
suis	0.0567	0.3841	am
l'idée	0.0447	0.3548	the idea
heureux	0.0334	0.3222	happy
j'étais	0.0324	0.3190	I was
étions	0.0295	0.3091	were
m'enthousiasme	0.0248	0.2918	excites me
de	0.0167	0.2556	of
par	0.0145	0.2441	by
excitée	0.0137	0.2394	excited
		wounds	
blessures	0.4156	0.7463	wounds
plaies	0.1764	0.5608	wounds
lors	0.0189	0.2666	during
plaies	0.0185	0.2646	wounds
les	0.0184	0.2640	the
dès	0.0170	0.2570	of the
soigne	0.0128	0.2337	neat
tenant	0.0128	0.2337	holding
envisage	0.0128	0.2337	consider
réactions	0.0127	0.2334	reactions
milieu	0.0125	0.2322	medium
nos	0.0123	0.2309	our

Table 2. Hyperparameters used during BERT fine-tuning.

Hyperparameter	**Value**
Max sequence length	128
Batch size	32
Learning rate	5e-5
Weight decay	0.01
Warmup ratio	0.25
Epochs	3 or 4 (task dependent)
Learning rate scheduler	Linear decay

Turkish, Hungarian, Finnish, Swahili and Xhosa. This selection covers a broad range of linguistically diverse languages, which allows us to investigate the impact of code-switching more generally.

All data is sourced from mC4. For each pair we use 1GB of English text and 100MB of text in the corresponding target-language. While we also explored other sampling ratios, such as increasing the English or reducing the low-resource portions, we observed that this often leads to significant degradation in performance for the target language. To avoid biasing the shared vocabulary towards English, we constructed the vocabulary using only 100MB of each language.

Our selected language pairs span various families and have varying amounts of shared lexicon with English. For instance, English and German are both West Germanic languages, while French, though Romance, shares substantial lexical overlap with English. Portuguese, also a Romance language, is arguably the least related Indo-European language in our selection. Hungarian and Finnish represent the Uralic family and are not closely related to any of the Indo-European language in the group. Turkish belongs to the Turkic family and is also unrelated to English. We also experiment with two Bantu languages: Xhosa and Swahili. The nature of the language combinations help us to infer whether lexical and grammatical similarity, either due to genealogical relatedness or extensive borrowing, can substitute for or enhance the efficacy of code-switch pretraining.

3.3 Models

We trained BERT models using the standard pretraining hyperparameters employed by Devlin et al. [11]. However, we reduced the maximum sequence length from 512 to 128 tokens due to computational constraints. To maintain training stability with shorter sequences, we doubled the batch size to 512.

Model sizes were selected to match the corresponding text volumes. Specifically, we used base-sized models, as recommended by Visser et al. [23], for text volumes of approximately 250MB to 1000MB. Additionally, Visser et al. [23] observed that training for thousands of epochs on relatively small datasets can be beneficial. Since incorporating code-switching introduces further variation during training, we trained each model for 1 million steps to fully leverage this

method. Pretraining was conducted on TPUv3 instances, with each model taking approximately 2 days to complete.

3.4 Tasks

We evaluated the cross-lingual capabilities of each model by fine-tuning on English data and assessing the zero-shot performance on the target language. Our evaluation spans two tasks, namely natural language inference (NLI) and named entity recognition (NER). For the NLI task we used XNLI dataset while for NER task, we used the PAN-X dataset. XNLI is a multiclass task which prompts the model to identify the relationship between two sentences, a premise and a hypothesis. There are three possible relationships, namely entailment, contradiction or neutrality. Entailment is defined as the hypothesis follows the premise. Contradiction occurs when two sentences are in disagreement. Neutrality denotes cases where neither entailment nor contradiction is observed. We followed the standard XNLI evaluation setup and used the English MultiNLI (MNLI) dataset for training. MNLI contains approximately 393,000 sentence pairs labeled for entailment, contradiction or neutrality. The XNLI validation sets comprises 15 languages and are then used for cross-lingual evaluation. The XNLI dataset is sampled from 10 distinct domains to ensure diversity across genres.

Since not all languages are available across both tasks, we only evaluated on languages present in the respective datasets. For XNLI, we additionally made use of AfriXNLI [1], which provides translated test sets for 16 African languages. Although these testsets are smaller, containing 600 instances instead of the 5010 found in XNLI, it is still sampled equally from the 10 XNLI domains before translation. We used the Xhosa portion for AfriXNLI to evaluate our English-Xhosa model.

For the NER tasks, we used the PAN-X dataset, introduced as part of the XTREME benchmark [14]. It consists of NER sentences derived from the Wikipedia WikiANN corpus and covers 40 languages. Each instance consists of a sentence where each token or word is annotated using BIO tagging for entity types such as person (PER), organization (ORG), and location (LOC). This dataset is constructed automatically using interlanguage links and cross-lingual Wikipedia alignment. For each language, PAN-X contains around 20,000 to 70,000 training examples, 5000 validation, and 5000 test sentences.

For XNLI, we report average accuracy across 5 runs, and for NER, we report average macro-F1 scores across 10 runs. While we adopted standard hyperparameters for fine-tuning, we performed a grid search for each task over the following hyperparameter configurations: learning rate 2e-5, 5e-5, 1e-4 and epochs 3, 4, 5. Both NER and NLI performed best with a learning rate of 5e-5. NER achieved the best performance when running for 3 epochs while NLI performed best after 4 epochs. Additionally, to increase run stability we used a weight decay of 0.01 and a warmup ratio of 0.25. The exact hyperparameters used for each task are listed in Table 2. Lastly, we also evaluated the models MLM performance on held-out test data. The test data we used was 50MB of the target language

text sourced from mC4, except for Xhosa, where we used 30MB due to limited available data.

4 Results

Table 3 reports the test MLM performance of our models that use the IBM and MUSE translation approaches. We also show the results of the baseline model trained on the same bilingual corpora without code-switching, shown in the column labeled No CS. We find that for six out of eight languages both code-switching models outperformed the baselines, with Turkish and German being the only instances where one or both code-switching techniques performed worse. This indicates that incorporating code-switched input generally improves MLM accuracy while maintaining consistent representations, suggesting the additional variety and including more anchor points, benefits cross-lingual alignment.

Table 3. Masked language modeling (MLM) performance for the target languages, reported as test set accuracy. Bold indicates the highest score for each language. The symbol '–' indicates that no MUSE translations are available for associated languages.

Language	No CS	IBM	MUSE
French	0.6779	0.6787	**0.6808**
German	0.6412	0.6367	**0.6418**
Hungarian	0.6489	**0.6516**	0.6505
Portuguese	0.6544	**0.6610**	0.6583
Turkish	**0.6428**	0.6353	0.6333
Finnish	0.6430	0.6433	**0.6437**
Swahili	0.6328	**0.6417**	-
Xhosa	0.6414	**0.6425**	-

For the cross-lingual XNLI performance results, we report the average accuracy of five runs together with associated standard deviations in Table 4. We observe that the IBM model consistently outperforms the other models across all languages. This advantage in cross-lingual NLI may stem from the multiple views provided by the different IBM translations. Such diversity could be particularly beneficial for agglutinative languages such as Swahili, Xhosa and Turkish, as well as for languages that frequently form compound words such as German, where single English translations often fail to convey the full meaning. Notably, IBM models also outperform both No CS and MUSE for French (a language with fewer compound words), although the margin of improvement is relatively small.

Table 5 lists the results of our models when evaluated on NER downstream tasks using the PAN-X datasets. We report the average macro-F1 performances of 10 runs across seven target languages. In contrast to the internal MLM performances and downstream XNLI results, no clear best strategy emerges across

Table 4. Cross-lingual XNLI accuracy for models finetuned on English XNLI and evaluated zero-shot on each target language.

Language	No CS	IBM	MUSE
German	0.7265 ± 0.0075	**0.7382** ± 0.0023	0.7290 ± 0.0032
French	0.7433 ± 0.0032	**0.7452** ± 0.0089	0.7442 ± 0.0053
Swahili	0.6952 ± 0.0044	**0.6968** ± 0.0031	-
Turkish	0.3759 ± 0.0108	**0.3931** ± 0.0059	0.3717 ± 0.0078
Xhosa	0.6169 ± 0.0195	**0.6293** ± 0.0089	-

the majority of languages. While the IBM approach offers greater translation variety, some translations may differ in part of speech types or named entity categories compared to the original tokens. In contrast, the single translation chosen by MUSE is more likely to match the original part of speech and named entity types, which could explain its stronger NER performance.

Table 5. Cross-lingual PAN-X macro-F1 for models finetuned on English and evaluated zero-shot on each target language. Bold indicates the highest score for each language.

Language	No CS	IBM	Muse
German	0.7223 ± 0.0038	**0.7348** ± 0.0037	0.7236 ± 0.0034
Finnish	0.6678 ± 0.0068	0.6665 ± 0.0096	**0.6748** ± 0.0067
French	0.7357 ± 0.0133	0.7340 ± 0.0086	**0.7453** ± 0.0111
Hungarian	0.5625 ± 0.0100	**0.5812** ± 0.0080	0.5748 ± 0.0072
Portuguese	**0.7475** ± 0.0052	0.7235 ± 0.0083	0.7237 ± 0.0032
Swahili	0.7296 ± 0.0140	**0.7320** ± 0.0083	-
Turkish	0.6032 ± 0.0085	0.5952 ± 0.0049	**0.6126** ± 0.0110

5 Conclusion

In this work, we proposed two code-switch pretraining strategies for improving cross-lingual representation learning in encoder-based language models. Both approaches inject translated tokens into bilingual text corpora during pretraining, where the first uses one-to-one word translation mappings from MUSE and the second uses a multiview probabilistic strategy based on on an IBM model's word alignments.

We evaluated our approaches using eight language pairs on two downstream tasks. We found that both methods consistently improved the zero-shot cross-lingual performance over models without code-switching in the majority of languages. Notably, the multiview translation strategy yields the largest improvements in natural language inference tasks and outperforms baselines across all

evaluated languages. This suggests that exposing the model to diverse translation mappings during pretraining enhances its cross-lingual understanding.

These findings highlight the potential of utilizing code-switch pretraining to improve models' cross-lingual natural language understanding. While both code-switching pretraining methods improve model performances for most languages, we find it to be particularly effective when multiple diverse translations are used during pretraining. This effect appears stronger for agglutinative or compound-forming languages, though further testing is needed to identify which languages benefit most.

6 Limitations

The evaluation in this study was restricted in terms of both languages and tasks. Although the proposed approach shows gains on the evaluated benchmarks, the lack of high-quality cross-lingual datasets for a wider range of low-resource languages and NLU tasks remains a limitation. Developing broader benchmarks will be crucial to enable more comprehensive evaluations in future work.

7 Acknowledgements

This work was supported by Google's TPU Research Cloud program.

References

1. Adelani, D.I., Ojo, J., Azime, I.A., Zhuang, J.Y., Alabi, J.O., He, X., Ochieng, M., Hooker, S., Bukula, A., Lee, E.S.A., Chukwuneke, C., Buzaaba, H., Sibanda, B., Kalipe, G., Mukiibi, J., Kabongo, S., Yuehgoh, F., Setaka, M., Ndolela, L., Odu, N., Mabuya, R., Muhammad, S.H., Osei, S., Samb, S., Guge, T.K., Sherman, T.V., Stenetorp, P.: Irokobench: A new benchmark for african languages in the age of large language models (2025), https://arxiv.org/abs/2406.03368
2. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics **7**, 597–610 (2018), https://api.semanticscholar.org/CorpusID:56895585
3. Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., Roossin, P.: A statistical approach to language translation. In: Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics (1988), https://aclanthology.org/C88-1016/
4. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. Comput. Linguist. **19**(2), 263–311 (Jun 1993)
5. Cao, S., Kitaev, N., Klein, D.: Multilingual alignment of contextual word representations. ArXiv **abs/2002.03518** (2020), https://api.semanticscholar.org/CorpusID:211069110
6. Chaudhary, A., Raman, K., Srinivasan, K., Chen, J.: Dict-mlm: Improved multilingual pre-training using bilingual dictionaries **abs/2010.12566** (2020), https://api.semanticscholar.org/CorpusID:225062397

7. Conneau, A.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
8. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. arXiv preprint arXiv:1710.04087 (2017)
9. Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S.R., Schwenk, H., Stoyanov, V.: Xnli: Evaluating cross-lingual sentence representations. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2018)
10. Conneau, A., Wu, S., Li, H., Zettlemoyer, L., Stoyanov, V.: Emerging cross-lingual structure in pretrained language models. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6022–6034. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.536, https://aclanthology.org/2020.acl-main.536/
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423/
12. Hämmerl, K., Libovický, J., Fraser, A.: Understanding cross-lingual Alignment—A survey. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 10922–10943. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.findings-acl.649, https://aclanthology.org/2024.findings-acl.649/
13. Heffernan, K., Çelebi, O., Schwenk, H.: Bitext mining using distilled sentence representations for low-resource languages. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 2101–2112. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.findings-emnlp.154, https://aclanthology.org/2022.findings-emnlp.154/
14. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization (2020), https://arxiv.org/abs/2003.11080
15. Lample, G., Conneau, A.: Cross-lingual language model pretraining. Advances in Neural Information Processing Systems (NeurIPS) (2019)
16. Nivre, J., de Marneffe, M.C., Ginter, F., Hajič, J., Manning, C.D., Pyysalo, S., Schuster, S., Tyers, F., Zeman, D.: Universal Dependencies v2: An evergrowing multilingual treebank collection. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 4034–4043. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.497/
17. Parovic, M., Glavas, G., Vulic, I., Korhonen, A.: Bad-x: Bilingual adapters improve zero-shot cross-lingual transfer. In: North American Chapter of the Association for Computational Linguistics (2022), https://api.semanticscholar.org/CorpusID:250390710

18. Pfeiffer, J., Vulić, I., Gurevych, I., Ruder, S.: MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7654–7673. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.617, https://aclanthology.org/2020.emnlp-main.617/
19. Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual bert? arXiv preprint arXiv:1906.01502 (2019)
20. Qin, L., Ni, M., Zhang, Y., Che, W.: Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. pp. 3853–3860. International Joint Conferences on Artificial Intelligence Organization (7 2020). https://doi.org/10.24963/ijcai.2020/533, https://doi.org/10.24963/ijcai.2020/533, main track
21. Ruder, S., Peters, M.E., Swayamdipta, S., Wolf, T.: Transfer learning in natural language processing. In: Sarkar, A., Strube, M. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials. pp. 15–18. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-5004, https://aclanthology.org/N19-5004/
22. Tang, H., Deshpande, A., Narasimhan, K.: Align-mlm: Word embedding alignment is crucial for multilingual pre-training. ArXiv **abs/2211.08547** (2022), https://api.semanticscholar.org/CorpusID:253553573
23. Visser, R., Grobler, T., Dunaiski, M.: Scaling behaviour of encoder language models in low-resource settings, manuscript submitted for review
24. Wang, Y., Che, W., Guo, J., Liu, Y., Liu, T.: Cross-lingual bert transformation for zero-shot dependency parsing. ArXiv **abs/1909.06775** (2019), https://api.semanticscholar.org/CorpusID:202578048
25. Wu, S., Dredze, M.: Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. arXiv preprint arXiv:1904.09077 (2019)
26. Yang, Z., Hu, B., Han, A., Huang, S., Ju, Q.: CSP:code-switching pre-training for neural machine translation. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2624–2636. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.208, https://aclanthology.org/2020.emnlp-main.208/