PREALIGN: Boosting Cross-Lingual Transfer by Early Establishment of Multilingual Alignment

Anonymous ACL submission

Abstract

Large language models demonstrate reasonable multilingual abilities, despite predominantly English-centric pretraining. However, the spontaneous multilingual alignment in these models is shown to be weak, leading to unsatisfactory cross-lingual transfer and knowledge sharing. Previous works attempt to address this issue by explicitly injecting multilingual alignment information during or after pretraining. Thus for the early stage in pretraining, the alignment is weak for sharing information or knowledge across languages. In this paper, we propose PREALIGN, a framework that establishes multilingual alignment prior to language model pretraining. PREALIGN injects multilingual alignment by initializing the model to generate similar representations of aligned words and preserves this alignment using a code-switching strategy during pretraining. Extensive experiments in a synthetic English to English-Clone setting demonstrate that PREALIGN significantly outperforms standard multilingual joint training in language modeling, zero-shot cross-lingual transfer, and cross-lingual knowledge application. Further experiments in real-world scenarios further validate PREALIGN's effectiveness across various model sizes.

1 Introduction

Large language models (Brown et al., 2020; Touvron et al., 2023a,b) have drastically changed the research paradigm of multilingual language processing. Despite being trained on mainly English texts, they still exhibit reasonable ability for other languages (Touvron et al., 2023a,b; Wang et al., 2024), and have established multilingual alignment to some extent (Devlin et al., 2019; Conneau and Lample, 2019; Lin et al., 2022). However, researchers (Wang et al., 2024; Gao et al., 2024; Zhang et al., 2023; Qi et al., 2023) have found the spontaneous alignment between languages in these model is still relatively weak, leading to weak cross-lingual factual knowledge retrieval (Wang et al., 2024; Gao et al., 2024) and inconsistency behaviors given the same input (Qi et al., 2023; Zhang et al., 2023).

A handful of works (Reimers and Gurevych, 2020; Cao et al., 2020; Wu and Dredze, 2020; Chaudhary et al., 2020; Yang et al., 2021; Tang et al., 2022; Feng et al., 2022; Gao et al., 2024) try to mitigate the problem by explicitly injecting alignment information using existing supervision data. They either construct cross-lingual prediction tasks (Chaudhary et al., 2020; Yang et al., 2021) or train models to produce similar representations of aligned words or sentences (Tang et al., 2022; Wu and Dredze, 2020; Reimers and Gurevych, 2020). Although these methods can bring reasonable improvements, the establishment of multilingual alignment requires a long training process either *during* or *after* pretraining (Dufter and Schütze, 2020), which prevents the model from effectively performing cross-lingual transfer at earlier stage in pretraining.

In this paper, we introduce PREALIGN, a framework designed to enhance the alignment of pretrained language models. PREALIGN differs from prior methods by integrating the multilingual alignment information *before* extensive language pretraining and maintaining it throughout the pretraining process. This proactive alignment effectively advances cross-lingual transfer, which enhances the model's proficiency in target languages early in its training, therefore improving the model's ability to acquire knowledge at that stage.

More specifically, before large-scale language pretraining, PREALIGN first collects multilingual translation pairs between English and languages to be transferred, and inject this information into the model by initializing it to produce similar representations of aligned words. In order to maintain the established multilingual alignment across the pretraining phase, we propose an input-only codeswitching strategy, which only substitutes words in the input text to its aligned words, and optimizes model using language modeling objective.

We firstly conduct experiments on a English to English-Clone settings (K et al., 2020; Dufter and Schütze, 2020; Schäfer et al., 2024). English-clone is a synthetic language that shares identical grammar and vocabulary distribution with English, but no vocab overlap. This allows us to study cross-lingual transfer on a more controlled environment. Experiments demonstrate that PREALIGN significantly improves models' ability of languages to be transferred, and strengthens cross-lingual transfer of downstream task abilities and knowledge. Further analysis shows that the early established multilingual alignment can be kept throughout large-scale language pretraining and generalize to other words. We further experiment with our methods on real-world settings, and validates the effectiveness of PREALIGN across different model scales.

2 Related Work

2.1 Understanding Cross-lingual Ability of Pretrained language models

Many works attempt to analyze the cross-lingual ability of LLMs. Dufter and Schütze (2020); Conneau et al. (2020) try to explain factors that contributes to spontaneous multilingual alignment developed in pretrained language models, including under-parameterization, shared model architectures and pivot words across languages. Other works investigate the working mechanism of multilingual representations. Wendler et al. (2024) find that English-centric models works on a concept space that is close to English when processing other languages. Recently, Gao et al. (2024); Qi et al. (2023) analyze multilingual knowledge alignment in existing LLMs, and find that multilingual training and instruction tuning can only lead to shallow alignment, i.e. LLMs can achieve similar task performances and consistent responses across languages, yet cannot apply knowledge across languages.

Our paper differs from theirs in that we focus on improving models' cross-lingual ability and successfully unlocks the ability of cross-lingual knowledge transferring.

2.2 Enhancing Cross-lingual Ability of Pretrained Language Models

Other studies also seek to enhance the cross-lingual capabilities of pretrained language models. These typically utilize explicit alignment signals, such as parallel sentences and dictionaries. They can be categorized based on when the alignment occurs: during pretraining or post-pretraining.

On the first category, Yang et al. (2020); Chaudhary et al. (2020) perform codeswitching on the monolingual data to make model better capture cross-lingual relation and dependency. Hu et al. (2021) train the model to produce consistent word alignment matrices between source and target language and similar representations for parallel sentences. Chi et al. (2022) explores multilingual replaced token detection and translation replaced token detection task. Tang et al. (2022) further maximize the cosine similarity of aligned word embeddings to explicitly inject multilingual alignment.

On the second category, researchers enhance the multilingual alignment after pretraining. Earlier works either optimizes pretrained models to produce similar representations for parallel sentences (Reimers and Gurevych, 2020; Pan et al., 2021; Feng et al., 2022) or parallel words (Cao et al., 2020; Wu and Dredze, 2020). Recent works on large language models typically train the model to produce consistent responses (She et al., 2024) or performing cross-lingual instruction-following tasks (Zhu et al., 2024b,a).

PREALIGN differs from all above works in that it establishes multilingual alignment before language pretraining, therefore facilitating the cross-lingual transfer at early pretraining stage.

3 Methodology

In this section, we present PREALIGN, a simple and effective framework that advances the establishment of multilingual alignment before language pretraining.

3.1 Injecting Multilingual Alignment before Language Pretraining

PREALIGN aims to inject multilingual alignment information before large-scale language model pretraining, which facilitates cross-lingual transfer as soon as possible. This involves two stages: *collection of multilingual alignment table* and *alignment injection via contrastive learning*.

Collection of multilingual alignment table Given an English monolingual corpus $\mathcal{D}$, PREALIGN extracts from $\mathcal{D}$ the collections of all unique words $\mathcal{W} = \{w\}_i^N$, where N is the number of unique words. For each word w, we translate it to all considered target languages, and denote the translation results as $T(w)$. Since there exist complex many-to-many alignment relationships between languages, PREALIGN needs to collect all possible translations. We rely on GPT-4 to collect the corresponding translations in this paper.

Alignment injection via contrastive learning After the multilingual alignment table is collected, PREALIGN initializes models' parameters using a contrastive alignment objective, which optimizes the model to produce similar representations for aligned words. Specifically, given an English word w_i and its translations across all other languages $T(w_i)$, PREALIGN firstly obtains representations of each layer for each $w \in \mathcal{S}_{w_i}$:

$$h_w^l = \text{MeanPool}(f(w, l)) \tag{1}$$

where $l = 0, 1, \cdots, L, L+1$. $f(w, l), 1 \le l \le L$ denotes of the l-th Transformer layer representations of the model's encoding of w. $f(w, 0)$ and $f(w, L+1)$ denotes the word embedding and output embedding of w, respectively. Note that since w could be tokenized to multiple subwords, PREALIGN aggregates them into a single representation using mean-pooling operator.

PREALIGN then leverages a contrastive learning objective (Khosla et al., 2021) to establish alignments between words in different languages:

$$\mathcal{L}_{\text{align}}^l = \sum_{\substack{w_j \in \mathcal{W} \\ w_i \in T(w_j)}} \log \frac{\exp(cos(h_{w_i}^l, h_{w_j}^l)/\tau)}{\sum_{w_k \in \mathcal{B}} \exp(cos(h_{w_j}^l, h_{w_k}^l)/\tau)} \tag{2}$$

where $\mathcal{B}$ is the set of all words in current mini-batch, τ is the temperature parameter. $cos(\cdot, \cdot)$ is the cosine similarity function. The final learning objective is the sum of contrastive loss of all layers:

$$\mathcal{L}_{align} = \sum_{l=0}^{L+1} \mathcal{L}_{align}^l \tag{3}$$

To prevent the initialization from being trapped in a local minima that is not suitable for the subsequent language modeling, we also add an auxiliary language modeling loss beside the contrastive objective in practice:

$$\mathcal{L}_{joint} = \mathcal{L}_{align} + \mathcal{L}_{LM} \tag{4}$$

Note that, the $\mathcal{L}_{LM}$ objective in the pre-alignment stage only serves to regularize the optimization process, rather than performing large-scale pretraining. In practice, this stage only consumes 5% pretraining data.

3.2 Maintaining Multilingual Alignment via Input-only Codeswitching

The method described previously introduces multilingual alignment information before language pretraining. However, this information may be quickly forgotten if not continuously reinforced. Inspired by prior research (Chaudhary et al., 2020; Yang et al., 2021) demonstrating that code-switching effectively promotes multilingual alignment, we propose using the code-switching technique to sustain this alignment throughout the pretraining process.

Originally, code-switching was applied to both the input sequence and the target tokens in raw data, posing no issues for pretraining encoder-only models. However, this approach exacerbates the issue of multilingual script mixing in the outputs of decoder-only models. To address this, we propose an input-only codeswitching strategy that affects only the input. The distinction between the traditional codeswitching and our input-only codeswitching is illustrated in Figure 1.

Formally, given a subword sequence $X_{<i} x_i^1 \cdots x_m^i X_{>i}$, where $X_{<i}$ and $X_{>i}$ are the subword sequences before and after the i-th word, respectively. $x_i^1 \cdots x_i^m$ is the subword sequence of the i-th words. Suppose the i-th word is substituted by $y_i^1 \cdots y_i^n$ after codeswitching, then the language modeling objective after the original codeswitching is

$$\begin{aligned} & p(X_{<i}) \cdot p(X_{>i} | y_i^1 \cdots y_i^n) \\ & \quad \cdot p(y_i^1 | X_{<i}) \cdot \prod_{j=2}^{n} p(y_i^j | X_{<i} y_i^1 \cdots y_i^{j-1}) \end{aligned} \tag{5}$$

In Equation 5, the item $p(y_i^1 | X_{<i})$ requires the model to generate words in another language given prefixes in one language. To mitigate this, input-only codeswitching modifies the objective to be

$$p(X_{<i}) \cdot p(X_{>i} | y_i^1 \cdots y_i^n) \cdot p(x_i^1 | X_{<i}). \tag{6}$$

Equation 6 changes the prediction objective of subwords in the word after codeswitching ($p(y_i^1 | X_{<i})$) to subwords in the word before codeswitching ($p(x_i^1 | X_{<i})$), therefore preventing the generation results contain scripts from other languages. In this paper, we use a codeswitching ratio of 5%.

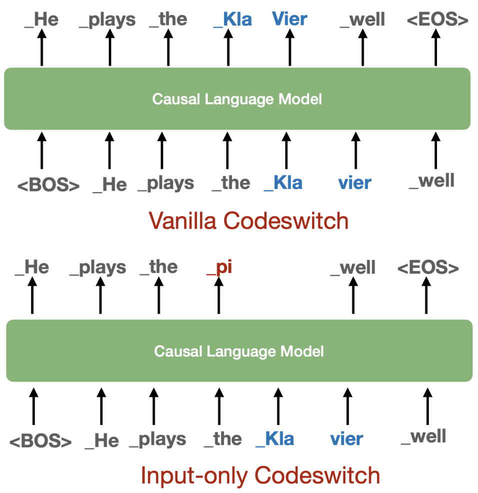

Figure 1: Comparison between vanilla codeswitching and the proposed input-only codeswitching. The original English sentence is *He plays the piano well*, and *Klavier* is the German translation of *piano*.

4 Experimental Settings

4.1 Datasets and Models

Model Configuration We adopt the GPT-2 style Transformer architecture for our model. At the defaulting setting, our model contains 12 Transformer layers with a hidden dimension of 1024. The number of total non-embedding parameters is about 150 million. We use AdamW (Kingma and Ba, 2017) optimizer with a global batch size of about 1 million tokens. The learning rate is decayed from $3e-4$ to $3e-5$ following a cosine scheduler.

Pretraining Dataset We adopt CulturaX (Nguyen et al., 2023) as the pretraining dataset. CulturaX is a multilingual pretraining corpus that has been rigorously cleaned. Due to the non-affordable computational cost to use all data for experiments, we only consider English as the source language, and Chinese (Zh), German (De), Russian (Ru), Arabia (Ar) as the target language. For English, we randomly select 10 billion tokens from CulturaX as the pretraining data. For each language to be transferred to, we randomly select 100 million tokens.

4.2 Evaluation Protocol

Target Language Modeling (LM) The first evaluation metric is the language modeling performance of target language. Given the same amount of target language data, this can reflect how well cross-lingual transfer is.

Zero-shot Cross-lingual Transfer (ZS-CLT) Another common way to evaluate model's cross-lingual ability is zero-shot cross-lingual transfer, where we finetune models on the task data in source languages, and test model's ability on the same task in target languages. We use the commonly-used XNLI (Conneau et al., 2018) dataset for ZS-CLT evaluation.

Cross-lingual Knowledge Application (CLKA) Large language models acquire extensive world knowledge from their pretraining corpora. However, significant portions of knowledge exist exclusively in texts of specific languages. It is crucial for LLMs to learn knowledge from texts in one language and apply it across other languages.

In order to evaluate models' ability to perform such cross-lingual knowledge application, we propose a setting where we attach English texts describing synthetic knowledge to the pretraining corpus, and test models' completion accuracy of the injected knowledge in the target language. Each synthetic knowledge is a triplet like (subject, relation, object), where relations are extracted from WikiData (Vrandečić and Krötzsch, 2014), and subjects and objects are artificial entities.

To better monitor the model's learning dynamics, we segmented the pretraining process into shorter periods, each consisting of 250 training steps. During each period, we incorporate various knowledge triplets into predefined templates to create sentences that encapsulate specific knowledge, which are then added to the pretraining data exclusively during that period. Following each learning period, we assess the model's knowledge retention by introducing three distractors—random named entities substituted for the original object in the knowledge statement—and evaluate the model's ability to correctly assign the highest likelihood to the correct statement. This assessment occurs immediately after each training period using the corresponding model checkpoint.

5 Experiments on Synthetic Transferring Settings

We start our evaluation on a English to Synthetic language transferring setting, which allows us to better control the relationship between the source language and target language. We first describe the construction of synthetic language and implication of the setting in Section 5.1. We then present experimental results in Section 5.2 and Section 5.3.

	#Tokens		LM (ppl. ↓)		ZS-CLT (acc. ↑)		CLKA (acc. ↑)
	En	En-Clone	En	En-Clone	En	En-Clone	En-Clone
Only Tgt	-	0.1B	-	47.2	-	-	-
Full Tgt	-	10B	-	16.2	-	-	-
Joint Training	10B	0.1B	16.1	21.6	79.8	74.9	27.7
PREALIGN	10B	0.1B	**15.9**	**16.5**	**80.1**	**79.3**	**64.6**

Table 1: Performance of PREALIGN and other methods on language modeling, ZS-CLT and CLKA. The performance of CLKA is averaged over each learning period.

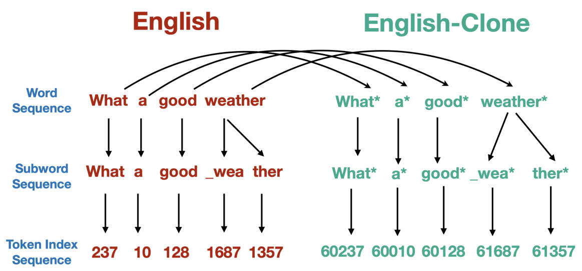

Figure 2: Illustration of the creation of English-Clone.

Finally, in-depth discussions are presented in Section 5.4, 5.5 and 5.6.

5.1 Investigating Cross-Lingual Transfer based on Cloned English

We construct a synthetic language called En-Clone, by cloning all English words by a one-to-one mapping. En-Clone shares the same linguistic properties with English, such as vocabulary distribution, grammar and syntax, yet they have no word overlapping. See Figure 2 for an illustration of the creation of En-Clone.

This synthetic setting provides many benefits. Firstly, the English to En-Clone setting arguably forms the easiest setting for testing the cross-lingual transferring ability of LLMs, since it does not involve the discrepancy of word ordering and possibly complex one-to-many/many-to-one alignments between real-world languages. Therefore, this setting can serve as a sanity-check for cross lingual transferring methods.

Secondly, since the golden alignment between English and En-Clone is trivial to get, we can easily achieve *perfect* alignment at the initialization stage by setting the input and output embedding of aligned tokens to be identical. In this way, hidden states of all intermediate layers would also be identical. This provides us a chance to analyze the upper-bound performance of our method.

5.2 Experimental Results

We present the results on LM, ZS-CLT and CLKA in Table 1. Beside Joint Training and PREALIGN, we also list the performance of Only-Tgt, where we only train the model on the same amount of En-clone data, and Full-Tgt, where we train the model on the En-clone data with the same size as full English data.

Joint Training achieves spontaneous multilingual transfer to some extent. It can be seen from Table 1 that compared to Only Tgt, Joint training achieves notable improvements on LM despite there are neither parallel signal or pivot words between English and English-clone. However, this transfer does not work well on CLKA, which is consistent with previous findings (Gao et al., 2024) that CLKA cannot be improved by multilingual pretraining.

PREALIGN improves over Joint Training on all evaluation tasks. We can also see that PREALIGN significantly outperforms Joint Training on all three evaluation tasks. On the LM evaluation, PREALIGN even achieves performance comparable to Full Tgt, despite it only uses 1% En-Clone data. This demonstrates the effectiveness of PREALIGN for facilitating cross-lingual transfer.

5.3 An in-depth investigation of CLKA

In order to better investigate the dynamic of cross-lingual knowledge transfer during pretraining, we plot the accuracy of knowledge completion of different training period. Figure 3 presents the results.

Language ability affects the rate of knowledge learning. Firstly, we can see from the top-left of Figure 3, where we test English knowledge in English language, models' knowledge completion accuracy after each learning period rapidly grows as the pretraining goes on. This indicates that the rate of knowledge learning strongly correlates with

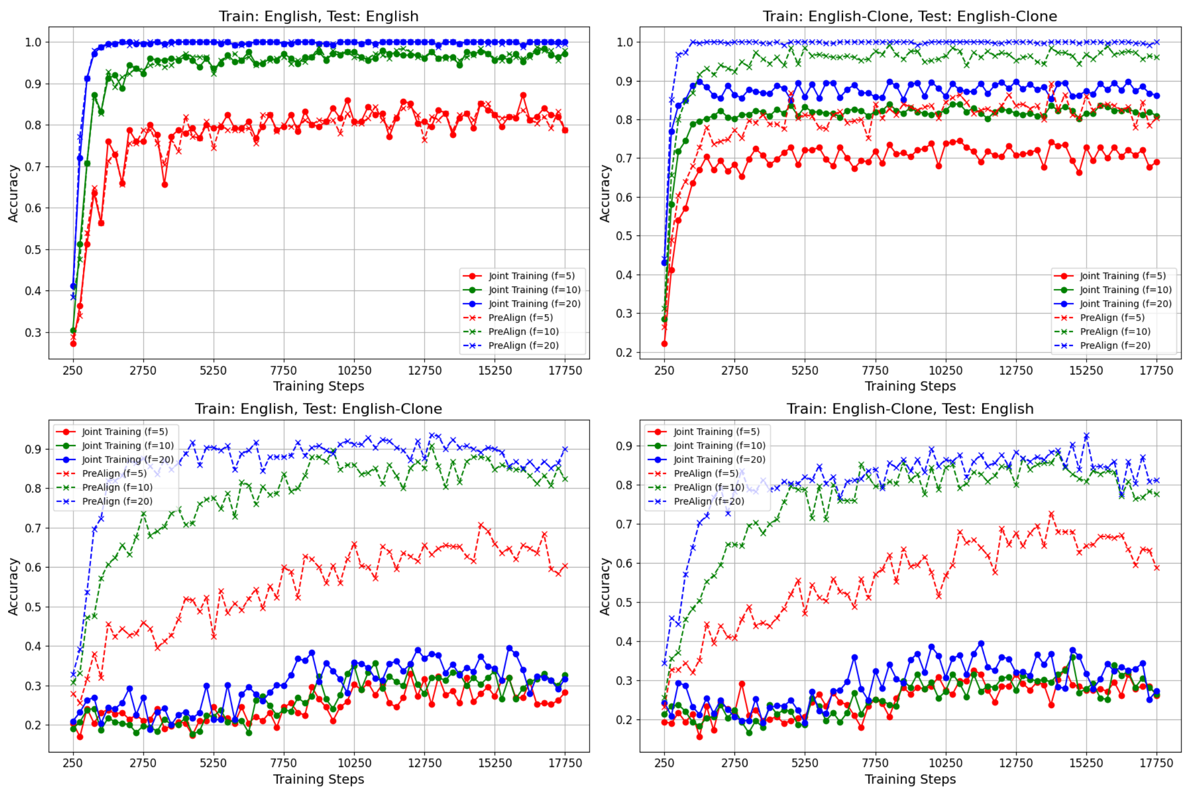

Figure 3: Knowledge application accuracy at each training period of different models. f indicates the frequency of the test knowledge.

	Joint Training	Multi-Align Init	Input-only CS	LM (ppl. ↓)	ZS-CLT (acc. ↑)	CLKA (acc. ↑)
#1	✓			21.6	74.9	27.7
#2	✓		✓	19.7	76.1	32.6
#3	✓	✓		17.1	77.8	54.5
#4	✓	✓	✓	**16.5**	**79.3**	**64.6**

Table 2: Ablations of PREALIGN. Multi-Align Init: using multilingual alignment objective to initialize LM. Input-only CS: the proposed data augmentation method by only codeswitching the input words. All reported performance are evaluated in English-Clone.

models' language modeling ability. The final performance also correlates with the knowledge frequency in the learning period as expected.

Early cross-lingual transfer enhance target language ability, facilitating knowledge learning. In the top-right of Figure 3 where we test English-Clone knowledge in English-clone language, we observe a similar trend with the top-left figure. However, the growing rate of Joint Training is slower compared to PREALIGNespecially when the frequency of knowledge is low, indicating the early alignment introduced by PREALIGN can boost target language modeling ability, therefore improving the learning of target language knowledge.

PREALIGN unlocks cross-lingual knowledge transfer. From the bottom two figures in Figure 3, we can see the CLKA ability of Joint Training is significantly weaker than PREALIGN. This renders PREALIGN a promising method for learning truly multilingual knowledge alignment.

5.4 Ablation Study

In this section, we present an ablation study of the proposed methods, including the multilingual alignment initialization and the input-only codeswitching strategy. The results is presented in Table 2.

Solely input-only CS helps LM and ZS-CLT, but not CLKA. Comparing Line #1 and Line #2, we can see that adding input-only CS to the pretraining stage can bring improvements to language

	LM	Codeswitching Ratio
Original CS	17.1	4.17%
Input-only CS	16.5	0.02%

Table 3: Comparison of the original codeswitching strategy and the proposed input-only codeswitching strategy. Not the codeswitching ratio in the table refers to the portion of random English samples that contains English-clone scripts during inferencing.

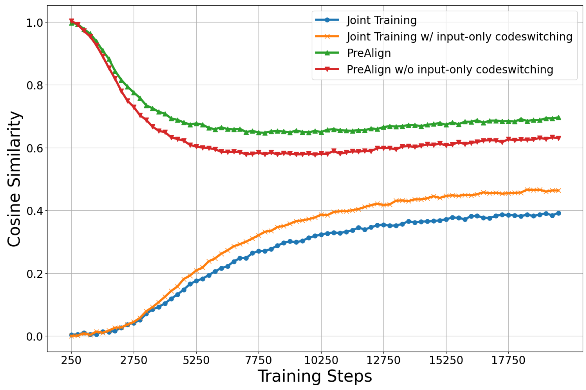

Figure 4: The evolution of word embeddings' cosine similarity between aligned words from different models.

modeling and downstream cross-lingual transferring performance, which is consistent with findings in previous works (Chaudhary et al., 2020; Yang et al., 2021). However, the improvement on CLKA is much smaller (27.7 → 32.6).

Multilingual alignment initialization significantly facilitates CLT, especially CLKA. By establishing multilingual alignment before language model pretraining, all considered metrics that evaluating cross-lingual transfer are significantly improved (Line #1 vs. Line #3 and Line #2 vs. Line #4). Notably, this brings a much better CLKA performance, highlighting the importance of early multilingual alignment for knowledge transferring.

Combining Multi-Align Init with input-only codeswitching achieves the best performance. Finally, by comparing Line #4 vs. Line #2 and Line #3, we can see the proposed two strategies all contributes to the good performance that PREALIGN achieves.

We also compare the proposed input-only codeswitching strategy with the vanilla codeswitching strategy in Table 3, in terms of both English language modeling performance and the ratio that generation results contains En-clone tokens. It can be seen that when the training time codeswitching ratio is to 5%, adopting vanilla codeswitching strategy would result in 4.17% sentences contains En-clone tokens, which would significantly decrease the generation quality in real-world settings. However, the input-only codeswitching strategy proposed in this paper effectively decrease the ratio to 0.02%, and achieves better English LM perplexity.

	LM	ZS-CLT	CLKA
Joint Training	21.6	74.9	27.7
PREALIGN			
$\beta = 25\%$	17.0	78.2	58.5
$\beta = 50\%$	16.8	78.6	60.9
$\beta = 75\%$	16.6	78.8	62.1
$\beta = 100\%$	**16.5**	**79.3**	**64.6**

Table 4: Performance of PREALIGN when using different portion of aligned word pairs. For reference, we also list the performance of Joint Training.

5.5 Multilingual alignment is maintained across pretraining.

In order to understand how the injected multilingual alignment information before pretraining evolves, we compute the similarity of aligned word embeddings at different training steps. Figure 4 illustrates the results.

Firstly, we can see that despite there are no vocabulary overlap between English and English-clone, the embedding similarity of aligned words still grows during pretraining, which is consistent with findings in previous works (Dufter and Schütze, 2020). This indicates the ability of spontaneous establishment of multilingual alignment of language models. Secondly, the aligned similarity score of PREALIGN is near perfect as designed, and despite the score decreases at the beginning of pretraining, it maintains to be significantly higher than Joint Training throughout the pretraining process. Finally, the codeswitching strategy is helpful for both Joint Training and PREALIGN, as it accelerates the increment of Joint Training's aligning similarity score, and helps slow down the decrement of PREALIGN's aligning similarity score.

5.6 Generalization to Unseen Word Pairs

In previous experiments, we assumes that we can collect translations for all words in the pretraining corpus. However, in real-world settings, this might

	LM(ppl. ↓)					ZS-CLT(acc. ↑)					CLKA(acc. ↑)				
	En	Zh	De	Ar	Ru	En	Zh	De	Ar	Ru	En	Zh	De	Ar	Ru
150M															
Joint Training	25.7	99.7	43.5	46.9	49.8	**80.6**	24.6	63.5	58.3	62.0	-	25.7	25.4	25.8	26.8
PREALIGN	**25.4**	**91.1**	**39.8**	**40.7**	**44.6**	**80.6**	**69.2**	**67.5**	**60.8**	**65.1**	-	45.7	48.2	43.4	46.0
400M															
Joint Training	20.3	79.8	32.5	34.8	39.6	82.3	65.8	65.3	56.9	63.7	-	31.2	30.5	34.1	29.7
PREALIGN	**19.9**	**75.2**	**28.3**	**30.7**	**33.6**	**82.4**	**70.0**	**69.3**	**65.6**	**68.2**	-	**50.2**	**51.0**	**49.3**	**48.9**
1.3B															
Joint Training	**15.8**	62.2	24.0	27.7	31.2	**84.3**	70.8	70.6	63.7	68.6	-	36.7	35.6	36.4	33.0
PREALIGN	16.1	**58.0**	**23.3**	**25.3**	**29.4**	83.9	**74.0**	**72.9**	**68.2**	**71.4**	-	**54.3**	**53.1**	**52.4**	**50.1**

Table 5: Performance on LM, ZS-CLT and CLKA of Joint Training and PREALIGN across different scale of models.

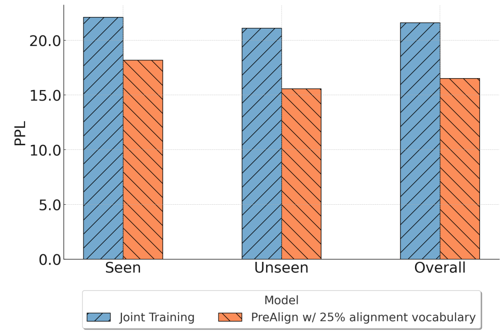

Figure 5: Language modeling perplexity on Seen and Unseen words categorized according to multilingual alignment stage.

be impractical. Therefore we present an investigation on whether we can only collect alignment table of high-frequency words, and generalize the alignment to words unseen in the alignment table.

Specifically, we sort words in our unique word set according to their frequency, and only train PREALIGN model based on the top β word alignment. Table 4 shows the results. We can see that when using the most frequent 25% words for multilingual alignment, PREALIGN can already achieve significant improvements over Joint Training. This indicates the alignment information can be generalize between words.

To better validate this, we split all words into Seen and Unseen according to their appearance during the multilingual alignment phase. We then compute the test LM perplexity of seen words and unseen words, and present the results in Figure 5. It can be seen that PREALIGN not only can effectively leverage seen words to enhance the language modeling ability, but only can generalize the alignment information to unseen words.

6 Experiments on Real-world Settings

We have presented experiments on a synthetic English to English-Clone settings. In this section, we aim to validate the effectiveness of PREALIGN under real-world settings. Specifically, we consider the transfer from English to Chinese, Russian, German and Arabia. The target languages spans four different language families and serves as good representatives of world languages. Performances of LM, ZS-CLT and CLKA is shown in Table 5.

PREALIGN are also effective under real-world scenarios. It can be seen from Table 5 that PREALIGN can still achieve substantially better performance compared to the original Joint Training method. This improvements is consistent across different model scales, rendering the effectiveness of PREALIGN in real-world scenarios.

Enlarging models is beneficial for CLKA. We can also see that although Joint Training gets near-random performance at the small scale, the performance grows with the scale of model parameters. This indicates that the ability of spontaneous multilingual alignment only appears on larger models, which is consistent with finding in Qi et al. (2023).

7 Conclusion

We present the PREALIGN framework in this paper. It advances the establishment of multilingual alignment prior to language pretraining, and maintain it throughout pretraining using an input-only codeswitching strategy. Through extensive experiments and analysis, both on synthetic and real-world settings, we demonstrate the effectiveness of PREALIGN for facilitating cross-lingual ability and knowledge transfer.

Limitations

The main limitation of this paper is scale of studied models and datasets. Although we proved the effectiveness of PREALIGN up to 1.3B models, it is still very small compared to LLMs nowadays. Whether the findings in the paper holds on larger settings still remains to be explored.

Another limitation is that we only test simple factual knowledge in this paper. In real worlds, knowledge may take more complex forms, and the effectiveness of PREALIGN on these settings need to examined.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Aditi Chaudhary, Karthik Raman, Krishna Srinivasan, and Jiecao Chen. 2020. Dict-mlm: Improved multilingual pre-training using bilingual dictionaries. *Preprint*, arXiv:2010.12566.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. XLM-E: Cross-lingual language model pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6170–6182, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. 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. 2020. 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.

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

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. *ArXiv*, abs/2404.04659.

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.

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*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning. *Preprint*, arXiv:2004.11362.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *Preprint*, arXiv:2309.09400.

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual BERT post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 210–219, Online. Association for Computational Linguistics.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Anton Schäfer, Shauli Ravfogel, Thomas Hofmann, Tiago Pimentel, and Imanol Schlag. 2024. Language imbalance can boost cross-lingual generalisation. *Preprint*, arXiv:2404.07982.

Shuaijie She, Shujian Huang, Wei Zou, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *ArXiv*, abs/2401.06838.

Henry Tang, Ameet Deshpande, and Karthik Narasimhan. 2022. Align-mlm: Word embedding alignment is crucial for multilingual pre-training. *Preprint*, arXiv:2211.08547.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy F. Chen. 2024. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *Preprint*, arXiv:2309.04766.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *Preprint*, arXiv:2402.10588.

Shijie Wu and Mark Dredze. 2020. 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.

Jian Yang, Shuming Ma, Dongdong Zhang, ShuangZhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9386–9393.

Ziqing Yang, Wentao Ma, Yiming Cui, Jiani Ye, Wanxiang Che, and Shijin Wang. 2021. Bilingual alignment pre-training for zero-shot cross-lingual transfer. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 100–105, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Wenhao Zhu, Shujian Huang, Fei Yuan, Cheng Chen, Jiajun Chen, and Alexandra Birch. 2024a. The power of question translation training in multilingual reasoning: Broadened scope and deepened insights. *Preprint*, arXiv:2405.01345.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024b. Question translation training for better multilingual reasoning. *Preprint*, arXiv:2401.07817.