Zero-shot Cross-lingual Alignment for Embedding Initialization

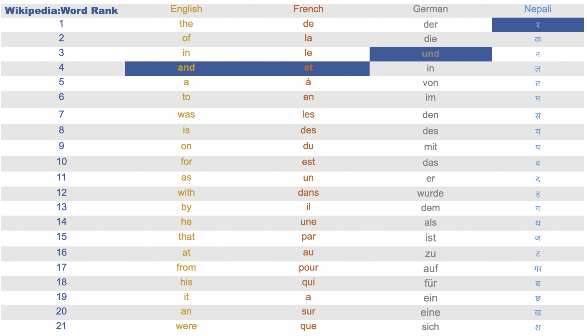
Xi Ai **Zhiyong Huang**
National University of Singapore, NUS Research Institute in Chongqing
barid.x.ai@gmail.com huangzy@comp.nus.edu.sg

Abstract

For multilingual training, we present CrossInit, an initialization method that initializes embeddings into similar geometrical structures across languages in an unsupervised manner. CrossInit leverages a common cognitive linguistic mechanism called Zipf's law, which indicates that similar concepts across languages have similar word ranks or frequencies in their monolingual corpora. Instead of considering point-to-point alignments based on ranks, CrossInit considers the same span of consecutive ranks in each language as the Positive pairs for alignment, while others out of the span are used as Negative pairs. CrossInit then employs Contrastive Learning to iteratively refine randomly initialized embeddings for similar geometrical structures across languages. Our experiments on Unsupervised NMT, XNLI, and MLQA showed substantial gains in low-resource and dissimilar languages after applying CrossInit [1].

1 Introduction

Zipf's law suggests that words with similar meanings and senses in different languages may have a similar word rank or frequency in the individual language [2]. The starting point is how Zipf's law reflects on the multilingual corpus we use. To observe it from an inspiring example, we counted words appearing in Wikipedia dumps † [3] and ranked them in order of frequency, presenting the result in Figure 1. In a short conclusion, as supported by the literature, languages are motivated by common cognitive mechanisms to form similar structural patterns across languages, thus conforming to Zipf's law (Zipf, 1949, 2013; Divjak and Caldwell-Harris, 2019).



Wikipedia:Word Rank	English	French	German	Nepali
1	the	de	der	र
2	of	la	die	क
3	in	le	und	न
4	and	et	in	ल
5	a	à	von	त
6	to	en	im	म
7	was	les	den	स
8	is	des	des	य
9	on	du	mit	प
10	for	est	das	व
11	as	un	er	द
12	with	dans	wurde	ह
13	by	il	dem	ग
14	he	une	als	थ
15	that	par	ist	ज
16	at	au	zu	ट
17	from	pour	auf	गर
18	his	qui	für	ब
19	it	a	ein	छ
20	an	sur	eine	ख
21	were	que	sich	श

Figure 1: Word Ranks of Wikipedia dumps. Words representing the meaning "and" are in a similar rank across languages.

One idea we can derive from Zipf's law is how to align words in different languages before multilingual training, e.g., aligning words in initialization. However, directly aligning words across languages based on word ranks is still challenging because it is impossible to use point-to-point alignments based on word ranks in practice. On multilingual corpus, although we use a shared vocabulary for all the languages, each language has a different local vocabulary, and words with similar concepts and meanings only have a similar rank (not identical) in different languages, as observed in Figure 1. To tackle these challenges, we can approximately separate irrelevant words to some extent. Intuitively, we can consider Positive pairs between the same spans of consecutive word ranks and Negative pairs between different ones in different languages. In this way, the multilingual model is encouraged to understand possible and impossible alignments between words across languages.

Another motivation comes from self-inference multilingual models (Ai and Fang, 2023b). Existing works have shown that a pre-trained multilingual model can infer translations for input words, where translations and input words have similar word ranks or frequencies on their monolingual corpora. If the model is more likely to under-

[1] https://github.com/baridxiai/crossInit_trial

[2] Suppose f is the frequency of a word in the corpus and r is the rank. Zipf's law indicates $f = \frac{k}{r^{\beta}}$, where k and β are constants for the corpus.

[3] Sources, scripts, and tools marked with † are listed in Table 8. Source code will be publicly available.

stand words with similar ranks across languages as cross-lingual transferable entries, we can align these words in the initialization phase to provide meta-learning supervision.

In this work, we present CrossInit, a method to iteratively initialize an embedding space for a multilingual model before formal training or pre-training on a multilingual corpus. In each initialization step, according to word ranks in each language, we randomly sample a span of consecutive ranks and use all words in this span across languages for Positive pairs. In contrast, we create Negative pairs between words inside and outside this span across languages. We show the idea in Figure 2. We experimented with Contrastive Learning to train these Positive and Negative pairs in each initialization step, but we believe there is a significant potential for the development of new alternatives. Our experimental results demonstrated that CrossInit can improve results in low-resource and dissimilar languages on unsupervised NMT, XNLI, and MLQA. We summarize contributions and findings as follows:

- We introduce CrossInit, a method for initializing embeddings to form similar geometrical structures across languages in an unsupervised manner.
- Previous works like (K et al., 2020) have provided some evidence that word frequencies alone do not contain enough information for cross-lingual learning. However, we found that words with similar frequencies might help the model in forming a similar geometric structure across languages.
- We observed that CrossInit was able to predict a possible geometric structure of the embedding space for cross-lingual transfer during the initialization phase.
- In experiments, CrossInit improved zero-shot cross-lingual transfer for low-resource and dissimilar languages.

2 Cross-lingual Initialization

CrossInit aims to iteratively initialize random embeddings before any multilingual pre-training for downstream tasks.

2.1 Step 1: Sorting

CrossInit requires word ranks in order of their frequencies/counts in each target language. To obtain these resources, we count the occurrences of each word on each monolingual corpus and sort word counts in descending order. For our implementation, we collected word counts from the monolingual Wikipedia dumps.

2.2 Step 2: Pairing

When creating Positive and Negative pairs, we randomly sample a span of consecutive ranks. We create Positive pairs inside this span using words in different languages on both sides. For Negative pairs, we use words in different languages inside and outside this span on each side. Suppose the span width is n, the language id is L_i, and the word ranks for each language are V^{L_i}. We zip pairs:

- **Positive**: $\forall i, j : \{V^{L_i}_{span}, V^{L_j}_{span}\}$
- **Negative**: $\forall i, j : \{V^{L_i}_{span}, V^{L_j}_{\notin span}\}$

where $V^{L_i}_k$ stands for the word with rank k in V^{L_i}, $span$ is a span of consecutive ranks $[(k - n/2), \ldots, (k + n/2)]$. In our experiments, we considered a quick dev experiment to find the key hyper-parameter n. We will discuss this later.

2.3 Step 3: Contrastive Learning

CrossInit follows the random initialization of embedding space for a multilingual model. For example, in our experiments with XLM (Lample and Conneau, 2019), we randomly initialized XLM and then ran CrossInit. In each CrossInit step, we execute **Step 2: Pairing** to acquire Positive and Negative pairs for Contrastive Learning. We compute dot products for paired word embeddings in Positive and Negative pairs, respectively. The two dot products are classified using labels $\{1, 0\}$ (Positive and Negative) with a binary cross-entropy loss. Let $E_{V^L_k}$ denote the embedding of word V^L_k. In practice, if word V^L_k is split into 1+ sub-tokens, we average all embeddings for this word. CrossInit

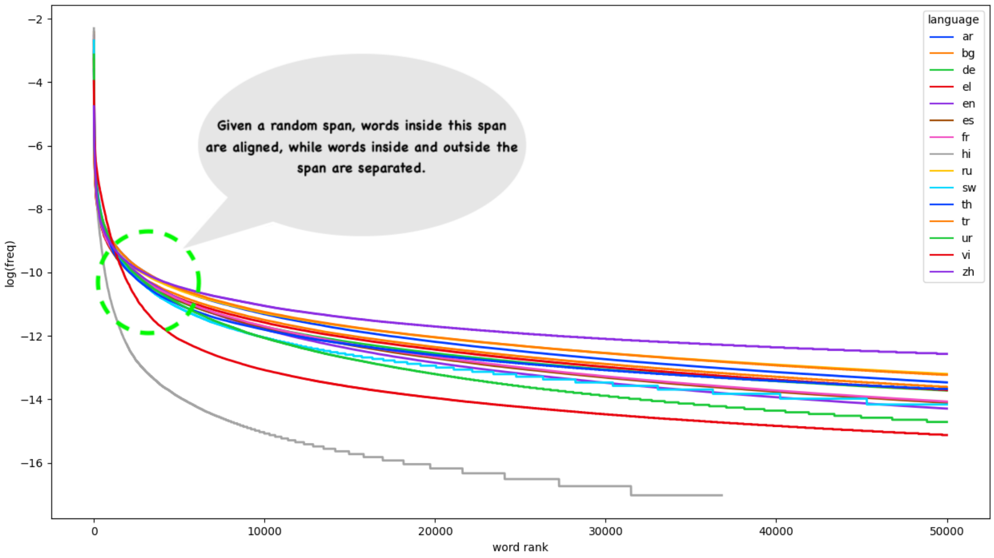


Figure 2: Example of CrossInit in each step. We assign labels 1 and 0 to Positive and Negative pairs, respectively. Then, we leverage Supervised Contrastive Learning (SCL) to train embeddings from these pairs. We suggest frequent words in creating Positive pairs because frequencies across languages differ significantly in long tails.

requires an initialization objective:

$$
\begin{aligned}
\mathcal{L}_{CrossInit} &= -\log P(1|E_{V^{L_i}_{span}} E^T_{V^{L_j}_{span}}) \\
&- \log P(0|E_{V^{L_i}_{span}} E^T_{V^{L_j}_{\notin span}}), \\
E_{V^{L_i}_{span}} &= \frac{1}{n} \sum_{k}^{V^{L_i}_{span}} E_{V^{L_i}_{k}}, \\
E_{V^{L_j}_{span}} &= \frac{1}{n} \sum_{k}^{V^{L_j}_{span}} E_{V^{L_j}_{k}}, \\
E_{V^{L_j}_{\notin span}} &= \frac{1}{|V^{L_j}_{\notin span}|} \sum_{k}^{V^{L_j}_{\notin span}} E_{V^{L_j}_{k}}.
\end{aligned}
\tag{1}
$$

Additionally, in each CrossInit step, we randomly select the span center k in **Step 2: Pairing** for different spans and i and j for different languages. We refine embeddings until the flatness of $\mathcal{L}_{CrossInit}$.

3 Analysis and Discussion

3.1 Word Ranks and Conceptions

Before analyzing CrossInit, we attempted to understand the correspondence and relevance between word ranks and conceptions. Concretely, we calculated word ranks on En and De Wikipedia dumps † and downloaded conception mappings from CLLD (List et al., 2022)† s. These conception mappings associate words to conceptions or semantics, e.g., "Etwas" (De), "Wenig" (De), "bit", and "little" are associated with the same ID 2949, and the shared conception is "A LITTLE". Our results, presented in Figure 3, demonstrated similar patterns across en and de, supporting Zipf's law.

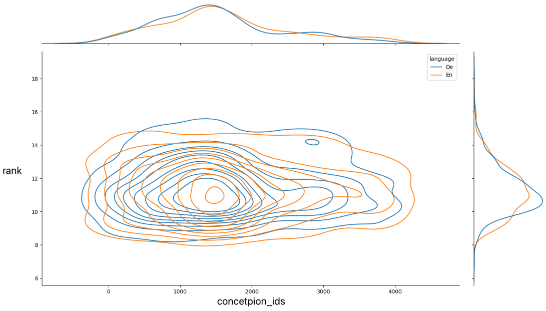


Figure 3: Relevence between word ranks and conceptions. We map words to conceptions via conception mappings from CLLD (List et al., 2022)†.

3.2 Analysis Setup

To analyze CrossInit quickly, we configured an XLM model (Lample and Conneau, 2019) and made 3 significant modifications:

- trained the model on 3 languages $\{En, De, Hi\}$.

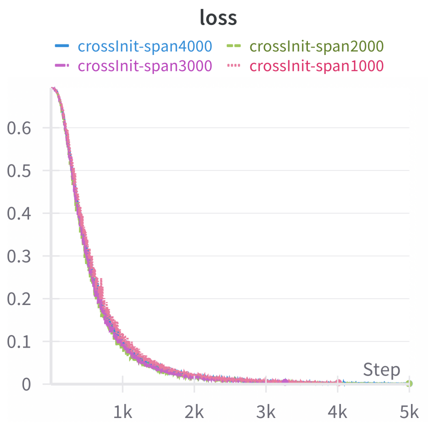


Figure 4: $\mathcal{L}_{CrossInit}$ with different span settings.

- adjusted the number of layers to 6.
- ran CrossInit for the randomly initialized embeddings.

Other settings were identical to the XLM model. In this scenario, the model is not overly parameterized for these three languages, allowing for successful unsupervised cross-lingual transfer. Additionally, Hi is distant from $\{En, De\}$, which can be potentially used to test the effectiveness of distant and low-resource languages. We used Adam optimizer with learning rate $1e-4$ for CrossInit. We marked these settings as **XLM-tiny-3** in our work.

3.3 Hyper-parameter

In **Step 2: Pairing**, there are two important hyperparameters. The first one is the random bound of the span center k. As shown in Figure 2, we observed that word frequencies are divergent and not ideally comparable in the long tail area. Therefore, in our experiments, we only considered the first 20k most frequent words in each language as candidates for Positive pairs, i.e., words ranked between 1 and 20k in each language. Those words contribute to over 80% of total word frequency in the training corpus. Note that Negative pairs are still generated from all the ranks. The second one is the width of span n in Eq. 1. We experimented with 4 settings $span = \{1000, 2000, 3000, 4000\}$ for $\mathcal{L}_{CrossInit}$. As shown in Figure 4, we found that $\mathcal{L}_{CrossInit}$ can converge for all settings.

3.4 Initialized Structure

We examined the embedding space structures CrossInit initializes in an unsupervised manner. We demonstrated PCA visualizations in Figure 5. We found that compared to Random initialization, CrossInit successfully formed some consistent patterns across languages, showing different distributions while sharing a similar geographic structure. Note that for shared tokens, we randomly chose colors for the scattered points. We observed that CrossInit was inclined to move shared tokens into a dense area.

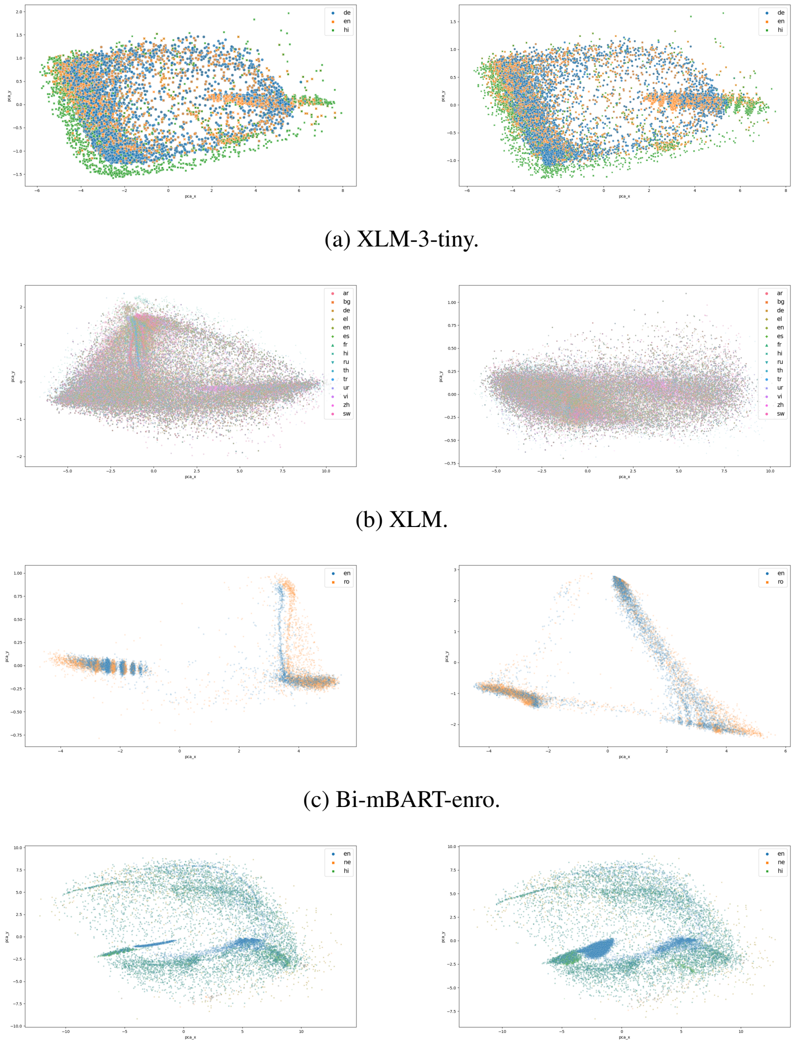
3.5 Cross-lingual Analogy Test

Ai and Fang (2023a) used the classic analogy test: "English: *King* - *Man* + *Woman* = *Queen* and German: *König-Mann+Frau* = *Königin*" to observe cross-lingual analogical phenomenon. We show the results of 3 runs in Table 1. Compared to Random initialization, CrossInit obtained positive scores for mixed languages *multi*, indicating the potential for enhancing cross-lingual transferability. This test suggested that CrossInit might improve cross-lingual transferability due to cross-lingual analogy.

In addition to the above toy example, we developed more cross-lingual analogy tests using statements from mLAMA (Kassner et al., 2021). Specifically, mLAMA offers triples in the form of (object, relation, subject), e.g., (Paris, capital, France). Similar to the toy example, we created analogy tests in the form of $object_{lang1} - subject_{lang1} + subject_{lang2} = object_{lang2}$. As shown in Table 2, we observed that CrossInit can initialize cross-lingual analogy information.

3.6 Fast XNLI Experiment

So far we have analyzed CrossInit from the perspective of cross-lingual analogies. Before applying it to standard multilingual experiments, we rendered an analysis of XNLI. We pre-trained XLM-tiny-3 for 200k steps with batch size 128 on Wikipedia dumps. We used Adam optimizer with learning rate $1e-4$. For temperature sampling (Lample and Conneau, 2019), we set $\alpha = 0.5$. After pretraining, we evaluated XLM-tiny-3 on the XNLI dataset with zero-shot settings (only fine-tuning on the English dataset). We ran experiments *3 times* and showed the average results in Table 3. Due to language diversities (i.e., Hi is distant), the model is difficult to learn zero-shot cross-lingual transferability in zero-shot settings because of "the curse of multilinguality" (Conneau et al., 2020). However, we still observed significant gains for the distant and low-resource language in all settings, which means the gain is independent of shared tokens and language similarities. We attributed this to a similar



(a) XLM-3-tiny.

(b) XLM.

(c) Bi-mBART-enro.

(d) mBART-ennehi .

Figure 5: PCA visualization for "CrossInit vs After Multilingual Pre-training". CrossInit is derived from the fact that languages are motivated by common cognitive mechanisms and results in Zipf's law with similar structural patterns, as reported in the literature (Zipf, 1949, 2013; Divjak and Caldwell-Harris, 2019). This might be the main reason that CrossInit predicts a possible structure of the embedding space for multilinguality and shows a long-lasting effect from beginning to ending.

X	cos (X , Queen)				cos(X , Königin)			
	Random	span=1000	span=2000	span=3000	Random	span=1000	span=2000	span=3000
mono: King-Man+Woman	0.00	0.90	0.90	0.93	-0.04	0.92	0.93	0.91
mono: König-Mann+Frau	-0.05	0.93	0.96	0.95	0.24	0.92	0.93	0.91
multi: King-Man+Frau	-0.08	0.85	0.85	0.91	-0.10	0.92	0.93	0.93
multi: King-Mann+Woman	0.04	0.97	0.98	0.96	0.24	0.91	0.91	0.89
multi: King-Mann+Frau	-0.05	0.96	0.97	0.96	0.16	0.92	0.93	0.91
multi: König-Man+Woman	0.00	87	0.87	0.92	0.04	0.92	0.93	0.92
multi: König-Man+Frau	-0.09	-0.78	0.80	0.89	-0.03	0.89	0.91	0.93
multi: König-Mann+Woman	0.04	0.96	0.98	0.96	0.32	0.91	0.92	0.90

Table 1: Word analogy: King - Man + Woman = Queen (German: König-Mann+Frau = Königin).

CrossInit	Lang1, 2=En, Hi	Lang1, 2=En, De	Lang1, 2=Hi, De	avg.
span=1000	0.323/4.427	0.638/3.036	0.400/4.043	0.453/3.835
span=2000	0.323/5.080	0.641/3.473	0.408/4.518	0.457/4.357
span=3000	0.341/4.060	0.650/2.795	0.411/3.686	0.467/3.514
span=4000	0.346/4.503	0.652/3.088	0.431/3.967	0.476/3.853

Table 2: Word analogy from mLAMA statements. We create analogy tests in the form of $object_{lang_1} - subject_{lang_1} + subject_{lang_2} = object_{lang_2}$ and $object_{lang_2} - subject_{lang_2} + subject_{lang_1} = object_{lang_1}$ from triples (object, relation, subject) in 3 languages. We report $cos(lang_1, lang_2)$ and $L2(lang_1, lang_2)$. This is our dev test in searching for an optimum $span$.

CrossInit	en	de	hi	avg.
Random initialization	72.69	54.05	41.99	56.24
CrossInit	74.17	58.66	47.58	60.14

Table 3: Fast XNLI Experiment. Results are reported by averaging 3 runs.

geometric structure that CrossInit forms in initialization across languages. *We skip the introduction of both XLM and XNLI here and will introduce them properly in §Experiment.*

3.7 Predictable Structure

Recall that, in §Introduction, we justified our motivation for CrossInit from cognitive mechanisms and statistics on Wikipedia. However, if there is a sufficient amount of training corpora available, the effectiveness of well-organized embeddings in initialization might be washed out due to extensive mappings between these trainable embeddings throughout training. There is an interesting question: *Can we predict a possible structure for the embedding space?* To answer this question, we compared the embedding space at CrossInit with the one after multilingual training. The idea is, if the two structures are similar, it is possible to predict an optimal structure for embeddings at initialization. To set up experiments, we considered XLM (encoder-based) and mBART (encoder-decoder-based). Figure 5 demonstrates the difference between "CrossInit" and "After Multilingual Training", showing that the embedding space keeps its shape throughout training. This suggests that CrossInit has a long-lasting effect and predicts a possible structure of the embedding space for all the languages at initialization.

4 Experiment

Our analysis, including fast XNLI experiments, demonstrated the effectiveness of CrossInit. In scaled experiments, we evaluated our method in larger-scale settings.

	Model Card
XLM	facebook/xlm-mlm-xnli15-1024
mBART	facebook/mbart-large-en-ro
Wiki	wikimedia/wikipedia (version: 20231101)
CC	cc100
wmt16	wmt/wmt16
XNLI	facebook/xnli
MLQA	facebook/mlqa
FLoRes	facebook/flores

Table 4: List of model cards.

4.1 Model and dataset cards

We used pre-configured models (including the tokenizers) and training datasets from Huggingface. Model cards are listed in Table 4.

4.2 Multilingual Training with CrossInit

Following the previous work, we set up identical XLM and mBART using the model cards and the same corpora. Instead of using pre-trained weights, we randomly initialized these models and applied our CrossInit to embeddings. For pre-training, we used the Adam optimizer (Kingma and Ba, 2015) with hyperparameters $\beta_1 = 0.9$,$\beta_2 = 0.99$, $\epsilon = 10^{-6}$,$lr = 1e-4$, and learning warm-up step 30k. We set dropout regularization with a drop rate $rate = 0.1$. The batch size was 256. We trained the model until no improvements were observed in the dev sets.

4.3 Multilingual Task

XNLI We experimented with the cross-lingual classification task on XNLI † (Conneau et al., 2018) including all 15 languages to test the general cross-lingual capabilities our method could impact. The model with an additional layer deployed on top is only fine-tuned on the En NLI dataset for the English classification, aiming at making zero-shot classification for other languages.

MLQA We experimented with MLQA† (Lewis et al., 2020) for a cross-lingual question-answering task. Given a question and a passage containing the answers, the goal is to predict the answer text span in the passage. This task involves identifying the answer to a question as a span in the corresponding paragraph. Similar to XNLI, the model is fine-tuned on the English dataset and makes zero-shot predictions for 6 other languages.

Unsupervised NMT UNMT (Lample and Conneau, 2019; Lample et al., 2018b; Liu et al., 2020) tackles bilingual translation (Bahdanau et al., 2015; Vaswani et al., 2017) on non-parallel bilingual corpora without access to any parallel sentence. In the pre-training phase, UNMT models are trained on monolingual corpora for the two languages. In the training phase, on-the-fly back-translation (Sennrich et al., 2016) performs to generate synthetic parallel sentences that can be used for training of translation as NMT (neural machine translation) is trained on genuine parallel sentences in a supervised manner.

5 Result

5.1 XNLI

Setup and Fine-tuning After multilingual training with CrossInit, we fine-tuned the models on the English NLI dataset with mini-batch size 8. We used Adam optimizer (Kingma and Ba, 2015) with $lr = 5e - 6$. Categorical cross-entropy was employed with three labels: entailment, contradiction, and neutral. Following fine-tuning, we made zero-shot predictions for the other 14 languages.

Performance We report the result in Table 5. Our method consistently improved baseline models by 1.8% (Avg). As discussed in previous models (Conneau et al., 2018; K et al., 2020; Wu and Dredze, 2019; Pires et al., 2019; Dufter and Schütze, 2020), multilinguality is crucial for this task. Therefore, we obtained some evidence that CrossInit can enhance multilinguality, helping cross-lingual transfer. Additionally, the result was consistent with our fast experiment on XNLI as we observed more gains in low-resource and dissimilar languages than in rich-resource languages. For rich-resource languages, the result was slightly improved. In this way, CrossInit is suitable for low-resource and dissimilar languages, which improves the fairness of multilingual models when considering all languages.

5.2 MLQA

Setup and Fine-tuning The setup was similar to the experiment on XNLI. We used Adam optimizer (Kingma and Ba, 2015) with $lr = 5e-5$ and linear decay of lr. Meanwhile, as suggested, we fine-tuned the model on the SQuAD v1.1 (Rajpurkar et al., 2016) dataset and then made zero-shot predictions for the 7 languages of MLQA.

Performance The results are represented in Table 6. CrossInit substantially improved the overall performance (Avg) in terms of both F1 and EM metrics, respectively. In addition, CrossInit yielded more improvements for low-resource and dissimilar languages, which was consistent with fast experiments and XNLI. Meanwhile, answers across languages are most likely to consist of analogous nouns and terms with comparable frequencies in Wikipedia. CrossInit can prompt similar embeddings for them at initialization, which could be observed from our word analogy tests in Table 2.

5.3 UNMT

Setup and Training We considered 2 language families. Specifically, we considered low-resource language pairs $Ro \leftrightarrow En$ on *newstest2016*. Meanwhile, we shared the FLoRes† (Guzmán et al., 2019) task to evaluate a dissimilar language pair $Ne \leftrightarrow English$. In the translation training phase, we used Adam optimizer (Kingma and Ba, 2015) with parameters $\beta_1 = 0.9$,$\beta_2 = 0.997$, $\epsilon = e - 9$, $warm\_up = 8000$ and $lr = 7e - 4$ (Vaswani et al., 2017). We set dropout regularization with a drop rate $rate = 0.1$ and label smoothing with $gamma = 0.1$ (Mezzini, 2018). On-the-fly back-translation (Sennrich et al., 2016) (the inference mode of the model) performed to generate synthetic parallel sentences that can be used for translation training as NMT (neural machine translation) is trained on genuine parallel sentences in a supervised manner. We reported *scareBLEU*† with default rules.

Performance In Table 7, we report *sacreBleu* † to compare with mBART (Liu et al., 2020). Given Ne's extremely low resources, we additionally included its similar language Hi in our multilingual training. We observed that CrossInit substantially improved translation for low-resource and dissimilar languages. These findings were consistent with results from other experiments. CrossInit initializes geometrical alignments and multilingual analogies,

Model	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
mBERT (Devlin et al., 2019)	81.4		74.3	70.5					62.1			63.8			58.3	
XLM	83.1	76.4	76.3	74.2	73.1	74.0	73.1	67.6	68.3	71.1	69.1	71.6	65.6	64.5	63.4	71.5
+ CrossInit	83.2	77.1	76.6	74.6	73.7	74.8	73.7	70.7	70.6	73.2	71.2	73.7	69.1	68.6	67.6	73.3

Table 5: Performance of cross-lingual classification on XNLI. Models are initialized by CrossInit and trained on 15 languages.

Model	en	es	de	ar	hi	vi	zh	Avg
mBERT	77.7 / 65.2	64.3 / 46.6	57.9 / 44.3	45.7 / 29.8	43.8 / 29.7	57.1 / 38.6	57.5 / 37.3	57.7 / 41.6
XLM	74.9 / 62.4	68.0 / 49.8	62.2 / 47.6	54.8 / 36.3	48.8 / 27.3	61.4 / 41.8	61.1 / 39.6	61.6 / 43.5
+CrossInit	75.2 / 63.1	68.8 / 50.4	63.7 / 48.1	57.5 / 39.1	51.5 / 29.4	64.5 / 42.5	63.9 / 41.2	63.4 / 44.8

Table 6: Performance of cross-lingual question answering on MLQA. We report the F1 and EM (exact match) scores for zero-shot prediction. Models are initialized by CrossInit and trained on 15 languages.

helping the model preserve language characteristics based on a similar geometric spatial structure across languages (Vulić et al., 2020). As a result, we suspected that translation might be more fluent due to the initialized language characteristics and dependencies of each language.

6 Related Work and Other Inspiration

Structural Similarity and Zipf's Law Zipf's law (Zipf, 1949, 2013; Søgaard, 2020) indicated that words or phrases appear with different frequencies, and one may suggest analogical words or phrases appear with relatively similar frequencies in other languages. In multilingual training, Wu et al. (2020); K et al. (2020); Pires et al. (2019); K et al. (2020); Sinha et al. (2021) studied structural information and found that structural similarities across languages are essential for multilinguality, where in this paper, structural similarities referred to similar ranks as Zipf's law indicated. Another interesting work was from Ai and Fang (2023a). They used translation pairs to show that phrases with similar meanings have similar (not identical) frequencies in comparable corpora. We considered spans of word ranks to alleviate the non-identical problem reported by Ai and Fang (2023a) with Contrastive Learning. Moreover, Artetxe et al. (2020) demonstrated that monolingual models trained individually on monolingual corpora eventually result in a similar structure including the embedding space. In our case, due to Positive and Negative pairs used for Contrastive Learning, embeddings were refined to a similar geometric structure in the embedding space across languages.

6.1 Language Adaptation

Lample et al. (2018a) demonstrated that bi-lingual embedding space could be obtained by merging all high-frequent words in two languages through domain-adversarial training (Ganin et al., 2016). Our method differs from these that we dynamically search for principled words available to merge and separate based on Zipf's law. Some existing works leveraged principled information from multiple words, such as dynamic cross-lingual prototypes with multiple target words (Ai and Fang, 2023b), information theory with word context (Chi et al., 2021), language domain adaptation with language-specific words (Ai and Fang, 2022).

6.2 Pre-trained Embeddings for Initialization

Qi et al. (2018) showed an effective initialization from pre-trained embeddings for downstream multilingual tasks. (Dufter and Schütze, 2020; Ai and Fang, 2023b) considered pre-trained embeddings in initialization for multilingual training with the MLM objective. We shared the same goal. Compared to those existing works, which considered embedding similarity, CrossInit used word ranks as implicit signals to align and refine embeddings. Another interesting line was initializing embeddings for transfer learning (Minixhofer et al., 2022; Kim et al., 2019), where new embeddings were properly initialized in order to be merged with pre-trained embeddings. In contrast, we focused on initializing embeddings before training. However, CrossInit might be further explored for a similar application.

7 Conclusion

We present CrossInit, a method for initializing embeddings to create similar geometric structures across languages in an unsupervised man-

Language pair	$Ro \leftrightarrow En$		$Ne \leftrightarrow En$	
mBART25	30.5	35.0	10.0 (+cc25)	4.4 (+cc25)
bi-mBART $\star$	31.5	32.9	2.3 (+Hi)	0.9 (+Hi)
bi-mBART + CrossInit	32.2	35.1	4.2 (+Hi)	2.1 (+Hi)

Table 7: Performance of UNMT. Models are initialized by CrossInit and trained on monolingual corpora in paired languages. Given Ne's extremely low resources, we use its similar language Hi in our multilingual training (+Hi). $\star$ denotes models we reimplement with model cards. +cc25 stands for using cc25 corpora.

ner. CrossInit is based on Zipf's law, a common cognitive mechanism, indicating similar concepts across languages have similar word ranks or frequencies in their monolingual corpora. To address the issue of non-identical ranks across languages, CrossInit considers a span of consecutive ranks in each language as the Positive pairs for alignment, while words inside and outside the span are Negative pairs. CrossInit further employs Contrastive Learning for Positive and Negative pairs to refine embeddings. In our analysis, we observed that CrossInit can predict a possible structure of the embedding space for cross-lingual transfer and show a long-lasting effect throughout multilingual training. In our experiments on UNMT, XNLI, and MLQA, we observed significant gains in low-resource languages and dissimilar languages after applying CrossInit.

8 Limitation

We did not conduct experiments on incomparable corpora. Incomparable corpora across languages might have different domains, which results in significant differences in word ranks as Zipf's law might be satisfied only for similar domains in practice. This might limit the scope of our method. However, multilingual models are commonly pre-trained on comparable corpora, e.g., Wikipedia and CC.

Acknowledgements

We thank several group members from the Web Information Retrieval / Natural Language Processing Group (WING) at NUS for their research discussions and proofreading of our drafts, especially Professor Min-Yen Kan.

References

Xi Ai and Bin Fang. 2022. Vocabulary-informed Language Encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4883–4891.

Xi Ai and Bin Fang. 2023a. Multilingual pre-training with self-supervision from global co-occurrence information. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7526–7543, Toronto, Canada. Association for Computational Linguistics.

Xi Ai and Bin Fang. 2023b. On-the-fly cross-lingual masking for multilingual pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 855–876, Toronto, Canada. Association for Computational Linguistics.

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

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

Table 8: Links of source.

Item	Links
WMT 2016	http://www.statmt.org/wmt16/translation-task.html
XTREME	https://github.com/google-research/xtreme
WikiExtractor	https://github.com/attardi/wikiextractor
Stanford Word Segmenter	https://nlp.stanford.edu/software/segmenter.html
HuggingFace	https://huggingface.co
Wiktionary:Frequency lists	https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists

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*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dagmar Divjak and Catherine L Caldwell-Harris. 2019. Frequency and entrenchment. *Cognitive Linguistics, eds E. Dabrowska and D. Divjak (Berlin: De Gruyter)*, pages 61–86.

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*, pages 4423–4437, Online. Association for Computational Linguistics.

Yaroslav Ganin, Hugo Larochelle, and Mario Marchand. 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17:1–35.

Francisco Guzmán, Peng Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The Flores evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6098–6111.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *8th International Conference on Learning Representations, ICLR 2020 - Conference Track Proceedings*.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *to appear in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

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

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Johann Mattis List, Annika Tjuka, Christoph Rzymski, Simon Greenhill, and Robert Forkel, editors. 2022. *CLLD Concepticon 3.0.0*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Mauro Mezzini. 2018. Empirical study on label smoothing in neural networks. In *WSCG 2018 - Short papers proceedings*.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. Wechsel: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? pages 4996–5001.

Ye Qi, Devendra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, and Graham Neubig. 2018. When and why are pre-trainedword embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2, pages 529–535.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*.

Anders Søgaard. 2020. Some languages seem easier to parse because their treebanks leak. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2765–2770.

Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in neural information processing systems*, pages 5998–6008.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2020. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4407–4418.

Shijie Wu, Alexis Conneau, 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*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

George Kingsley Zipf. 1949. Human behavior and the principle of least effort: an introd. to human ecology.

George Kingsley Zipf. 2013. *The psycho-biology of language: An introduction to dynamic philology*. Routledge.