Multilingual Word Embeddings for Low-Resource Languages using Anchors and a Chain of Related Languages

Viktor Hangya[1,2]**, Silvia Severini**[1]**, Radoslav Ralev**[3]**,**
Alexander Fraser[1,2] **and Hinrich Schütze**[1,2]
[1]Center for Information and Language Processing, LMU Munich, Germany
[2]Munich Center for Machine Learning,
[3]Technical University of Munich
{hangyav, silvia, fraser}@cis.lmu.de,
radoslav.ralev@tum.de

Abstract

Very low-resource languages, having only a few million tokens worth of data, are not well-supported by multilingual NLP approaches due to poor quality cross-lingual word representations. Recent work showed that good cross-lingual performance can be achieved if a source language is related to the low-resource target language. However, not all language pairs are related. In this paper, we propose to build multilingual word embeddings (MWEs) via a novel language chain-based approach, that incorporates intermediate related languages to bridge the gap between the distant source and target. We build MWEs one language at a time by starting from the resource rich source and sequentially adding each language in the chain till we reach the target. We extend a semi-joint bilingual approach to multiple languages in order to eliminate the main weakness of previous works, i.e., independently trained monolingual embeddings, by anchoring the target language around the multilingual space. We evaluate our method on bilingual lexicon induction for 4 language families, involving 4 very low-resource ($\leq$ 5M tokens) and 4 moderately low-resource ($\leq$ 50M) target languages, showing improved performance in both categories. Additionally, our analysis reveals the importance of good quality embeddings for intermediate languages as well as the importance of leveraging anchor points from all languages in the multilingual space.

1 Introduction

Cross-lingual word representations are shared embedding spaces for two – *Bilingual (BWEs)* – or more languages – *Multilingual Word Embeddings (MWEs)*. They have been shown to be effective for multiple tasks including machine translation (Lample et al., 2018c) and cross-lingual transfer learning (Schuster et al., 2019). They can be created by jointly learning shared embedding spaces (Lample et al., 2018a; Conneau et al., 2020) or via

mapping approaches (Artetxe et al., 2018; Schuster et al., 2019). However, their quality degrades when low-resource languages are involved, since they require an adequate amount of monolingual data (Adams et al., 2017), which is especially problematic for languages with just a few millions of tokens (Eder et al., 2021).

Recent work showed that building embeddings jointly by representing common vocabulary items of the source and target languages with a single embedding can improve representations (Wang et al., 2019; Woller et al., 2021). On the other hand, these approaches require the source and target to be related, which in practice means high vocabulary overlap. Since for many distant language pairs this requirement is not satisfied, in this paper, we propose to leverage a chain of intermediate languages to overcome the large language gap. We build MWEs step-by-step, starting from the source language and moving towards the target, incorporating a language that is related to the languages already in the multilingual space in each step. Intermediate languages are selected based on their linguistic proximity to the source and target languages, as well as the availability of large enough datasets.

Since our main targets are languages having just a few million tokens worth of monolingual data, we take static word embeddings (Mikolov et al., 2013a) instead of contextualized representations (Devlin et al., 2019) as the basis of our method, due to the generally larger data requirements of the latter. Additionally, the widely used mapping-based approaches (Mikolov et al., 2013b), including multilingual methods (Kementchedjhieva et al., 2018; Jawanpuria et al., 2019; Chen and Cardie, 2018), require good quality monolingual word embeddings. Thus, to incorporate a single language to the multilingual space in each step we rely on the anchor-based approach of Eder et al. (2021). We refer to this method as ANCHORBWEs. It builds the

target embeddings and aligns them to the source space in one step using anchor points, thus not only building cross-lingual representations but a better quality target language space as well. We extend this bilingual approach to multiple languages. Instead of aligning the target language to the source in one step, we maintain a multilingual space (initialized by the source language), and adding each intermediate and finally the target language to it sequentially. This way we make sure that the language gap between the two spaces in each step stays minimal.

We evaluate our approach (CHAINMWES) on the Bilingual Lexicon Induction (BLI) task for 4 language families, including 4 very ($\leq$ 5 million tokens) and 4 moderately low-resource ($\leq$ 50 million) languages and show improved performance compared to both bilingual and multilingual mapping based baselines, as well as to the bilingual ANCHORBWES. Additionally, we analyze the importance of intermediate language quality, as well as the role of the number of anchor points during training. In summary, our contributions are the following:

- we propose to strengthen word embeddings of low-resource languages by employing a chain of intermediate related languages in order to reduce the language gap at each alignment step,

- we extend ANCHORBWES of Eder et al. (2021) to multilingual word representations which does not take the distance between the source and target languages into consideration,

- we test our approach on multiple low-resource languages and show improved performance,

- we make our code available for public use.[1]

2 Related Work

Bilingual lexicon induction is the task of inducing word translations from monolingual corpora in two languages (Irvine and Callison-Burch, 2017), which became the de facto task to evaluate the quality of cross-lingual word embeddings. There are two main approaches to obtain MWEs: mapping and joint learning. Mapping approaches aim at computing a transformation matrix to map the

embedding space of one language onto the embedding space of the others (Ravi and Knight, 2011; Artetxe et al., 2017; Lample et al., 2018b; Artetxe et al., 2018; Lample et al., 2018a; Artetxe et al., 2019, inter alia). Alternatively, joint learning approaches aim at learning a shared embedding space for two or more languages simultaneously. Luong et al. (2015) learn sentence and word-level alignments jointly and create BWEs by modifying the Skip-gram model. The Skip-gram model is also used by Vulic and Moens (2015) who train it on a pseudo-bilingual corpus obtained by merging two aligned documents. Artetxe and Schwenk (2019) use a large parallel corpus to train a bidirectional LSTM and jointly learn representations for many languages. Most recently, transformer based large LMs are trained jointly on multiple languages using a shared subword vocabulary to obtain contextualized cross-lingual representations (Devlin et al., 2019; Conneau et al., 2020). However, large LMs require more training data than static word embeddings, thus we focus on the latter in our work.

Ruder et al. (2019) provided a survey paper on cross-lingual word embedding models and identified three sub-categories within static word-level alignment models: mapping-based approaches, pseudo-multilingual corpus-based approaches and joint methods, highlighting their advantages and disadvantages. To combine the advantages of mapping and joint approaches Wang et al. (2019) proposed to first apply joint training followed by a mapping step on overshared words, such as false friends. Similarly, a hybrid approach was introduced in (Woller et al., 2021) for 3 languages, which first applies joint training on two related languages which is then mapped to the distant third language. A semi-joint approach was introduced in (Ormazabal et al., 2021) and (Eder et al., 2021), which using a fixed pre-trained monolingual space of the source language trains the target space from scratch by aligning embeddings close to given source anchor points. We utilize (Eder et al., 2021) in our work, since it is evaluated on very low-resource languages which is the main interest of our work.

Most work on cross-lingual word embeddings is English-centric. Anastasopoulos and Neubig (2019) found that the choice of hub language to which others are aligned to can significantly affect the final performance. Other methods leveraged multiple languages to build MWEs (Kementched-

jhieva et al., 2018; Chen and Cardie, 2018; Jawan-
puria et al., 2019), showing that some languages
can help each other to achieve improved perfor-
mance compared to bilingual systems. However,
these approaches rely on pre-trained monolingual
embeddings, which could be difficult to train in
limited resource scenarios. In our work we also
leverage multiple languages, but mitigate the issue
of poor quality monolingual embeddings.

Søgaard et al. (2018) showed that embedding
spaces do not tend to be isomorphic in case of dis-
tant or low-resource language pairs, making the
task of aligning monolingual word embeddings
harder than previously assumed. Similarly, Patra
et al. (2019) empirically show that etymologically
distant language pairs are hard to align using map-
ping approaches. A non-linear transformation is
proposed in (Mohiuddin et al., 2020), which does
not assume isomorphism between language pairs,
and improved performance on moderately low-
resource languages. However, Michel et al. (2020)
show that for a very low-resource language such as
Hiligaynon, which has around 300K tokens worth
of available data, good quality monolingual word
embeddings cannot be trained, meaning that they
can neither be aligned with other languages. Eder
et al. (2021) found that mapping approaches on lan-
guages under 10M tokens achieve under 10% P@1
score when BLI is performed. In our work, we
focus on such low-resource languages and propose
to combine the advantages of related languages
in multilingual spaces and hybrid alignment ap-
proaches.

3 Method

The goal of our approach is to reduce the distance
between two languages which are being aligned
at a time. Thus instead of directly aligning the
source and target languages we incorporate a chain
of intermediate related languages in order for a
reduced distance. Our approach starts from the
source language as the initial multilingual space
and iteratively adds the languages in the chain till
it reaches the target language. We build upon the
bilingual ANCHORBWEs algorithm presented in
(Eder et al., 2021) by extending it to multilingual
setting. First, we discuss the ANCHORBWEs ap-
proach, followed by our proposed intermediate
language-based CHAINMWEs method.

3.1 ANCHORBWEs

The anchor-based method assumes that the source
language is high-resource, thus starts by training
source monolingual word embeddings with a tradi-
tional static word embedding approach, more pre-
cisely *word2vec* (Mikolov et al., 2013a). Using this
vector space it trains an embedding space for the
low-resource target language by aligning them at
the same time, this way the properties of the good
quality source space, such as similar embeddings
for words with similar meaning, is transferred to the
target space. Given a seed dictionary defining word
translation pairs, the source side of the pairs are
defined as the anchor points. Instead of randomly
initializing all target language words at the begin-
ning of the training process, the method initializes
target words in the seed dictionary using their re-
lated anchor points. The rest of the training process
follows the unchanged algorithm of either CBOW
or Skip-gram on the target language corpus. This
approach significantly outperforms previous meth-
ods in low-resource bilingual settings, as demon-
strated by strong results on both simulated low-
resource language pairs (English-German) and true
low-resource language pairs (English-Hiligaynon).
Additionally, Eder et al. (2021) shows that not only
the cross-lingual performance is improved, but the
monolingual space is of better quality compared
when the target space is trained independently of
the source language.

3.2 CHAINMWEs

We extend ANCHORBWEs by first defining a chain
of languages $C = [c_1, c_2, ..., c_n]$, starting from
the high-resource source language (c_1) and ending
at the low-resource target language (c_n), includ-
ing intermediate languages that are related to the
preceding and following nodes. As described in
Section 4, we define chains in which the lower-
resource languages are of the same language fam-
ily. The intuition is to interleave the source and
target with languages that are similar in terms of
linguistic properties. After selecting the intermedi-
ate languages, our method comprises five steps as
depicted in Figure 1:

1. As the first step ($i = 1$), we construct the ini-
tial monolingual embedding space (E_1) for
the source language (c_1) using its monolin-
gual corpus (D_1), by training a Word2Vec
(Mikolov et al., 2013a) model. We consider
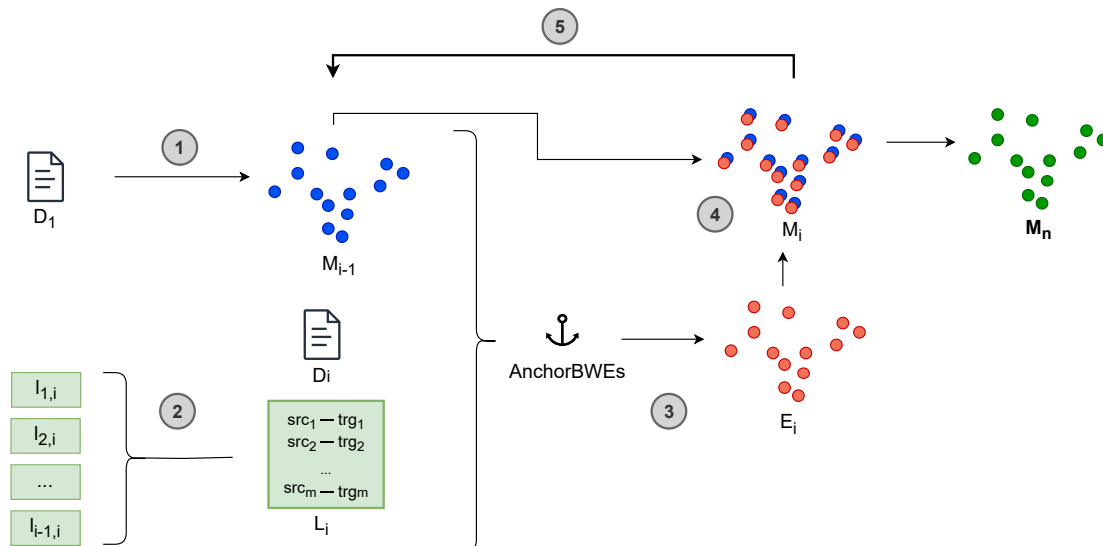this space as the initial multilingual space

Figure 1: Visual depiction of our CHAINMWES method. The resulting embedding (M_n in green) is multilingual involving all languages in the chain.

($M_1 := E_1$) which we extend in the following steps.

2. In the next step ($i = i + 1$), we collect the seed lexicon (L_i) for training embeddings for the next language in the chain (c_i) by concatenating the seed lexicons of all the languages before c_i in the chain paired with c_i. More precisely:

$$L_i = \bigcup_{k=1}^{i-1} l_{k,i}$$

where $l_{k,i}$ is the seed lexicon between languages k and i. Since Eder et al. (2021) showed that ANCHORBWES performs better as the number of available anchor points increase, our goal is to take all available anchor points already in M_{i-1}.

3. Apply ANCHORBWES using M_{i-1} as the source embedding space, D_i as the training corpus and L_i as the anchors to build embeddings (E_i) for c_i.

4. Since ANCHORBWES builds embeddings for c_i which are aligned with the maintained multilingual space, we simply concatenate them $M_i = M_{i-1} \cup E_i$.

5. Goto step 2 until the target language is reached.

By strategically integrating intermediate languages, we enrich the quality of the multilingual space by making sure that the distance between two languages at any alignment step is minimal. Our experiments show that without the intermediate languages the quality of the embeddings built by ANCHORBWES is negatively affected by the large gap between the source and target.

4 Experimental Setup

In this section, we describe the experimental setup, including the selection of languages, datasets, and model parameters used in our study.

4.1 Data

We select four language families of different geographic locations for evaluation. Figure 2 depicts the language similarities in 2D using *lang2vec* language embeddings based on their syntactic features (Malaviya et al., 2017). We discuss their relevance on the final results in Section 5. Although, we selected low-resource target and intermediate languages based on language families, we stepped over their boundaries in order to have intermediate languages related to the source language as well by considering the influence some languages had on others, e.g., during the colonial era. Our source language is English in each setup, and sort the intermediate languages based on their monolingual corpora sizes. We present the exact chains of these languages in section 5.

Austronesian We select two languages spoken in the Philippines: Tagalog as moderately and Hiligaynon as very low-resource target languages, with Indonesian and Spanish as the intermediates. Spanish being an Indo-European language is related to English. Additionally, due to colonization, it influenced the selected Austronesian languages to a varying degree. Furthermore, Indonesian, Tagalog and Hiligaynon show similarities, especially the two languages of the Philippines, due to their close proximity.

Turkic languages using the Cyrillic script. We take Kazakh as moderately, and Chuvash and Yakut as very low-resource languages. Since they use the Cyrillic alphabet and mostly spoken in Russia, we use Russian as the intermediate language. Due to Russian being high-resource, it can be well aligned with English.

Scandinavian We select Icelandic and Faroese as two very low-resource languages, with Norwegian and Swedish as the intermediates that are related to both of them and to English.

Atlantic-Congo Finally, we select Swahili as a moderately low-resource language, which has a high number of loanwords from Portuguese and German which we take as the intermediate languages. We note that we experimented with the very low-resource Zulu and Xhosa languages as well, however due to difficulties acquiring good quality lexicons for training and evaluation, we achieved near zero performance, thus we do not present them in this paper.

The embeddings were trained on Wikipedia dumps for all languages except Hiligaynon, which was trained on the corpus used in (Michel et al., 2020) due to comparison reasons. Hiligaynon is extremely low-resource, having 345K tokens in its monolingual corpus. Corpus sizes for each language are presented in Table 1. Bilingual dictionaries for training and testing are taken from the Wiktionary based resource released in (Izbicki, 2022). As mentioned in the previous section, at each iteration of our approach we take training dictionaries between the current language and all languages which are already in the multilingual vector space. Since, Izbicki (2022) only release resources for English paired with various target languages, we build dictionaries for the other language pairs through
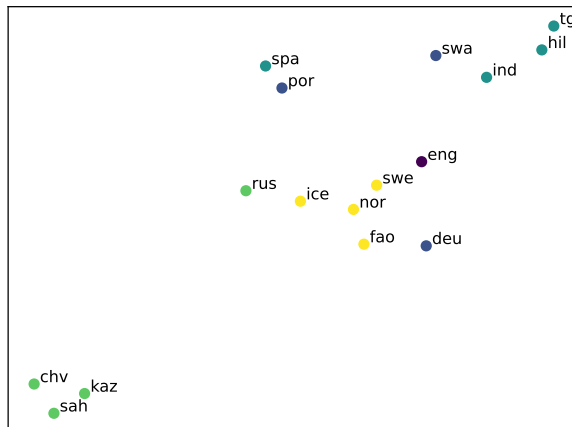

Figure 2: Visualization of language embeddings using *lang2vec* syntax features. Colors indicate different language families: Austronesian in turquoise, Turkic in green, Scandinavian in yellow and Atlantic-Congo in blue.

pivoting, more precisely:

$$
\begin{aligned}
l_{k,i} = \{(trg_{e,k}, trg_{e,i}) \mid \\
(src_{e,k}, trg_{e,k}, src_{e,i}, trg_{e,i}) \in l_{e,k} \times l_{e,i}, \\
src_{e,i} = src_{e,k}\}
\end{aligned}
$$

where $l_{e,x}$ is a dictionary between English (e) and an arbitrary language (x), while $src_{x,y}$ and $trg_{x,y}$ is a source (x) and target (y) language translation pair. Number of dictionary entries for each language pair is presented in Table 2.

4.2 Baselines and Model Parameters

We compare our approach to the mapping-based bilingual *VecMap* (Artetxe et al., 2018) and multilingual *UMWE* (Chen and Cardie, 2018) approaches. Additionally, we run ANCHORBWEs (Eder et al., 2021) as our joint alignment baseline.

We trained word2vec embeddings (Mikolov et al., 2013a) with a maximum vocabulary size of 200 000 in every setup, i.e., for the mapping-based baselines as well as in ANCHORBWEs and CHAINMWEs. The training was performed using standard hyperparameters included in the Gensim Word2Vec package (Řehůřek and Sojka, 2010): context window of 5, dimensionality of 300 and for 5 epochs, with the exception that we used minimum word frequency of 3 due to the small corpora for the target languages. Additionally, since Eder et al. (2021) showed that CBOW outperforms SG in ANCHORBWEs, we used the former in our experiments.

	Language	ISO	# tokens (M)
intermediate	English	eng	3 044
	German	deu	1 124
	Spanish	spa	836
	Russian	rus	717
	Portuguese	por	377
	Swedish	swe	252
	Indonesian	ind	128
	Norwegian	nor	127
moderate	Kazakh	kaz	32
	Tagalog	tgl	11
	Icelandic	ice	10
	Swahili	swa	9
very-low	Chuvash	chv	4
	Yakut	sah	3
	Faroese	fao	2
	Hiligaynon	hil	0.35

Table 1: Selected intermediate as well as moderately and very low-resource languages. Monolingual corpora sizes are shown in millions.

lang.	train	test	lang.	train
en-de	65 120	-	es-id	19 952
en-es	88 114	-	es-tl	26 088
en-ru	67 397	-	es-hil	4 661
en-pt	53 336	-	ru-kk	21 147
en-sv	25 214	-	ru-cv	1 212
en-id	9 868	-	ru-sah	6 913
en-no	18 916	-	pt-sw	13 197
en-kk	8 990	2 358	sv-no	15 843
en-tl	15 242	2 597	sv-is	13 749
en-is	17 004	2 568	sv-fo	6 425
en-sw	5 203	2 132	id-tl	6 089
en-cv	170	823	id-hil	1 575
en-sah	1 202	2 065	no-is	10 759
en-fo	4 505	1 786	no-fo	4 917
en-hil	1 132	200	kk-cv	160
de-pt	44 791	-	kk-sah	1 000
de-sv	34 659	-	tl-hil	1 683
de-sw	14 818	-	is-fo	5 587

Table 2: Number of unique words in the train and test dictionaries of the used language pairs.

We use the MUSE evaluation tool (Lample et al., 2018b) to report precision at 1, 5, and 10, using the nearest neighbor search. For the mapping based approaches we leverage the CSLS similarity score as it was shown to perform better by handling the hubness problem (Lample et al., 2018b). However, similarly to (Woller et al., 2021) we found that jointly trained embeddings do not benefit from the CSLS method, thus we use simple cosine similarity (NN) based search for both ANCHORBWES and CHAINMWES.

5 Results

We present our results in Table 3 split into the moderately and very low-resource language groups and sorted based on the size of available monolingual data for each target language (Table 1). Overall, the results show the difficulties of building cross-lingual word embeddings for the selected target languages, since the performance is much lower compared to high resource languages in general, which for example is around 50% P@1 for English-German on the Wiktionary evaluation set (Izbicki, 2022). Comparing the multilingual UMWE approach to the bilingual VecMap the results support the use of related languages, since they improve the performance on most source-target language pairs. However, this is most apparent on the moderately low-resource languages. The results on the very low-resource languages are very poor for the mapping-based approaches, which as discussed depend on the quality of pre-trained monolingual embeddings. In contrast, the semi-joint anchor-based approaches can significantly improve the embedding quality showing their superiority in the very low-resource setups.

Our proposed CHAINMWES method outperforms mapping-based approaches on 7 out of 8 target languages, and ANCHORBWES on 6 target languages, which is most apparent when retrieving more than one translation candidate (P@5 and P@10). Interestingly when looking at P@1, the systems are close to each other, indicating that our method improves the general neighborhood relations of the embedding space instead of just improving the embeddings of a few individual words. This is further supported in the case of Kazakh and Icelandic where UMWE outperforms CHAINMWES in terms of P@1, however it performs lower when a larger neighborhood is leveraged for the translation. This property is caused by the combination of the semi-joint anchor-based training, instead of relying on independently trained monolingual spaces, and the smaller distances between aligned languages.

When comparing moderately and very low-resource languages, we found similar trends in the two groups. In both cases CHAINMWES outperforms ANCHORBWES on 3 out of 4 languages, however in case of Hiligaynon, which has less than 1 million tokens, the results are mixed, i.e., ANCHORBWES tends to perform better when the smaller neighborhood of P@5 is considered, but it is the opposite when P@10 is measured.

	Method	Intermediate	P@1	P@5	P@10
	Moderately low-resource				
Kazakh	VecMap	-	12.37	23.06	29.42
	UMWE	rus	**14.58**	25.18	29.95
	AnchorBWEs	-	12.79	24.51	31.22
	ChainMWEs	rus	14.37	**26.90**	**33.16**
Tagalog	VecMap	-	7.63	14.94	17.76
	UMWE	esp - ind	15.59	24.69	29.08
	AnchorBWEs	-	15.38	26.57	32.01
	ChainMWEs	esp - ind	**15.90**	**28.66**	**33.79**
Icelandic	VecMap	-	4.48	9.26	12.68
	UMWE	swe - nor	**12.35**	18.23	21.02
	AnchorBWEs	-	8.77	17.94	21.67
	ChainMWEs	swe - nor	8.17	**18.75**	**23.19**
Swahili	VecMap	-	2.29	7.08	10.68
	UMWE	deu - por	**13.38**	**24.05**	**28.07**
	AnchorBWEs	-	10.23	21.44	26.22
	ChainMWEs	deu - por	10.99	20.78	25.90
	Very low-resource				
Chuvash	VecMap	-	0.00	0.00	0.00
	UMWE	rus	0.00	0.30	0.30
	AnchorBWEs	-	**0.31**	0.61	1.53
	ChainMWEs	rus	**0.31**	**0.92**	**2.75**
Yakut	VecMap	-	0.00	0.25	0.38
	UMWE	rus	0.76	1.78	2.42
	AnchorBWEs	-	**2.92**	**7.49**	**9.90**
	ChainMWEs	rus	2.03	6.98	9.14
Faroese	VecMap	-	0.00	0.51	0.63
	UMWE	swe - nor	1.01	3.42	3.93
	AnchorBWEs	-	4.09	9.20	12.26
	ChainMWEs	swe - nor	**4.21**	**9.96**	**13.67**
Hiligaynon	VecMap	-	0.00	0.00	0.00
	UMWE	esp - ind	0.00	0.00	0.00
	AnchorBWEs	-	**5.08**	**7.63**	8.47
	ChainMWEs	esp - ind	**5.08**	6.78	**10.17**

Table 3: Precision at $k \in \{1, 5, 10\}$ values for the target languages paired with English as the source in each case. The *Intermediate* column shows the languages in between the source and target (e.g., line 2 shows the chain *English→Russian→Kazakh*

Furthermore, UMWE tends to be more competitive with AnchorBWEs on the moderately low-resource languages, e.g., it performs better in case of Kazakh, while it does not improve over ChainMWEs. Overall however, we found no strong correlation between the available monolingual resources for a given language and on which target language ChainMWEs achieved the best results, since the two cases where it did not improve over the baselines are the 3rd (Yakut) and 5th (Swahili)

lowest resource languages. Looking at the visualization of language embeddings in Figure 2, the negative results on Swahili can be explained by the relatively large distance between its two intermediate pairs. Although Swahili has a large number of German and Portuguese loan words, the syntactic properties of the languages seem to be too different. Similarly, Yakut (sah) is the furthest away from Russian which could explain our negative results.

	Method	Inter.	P@1	P@5	P@10
sah	CHAINMWES	rus	2.03	6.98	9.14
	CHAINMWES	rus - kaz	1.78	5.58	8.12
fao	CHAINMWES	swe - nor	4.21	9.96	13.67
	CHAINMWES	swe - nor - ice	3.83	7.15	8.81
hil	CHAINMWES	esp - ind	5.08	6.78	10.17
	CHAINMWES	esp - ind - tgl	5.08	6.78	7.63

Table 4: Experiments on adding related moderately low-resource languages to the language chains of very low-resource languages.

5.1 Adding Moderate Resource Languages

Since some moderately low-resource languages are related to the very low-resource ones (Kazakh to Yakut[2], Icelandic to Faroese and Tagalog to Hiligaynon), we add them to the language chain in the experiments presented in Table 4. The results show, that although these languages are closely related, they do not contribute positively to the quality of the resulting MWEs. These results indicate, that the languages involved in the language-chains as intermediate steps should have good quality embeddings (the BLI performance P@5 for the Russian, Swedish, Norwegian and Spanish range between 45% and 65%), thus embedding quality is more important than language closeness. Additionally, Figure 2 shows that Tagalog is less similar to Indonesian and Spanish than to Hiligaynon, and Icelandic is less similar to Faroese than to Norwegian or Swedish.

5.2 Ablation Study

An advantage of the sequential nature of our approach is that as we add more languages to the multilingual space step-by-step, the number of potential anchor points for aligning the language next in line increases. We exploit this by accumulating all word translation pairs from the dictionaries between all languages already in the multilingual space and the currently trained language (Step 2). Although this requires dictionaries between all language pairs, we mitigated this requirement by pivoting through English. In Table 5 we present an ablation study, where we turn dictionary accumulation off, by using dictionaries only between the trained language and its preceding neighbor. The results show that this has a sizable impact on the performance. Although there are a few cases where P@1 is marginally improved (Icelandic, Swahili,

[2]Kazakh is also related to Chuvash which we omitted in these experiments due to low results on Chuvash in general.

	Method	Inter.	P@1	P@5	P@10
	Moderately low-resource				
kaz	CHAINMWES	rus	14.37	26.90	33.16
	CHAINMWES*	rus	13.67	26.19	31.22
tgl	CHAINMWES	esp - ind	15.90	28.66	33.79
	CHAINMWES*	esp - ind	13.28	23.43	28.66
ice	CHAINMWES	swe - nor	8.17	18.75	23.19
	CHAINMWES*	swe - nor	8.27	15.42	19.96
swa	CHAINMWES	deu - por	10.99	20.78	25.90
	CHAINMWES*	deu - por	11.21	20.67	24.92
	Very low-resource				
chv	CHAINMWES	rus	0.31	0.92	2.75
	CHAINMWES*	rus	0.61	1.53	3.67
sah	CHAINMWES	rus	2.03	6.98	9.14
	CHAINMWES*	rus	2.28	6.85	9.01
fao	CHAINMWES	swe - nor	4.21	9.96	13.67
	CHAINMWES*	swe - nor	3.96	8.56	12.52
hil	CHAINMWES	esp - ind	5.08	6.78	10.17
	CHAINMWES*	esp - ind	4.24	5.93	8.47

Table 5: Results of the ablation experiments, where we turn training dictionary accumulation off in CHAINMWES*, by using only the dictionary between a given language and its preceding neighbor.

Chuvash and Yakut), both P@5 and P@10 are decreased in most cases even where P@1 is improved except Chuvash. The least impacted by the accumulated dictionaries are Turkic languages which indicates their strong relation to Russian and distance from English which could stem from their different scripts. Overall, these findings align with the results of (Eder et al., 2021), who showed that the embedding quality improves as more dictionary entries are available.

6 Conclusion

In this paper we proposed CHAINMWES, a novel method for enhancing multilingual embeddings of low-resource languages by incorporating intermediate languages to bridge the gap between distant source and target languages. Our approach extends ANCHORBWES, the bilingual approach of Eder et al. (2021) to MWEs by employing chains of related languages. We evaluate CHAINMWES on 4 language families involving 4 moderately and 4 very low-resource languages using bilingual lexicon induction. Our results demonstrate the effectiveness of our method showing improvements on 6 out of 8 target languages compared to both bilingual and multilingual mapping-based, and the ANCHORBWES baselines. Additionally, we show

the importance of involving only those intermediate languages for which building good quality embeddings is possible.

Limitations

One limitation of our work is the manual selection of intermediate languages. Although, the selection and ordering of languages in the chains was straightforward based on language family information, such as Glottolog (Nordhoff and Hammarström, 2011), and available data size, it could be possible that other languages which we did not consider in our experiments are also helpful in improving the quality of MWEs. Additionally, we did not consider all possible ordering of intermediate languages, such as the order of English→Norwegian→Swedish→Faroese instead of English→Swedish→Norwegian→Faroese, in order to save resources. Thus, a wider range of chains could uncover further improvements.

Acknowledgements

References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947, Valencia, Spain. Association for Computational Linguistics.

Antonios Anastasopoulos and Graham Neubig. 2019. Should All Cross-Lingual Embeddings Speak English? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8658–8679.

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

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270, Brussels, Belgium. 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.

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.

Tobias Eder, Viktor Hangya, and Alexander Fraser. 2021. Anchor-based bilingual word embeddings for low-resource languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 227–232.

Ann Irvine and Chris Callison-Burch. 2017. A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics*, 43(2):273–310.

Mike Izbicki. 2022. Aligning word vectors on low-resource languages with wiktionary. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 107–117.

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2019. Learning multilin-

---

gual word embeddings in latent metric space: a geometric approach. *Transaction of the Association for Computational Linguistics (TACL)*, 7:107–120.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing Procrustes analysis for better bilingual dictionary induction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 211–220, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018c. 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.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st workshop on vector space modeling for natural language processing*, pages 151–159.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535.

Leah Michel, Viktor Hangya, and Alexander Fraser. 2020. Exploring bilingual word embeddings for Hiligaynon, a low-resource language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2573–2580, Marseille, France. European Language Resources Association.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation.

Tasnim Mohiuddin, M Saiful Bari, and Shafiq Joty. 2020. Lnmap: Departures from isomorphic assumption in bilingual lexicon induction through non-linear mapping in latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2712–2723.

Sebastian Nordhoff and Harald Hammarström. 2011. Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In *First International Workshop on Linked Science 2011-In conjunction with the International Semantic Web Conference (ISWC 2011)*.

Aitor Ormazabal, Mikel Artetxe, Aitor Soroa, Gorka Labaka, and Eneko Agirre. 2021. Beyond offline mapping: Learning cross-lingual word embeddings through context anchoring. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6479–6489, Online. Association for Computational Linguistics.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, Portland, Oregon, USA. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

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

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Ivan Vulic and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction.

In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, volume 2, pages 719–725. ACL; East Stroudsburg, PA.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G Carbonell. 2019. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *International Conference on Learning Representations*.

Lisa Woller, Viktor Hangya, and Alexander Fraser. 2021. Do not neglect related languages: The case of low-resource Occitan cross-lingual word embeddings. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 41–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.