Do not neglect related languages: The case of low-resource Occitan cross-lingual word embeddings

Lisa Woller and **Viktor Hangya** and **Alexander Fraser**
Center for Information and Language Processing
LMU Munich
lisa_woller@web.de,{hangyav,fraser}@cis.lmu.de

Abstract

Cross-lingual word embeddings (CLWEs) have proven indispensable for various natural language processing tasks, e.g., bilingual lexicon induction (BLI). However, the lack of data often impairs the quality of representations. Various approaches requiring only weak cross-lingual supervision were proposed, but current methods still fail to learn good CLWEs for languages with only a small monolingual corpus. We therefore claim that it is necessary to explore further datasets to improve CLWEs in low-resource setups. In this paper we propose to incorporate data of related high-resource languages. In contrast to previous approaches which leverage independently pre-trained embeddings of languages, we (i) train CLWEs for the low-resource and a related language jointly and (ii) map them to the target language to build the final multilingual space. In our experiments we focus on Occitan, a low-resource Romance language which is often neglected due to lack of resources. We leverage data from French, Spanish and Catalan for training and evaluate on the Occitan-English BLI task. By incorporating supporting languages our method outperforms previous approaches by a large margin. Furthermore, our analysis shows that the degree of relatedness between an incorporated language and the low-resource language is critically important.

1 Introduction

Cross-lingual word embeddings (CLWEs) are important for a wide range of NLP tasks including bilingual lexicon induction (BLI) (Vulić and Korhonen, 2016; Patra et al., 2019), Machine Translation (Lample et al., 2018), and cross-lingual transfer learning (Xiao and Guo, 2014; Schuster et al., 2019). Two main types of approaches to learn CLWEs are mapping methods, where a set of pre-trained monolingual embeddings is projected into another monolingual space (Mikolov et al., 2013), and joint methods, where the monolingual and cross-lingual objectives are optimized jointly (e.g., Klementiev et al., 2012; Lample et al., 2018).

Since recent research is more and more interested in dealing with low-resource languages, learning multilingual representations for low-resource languages is important as well (Conneau et al., 2018; Kementchedjhieva et al., 2018; Vulić et al., 2019). However, a lack of parallel data impairs the performance of existing strongly supervised models, which is why a lot of recent research focuses on reducing the need for parallel data (Artetxe et al., 2017; Smith et al., 2017; Artetxe et al., 2018; Conneau et al., 2018). Mapping methods are sensitive to the approximate isomorphism of embedding spaces, which is not the case for many languages (Søgaard et al., 2018). The low isomorphism of distant language pairs was tackled by learning CLWEs jointly (Lample et al., 2018; Ormazabal et al., 2019; Devlin et al., 2019). However, they rely on large monolingual corpora which are not available for many languages. Furthermore, the lack of large data leads to low isomorphism as well, since it results in low-quality monolingual embedding spaces (Michel et al., 2020). Hence, mapping methods, which rely on the assumption of approximate isomorphism cannot be fruitfully applied in many cases.

However, as there are still only poor CLWEs for many low-resource language pairs (Vulić et al., 2019), we argue that in addition to reducing requirements for training data, methods which offer opportunities precisely for low-resource setups, like leveraging data from linguistically related high-resource languages, should be considered as well. While there exist NLP systems that make use of related languages, e.g., in Machine Translation (Nakov and Ng, 2012; Nguyen and Chiang, 2017), only few work focuses on including them directly into CLWEs. An approach considering a related language in order to improve CLWEs for low-resource language pairs, including English-

Proceedings of the 1st Workshop on Multilingual Representation Learning, pages 41–50
November 11, 2021. ©2021 Association for Computational Linguistics

Occitan, has been proposed by Kementchedjhieva et al. (2018). However, using pre-trained monolingual embedding spaces, they do not take into account that monolingual representations of low-resource languages might be of poor quality, which can impede mapping performance.

In this paper, we propose a method where, in contrast to previous work, we consider both addressing the issue of monolingual embedding quality and leveraging information from a supporting language. To this end, we learn multilingual representations for a low-resource source language, a related language, and a target language in two steps: First, we train CLWEs for the low-resource language and the related higher-resource language using the *joint-align* approach by Wang et al. (2020). In that manner, the internal structure of the low-resource embeddings becomes more similar to the structure of the higher-quality related language embeddings. In the second step, we map the resulting CLWE space to the target space using the supervised *MUSE* model (Conneau et al., 2018). Since the first step results in a higher-quality embedding space for the source language, a better mapping to the target space can be found due to their higher isomorphism.

In our experiments, we learn representations for Occitan together with a related language and English. Occitan is a low-resource Romance language, which is related to high-resource languages like French and Spanish, and especially closely related to Catalan. Since particularly good CLWEs exist for each of the three related languages paired with English, we make use of monolingual data from these languages in order to obtain better representations for Occitan and English.

By evaluating our final multilingual embedding space on the Occitan-English BLI task, we show that our method improves CLWEs for these languages compared to all the baseline settings. Furthermore, we find that there are significant differences in how much of an improvement is achieved with each of the supporting languages. Investigating the impact of multiple factors, such as the pairwise linguistic relatedness of the source, target and the related languages, their BLI performance and the dataset sizes of the individual languages, we found the relatedness of the low-resource and the related language to be most influential.

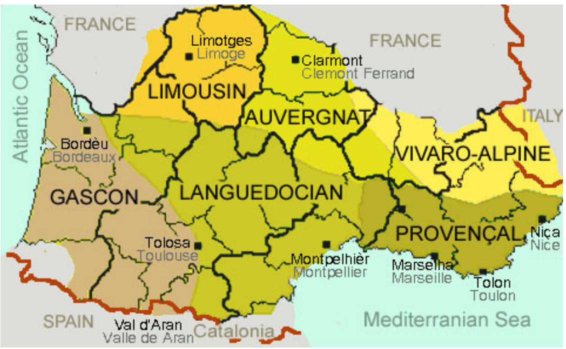

Figure 1: The Occitan-speaking area and its dialects.

2 Related work

The Occitan language Occitan is a Romance language which is spoken in the south of France, in the Aran Valley (a part of Catalonia, Spain), in a small region in Italy at the French border and in Monaco (see Figure 1[1], where the ensemble of all colored areas represents the Occitan-speaking territory). However, it is not used as a primary language in any of these countries and it only has an official status in Catalonia.

The language the closest related to Occitan is Catalan and they both belong to the Occitano-Romance languages (Bec, 1970). It is also closely related to other Romance languages, e.g., French and Spanish. Occitan is (like all Romance languages) an inflectional language which is morphologically richer than English: there is no case inflection, but it has a rather complex inflectional system for verbs. Occitan word order follows the subject-verb-object regularities and it is therefore syntactically very similar to English. However, like Spanish and Catalan, but unlike French and English, Occitan is a so-called *pro-drop* language, i.e., a conjugated verb can be used without a personal pronoun and hence the subject position does not necessarily have to be filled in an Occitan sentence.

The exact number of speakers of Occitan is not known for certain. Most sources report numbers between 1 and 10 millions, and there are significantly more people with passive knowledge of Occitan than active speakers (Cichon, 2002, pp. 19f). Furthermore, rather than one Occitan language, there are many different dialects (see Figure 1). However, the Languedocian variant is mostly used in written Occitan and thus in the Occitan Wikipedia, which we use for our experiments. Due to these factors

[1] The illustration is available at `http://lowlands-l.net/anniversary/images/occitania.jpg`.

the amount of available written digital resources is low.

CLWEs for low-resource setups A lot of research on CLWEs for low-resource languages focuses on reducing the need for cross-lingual data. Zhang et al. (2017) use adversarial training for aligning monolingual vector spaces without any bilingual signal. Conneau et al. (2018) propose an unsupervised mapping method where they combine adversarial training with a Procrustes Analysis refinement step in every iteration. Lample et al. (2018) learn CLWEs jointly for their unsupervised neural machine translation model by concatenating corpora of source and target languages and training fastText skipgram embeddings (Bojanowski et al., 2017) on this corpus. In order to combine the benefits of joint and mapping methods, Wang et al. (2020) propose an approach where they combine both methods. First, CLWEs are trained jointly on a concatenated corpus containing monolingual source and target language data. Oversharing among source and target language vocabularies is then reduced by a vocabulary reallocation step, and finally, source embeddings are mapped to the target embeddings.

However, despite the progress of unsupervised CLWE models, multiple surveys argue against focusing on fully unsupervised approaches. Firstly, giving up on every supervision signal is not necessary, since there is always a small amount of parallel data available if monolingual data is abundant (Artetxe et al., 2020). Secondly, Vulić et al. (2019) show that even the most robust unsupervised approach (Artetxe et al., 2018) cannot deal properly with multiple distant and low-resource languages.

Nevertheless, there are still a lot of languages for which even monolingual data is extremely scarce. For these languages, monolingual embeddings are usually of poor quality (Michel et al., 2020). Consequently, mapping methods are not fruitfully applicable, since they rely on high-quality monolingual embedding spaces. Adams et al. (2017) show that monolingual embedding quality of extremely low-resource languages can be improved if CLWEs for a low- and a high-resource language are trained jointly. Eder et al. (2021) propose a method for better CLWEs by using a small bilingual seed dictionary together with pre-trained monolingual embeddings of the higher-resource language for initialization. On the other hand, these approaches rely only on the source and target languages, while we show the benefits of incorporating further related languages into a multilingual space.

Leveraging related languages Besides reducing data requirements, it is also helpful to explore information from linguistically related high-resource languages in low-resource setups. This idea has, for example, been considered in Machine Translation (MT). Nakov and Ng (2012) propose a statistical MT model which requires only a small parallel corpus of the low-resource source and the high-resource target languages, and additionally a larger parallel corpus of a related high-resource language and the target language. Nguyen and Chiang (2017) introduce a transfer learning model for neural MT (NMT) where embeddings of shared words are kept when transferring the model from the original to a related low-resource language. Gu et al. (2018) train a NMT model where embeddings learned during training are computed from a universal embedding space which embed multiple languages. Thus, high-resource languages can provide support for related low-resource languages.

Leveraging information from related high-resource languages to build CLWEs for low-resource setups has only been considered in a few works until now. Multiple approaches were proposed to build representations involving more than two languages, but they either rely on pre-trained monolingual embeddings (Ammar et al., 2016; Heyman et al., 2019; Chen and Cardie, 2018; Alaux et al., 2018) or large training corpora (Devlin et al., 2019), and are thus not well suited for low-resource setups. Kementchedjhieva et al. (2018) proposed Multi-support Generalized Procrustes Analysis (MGPA) to directly incorporate related languages into CLWEs by learning a three-way alignment among English, a low-resource language, and a supporting language. They improve CLWE quality for multiple low-resource language pairs, including Occitan-English. However, unlike our method, MGPA does not consider the internal structure of the monolingual low-resource language space (since it relies on pre-trained monolingual embeddings).

3 Approach

To improve CLWEs for low-resource setups, we incorporate a related language by learning representations in two steps: First, we train CLWEs for the low-resource and a related language jointly. Subsequently, we use the resulting joint space together with a set of monolingual target language embeddings to learn the final multilingual space including the low-resource, the supporting, and the target languages. We detail the two steps below.

Joint alignment In the first step of our model, we train CLWEs for a low-resource and a related language jointly. This helps to make the internal structure of the low-resource embeddings more similar to the structure of the related language space. Since isomorphism of vector spaces is correlated with mapping performance (Søgaard et al., 2018; Ormazabal et al., 2019) and given that high-quality alignments among English and the supporting language exist, joint training of a low-resource and a related language allows for achieving a better mapping among the low-resource language and English as well.

Instead of simply building embeddings on the concatenated corpora of the two languages (Lample et al., 2018), we use the *joint-align* model proposed by Wang et al. (2020). In their approach, CLWEs are learned in three steps, which we outline in the following. First, unsupervised joint training is performed by running fastText skip-gram (Bojanowski et al., 2017) on the concatenated corpus consisting of monolingual data from both languages (L_1 and L_2). Since related languages share part of their vocabulary, these words act as a cross-lingual signal to automatically align the vectors of the two languages. However, this step suffers from vocabulary oversharing, i.e., the corpus of L_1 contains words which are only part of the vocabulary of L_2 due to noise and vice-versa, which leads to errors. To mitigate the issue, vocabulary reallocation is performed in the second step, where words are assigned to one of three sets: the vocabulary of only L_1, only L_2 or the so-called shared vocabulary. The reallocation is decided based on the frequency ratio of a given word in the two corpora. Using a threshold value, if a word is mainly appearing in the corpus of L_1 or L_2, it is allocated to the language specific vocabulary, otherwise it is kept in the shared vocabulary. Finally in step three, the language specific embeddings are refined by mapping word embeddings of L_1 to L_2 in order to improve the final CLWE quality. The resulting CLWE space thus consists of embeddings of shared words and aligned embeddings of non-shared words among the two languages.

	Occitan	French	Spanish	Catalan
Tokens	15.00	985.38	745.46	246.07
Types	0.50	4.89	4.14	2.35

Table 1: Corpora and vocabulary sizes of the extracted Wikipedia corpora (in millions).

	Oc/Fr	Oc/Es	Oc/Ca
Types overall	5.08	4.36	2.57
Types shared	0.31 (6.10%)	0.28 (6.42%)	0.28 (10.89%)

Table 2: Vocabulary sizes of the joint corpora (in millions). 'Types shared' indicates the number of shared words among the two languages; the percentage of shared words per corpus is reported in parentheses.

Mapping In the second component of our approach, we use MUSE (Conneau et al., 2018) to map the embeddings resulting from joint-align training with the monolingual target language embeddings. We use the supervised version of the MUSE model, which we find to work better for our embeddings than the unsupervised version. In addition, supervised MUSE yields good results when training with identical character strings as a supervision signal (Kementchedjhieva et al., 2018). We consider this supervision method in our experiments as well to ensure that a small training dictionary is not holding back performance.

4 Experimental Setup

Corpora and vocabulary We pursue our experiments for the low-resource Occitan language and we choose French, Spanish, and Catalan as supporting languages. Like Occitan, they are all Romance languages and hence they all have a partly shared vocabulary with Occitan as well as some similarities in morphology and syntax. French and Spanish have been chosen because they are very high-resource. Catalan has been chosen because it is the language the closest related to Occitan. Furthermore, it has been shown that for all three languages, very good CLWEs together with English can be obtained (Conneau et al., 2018).

We extract Occitan, French, Spanish, and Catalan corpora from respective Wikipedia dumps.[2] Corpora and vocabulary sizes are listed in Table 1.

[2] Available at https://dumps.wikimedia.org/. They are preprocessed using the tools available at: https://www.kdnuggets.com/2017/11/building-wikipedia-text-corpus-nlp.html.

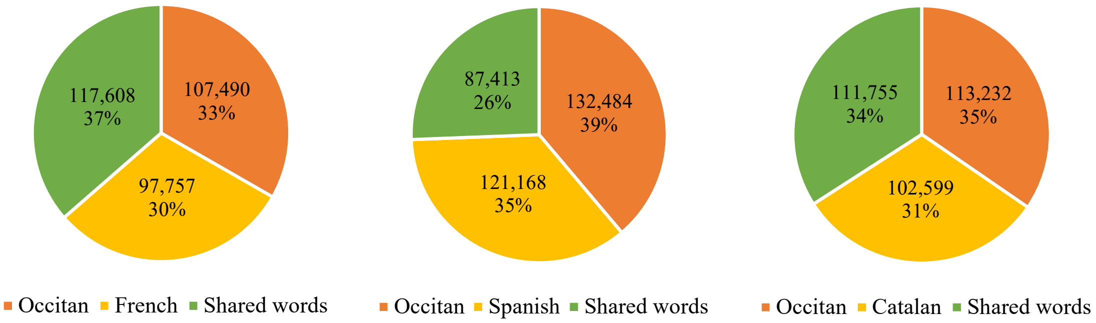

Figure 2: Number of embeddings resulting from joint-align training with Occitan and a related language.

Furthermore, in Table 2, we report vocabulary sizes of the joint corpora used for training the Occitan-related language CLWEs. We also include the number and proportion of shared words per language pair in this table.

Embeddings In all our experiments, we used the pre-trained English fastText wiki word vectors released by Bojanowski et al. (2017).[3] For Occitan, French, Spanish, and Catalan, we train our own monolingual embeddings using the Gensim version of fastText skipgram (Řehůřek and Sojka, 2010) with the same parameters used for the pre-trained English embeddings. This is to ensure that they are learned on the same corpora than the embeddings in our proposed model. The monolingual Occitan embedding space used for our baselines contains 111,353 word vectors. All the other monolingual spaces are restricted to the most frequent 200,000 words for training. The smaller number of Occitan embeddings is due to the small corpus and the threshold of at least five occurrences for a word to be considered when training fastText embeddings. The number of embeddings resulting from joint-align training with Occitan and each of the supporting languages is shown in Figure 2. Here, the proportion of Occitan, related language, and shared word vectors is illustrated.

Parameters We compare the performance of our model against multiple baselines. We use supervised MUSE (Conneau et al., 2018) and Generalized Procrustes Analysis, an extension of MUSE (GPA; Kementchedjhieva et al., 2018), as baseline models where a mapping between monolingual Occitan and monolingual English embeddings is performed. In addition, we train three baselines using Multi-support GPA (MGPA; Kementchedjhieva et al., 2018) where pre-trained monolingual embeddings from either French, Spanish or Catalan are incorporated. We use all baseline models with default parameters except the threshold for ranking candidate translation pairs, which we set to 15,000 instead of default 10,000 in all models, since it results in a better alignment.

In the first step of our proposed model, we use the joint-align model (Wang et al., 2020) for Occitan and a related language with default parameters. The only exception is that we use supervised MUSE (Conneau et al., 2018) for mapping instead of default RCSLS (Joulin et al., 2018) in order to stay consistent with the second mapping step in our model. We tested using RCSLS in both steps instead, but it did not yield a good mapping for Occitan and English. We use supervised MUSE (Conneau et al., 2018) with the same parameters as in our baseline, both within joint-align training and in the second step of our proposed model.

Evaluation task Our evaluation task is bilingual lexicon induction (BLI). We use it to evaluate the quality of our final multilingual embedding spaces, translating from Occitan to English. We also use it for evaluating the shared Occitan and related language spaces resulting from the first step of our model. For this purpose, we run the MUSE evaluation script (Conneau et al., 2018) and we report scores achieved with CSLS retrieval.

Bilingual dictionaries We extract training dictionaries for English → Occitan (En-Oc), Occitan → English (Oc-En), Occitan → French (Oc-Fr), and Occitan → Spanish (Oc-Es) from *freelang*.[4] Test dictionaries for these language pairs are ex-

[3] Available at `https://fasttext.cc/docs/en/pretrained-vectors.html`.

[4] They are based on the dictionaries available at `https://www.freelang.net/dictionary/occitan.php` (for Oc-Fr and Oc-Es see linked French and Spanish versions of freelang).

	train		test	
En → Oc	738	(580)	1,225	(1,043)
Oc → En	894	(784)	1,225	(1,027)
Oc → Es	1,638	(1,539)	1,115	(1,065)
Oc → Ca	5,511	(4,118)	1,000	(753)
Oc → Fr	8,082	(7,650)	1,086	(1,055)

Table 3: Number of word pairs in our bilingual dictionaries (number of unique source words in parentheses).

tracted from an Occitan website.[5] For English → Occitan, we use the test dictionary that Kementchedjhieva et al. (2018) extracted from this website. We clean all the dictionaries manually in a manner that they only contain 1-to-1 pairs and that source words appearing in both the initial training and test dictionary of a certain source language → target language pair are discarded from the training dictionary. For the Occitan-Catalan (Oc-Ca) language pair, there is to our knowledge no comparable bilingual dictionary available online. We therefore create our own training and test dictionaries by extracting a Catalan → French dictionary from *freelang*[6] and using it together with our Occitan → French dictionaries for mapping Occitan and Catalan words that have the same translation into French. In addition, we check the resulting dictionaries manually to avoid improper translation pairs. Dictionary sizes are reported in Table 3. Note that especially our training dictionaries vary significantly in size, since we use all the words available from our sources for every language pair. Unfortunately, due to copyright restrictions, we are not able to release dictionaries based on *freelang*. Please follow the above instructions to recreate them.

Furthermore, we use French-English, Spanish-English, and Catalan-English training dictionaries available from MUSE (Conneau et al., 2018).

5 Results

We show the results for Occitan → English BLI yielded by the baselines and our model in Table 4, Settings a-f and 1-6, respectively. Note that as the mapping direction in case of MUSE and GPA, Occitan was taken as the source and English as the target language. MGPA, however, can only be trained with the low-resource and the related language on the target language side. We evaluated the resulting CLWEs for Occitan → English afterwards. For MGPA and our model, results for incorporating either French, Spanish, or Catalan are listed separately in different columns. Furthermore, Settings 1-6 of our model vary in two more dimensions. Firstly, we employ two different subsets of the shared Occitan-related language space as source embeddings: In Settings 1-4, we use the 'full space' containing vectors of words contained in the shared and language specific (Occitan and the given related language) vocabularies. In Settings 5-6, we use a 'reduced space' containing only the vectors of shared and Occitan vocabularies. Secondly, we experiment with various bilingual supervision signals: the Occitan-English training dictionary (oc-en), the dictionary of the respective incorporated related language and English (rel-en), both training dictionaries concatenated (full), or identical character string supervision (id char). In settings where the reduced source embedding space is used, we omit training with the 'rel-en' and 'full' dictionaries, since the related language words are excluded from the embedding space.

It can be seen from Table 4 that all 18 settings of our model outperform all the baseline models, i.e., regardless of which language we use for support, which subset of the shared Occitan-related language space we employ, and which initial supervision signal we use. However, there are significant differences in performance across the various settings: Relative improvements compared to the strongest baseline (MGPA ca) are between 2.78% and 15.47%. We discuss these differences in the following.

Support from related language words Having a closer look at the numbers in Table 4, it becomes obvious that for every incorporated language, Settings 1-4 (full space) yield better scores than Settings 5-6 (reduced). The only exception is Setting 5 in the experiments with Spanish. More precisely, if related language words are considered during training, P@1 for Occitan-English BLI is up to 4.4% higher than in settings where only Occitan and shared words are included. This shows that in terms of representing the low-resource language together with English, the multilingual embedding space containing low-resource, related language, and English words is of higher quality than the embedding space with only low-resource language

[5] http://www.occitania.online.fr/aqui.comenca.occitania/dicolist.html

[6] Available at https://www.freelang.com/dictionnaire/catalan.php.

No.	Model	Src. emb.	Train dict.	P@1	P@10
a	MUSE	Oc	oc-en	15.47	31.05
b			id char	15.91	30.94
c	GPA	Oc	oc-en	15.69	32.38
d			id char	15.91	31.71

				French		Spanish		Catalan	
				P@1	P@10	P@1	P@10	P@1	P@10
e	MGPA	Oc	full	20.77	35.69	19.01	34.59	20.66	36.69
f			id char	21.10	35.58	19.34	33.26	21.33	37.68
1	Our	full space	full	27.17	44.54	27.51	46.55	36.69	**55.09**
2			oc-en	27.17	42.09	28.29	46.44	34.83	53.78
3			rel-en	27.73	43.76	27.73	46.88	**36.80**	52.68
4			id char	26.73	41.98	27.95	45.99	34.39	52.46
5		reduced	oc-en	26.05	43.16	28.63	46.50	34.06	53.43
6			id char	24.76	41.44	24.11	43.27	32.43	51.79

Table 4: Results for Occitan → English BLI achieved by various baselines and our model. The best P@1 and P@10 scores per incorporated language are underlined, while **bold** indicates the overall best. 'Full space' denotes using the ensemble of Occitan + related language + shared source embeddings for mapping, while the 'reduced' space only consists of Occitan + shared words. The 'full' training dictionary is a concatenation of the Occitan → English (oc-en) and the incorporated related language → English (rel-en) dictionaries.

and English words. The reason for this is that the related language does not only help to build better representations for the low-resource language in step 1 (joint-alignment) of our model, but it also helps to build a better mapping in step 2. This is due to the iterative refinement of MUSE which can update the initial training dictionary with good-quality related language-English word pairs as well in addition to the Occitan-English pairs.

Differences across incorporated languages Comparing performance across the different supporting languages shows that incorporating Catalan leads by far to the largest improvements (up to 15.5% P@1 compared to the strongest baseline), while French and Spanish only contribute to smaller improvements (up to 6.4% and 7.3% P@1, respectively).

We investigated multiple factors to find out where these differences come from: the quality of the Occitan-related language CLWEs, the quality of the related language-English CLWEs, and the linguistic relatedness of Occitan and an incorporated language, among others. For this purpose, we evaluate the Occitan-related language CLWEs resulting from the first step of our model as well as the embedding spaces resulting from the second step of our model on the BLI task for the respective language pairs.

No.	Language pair	P@1	P@10
1	Occitan → French	54.83	67.66
2	Occitan → Spanish	48.66	62.79
3	Occitan → Catalan	45.17	58.49
4	French → English	76.55	89.99
5	Spanish → English	77.23	90.42
6	Catalan → English	67.97	83.58

Table 5: Results for BLI. 1-3: Occitan-related language CLWEs resulting from the first step of our model. 4-6: Multilingual space resulting from the second step of our model.

The numbers in Table 5 show that the quality of the CLWE spaces mentioned above cannot explain that Catalan provides the best support for Occitan-English CLWEs. This is because results for French and Spanish are even better than scores for Catalan. Note, however, that Settings 1-3 in Table 5 are not completely comparable, since our test dictionaries do not contain the exact same word pairs for every language pair. Nevertheless, by evaluating the Occitan-related language CLWEs we show that the shared Occitan-Catalan space is not clearly better than the other two CLWE spaces in terms of BLI performance and thus this aspect is not responsible for the better quality of the final multilingual embeddings resulting from our model. The same holds for the evaluation of the related language-English

Approach		P@1	P@10
Baselines	MUSE	17.74	31.40
	GPA	17.85	32.15
	MGPA fr	21.61	34.95
	MGPA es	19.46	33.44
	MGPA ca	**22.47**	36.88
Our model	French	4.76	32.65
	Spanish	5.84	31.60
	Catalan	15.34	**43.34**

Table 6: Results for English → Occitan BLI. (Parameters for training our model are the same as in Table 4, Setting 1.)

Source word	MUSE	Our model
age	**edat**	<u>âge</u>
bird	**aucèl**	<u>oiseau</u>
bank	**banca**	bank

Table 7: Examples of English source words and their nearest neighbors in the Occitan embeddings before and after incorporating French (**bold**: correct Occitan translation; <u>underlined</u> correct French translation).

CLWEs in Settings 4-6.

The degree of linguistic relatedness to Occitan, however, is the only factor where Catalan is clearly more favorable than French and Spanish (as described in Section 2). Consequently, we can infer that it is the decisive factor for how much support an incorporated language provides for learning better Occitan-English CLWEs.

English → Occitan direction In another set of experiments, we switch source and target languages to examine how our model performs when translating from English to Occitan. For completeness, we do not only reverse the evaluation direction but the mapping direction of the used CLWEs in step 2 of our approach as well, i.e., we use the pre-trained monolingual English embeddings as source and map them to the shared Occitan-related language space resulting from the first step of our model as before. We show our results for English → Occitan in Table 6, including the results of our baseline models for the same mapping direction.

We find that, contrarily to our experiments for the Occitan → English direction, our approach cannot clearly improve P@1 on the English → Occitan BLI task. Checking the nearest neighbors of English test source words in our shared Occitan-French space reveals that it is very French-centric. In many cases, a French word is retrieved as the nearest neighbor of an English word, as shown in Table 7. This problem does not occur in the baselines due to no shared embeddings between languages. On the other hand, the phenomenon affects other multilingual models with shared vocabularies as well, such as mBERT (Devlin et al., 2019), which are mainly used for downstream tasks, e.g., zero-shot cross-lingual transfer learning. To mitigate the issue, we experimented with excluding either French only or French only and shared words from the translation candidates, respectively. However, it did not solve the issue, since the shared vocabulary includes a large number of relevant French and Occitan words, which leads to either noise or missing Occitan words depending on their inclusion as translation candidates.

On the other hand, P@10 scores achieved by our model are comparable and even significantly higher in case of Catalan than the baseline scores. This indicates that although not being the top 1 retrieved translation, the correct Occitan translation can be found in the near neighborhood of an English source word, indicating the good quality of our CLWEs. Consequently, our embeddings are still useful for various downstream tasks in the English → Occitan direction. For instance, when using them for cross-lingual transfer learning, e.g., classifying Occitan texts using a model trained on English, noise in the Occitan target space stemming from the related language vocabulary is not an issue, since the inputs to be classified are well-formed Occitan sentences.

6 Conclusion

In this paper, we presented a model for improving CLWE quality in low-resource setups by learning multilingual embedding spaces with a related language. To this end, a multilingual embedding space containing the low-resource source language, a related language, and the target language words is learned in two steps: first joint training of low-resource and related language embeddings; and second mapping the resulting CLWEs to a target language space. We pursued our experiments for the low-resource language Occitan with support from French, Spanish, or Catalan in different settings. We showed that our method improves the quality of CLWEs for these languages compared to both bilingual and multilingual baselines, especially when Catalan, the closest related language to Occitan, is incorporated (up to 15.5% P@1 improvement). Investigating multiple factors, we found that the degree of linguistic relatedness of the low-resource and the incorporated language is the most decisive for how much support a language provides. Our work indicates that novel approaches should not only focus on learning better representations using small corpora but also on incorporating data from related languages.

Acknowledgments

The work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 640550) and by German Research Foundation (DFG; grant FR 2829/4-1).

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 EACL*, pages 937–947.

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2018. Unsupervised Hyper-alignment for Multilingual Word Embeddings. In *Proceeding of IRLC*.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR*, abs/1602.01925.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL*, pages 451–462.

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 ACL*, pages 789–798.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of ACL*, pages 7375–7388.

Pierre Bec. 1970. *Manuel pratique de philologie romane*. Picard.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Xilun Chen and Claire Cardie. 2018. Unsupervised Multilingual Word Embeddings. In *Proceedings of EMNLP*, pages 261–270.

Peter Cichon. 2002. *Einführung in die Okzitanische Sprache*. Romanistischer Verlag.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In *Proceedings of ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL: HLT*, pages 4171–4186.

Tobias Eder, Viktor Hangya, and Alexander Fraser. 2021. Anchor-based Bilingual Word Embeddings for Low-Resource Languages. In *Proceedings of ACL-IJCNLP*, pages 227–232.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of NAACL: HLT*, pages 344–354.

Geert Heyman, Bregt Verreet, Ivan Vulić, and Marie-Francine Moens. 2019. Learning Unsupervised Multilingual Word Embeddings with Incremental Multilingual Hubs. In *Proceedings of NAACL-HLT*, pages 1890–1902.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of EMNLP*, pages 2979–2984.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing Procrustes analysis for better bilingual dictionary induction. In *Proceedings of CoNLL*, pages 211–220.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING*, pages 1459–1474.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of EMNLP*, pages 5039–5049.

Leah Michel, Viktor Hangya, and Alexander Fraser. 2020. Exploring bilingual word embeddings for Hiligaynon, a low-resource language. In *Proceedings of LREC*, pages 2573–2580.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Preslav Nakov and Hwee Tou Ng. 2012. Improving statistical machine translation for a resource-poor language using related resource-rich languages. *Journal of Artificial Intelligence Research*, 44:179–222.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of IJCNLP*, pages 296–301.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of ACL*, pages 4990–4995.

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 ACL*, pages 184–193.

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.

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 NAACL: HLT*, pages 1599–1613.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of ACL*, pages 778–788.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of EMNLP-IJCNLP*, pages 4407–4418.

Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of ACL*, pages 247–257.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *Proceedings of ICLR*.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of CoNLL*, pages 119–129.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of ACL*, pages 1959–1970.