Isomorphic Cross-lingual Embeddings for Low-Resource Languages

Anonymous ACL submission

Abstract

Recent research in cross-lingual representation learning has focused on offline mapping approaches due to their simplicity, computational efficacy, and ability to work with minimal parallel resources. However, they crucially depend on the assumption of embedding spaces being approximately isomorphic, which does not hold in practice, leading to poorer performance on low-resource and distant language pairs. In this paper, we introduce a framework to learn cross-lingual word embeddings, without assuming isometry, for low-resource pairs via joint exploitation of a related higher-resource language. Both the source and target monolingual embeddings are independently aligned to the related language, enabling the use of offline methods. We show that this approach successfully outperforms other methods on several low-resource language pairs in both bilingul lexicon induction as well as eigen value simialrity.

1 Introduction

In a world with over 7000 spoken languages, out of which nearly 43% are endangered, there is an acute need for accurate machine translation systems to ensure equal access of resources in a predominantly digital world. Although machine translation (MT) has shown remarkable progress over the last few years, propelled by advances in neural language modelling, this success has been mainly confined to major world languages. However, a significant proportion of languages are endangered or otherwise have a very scarce amount of digital resources which presents serious challenges for training MT systems. To ensure greater accessibility to these resources, there is therefore an acute need for MT methods that can deal with low-resource languages.

Rather than traditional expert-guided feature engineering, neural MT (NMT), like deep neural architectures more generally, require notoriously large data sets from which to extract features automatically in the context of hidden layers; for example with recurrent (Cho et al., 2014; Schmidhuber and Hochreiter, 1997), and attention mechanisms (Bahdanau et al., 2014). It is for this reason that the most impressive results (e.g., (Liu et al., 2020a; Barrault et al., 2019)) come from languages with large scale digital resources (and notably, parallel corpora) with which to train them. This is, however, not the case for most minority languages.

Recently, there have been significant improvements in semi and completely unsupervised NMT systems; notably with denoising auto-encoders (Cheng, 2019), iterative back-translation (Hoang et al., 2018), and initialisation via weak translation models sharing important cross-lingual information (Lample et al., 2017). In this work we focus on the third idea, namely Cross-Lingual Word Embeddings (CLWEs). As CLWEs represent words from multiple languages in a shared vector space, they are key in promoting language sharing across low and high-resource languages which would allow current systems to overcome the data scarcity problem. Most current methods fall into one of two categories: 1) *mapping methods* which independently map monolingual word embeddings by learning a linear transformation matrix to project them into a shared space with very little supervision (Artetxe et al., 2018a; Mikolov et al., 2013) or 2) *joint methods* which learn word representations jointly using parallel corpora thus requiring a strong cross-lingual signal (i.e. parallel resources). (Gouws et al., 2016; Luong et al., 2015)

As mapping methods use transformation matrices to align embedding spaces they make the crucial assumption that, regardless of domain or linguistic differences, these spaces are *approximately isomorphic* i.e. they share a similar structure. It has been shown (Søgaard et al., 2018; Vulić et al., 2020) that this assumption does not hold in general and therefore the benefit of mapping methods requiring little

to no cross-lingual signal in low-resource scenarios can no longer be taken advantage of directly.

In this paper, we address the limitations outlined above by proposing an alternative method to learn CLWEs for low-resource and distant languages. Unlike earlier methods, we combine the benefits from both mapping and joint-training methods to develop high-quality, isomorphic embeddings. In our proposed framework, we maintain the low level of supervision as obtained by mapping methods while still guarding the isomorphic embeddings achieved by joint-training by independently aligning source and target embeddings to a related higher-resource language. We apply our method in several low-resource settings and conduct evaluations on bilingual lexicon induction and eigenvalue similarity. Our experiments show that, despite no additional source-target parallel data, our approach outperforms conventional mapping and joint-training methods on both evaluation metrics. The main contributions of this work can be outlined as the following:

- We introduce a novel framework combining mapping and joint methods to learn isomorphic cross-lingual embeddings for low-resource language pairs.
- We successfully employ CLWEs in challenging, low-resource scenarios without the use of explicit source-target parallel data.
- We achieve significant gains over state-of-the-art methods in both bilingual word induction as well as eigenvalue similarity.

2 Related Work

Cross-Lingual Word Embeddings CLWEs aim to represent words from several languages into a shared embedding space which allows for several applications in low-resource areas such as transfer learning (Peng et al., 2021) and NMT (Artetxe et al., 2018c). Largely, there are two classes of approaches to learn CLWEs: **mapping** and **joint** methods. While the former aims to map molingually learnt embeddings together, the latter simultaneously learns both embedding spaces using some cross-lingual supervision (i.e. a cross-lingual signal). Common approaches to achieve this cross-lingual signal come from parallel corpora aligned at the word (Luong et al., 2015) or sentence level (Gouws et al., 2015). In addition to this, later methods proposed the use of comparable corpora (Vulić and Moens, 2016) or large bilingual dictionaries (Duong et al., 2016) as a form of supervision. For a more detailed survey of methods and limitations of CLWEs, the reader is referred to (Ruder et al., 2019).

Offline Mapping As mapping methods map monolingual embedding spaces together, instead of relying on a cross-lingual signal (such as in joint methods) they work by finding a transformation matrix that can be applied to the individual embedding spaces. In the case of supervised learning, a large bilingual dictionary would have been used as supervision however (Artetxe et al., 2018b) gets rid of this required via a self-learning strategy. Their approach is based on a robust iterative method combined with initialisation heuristics to get state-of-the-art performance using offline mapping. Most of these methods align spaces using a linear transformation- usually imposing orthogonality constraints- in turn assuming that the underlying structure of these embeddings are largely similar. Several works (Søgaard et al., 2018; Vulić et al., 2020) have shown that this assumption does not hold when working with non-ideal scenarios such as low-resource or typologically different language pairs. In order to mitigate this assumption, (Mohiuddin et al., 2020) learn a non-linear map in a latent space, (Nakashole, 2018) uses maps that are only locally linear, and (Glavaš and Vulić, 2020) propose to learn a separate map for each word. However these are supervised methods, meaning they suffer from limitations of hubbness and isomorphism as outlined in (Ormazabal et al., 2019). To address these limitations, (Ormazabal et al., 2021) proposes a method in which they fix the target language embeddings, and learn a new set of embeddings for the source language that are aligned with them using self-learning. Their method outperforms current mapping, joint, as well hybrid methods on the MUSE dataset (Conneau et al., 2018). Due to the unavailability of source code, we were not able to directly compare results obtained by their method but as we will report later, our method obtains strong performance across a number of low-resource language pairs.

Joint-Training The fundamental limitations of offline methods are not faced by joint-training methods if there is a strong cross-lingual signal available (Ormazabal et al., 2019). In practice, however, we don't always have access to such forms of

supervision therefore recent works have attempted to reduce the supervision level so as to preserve the isomorphism achieved by joint methods while still being as widely applicable as mapping methods. (Lample et al., 2018) use concatenated monolingual corpora in different languages and learn word embeddings over this constructed corpus, using identical words as anchor points. Further extending their work, (Wang et al., 2020) effectively combined joint and mapping based methods in their framework "joint-align" however their method was not tested on distant language low-resource pairs. In their work, they use fully unsupervised joint initialisation as the first step, vocabulary reallocation where they "unshare" some vocabulary to better align them, and lastly they perform a refinement step using off-the-shelf alignment methods. As our experiments will show, we supersede (Wang et al., 2020) in BLI across all low-resource language pairs considered.

3 Methodology

Given two embedding spaces, X and Y, our goal is to align them together without any direct parallel data between them and without assuming orthogonality/structural similarity. In order to do this, let us consider a third embedding space, Z, of a language related to the source X. Furthermore, let there also be sufficient parallel data between Y and Z to jointly learn their aligned embedding spaces. Our approach first aligns the spaces X and Z using an unsupervised offline mapping method (Artetxe et al., 2018b). (Vulić et al., 2020) find that for typologically similar languages that have in-domain monolingual corpora, isomorphism in their learnt vector spaces in preserved. To that end, due to the linguistic similarities between X and Z we may perform offline mapping. Figure 1 shows a visualisation of how these two embedding spaces are aligned using an induced seed dictionary as per (Artetxe et al., 2018b). For further details about the offline alignment, the reader is referred to read the original paper.

Once the spaces X and Z are aligned, we wish to align Y and Z as well. Due to the typological differences between the two languages, we can no longer assume isometry of their embedding spaces therefore can no longer use offline mapping methods. However, due to higher-resource nature of Z, we have access to parallel corpora between Y and Z. This allows us to apply joint-training approaches (Luong et al., 2015) to simultaneously learn their embeddings. As found in (Ormazabal et al., 2019), under ideal conditions of having parallel data, joint-training approaches produce isomorphic embeddings that perform better than their offline counterparts in bilingual lexicon induction. As shown in Figure 2, we can now produce two embedding spaces, Source aligned to Related and Target aligned to Related while preserving isomorphism. As a final step in our alignment framework, we use the $Z-$aligned embedding spaces, $\tilde{X}$ and $\tilde{Y}$, to induce the final cross-lingual word embedding spaces. Now that both X and Y are projected onto Z, they share structural similarity which permits the use of offline mapping on $\tilde{X}$ and $\tilde{Y}$. Figure 2 shows the complete alignment framework to produced the resultant isomorphic embedding spaces.

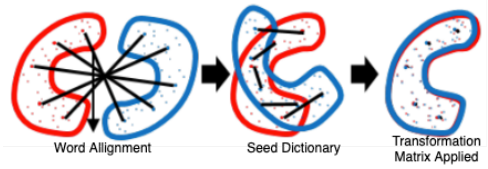


Figure 1: Toy visualisation of mapped cross lingual embedding spaces with red representing one language and blue the other

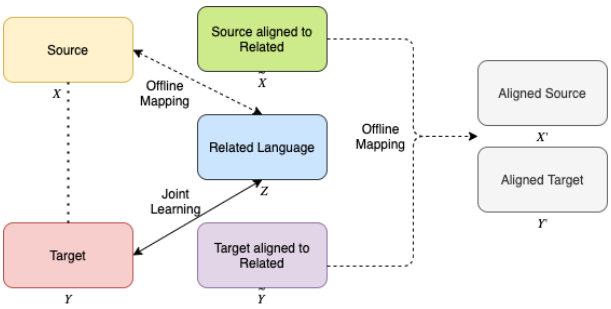
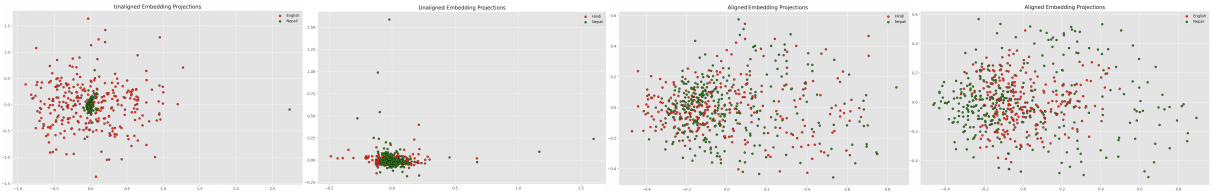


Figure 2: Visualisation of our proposed alignment method in context; dotted lines represent lack of parallel data between language pairs

Our proposed framework can be summarised in the following steps:

1. For a source-target pair, choose a related higher-resource language to the low-resource target such that there is sufficient source-related parallel data.

2. Use mapping to align related and target language into a shared embedding space. Due to their relatedness, these resultant embeddings



(a) Unaligned En,Ne embeddings (b) Unaligned Hi,Ne embeddings (c) Hi-En aligned using offline mapping (d) Ne-En aligned using our proposal

Figure 3: Embedding Projections using PCA for English-Nepali with Hindi as the related language. Green represents Nepali and Red represents English/Hindi depending on the figure

remain isomorphic as the assumption in mapping methods hold true.

3. Use joint training to map related and source language into a shared embedding space using the higher-resource parallel data between them. As this is the highest level of supervision possible, we ensure that the embedding spaces remain isomorphic.

4. Lastly, map the aligned-source and aligned-target embeddings using unsupervised mapping methods as they are now isometric in nature following the alignment to the related language for both the source and target.

This framework uses the low cross-lingual signal utilised by mapping techniques while still maintaining the isomorphism of the resultant embedding spaces as in joint approaches. This is achieved by exploiting the existing isomorphism between embeddings as much as possible by pre-aligning the spaces via a pivot-language. However, unlike pivot-based MT we do not compound errors across embedding spaces due to the final refinement step done by mapping the aligned embeddings into their shared cross-lingual space. In Figure 3, the embedding projections have been illustrated for the language pair English-Nepali with Hindi as the related language. Before performing any alignment on the monolingual embeddings, we note that Nepali and Hindi are far more structurally similar than English and Nepali as seen by Figures 3a and 3b. Upon using offline mapping on Hindi and Nepali embeddings, we obtain a well-aligned cross-lingual space as shown in Figure 3c. This allows us to construct the final alignment of Nepali and English embeddings in Figure 3d.
With this pipeline, we are able to target a large group of low-resource languages which belong to higher-resource language families for instance, English-Nepali via Hindi. Linguistically, Nepali and Hindi are quite similar as they share the same script and also have 80% of subword tokens in common when using a shared BPE vocabulary of 100k subword units (Lample and Conneau, 2019). In this work, we perform experiments on several low-resource language pairs to show the effectiveness of our approach in various language families-specifically we look at Uralic, Indo-European, and Romance languages.

Our goal is not to fully replace current methods of learning cross-lingual word representations but to aid them in the area of low-resource languages. As shown by (Ormazabal et al., 2019), depending on the type of resources available as well as the languages considered, different methods can be preferred. While current approaches perform well for several languages and resource levels (Ormazabal et al., 2021), their performance still leaves room for improvement in the low-resource, typologically diverse area. Despite the simplicity of our method, our experiments show that we perform competitively on quality as well as degree of isomorphism across all low-resource pairs considered. Due to the reliance on a sufficiently resourced related language, our method is not applicable to every low-resource pair however referring to the task of related-language NMT we see that there is indeed a large group of languages that could benefit from this approach.

4 Experimental Design

In this section we discuss the datasets used, training settings for different configurations used in our experiments, and lastly the evaluation metrics used to assess the embedding spaces produced by our framework.

4.1 Datasets

In our work, we train CLWEs between English and five other low-resource languages: Nepali (ne), Finnish (fi), Romanian (ro), Gujarati (gu), and Hungarian (hu). We use Wikipedia dumps for all languages and the FLoRes evaluation set (Guzmán et al., 2019) for Nepali. In addition to this, we use available parallel data between the following related language pairs respectively: English-Hindi (hi) for Nepali, English-Estonian (et), English-Italian (it), English-Hindi (hi) for Gujarati, English-Finnish (fi). We obtain the data from IIT Bombay [1] for En-Hi and from the WMT workshops [2]. We preprocess all the data using Moses scripts and tokenise using BPE, restricting to the 200 most frequent tokens. For the Indic languages, we use IndicNLP [3] for word segmentation. Table 1 details the statistics of the corpus sizes as well as their sources. For evaluation, we use the gold-standard bilingual dictionary from the MUSE dataset (Conneau et al., 2018) for Finnish and for the remaining language pairs, we use bilingual dictionaries published by (Pavlick et al., 2014). We also use the FLoRes evaluation set [4] (Guzmán et al., 2019) to conduct all our experiments in English-Nepali NMT. As it was the first large-scale effort to produce high-quality English-Nepali parallel data, it serves as a benchmark evaluation and allows for fair comparisons across several baselines.

Languages	Sentences	Tokens
Ne	92.3K	2.8M
Fi	6M	91M
Ro	88.6K	2.28M
Gu	382K	6M
Hu	1M	15M
En	67.8M	2.0B

Table 1: Monolingual Training Corpora sizes

4.2 Training Settings

Mapping: Using fasttext (Grave et al., 2018) with the default parametres [5], we first gather monolingual word embeddings for each of the respective languages. After this, we map the embeddings to a cross-lingual space using VecMap (Artetxe et al., 2018b) in the **unsupervised mode** as we do not have any bilingual dictionaries. In this mode an initial solution is found using heuristics and iteratively refined.

Language Pairs	Segments
Hi-En	1.5M
Et-En	1.7M
It-En	151M
Fi-En	6.2M

Table 2: Parallel Training Corpora sizes

Joint Training: To train the embeddings jointly, we use the BiVec tool proposed by (Luong et al., 2015) which is an extension of skip-gram algorithm aiming to predict the context around both the source and target word aligned to a given parallel corpus at the word level. We use the same hyperparameters as in the mapping methods. In both cases, we restrict the vocabulary to the most frequent 200000 words.

In addition to the mapping and joint-training methods trained as described earlier, we also train Joint Align (Wang et al., 2020). In order to this, we use the official implementation [6] on preprocessed tokenised data. We use the non-contextual model in specific as we are working on non-contextual word embeddings.

NMT Evaluation: Lastly, as a downstream task we consider supervised NMT for English-Nepali. Using a single GPU, we train several transformer (Vaswani et al., 2017) models with 5 encoder and 5 decoder layers where the number of attention heads, embedding dimension and inner-layer dimension are 2, 512 and 2048, respectively in the completely supervised setting. We utilise the OpenNMT library [7] (Klein et al., 2017) and Pytorch (Paszke et al., 2019) to build our models. In addition to the hyperparameter settings optimised in FLoRes, we also employ early stopping with patience 4 using

[1] http://www.cfilt.iitb.ac.in/iitb_parallel/

[2] http://www.statmt.org

[3] https://github.com/anoopkunchukuttan/indic_nlp_library

[4] https://github.com/facebookresearch/flores

[5] We learn 300-dimensional vectors with 10 negative samples, a sub-sampling threshold of 1e-5 and 5 training iterations

[6] https://github.com/thespectrewithin/joint_align

[7] https://github.com/OpenNMT/OpenNMT-py

the validation perplexity as the criterion to choose the best model and we use the devtest set to evaluate every 1000 training steps. We report BLEU4 scores (Papineni et al., 2002) on detokenised text following standard practice.

4.3 Evaluation Metrics

We evaluate our embeddings on two aspects: their quality, and the degree of isomorphism achieved between the source and target. As in (Ormazabal et al., 2019), we measure this by bilingual lexicon induction (BLI) and eigenvalue similarity respectively. Firstly, we induce the word-level translations by linking neighbouring source-target word translations in the resultant embeddings spaces (Nearest Neighbour with cosine similarity) and finally evaluate the induced dictionary against the English-Nepali bilingual dictionary released by (Pavlick et al., 2014) to compute precision scores for the BLI task. [8] Next, we measure eigenvalue similarity for the embeddings following the procedure in (Søgaard et al., 2018) on centralised and normalised embeddings. We perform the same evaluations across different cross-lingual alignment methods on all the considered language pairs, particularly we report the result of mapped alignment in the unsupervised mode (Artetxe et al., 2018c), Joint Align (Wang et al., 2020), and lastly our hybrid alignment method. Due to the unavailability of the source code, we were not able to report results of (Ormazabal et al., 2021) however for comparison we test on the Fi-En language pair for which they receive a score of 64.2 (ours 65.2).

5 Results and Discussion

In this section, we discuss our main experimental results on BLI and eignevalue similarity across the chosen language pairs. Furthermore, we also conduct ablation tests on our learnt embeddings at each step of our framework.

5.1 BLI

Results in Table 3 show that our method produces higher BLI scores than mapping, joint-training, and hybrid methods. In particular, Joint Align performs poorly on most language pairs, suggesting that it is inapplicable in a truly low-resource scenario. VecMap performs well overall, however, our approach performs best by a significant margin. Despite using VecMap and a purely joint-training based approach without any additional source-target supervision, the gains in the scores are substantial. Interestingly, our method performs well even in the case of fi $\rightarrow$ en where we use Estonian as the related language; Estonian is in fact lower-resource than Finnish, however our performance suggests that "pivoting" via Estonian was still helpful in learning Finnish-English word embeddings. Therefore, even if the embeddings learnt in the intermediate stages are not ideal, the structural alignments earned are ultimately helpful in obtaining better source-target embeddings.

5.2 Eigenvalue Similarity

In eigenvalue similarity, mapping methods perform much worse than joint training (Table 4). This finding is in line with the literature (Ormazabal et al., 2019), and is explained by the high linguistic divergence between English and source languages, resulting in embeddings that are far less isomorphic. Our hybrid approach performs even better than joint methods and achieved the best eigenvalue similarity score across all langauge pairs, showing that we do indeed obtain isometric embeddings while still not requiring the higher level of supervision in joint learning approaches. Although our proposed framework does not make any significant changes to the mapping and joint components, the combination of the two cross-lingual approaches leads to better embeddings both in terms of quality, shown by the performance in BLI, as well as structure, shown by the eigenvalue similarity scores.

5.3 Downstream Task: Supervised MT

To see the improvements afforded by our embedding initialisation, we report results on supervised NMT from Nepali (ne) to English (en) by initialising transformer models with embeddings obtained from our framework. In particular, we use the FLoRes evaluation set (Guzmán et al., 2019) to allow for a more accurate representation of the gains in performance.

As a baseline, we first train a transformer model (Vaswani et al., 2017) with random initialisation (marked **No Pretraining** in Table 5) following the 5-layer fully supervised model (Section 4.2). To further contexualise our results, we also present the **Mult.** system from FLoRes (Guzmán et al., 2019). This setting uses Hindi-English paral-

[8] https://cs.brown.edu/people/epavlick/data.html

	ne → en	fi → en	ro → en	gu → en	hu → en	**avg**
VecMap (Artetxe et al., 2018b)	52.3	61.9	61.6	45.4	53.2	54.8
Joint Align (Wang et al., 2020)	24.5	31.3	28.2	35.4	26.5	25.2
Ours	**58.4**	**65.2**	**64.5**	**48.4**	**56.3**	**58.6**

Table 3: Precision at 1 scores of proposed method and previous works on BLI (higher is better)

	ne → en	fi → en	ro → en	gu → en	hu → en	**avg**
Mapping (Artetxe et al., 2018b)	205.8	118.2	176.4	189.3	94.5	
Joint (Gouws et al., 2016)	48.6	30.3	41.2	42.5	35.6	
Ours	**37.5**	**23.4**	**32.7**	**33.2**	**26.6**	

Table 4: Eigenvalue Similarity Scores (lower is better)

Embedding Scheme	**DevTest**(↑)	**Test**(↑)
No Pretraining	4.2	4.3
Mult.	6.9	-
Monolingual	5.5	5.2
Mapped	6.4	6.1
Joint	6.3	6.0
Cross-Lingual∗	7.1	6.9
Shared Embeddings∗	7.3	7.1
mBART25	**7.4**	-

Table 5: Tokenized BLEU [%] scores on FLoRes Evaluation Set for ne → en- best score is in bold, ours marked with ∗, higher is better

lel data by concatenating available Nepali data with back-translated Hindi allowing for an augmented dataset. In order to isolate the improvements earned from cross-lingual word embeddings, we further compare monolingual embeddings and cross-lingual embeddings. The models marked as **Cross-Lingual** and **Shared Embeddings** represent models initialised with embeddings produced by our framework. In these models, we initialise the 5-layer transformer models with embeddings on the source and target side and in the case of **Shared Embeddings**, we tie the weights to share the emebddings across the encoder and decoder layers. In the case of **Monolingual**, we initialise the transformer model on the source and target side with English and Nepali fasttext embeddings without any prior alignment. To further understand the gains from our framework, we also report results by initialising the models with **Mapped** and **Joint** methods learnt using previously described methodology. Lastly, to provide a state-of-the-art comparison against our proposed system, we utilise **mBART25** (Liu et al., 2020b) which pre-trains using multilingual denoising on 25 languages.

Our results show that even a baseline supervised model achieves a very poor BLEU score on this task (Table 5). This indicates how challenging English-Nepali is for NMT, therefore improving this baseline result without using additional parallel training data and just a different embedding initialisation is a difficult task. Between the monolingual and cross-lingual embeddings, there are significant gains in the final NMT system which follows results published in (Lample et al., 2018). In addition to this, amongst the different CLWEs the best performance is observed by our proposal. This is indicative of the higher quality representation as shown by the BLI scores earlier. Furthermore, sharing these embeddings across the encoder and decoder layers lead to more improvements which we can attribute to the larger degree of isomorphism between the embeddings (allowing for better alignment when shared). Even though our goal is not to surpass state-of-the-art performance but rather to quantify the improvements chieved from our CLWES, our method performs competitively against the mBART setting. It is notable that we train on baseline transformer architectures of 5 layers whereas mBART is pre-trained on a much larger corpus using a 12-layer transformer thus making our method computationally cheaper with similar results. In all cases, mBART, FLoRes, and ours, a significant improvement from random initialisation is achieved when using a careful pre-training system. Especially in a language pair as difficult as English-Nepali, initialisation is a key component to obtaining good results.

5.4 Ablation Tests

To study where the improvements of the cross-lingual encoding method come from, we conduct several ablation tests (results in Table 6), assessing the contribution of different embedding schemes to the final quality of the embeddings: firstly, we look at the initial unaligned monolingual embeddings, next we look at the embeddings that are independently aligned to the related language, and lastly we look at the emebddings after the final offline map has been constructed. These embedding schemes allows us to verify the importance of the intermediate structural alignments via the related language. As expected the unaligned embeddings have a near 0 BLI score, suggesting that the initial embeddings do not have any linking however as the score is still non-zero we can attribute this to identical words across some language pairs. However, the intermediate embeddings obtained (Related-Aligned in Table 6) have a significant jump in performance even though there is no explicit alignment between the source and target at this stage. This intermediate performance is surprisingly close to the final performance obtained by Joint Align as well, which suggests that the related-language strategy allows for a better understanding of word associations even before performing the final step of offline mapping.

	BLI Score
Embeddings	
Our Method	
Unaligned	0.4
Related-Aligned	24.6
Full Alignment	58.6
Offline Mapping	
Unaligned	0.4
Mapped	54.8
Joint Align	
Unaligned	0.4
Aligned	25.2

Table 6: Ablation Tests on Different Embeddings, reporting average Precision @ 1 score

6 Conclusion and Future Work

In this work, we developed a framework to learn cross-lingual word embeddings in low-resource scenarios. We addressed limitations of both offline as well as joint training methods to develop high quality, isomorphic embeddings for several low-resource language pairs. In particular, we maintain the low cross-lingual signal as required by offline methods while still obtaining structurally sound/isomorphic embeddings as in joint-training based approaches. Our method works by exploiting a higher-resource related-language to jointly learn a cross-lingual space between the related-language and target while also learning a cross-lingual space between the source and the related language using offline mapping. Due to the pre-alignment with a related-language, the resultant cross-lingual spaces are now structurally similar and can be mapped to each other without breaking any orthogonality assumption. Whilst our approach does not change the individual components at all, we obtain far superior results in both BLI as well as eigenvalue similarity across all languages. On a high-level, the gains in our method can be attributed to incorporating more linguistic information in the low-resource language via the related language. This would in turn allow for better modelling of the structure of the embedding spaces without explicitly requiring additional source-target parallel data. As our ablation tests show, indeed the intermediate embeddings themselves have some performance gains even though the source and target embeddings are not aligned to each other yet.

Future work in this direction would include verifying how high-resource the related language needs to be to still see performance gains. In addition to this, we would like to explore how the relatedness of the pivot language affects the performance of the learnt embeddings. Specifically, we would like to discover to what extent isomorphism is preserved in related language pairs- permitting the use of offline methods in more distant languages. Studying this would allows us to suggest further generalisations of our approach to cover a wider range of language families.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. 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.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. Unsupervised neural machine translation.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Yong Cheng. 2019. Semi-supervised learning for neural machine translation. In *Joint training for neural machine translation*, pages 25–40. Springer.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295, Austin, Texas. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. 2020. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7548–7555, Online. Association for Computational Linguistics.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 748–756, Lille, France. PMLR.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2016. Bilbowa: Fast bilingual distributed representations without word alignments.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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. *arXiv preprint arXiv:1902.01382*.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based neural unsupervised machine translation.

Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. 2020a. Very deep transformers for neural machine translation. *arXiv preprint arXiv:2008.07772*.

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

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.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

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, Online. Association for Computational Linguistics.

Ndapa Nakashole. 2018. NORMA: Neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 512–522, Brussels, Belgium. Association for Computational Linguistics.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995, Florence, Italy. Association for Computational Linguistics.

Aitor Ormazabal, Mikel Artetxe, Aitor Soroa, Gorka Labaka, and Eneko Agirre. 2021. Beyond offline mapping: Learning cross lingual word embeddings through context anchoring.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Xutan Peng, Yi Zheng, Chenghua Lin, and Advaith Siddharthan. 2021. Summarising historical text in modern languages.

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

Jürgen Schmidhuber and Sepp Hochreiter. 1997. Long short-term memory. *Neural Comput*, 9(8):1735–1780.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. *arXiv preprint arXiv:1805.03620*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic?

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