Limitations and Challenges of Unsupervised Cross-lingual Pre-training

Martín Quesada Zaragoza mquesadazaragoza@gmail.com
Francisco Casacuberta fcn@prhlt.upv.es
Research Center of Pattern Recognition and Human Language Technology, Universitat Politècnica de València, Valencia, 46022, Spain

Abstract

Cross-lingual alignment methods for monolingual language representations have received notable attention in recent years. However, their use in machine translation pre-training remains scarce. This work tries to shed light on the effects of some of the factors that play a role in cross-lingual pre-training, both for cross-lingual mappings and their integration in supervised neural models. The results show that unsupervised cross-lingual methods are effective at inducing alignment even for distant languages and they benefit noticeably from subword information. However, we find that their effectiveness as pre-training models in machine translation is severely limited due to their cross-lingual signal being easily distorted by the principal network during training. Moreover, the learned bilingual projection is too restrictive to allow said network to learn properly when the embedding weights are frozen.

1 Introduction

Unsupervised cross-lingual embeddings (CLE) concern a group of methods that exploit latent similarities between word embeddings in different languages to generate their own bilingual dictionary or parallel corpus. Fully unsupervised cross-lingual mappings have received notable attention due to being solely reliant on the distributional similarities of language continuous vector representations.

However, semi-supervised cross-lingual pre-training methods, which require very small amounts of parallel corpora – and in some cases only a simple bilingual dictionary for initialization – tend to outperform their unsupervised counterparts (Vulic et al., 2019; Doval et al., 2019; Patra et al., 2019). It is only in situations where no bilingual data exists that unsupervised techniques are preferable. In the case of languages with extensive written records where resources are plentiful, there is little reason to use a fully unsupervised method rather than one that takes advantage of a small bilingual dataset (Artetxe et al., 2020). Even for low-resource languages for which a reduced number of corpora exist, it is exceedingly probable that some sort of translation dictionary that associates them with a more widely studied language is available. This leads us to the question: are unsupervised cross-lingual models useful at all?

While they might not be the best performing tools in a realistic use case, drawing a hard line between unsupervised and semi-supervised cross-lingual mappings does not make much sense, because they are extremely similar processes. Semi-supervised cross-lingual methods are just forfeiting the generation of a seed bilingual dictionary in favor of one provided by the user or other strategies based on back-translation. But their remaining steps are analogous to that of unsupervised cross-lingual projection methods: they both use the seed translation dictionary to align and project the monolingual subspaces in a hypothetical cross-lingual space. In the case of many unsupervised models, seed generation is actually an adaptation of these previous steps, where some assumption on the structures of the matrices is made in order to obtain an initial set of translation pairs. Given that unsupervised and semi-supervised cross-lingual strategies share so many traits, most improvements over unsupervised methods can be transferred to semi-supervised ones.

This work aims to uncover some of the limitations of pre-training strategies based on unsupervised cross-lingual embeddings. These methods are fully dependent on intrinsic language similarities to operate, and therefore constitute a great vehicle to explore how different continuous representations may capture distinct linguistic and structural features. Previous research has already taken advantage of this property to analyze the behavior of language representation spaces (Nakashole and Flauger, 2018). In this work, we consider three different approaches to unsupervised cross-lingual embedding projection, and explore their interaction with different linguistic characteristics and model features, such as subword encoding and vector space structure. The resulting evaluations are put into context to propose potential improvements to language representation strategies as a whole.

2 Related Work

Modern mapping-based cross-lingual embeddings, which are sometimes also called projection-based embeddings, have become very popular by outperforming earlier mapping methods and many supervised cross-lingual methods that are dependent on segment-level alignment. As many initial mapping-based approaches (Mikolov et al., 2013b; Faruqui and Dyer, 2014), they require monolingual embeddings, but often also a small parallel corpus or a bilingual seed translation dictionary to perform the initial alignment. However, some do not need any parallel signal at all, as they can also perform the original alignment relying only in estimations based on structural similarities between the monolingual vector spaces (Artetxe et al., 2018; Lample et al., 2017). Another family of cross-lingual word embeddings are the so-called pseudo-bilingual word embeddings (Ruder, 2017). These approaches use a concatenation of large monolingual corpora and integrate it with some amount explicit bilingual information – such as replacing translation pairs in the dataset Gouws and Søgaard (2015) or substituting tokens based on their semantic cluster (Ammar et al., 2016) – . Both mapping-based and pseudo-bilingual word embeddings are especially useful for prototyping in extremely low-resource language models when no parallel data is available.

However, cross-lingual embeddings have not been particularly effective in deep neural network pre-training, and often do not seem to represent a significant improvement over using off-the-shelf monolingual embeddings (Qi et al., 2018). In contrast, cross-lingual language models such as BERT (Devlin et al., 2019) have fared better. Some approaches try to create a universal neural language model for all the languages included in a final multilingual system (Ji et al., 2020; Lin et al., 2020). A very successful alternative to these strategies has also been proposed by Lample and Conneau (2019), who create a cross-lingual neural language model by training a masked language model (Devlin et al., 2019) using a shared vocabulary between languages and subsampling frequent outputs as per Mikolov et al. (2013c). Ren et al. (2019) refine this masked language model (MLM) by introducing an explicit cross-language training objective, creating a cross-lingual masked language model (CMLM). Recent research in Wang and Zhao (2021) has obtained top-of-the-line results by using a large-scale CMLM and training the final supervised model using a joint optimization objective (Sun et al., 2019) that aims to maintain the original distribution of the CMLM while maximizing translation performance.

3 Unsupervised Cross-lingual Pre-training

3.1 Unsupervised cross-lingual embeddings

This work explores cross-lingual pre-training through three particular unsupervised cross-lingual embedding methods: VecMap (Artetxe et al., 2018), MUSE (Lample et al., 2017) and embeddings trained over multilingual corpora.

Both VecMap and MUSE are projection-based cross-lingual embeddings. Projection-based CLE generate an alignment between monolingual word embeddings, subsequently projecting them into a common representation space that facilitates a direct mapping between the distribution of both embeddings. The initial alignment is commonly produced using a seed dictionary of translation pairs. However, in some cases these translation pairs are estimated by the cross-lingual method itself. This capacity to generate a common vector space of aligned embeddings with no bilingual signal defines fully unsupervised cross-lingual embeddings.

The specifics behind alignment and projection are also dependent on the general topology of the embeddings that are to be mapped. In this work, all experiments are performed over Word2Vec embeddings, which offers two possible topologies: continuous bag-of-words (CBOW) and skip-gram. The former trains a shallow network to predict a word given an input context, while the latter learns to predict a context window from an input word. In this work, only skip-gram is used for all experiments, particularly skip-gram with negative sampling (SGNS) (Mikolov et al., 2013c).

For VecMap, Artetxe et al. (2018) assume that word translations have approximately identical vectors of monolingual similarity distribution. The proposed method operates on top of this idea, adding empirically motivated enhancements that make the procedure more robust.

In contrast, MUSE (Lample et al., 2017) uses adversarial training to create a generator network able to project word vectors from each monolingual embedding in a way such that it is very difficult to distinguish the space to which they originally belonged, thus achieving a common mapping between both embeddings.

Another approach to cross-lingual embeddings explored in this work are embeddings trained over multilingual corpora. The model proposed in this section is trained on corpora with no alignment or contextual proximity. The procedure is remarkably simple: two monolingual corpora in different languages are concatenated, and a word embedding is trained over the resulting multilingual text. Just as in the previous two methods, the result is a bilingual vector space that tends to group translation candidates close to each other. Therefore, it can also be used as bilingual pre-training for a translation model in the task studied in this work. This approach serves as a baseline to determine how much of the alignment achieved from cross-lingual embeddings is innate to the distribution of both languages and can be extracted with no mapping procedures.

3.2 Cross-lingual Pre-training

The aforementioned cross-lingual embeddings are integrated as pre-training in the input and output embedding layers of a machine translation neural network, in substitution of the embedding transformation that would otherwise be initialized randomly. The model used follows the Transformer architecture proposed by Vaswani et al. (2017), with some slight modifications that are described in detail in section 4.3.4. Although Qi et al. (2018) report that orthogonal alignment was not helpful when pre-training embeddings for an attention-based neural machine translation model, in this work we aim to evaluate new strategies that may influence pre-training performance. Namely, by trying out other state-of-the-art cross-mapping strategies based on orthogonal mapping Lample et al. (2017); Artetxe et al. (2018) and considering techniques that help better transfer and maintain cross-lingual alignment during training, such as using a joint BPE vocabulary or freezing the embedding layers of the translation model.

4 Experimental framework

4.1 Corpora

The corpora used in this work were selected from the data collection provided in the WMT14 Machine Translation shared task[1] (Macháček and Bojar, 2014). This collection of corpora allows to study language similarity as a variable, since all of the language pairs available feature English data aligned with languages with which it presents varying degrees of phylogenetic relatedness and typological similarity. The language pairs chosen are French–English, German–English, Russian–English and Hindi–English. This selection was also motivated by the interesting properties of the relationship triangle formed by English, German and French. English and German are the closest genetic relatives, and both are included in the West Germanic family. While German and English are similarly phylogenetically distant to French according to their classification, French and English share many more typological features. Notably, an estimated 25% of English loanwords come from French (Cannon, 1989). This three-way relationship is interesting because it juxtaposes genetic and typological features, both of which can have different effects over cross-lingual mappings. Additionally, Russian and Hindi provide increasingly more distant languages that help study projection methods for cases with low cross-lingual similarity and different alphabets.

Language	Corpora	Sentences	Tokens
English	News Crawl 2011-2013	51M	1,167M
French	News Crawl 2007-2013	30M	696M
German	News Crawl 2012-2013	55M	970M
Russian	News Crawl 2007-2013	32M	576M
Hindi	News Crawl 2007-2013 + HindMonoCorp 0.5	43M	932M

Table 1: Breakdown of training monolingual corpora sources, number of sentences and total number of tokens by source.

Language pair	Corpora	Sentences	Tokens	
			Source	English
French–English	Europarlv7 + Common Crawl corpus	5M	117M	129M
German–English	Europarlv7 + Common Crawl corpus	4.1M	99M	94M
Russian–English	Common Crawl corpus + Yandex 1M corpus v1.3	1.8M	41M	39.5M
Hindi–English	HindiEnCorp 0.5	0.26M	2.6M	4.1M

Table 2: Breakdown of training parallel corpora sources, number of sentences and total number of tokens by source.

The monolingual training set, which is described in Table 1, also includes a corpus outside the WMT14 set, HindMonoCorp (Bojar et al., 2014) collections. This dataset was added in order to keep a roughly similar volume of training data between all monolingual corpora, which is especially important for embedding cross-mapping procedures. Similarly, parallel corpora have been built by combining different corpora in such a way that the volume of data is comparable across all languages, as shown in Table 2

For all datasets, the following pre-processing steps have been applied: 1. Normalization of unicode punctuation encoding; 2. Tokenization; 3. Clean and eliminate empty sentences, those containing more than 60 words, and sentences with a source-target ratio greater than 1-9.

[1] https://www.statmt.org/wmt14/translation-task.html

4.2 Evaluation Metrics

4.2.1 Bilingual Lexicon Induction

The quality of the different cross-lingual representations generated is evaluated according to their bilingual lexicon induction (BLI) performance. As in Artetxe et al. (2018), this is calculated as the average accuracy of the induced cross-lingual vector space in a word translation task for a ground-truth bilingual dictionary, which in this work will be referred to as word translation accuracy. The evaluation only considers the source language words from this bilingual ground-truth dictionary that are also included in the vocabulary of the source language embedding. For each of these source language words, the closest word vector in the target embedding is found and taken as the most likely translation. The procedure then compares the translations obtained with this method and those provided by the bilingual dictionary, taking as correct the translation induced from the cross-lingual space if the target language closest vector is included in the list of possible translations that appears in the bilingual dictionary. The distance between word vectors is calculated using cross-domain similarity local scaling (CSLS), proposed by Lample et al. (2017). The ground-truth bilingual dictionaries used are provided by the MUSE toolkit[2], specifically those belonging to the ”full” set. The purpose of this evaluation is not to obtain a state-of-the-art unsupervised BLI system. Instead, it is to assess the interactions of the different cross-mapping methodologies studied in this work with the different degrees of language similarities present in each of the language pairs.

4.2.2 Machine Translation

As mentioned previously, the generated cross-lingual embeddings are also assessed on their utility as pre-training embeddings. To this end, they are integrated in a Transformer-based machine translation model, which then receives limited training and is evaluated using multi-BLEU as provided by the Moses toolkit (Koehn et al., 2007).

4.3 Model Configuration

4.3.1 BPE

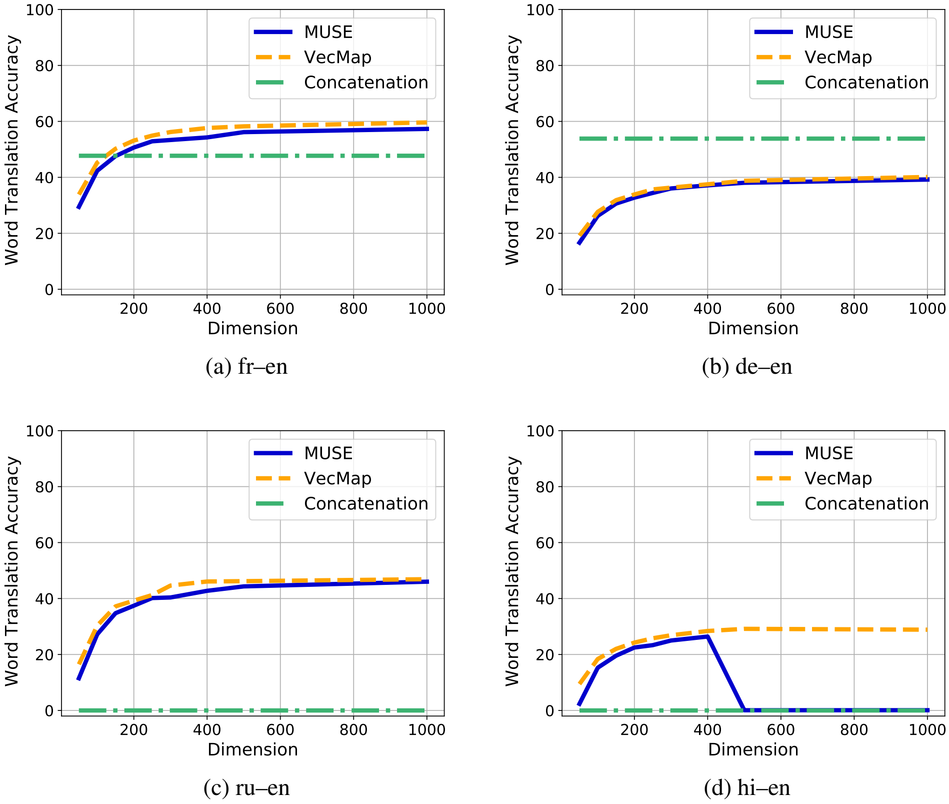
In some experiments, Byte Pair Encoding (BPE), particularly the subword-level adaptation of this method proposed by Sennrich et al. (2015), is applied to the corpora in order to study its effect in the performance of cross-lingual embedding mappings. BPE encodings build a shared subword-level vocabulary for the source and target language corpora, which reduces the total size of the vocabulary. Moreover, it allows the model to represent words not seen during training by combining subword units. As a result, some lexical information is transferred more effectively between languages, particularly in the case of certain word classes such as proper nouns, compounds, cognates and loanwords (Sennrich et al., 2015). The number of merge operations used in BPE is its single determining hyperparameter, which governs vocabulary size. Gowda and May (2020) show that large vocabulary sizes only maximize BLEU performance when using vast datasets with 4.5M sentences. However, since our monolingual datasets are all far bigger, and given that for Transformer-based neural machine translation it is recommended the largest possible BPE vocabulary (Gowda and May, 2020), we opt to guarantee a large vocabulary by using 48,000 merge operations.

4.3.2 Embeddings

All skip-gram word embeddings used in the experiments proposed in this work are trained during 5 epochs with a learning rate of 0.05. Changes in epoch size have not had any significant effect, as the corpora used to train the embeddings is sufficiently large and does not require of more training cycles. The main parameter to study when training the embeddings has been

[2] https://github.com/facebookresearch/MUSE

their dimension. A range of values between 100 and 1000, with steps of size 100, is used to examine the effect of embedding dimension in unsupervised cross-lingual methods. The results are displayed in Figure 1 as an evolution of the BLI performance of the model relative to the dimensionality of the embeddings.



(a) fr–en

(b) de–en

(c) ru–en

(d) hi–en

Figure 1: BLI performance evolution for each language pair and cross-mapping method according to the dimension of the monolingual embeddings. Each plot corresponds to a language pair, and the different series to one of the cross-mapping approaches considered.

In the case of the cross-mapping strategies of VecMap and MUSE, BLI increases rapidly with dimension up to close to 300 dimensions. From here on to 500 dimensions, performance still seems to be correlated to dimension, and it only sees very marginal growth from this point. This behavior is in line with the directives originally given for Word2Vec embeddings (Mikolov et al., 2013a). However, as seen in the section (d) of figure 1, relative to the hi–en language pair, it seems that increasing the dimension of the embeddings past a certain point may affect the viability of learning an effective cross-lingual projection for the adversarial approach in MUSE. In the absence of additional experiments with a larger number of distant languages, it is hypothesized that using too many dimensions while dealing with a complicated projection between very different languages that do not even share a common alphabet can lead to a failure in learning a reasonable projection matrix. VecMap is not affected by this phenomenon for the showcased experiments, which may be due to the use of ZCA whitening (Bell and Sejnowski, 1997), which encourages exploring dimensions that may not fit the current solution to help escape poor local optima.

Lastly, in the case of embeddings trained over concatenation of monolingual corpora, dimension does not affect significantly their BLI score as a cross-lingual model. This can be attributed to the fact that the implicit bilingual alignment in which this method relies is not influenced by any transformations or projections of the vector space, so long as the distance between points is measured with a dimension-invariant metric. This work uses CSLS to measure word translation accuracy between embeddings, which relies only on cosine similarity between the embedding vectors, a metric that remains unaffected by dimension scaling.

Since the embeddings used in this work will be integrated is a Transformer neural network in charge of machine translation, a dimension value of 512 has been chosen for them. This is in the range where cross-lingual performance is stabilized, while being a common encoder-decoder dimension value for Transformer-derived machine translation models (Vaswani et al., 2017; Lample and Conneau, 2019).

4.3.3 Cross-lingual mappings

Both the MUSE and Vecmap cross-mapping techniques have a number of parameters that dictate some of the characteristics of the alignment procedure. For VecMap we use the standard unsupervised configuration, which is equivalent to that of the models presented in VecMap, Artetxe et al. (2018). The maximum vocabulary is set to 20,000 words, the vocabulary used to generate the initial unsupervised translation table is limited to 4,000 words, the CSLS neighborhood used for vector distance calculations is of size 10 and the embeddings are normalized before the cross-mapping is initiated. In contrast, all MUSE cross-mappings are performed using the default unsupervised parameters. The only explicit adjustments made are the maximum size of vocabulary considered, which is set to 20,000, and the number of word vectors used for discrimination, which is set to the 7,500 more frequent words. Distance between vectors is calculated using a CSLS neighborhood of size 10, and the embeddings are normalized before the cross-mapping process begins. Memory limitations for the available setup meant that reproducing the VecMap benchmark was far easier than that of MUSE, and these changes were made to keep VecMap and MUSE running over with as similar of a set of parameters as possible.

4.3.4 Neural Machine Translation Pre-training

For the neural machine translation model that integrates cross-lingual embeddings for pre-training, we rely on an OpenNMT-py (Klein et al., 2017) model that mimics the original Transformer architecture proposed by Vaswani et al. (2017). It contains a stack of 6 encoder and 6 decoder layers . Each encoder layer includes positional encoding, 8 attention heads and a dense feed-forward network. The decoders use instead an initial masked multi-head attention layers that receives the shifted outputs, followed by another multi-head attention layer that is fed by the output of the encoder stack, and have a final feed-forward layer. All feed-forward layers and attention heads use a dropout probability of 0.1, and rely on the Adam optimization criterion (Kingma and Ba, 2015). The feed-forward layers have been changed to have a dimension of 1,024 units from the original 2,048 present in Vaswani et al. (2017). This accelerates model training and does not massively impact performance for models that are not trained extensively (Lample and Conneau, 2019). Similarly, the number of training steps is reduced from 200,000 to 20,000 maintaining a batch size of 4,096 tokens, as a compromise between training cost and minimum performance of the model to allow for pre-training integration.

5 Results and Discussion

5.1 Cross-lingual models

5.1.1 BPE

In the past, Lample et al. (2017) have shown that BPE (Sennrich et al., 2015) improves considerably the alignment of monolingual language spaces, particularly for cases where the languages share the same alphabet or anchor tokens (Smith et al., 2017). Anchor tokens are words with equivalent meaning that are written identically across languages and are therefore common vocabulary to both languages, such as proper nouns of places, organizations or people, acronyms, loanwords and digits.

Model	Dimension	BPE	Word Translation Accuracy			
			fr–en	de–en	ru–en	hi–en
MUSE	512	No	56.2	38.1	44.3	0.1
MUSE	512	Yes	62.2	41.4	0.1	42.3
VecMap	512	No	58.2	38.8	46.2	29.2
VecMap	512	Yes	**65.0**	46.2	**60.0**	**44.8**
Concatenation	512	No	47.7	**53.8**	0	0
Concatenation	512	Yes	46.6	45.3	0	0

Table 3: Results obtained for the best cross-lingual embeddings selected for neural model pretraining. Accuracy is measured comparing pairs from ground-truth bilingual dictionaries and employing CSLS as distance metric.

Table 3 illustrates the effect of BPE on the BLI performance of the generated cross-lingual embeddings. BPE usage seems to generally improve the score of the projection-based mappings, while having a slightly negative or non-existent influence on embeddings trained over concatenation of monolingual corpora. While the former result is expected (Lample et al., 2017), the latter phenomenon is more interesting, and can be explained by the fact that these embeddings are jointly learning both languages, but no cross-mapping is performed, so the relative position of words in the representation space should remain similar whether subword information is captured or not. In many cases there may be a certain loss of semantic information when creating a shared byte-paired encoding between languages (Ren et al., 2019). Since they will not will be compensated by the subword features that are retrieved – as the approach does not take advantage of them – , the overall effect of BPE tends to be negative.

The impact of BPE is especially significant for the MUSE mapping in language pairs ru–en and hi–en. In the case of ru–en, the use of this tokenization approach apparently does not allow for any sensible alignment, unlike the projection that uses non-BPE embeddings, which performs fine. A likely explanation for this is that Russian and English do not use the same alphabet or share many common words and anchor tokens. Therefore, almost no joint subword information is learned, while some semantic features may be diluted (Ren et al., 2019). However, the hi–en pair in Table 3 shows the opposite phenomenon, where the application of BPE has made possible a previously unavailable alignment. This case is especially puzzling, since Hindi also uses a completely different alphabet from that of English there should be very little transfer of information between the subword vocabularies. Upon closer inspection, both BPE vocabularies for ru–en and hi–en have a very similar size, which indicates that this behavior is not a function of semantic diversity. Moreover, the size of the common lexicon found in the training corpus between English and Hindi is an order of magnitude lower than that of English and Russian[3], which puts into question really how relevant anchor tokens are when it comes to generating a joint BPE vocabulary. A possible explanation for this phenomenon could be that the generated vector spaces simply have a slightly different distribution when using BPE, which can affect the chances of finding a good projection into a common bilingual space for the embeddings. The VecMap projection does not seem to be affected in the same way, which could be due to it having a more robust initialization and being able to escape local optima better than MUSE, as shown in (Vulic et al., 2019; Glavaš et al., 2019).

5.1.2 Language similarity

Table 3 showcases the performance of the considered cross-lingual models for all language pairs. Language similarity does seem to be somewhat indicative of BLI performance, although a weak signal at that.

The fr–en language pair is the best performing one across the board, especially for the projection-based alignments. This is expected, since these methods are reliant on semantic similarities. They also make great use of anchor tokens in their initial unsupervised dictionary induction (Lample et al., 2017), of which there are plenty in the English–French pair.

Although English and German are phylogenetically closer to each other, the performance for this pair is inferior to that of English and French for MUSE and VecMap. In contrast, embeddings trained over a concatenation of monolingual corpora surpass projection-based cross-maping methods, and their own BLI score for the fr–en pair. French and English share many anchor tokens, whereas German and English have a noticeably smaller common vocabulary, but show a greater degree of similarity in other typological features common in languages from the same family tree, such as word ordering or verbal categorization. Since for the concatenation strategy the pair de–en is actually performing better than fr–en, it can be hypothesized that the natural alignment resulting of training embedding over multilingual text is more sensible to other typological categories. This is especially likely for word ordering, since the skip-gram architecture is learning to predict contexts in a reduced local window (Mikolov et al., 2013a,c), which is sensible to large discrepancies in sentence structure.

For the ru–en and hi–en pairs, training embeddings over a multilingual corpus seems to produce no alignment whatsoever, as the selected languages do not share a common alphabet. However, both of the projection-based cross-lingual techniques, ru–en and hi–en are shown to be competitive with de–en. This casts some doubts on which are actually the typological features that govern explicit cross-linguality, since by all accounts German should be more semantically and grammatically similar to English than Russian or Hindi (Georgi et al., 2010). Further research that isolates typological features for cross-lingual evaluation is needed in order to produce meaningful guidelines on the adaptation of cross-lingual models according to language similarity, though the experiments show that semantic relatedness is not the only factor at play.

Overall, VecMap has been shown to be the best performing cross-lingual method and also the most robust one when dealing with distant languages, which is consistent with previous research (Vulic et al., 2019; Glavaš et al., 2019). MUSE appears to be generally weaker for cases where inducing an initial translation table is more difficult due to low language similarity, though seems to perform fine when this phase is completed successfully, which is a common trend in projection-based cross-lingual methods. Remarkably, training word embeddings over a concatenation of monolingual corpora outperforms projection-based methods for the de–en pair, although is not effective for distantly related languages. From this it can be inferred that some features learned by skip-gram word embeddings during training are valuable when it

[3] For 1M sentences considered, where numeric tokens have been discarded, there are around 11,800 common tokens for the ru–en pair, but only 1,670 for hi–en.

comes to producing an alignment and, like many other typological characteristics, are not being considered by current explicit cross-mapping methods but could prove to be valuable.

5.2 Neural machine translation pre-training

Pre-trained embeddings			Frozen embeddings	BLEU			
Cross-mapping	Dimension	BPE		fr–en	de–en	ru–en	hi–en
(None)	512	Yes	No	**34.1**	25.4	28.7	6.1
(None)	512	Yes	Yes	32.1	23.6	26.8	5.8
MUSE	512	Yes	No	33.9	26.0	29.4	6.7
MUSE	512	Yes	Yes	32.2	24.3	27.9	6.1
VecMap	512	Yes	No	33.4	**26.3**	30.0	**9.2**
VecMap	512	Yes	Yes	32.4	24.1	28.9	8.4
Concatenation	512	Yes	No	33.5	25.5	**30.2**	6.6
Concatenation	512	Yes	Yes	33.1	24.9	29.6	6.2

Table 4: Results obtained for the best Transformer translation models that use pre-trained vectors. Cross-mapping is indicated as (None) when no explicit cross-lingual technique is applied to the pre-trained word embeddings.

5.2.1 Freezing embeddings

As indicated in Sun et al. (2019) and Wang and Zhao (2021), embeddings used as the encoder-decoder pieces of an attention-based neural network tend to degenerate as the global model is fine-tuned for a particular task, which for this work corresponds to machine translation. For this reason, it has been decided to assess the impact of freezing the encoder and decoder embeddings during training. The results are shown in Table 4. Freezing the pre-trained embeddings does not improve BLEU performance in comparison with models that do modify the weights of their encoder and decoder during supervised training, which score slightly better. This effect is in line with prior work (Sun et al., 2019; Wang and Zhao, 2021), which shows that the integrated model needs to modify the pre-trained components during fine tuning to maximize its performance, but this behavior tends to break the cross-lingual alignment created previously. As a result, they propose to optimize supervised training based on two different objectives: maintaining the structural correspondence of the initial pre-trained components and maximizing the translation objective. Though the implementation of this strategy is not readily available for general use – which is the reason why they have not been considered for these experiments – , they have been shown to be the best current approach to transfer cross-lingual knowledge in pre-training.

5.2.2 Cross-linguality

The effect of cross-linguality in the pre-trained embeddings is relatively low across the board. We think that this could be attributed to two main factors. First, as shown by Qi et al. (2018), most of the increase in performance provided by pre-trained embeddings can usually be attributed to a better encoding of the source sentence. Since the cross-lingual alignment contained in the embeddings does not aid significantly in this task, the overall performance may not increase much. The second is the degeneration phenomenon (Sun et al., 2019), which distorts the structure of the embeddings during training, and therefore their cross-lingual alignment.

Still, the cross-ling projection-based cross-lingual methods showcase some amount of improvement, that seems to increase the more distant the languages are. This result can seem counter-intuitive at first, since BLI performance is generally a strong indicator of performance for a pre-training method in translation models. Such a behavior is shown in Lample and Conneau (2019), where cross-lingual language models outperform unsupervised cross-lingual embeddings both in terms of BLI performance and effectiveness as pre-training strategy. However, we should keep in mind that the scenarios involving distant language pairs correspond to the cases where cross-lingual alignments have the biggest impact on the global structure on the embeddings. For similar source and target languages, the projection learned by the neural network is of higher quality, and therefore initialization is less relevant. In contrast, translation of distant language pairs requires that the model learns a more complex projection, and therefore in this case it benefits more from a stronger initialization. This is in line with previous claims on the importance of initialization in attention-based encoder-decoder translation models (Devlin et al., 2019; Liu et al., 2020), and the results derived from cross-lingual language models such as that of Lample and Conneau (2019). It most importantly also suggests that initialization is a factor that can limit the quality of a translation model even when de facto unlimited training data and time is available. Recent advances on cross-lingual language models appear to follow this trend by consistently improving substantially on translation models for distant language pairs (Wang and Zhao, 2021).

6 Conclusions

This work makes use of fully unsupervised cross-lingual models to explore some of the factors that affect the performance of cross-lingual methods. The experimental results showcase agreement with prior publications regarding the usefulness of subword information in cross-linguality, and suggest that character-level encoding might be especially relevant for language pairs of low similarity. Moreover, reveal that phylogenetic relatedness should not be directly taken to dictate cross-lingual performance, and that purely semantic similarity is not the only typological feature captured by projection-based mappings.

Cross-lingual transfer remains ineffective even if the structure of the pre-trained encoder and decoder is fixed during fine-tuning, which is indicative of the degeneration of cross-lingual projections in supervised training, and the limited scope of pre-trained embeddings. Current pre-training approaches that rely on joint optimization appears to be the most promising approach going forward.

Future research may benefit from designing strategies that adapt cross-mapping methods to different language pairs according to their features, and that are able to capture typological information that is currently lost. Model initialization seems to be a fundamental factor that is able to limit machine translation ceiling, especially for long distance pairs.

Acknowledgements

This work received funds from the Generalitat Valenciana under project DeepPattern (PROMETEO 2019/121).

References

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. A. (2016). Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Artetxe, M., Labaka, G., and Agirre, E. (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.

Artetxe, M., Ruder, S., Yogatama, D., Labaka, G., and Agirre, E. (2020). A call for more rigor in unsupervised cross-lingual learning. *CoRR*, abs/2004.14958.

Bell, A. J. and Sejnowski, T. J. (1997). The ‘independent components’ of natural scenes are edge filters. *Vision Research*, 37(23):3327–38.

Bojar, O., Diatka, V., Rychlý, P., Straňák, P., Suchomel, V., Tamchyna, A., and Zeman, D. (2014). HindMonoCorp 0.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University.

Cannon, G. (1989). Historical change and english word-formation : recent vocabulary. *Language*, 65:880.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Doval, Y., Camacho-Collados, J., Anke, L. E., and Schockaert, S. (2019). On the robustness of unsupervised and semi-supervised cross-lingual word embedding learning. *CoRR*, abs/1908.07742.

Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Georgi, R., Xia, F., and Lewis, W. (2010). Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 385–393.

Glavaš, G., Litschko, R., Ruder, S., and Vulić, I. (2019). How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721.

Gouws, S. and Søgaard, A. (2015). Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390. Association for Computational Linguistics.

Gowda, T. and May, J. (2020). Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Ji, B., Zhang, Z., Duan, X., Zhang, M., Chen, B., and Luo, W. (2020). Cross-lingual pre-training based transfer for zero-shot neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):115–122.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 67–72.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Lample, G., Denoyer, L., and Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Lin, Z., Pan, X., Wang, M., Qiu, X., Feng, J., Zhou, H., and Li, L. (2020). Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2649–2663.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210.

Macháček, M. and Bojar, O. (2014). Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, Workshop Track Proceedings. CoRR*, abs/1301.3781.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting Similarities among Languages for Machine Translation. *CoRR*, abs/1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Nakashole, N. and Flauger, R. (2018). Characterizing departures from linearity in word translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 221–227.

Patra, B., Moniz, J. R. A., Garg, S., Gormley, M. R., and Neubig, G. (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.

Qi, Y., Sachan, D. S., Felix, M., Padmanabhan, S. J., and Neubig, G. (2018). When and why are pre-trained word embeddings useful for neural machine translation? *CoRR*, abs/1804.06323.

Ren, S., Wu, Y., Liu, S., Zhou, M., and Ma, S. (2019). Explicit cross-lingual pre-training for unsupervised machine translation. *CoRR*, abs/1909.00180.

Ruder, S. (2017). A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Smith, S. L., Turban, D. H. P., Hamblin, S., and Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859.

Sun, H., Wang, R., Chen, K., Utiyama, M., Sumita, E., and Zhao, T. (2019). Unsupervised bilingual word embedding agreement for unsupervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1235–1245.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Vulic, I., Glavas, G., Reichart, R., and Korhonen, A. (2019). Do we really need fully unsupervised cross-lingual embeddings? *CoRR*, abs/1909.01638.

Wang, R. and Zhao, H. (2021). Advances and challenges in unsupervised neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 17–21.