School of Informatics

Informatics Research Review
Unsupervised Cross-lingual Alignment of word embeddings for Bilingual Lexicon Induction

[redacted]

January 2021

Abstract

Word embeddings, which are non-sparse vector representations of words, have been influential in several Natural Language Processing (NLP) tasks. Bilingual Lexicon Induction (BLI) is the the NLP task of inducing translations from monolingual data in two different languages. In recent years, numerous models have obtained good results at BLI without any kind of supervision, just by realigning word embeddings.

In this IRR we will explain some of the most relevant unsupervised models that work on the cross-lingual alignment of word embeddings for BLI. We will also motivate this approach, and expose some of its limitations.

Date: Friday 22nd January, 2021

Supervisor: [redacted]

1 Introduction

Words are one of the first things we learn from a language. It comes naturally to us that many Natural Language Processing (NLP) systems work at word level and do not involve smaller or bigger units of languages such as characters, morphemes, sentences or paragraphs. Their usage for NLP requires us to represent them appropriately.

The simplest way to represent words is to consider each of them as a separate and different unit, erasing any possible similarity or relationship between them. Given a vocabulary set V, the i-th word $v_i \in V$ gets assigned a vector form where the i-th component is one and the others are zero $\delta ij = (0, \ldots, 0, 1, 0, \ldots, 0)$ in what is commonly known as *one-hot representation*. This one-to-one representation of words has evident drawbacks. One of them is that the size of the model is dependent on the size of the vocabulary. Another relevant problem is sparsity: words with small occurrences in the training data can challenge the NLP system.

In recent years there has been the surge of word embeddings, which are non-sparse vector representations of words. They generally involve real vectors of hundreds of dimensions, which can be then used in different NLP tasks achieving impressive results [Mikolov et al., 2013a]. For instance, the pioneer work of Mikolov et al. [2013a] trained a Neural Network to predict a missing part of text. They then observed that the latent space of such a network contained semantic or syntactic information from a given word. This representation achieved state-of-the-art results for tasks that measured word similarities, and could also capture analogies (*king* is to *queen* what *man* is to *woman*) and other language relations.

The success of word embeddings motivated their development [Pennington et al., 2014, Bojanowski et al., 2017] and now they are pretrained in over 150 languages [Grave et al., 2018]. The appearance of word embeddings in different languages motivated its usage for Bilingual Lexicon Induction (BLI). BLI is the NLP task of inducing translations from monolingual data in two different languages [Irvine and Callison-Burch, 2016]. There have been several attempts at BLI that used the alignment of word embeddings as a basis. These attempts have generally ranged from full supervision to some semi-supervised approaches [Artetxe et al., 2017]. Not much supervision is really required: for instance, Artetxe et al. [2017] obtained competitive results with as little supervision as a 25 word dictionary. Such resources are very easy to find for any pair of languages. However, several fully unsupervised approaches have been studied [Lample et al., 2018, Hoshen and Wolf, 2018].

How are fully unsupervised approaches motivated? There is usually a misconception when stating the main reasons for unsupervision, as Artetxe et al. [2020] pointed out. There are dictionaries readily available for most pairs of languages, and we have just seen that not a lot of supervision is required [Artetxe et al., 2017]. Even if it was not possible to use a dictionary, supervision could exploit parallel corpora, but the unsupervised approach is a bit *extreme*. A clear analogy would be trying to learn how to translate one language to another after reading one book in one language, and a different book in another language. The real motivations for such an approach are the following [Artetxe et al., 2020]:

- **The inherent scientific interest of the problem.** Being able to align two different languages without knowing how they are related is a fascinating task in itself. Unsupervised methods tell us a lot about the limits of the principles that inspire our language models (e.g. the distributional hypothesis). Unsupervised methods could also provide

insights on general properties of languages or the models that we assume.

Moreover, unsupervised methods might be useful in the event that we had to deal with an unknown or non-human language. Recent work [Luo et al., 2020] has dealt with undeciphered lost languages.

- **Simplicity.** Refusing to use available dictionaries or parallel data might not produce the best solution, but unsupervised models can be really simple and easy to train. Besides, some unsupervised models have shown a competitive performance when compared to their semi-supervised rivals [Glavaš et al., 2019], suggesting that the gap between supervised and unsupervised models is not that large.

- **Useful baseline.** Unsupervised methods can be useful as a good starting point for semi-supervised methods, especially when they mostly rely on monolingual corpora. Besides, semi-supervised or supervised methods are only justified if they really exploit the parallel data to improve the unsupervised models. In this way, research could benefit from competitive unsupervised models.

There are particular practical scenarios that have not been previously considered, but I believe should be taken into account:

- **Practical applications.** Sometimes, a dictionary cannot be used. For instance, that would be the case when X and Y do not share common words. As an example of that, we may want to merge word embeddings from two different areas of knowledge, such as Biology and Mathematics.

  This might look a bit counter-intuitive as the motivation for our alignments is BLI, but no matter what our primary motivation is, we are producing unsupervised alignments.

- **Useful for other domains.** Apart from words, we can also embed sentences, paragraphs, documents, images or sounds. Even if at word level we have plenty of dictionaries, we might find the ideas that arise from the rotation of word embeddings desirable.

  For instance, imagine that you have two sets of unnanotated images and their corresponding embeddings X and Y. You may want to find pairs of images (x_i, y_i) of similar things, but the embeddings were trained on different images and were seeded differently —or even used different parameters or algorithms—. In that case you would certainly benefit from the ongoing research in unsupervised BLI.

This Research Review explores some of the most relevant methods for the unsupervised cross-lingual alignment of word embeddings for Bilingual Lexicon Induction. We will explore some of the limitations of these models in the *Discussion* section. To limit the scope of this review, we will only consider orthogonal maps; that is, given two word embeddings X and Y, we find the orthogonal matrix W so that WX becomes the new projection in the space of Y. Intuitively speaking, we are rotating the source embeddings X so that they get closer to Y (see Figure 1).

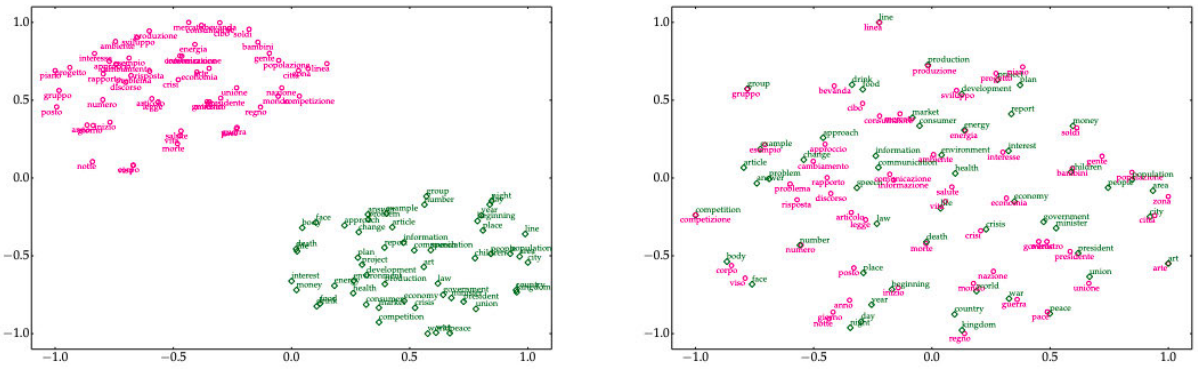

Figure 1: Monolingual word embeddings in English (green) and Italian (pink). On the left, word embeddings are unaligned. On the right, embeddings are projected into a cross-lingual space. Source: Ruder et al. [2019]

2 Bilingual Lexicon Induction

Bilingual Lexicon Induction is the task of inducing translations from monolingual data in two different languages. These two sources of monolingual data are not the translation of the same text in a different language, although they could be closely related (for example, a Wikipedia article on the same topic). Then, similarities between words are exploited to create a dictionary. There are many similarities that can be studied [Irvine and Callison-Burch, 2016]: words may be spelled similarly (orthographic similarity); they may appear in similar contexts (contextual similarity); they may have similar frequencies (frequency similarity); they may appear in the same types of documents (topic similarity) and they may appear and disappear over time (temporal similarity).

Before the appearance of word embeddings, BLI systems were built mainly combining such features. Edit distances, co-occurrence matrices or tf-idf were signals that provided valuable hints on which words in each language could be translations. Haghighi et al. [2008] transformed words from both the source and target language into feature vectors. Then vectors from the source and target language were matched together through a generative model, which was based on canonical correlation analysis (CCA).

Some years later, Irvine and Callison-Burch [2016] designed 18 different features and weighted them for a relatively simple discriminative model. The fact that they consistently outperformed Haghighi et al. [2008] may suggest that at the time an important line of research for BLI was designing multiple hand-engineered features and combining them. The situation dramatically changed with the irruption of word embeddings [Mikolov et al., 2013a] and their improvement [Pennington et al., 2014, Bojanowski et al., 2017, Devlin et al., 2019]. Soon it was more attractive to use them directly rather than manually crafting features and applying them in big datasets from many different languages.

Monolingual word embeddings are trained using a dataset in one language. The concept of cross-lingual word embedddings consists in projecting word embeddings into a common shared space across multiple languages. They can be useful to BLI as we can compare words in one language to words in the other; even if such words were not direct translations, the fact that

they reside in a common metric space is certainly useful. Moreover, cross-lingual word embeddings allow transfer learning when one language has far more resources than the other. There are many different ways to build cross-lingual word embeddings, which are not limited to the word-level methods that we will present in the following subsections. Ruder et al. [2019] provide an exhaustive survey of the topic.

The following methods will be closely related: they will learn an orthogonal matrix W that rotates the source embeddings X to Y. Smith et al. [2017] gave a self-consistency argument on why W should be orthogonal. First of all, let's define the similarity matrix $S = YWX^T$. The ij-th component of that matrix is

$$S_{ij} = y_i^T W x_j \tag{1}$$

which measures how similar the i-th word in Y and the j-th word in X are. Let's define an alternative similarity matrix $S' = XQY^T$, where Q is the inverse transformation. Looking at the ji-th component:

$$S'_{ji} = x_j^T Q y_i \tag{2}$$

this component also measures the similarity between i-th word in Y and the j-th word in X. So, if we want to be self consistent, both dot products are the same, which means $S_{ij} = Sji' \Rightarrow S' = S^T$. This implies that the inverse transformation Q is W^T. From the definition of inverse transformation, $WW^T = I$ (identity), which means that W must be orthogonal.

Notice that orthogonal matrices preserve vector norms, and therefore Equation 1 corresponds with the vector norm of words y_i and x_j -which is the cosine similarity if these vectors are unitary-. This is one of the reasons why orthogonal transformations are desirable. Mikolov et al. [2013b] suggested that monolingual word embeddings exhibit isomorphism across languages. Besides, there has been empirical support for such an orthogonal transformation [Xing et al., 2015, Zhang et al., 2016, Artetxe et al., 2016].

However, the orthogonal assumption does not come without limitations. Methods that achieved good results in similar pairs of languages fail when they are challenged with morphologically rich languages that are not dependent marking [Søgaard et al., 2018]. As Vulić et al. [2020] shown, not all word embedding spaces are isomorphic, and this is not only due to typological differences but also because of limited monolingual resources and undertraining of the embeddings. Patra et al. [2019] presented a semi-supervised method that relaxes the isomorphic assumption after estimating *how isomorphic* the studied languages are, achieving good results especially when the embedding spaces do not appear to be isomorphic.

In the following subsections, we divide the most relevant methods into three different categories: methods inspired by optimal transport problems, methods that use Generative Adversarial Networks and other methods.

2.1 Methods based on optimal transport problems

One of the first ways to realign word embeddings was to look at the geometry of the source and the target embedding spaces. This generally comes in hand with a minimisation of distances, which can generally be reinterpreted as an optimal transport problem.

Haghighi et al. [2008], at a time when word embeddings did not exist, was a predecessor of

methods that would appear in upcoming years. They elaborated feature vectors, and defined a generative model based on canonical correlation analysis (CCA). Matched words were close in a common latent space, which resembles the idea of cross-lingual embeddings. Additionally, they used the Hungarian algorithm [Tomizawa, 1971] for such mappings, an important algorithm for optimal transport. Ruder et al. [2018] developed these ideas and applied them to word embedding spaces. They used an expectation-maximization algorithm (Viterbi EM). First, they defined a probabilistic latent-variable model for BLI. Given the sets of embeddings X and Y, they defined the distribution

$$p_\theta(Y \mid X, m) = \prod_{(i,j)\in m} p_\theta(y_i \mid x_j) \prod_{i \in u_{trg}} p_\theta(y_i) \tag{3}$$

where m is a bipartite matching between X and Y. Note that the selected mapping m is not exhaustive, meaning that there is a set of words $u^*_{trg} = Y \setminus \{i : (i,j) \in m\}$ that is excluded from the matching. The distributions of probability are modelled as gaussians:

$$p_\theta(y_j \mid x_i, m) := \mathcal{N}(Wx_i, I) \propto \exp \|y_j - Wx_i\|_2^2 \tag{4}$$

$$p_\theta(x_i) := \mathcal{N}(\mu, I) \tag{5}$$

Where W is the orthogonal map. The optimization goes as follows. In the E-step, they find the map that maximizes the likelihood:

$$m^* = \arg\max_{m \in M} \log p_\theta(m \mid X, Y) \tag{6}$$

where M is the set of bipartite mappings between sets X and Y. In order to find the best map m^*, each arc gets assigned the weight $w_{ij} = \log p(y_i \mid x_j) - \log p(y_i) = \|y_i - Wx_j\|_2^2 - \|y_i - \mu\|_2^2$; then a version of the Jonker-Volgenant algorithm [Jonker and Volgenant, 1987] -a modification of the Hungarian algorithm optimised for sparse graphs- is employed to find the optimal map.

The M-step of the algorithm updates the parameters seen in Equations 4 and 5. Following from these equations, the joined probabilities have following form:

$$\log p(Y_{m^*} \mid X_{m^*}, m^*) = \|Y_{m^*} - WX_{m^*}\|_F^2 + C \tag{7}$$

$$log \prod_{i \in u_t rg} p(y_i) = \sum_{i \in u_{trg}} \|y_i - \mu\|_2^2 + D \tag{8}$$

where C and D are constants that can be ignored as we need to maximize both equations. Equation 7 corresponds with the orthogonal Procrustes problem [Schönemann, 1966] and has the closed form solution $W^* = UV^T$, where we have done the SVD decomposition $U\Sigma V^T = Y_m^T X_m$. Equation 8 is optimized at the centroid $\mu^* = \frac{1}{|u_{trg}|} \sum_{i \in u_{trg}} y_i$.

Zhang et al. [2017a] applied the concept of the Earth mover distance (EMD) to BLI. The EMD defines a distance between distributions, and could be interpreted as defining a set of holes (target language) and some ground that needs to fill them (source language). If we have a pair of distributions $P_1 = \sum_i u_i \delta_{x_i}$ and $P_2 = \sum_i v_i \delta_{y_i}$, the mathematical formulation is the following:

$$EMD(P_1, P_2, C) = \min_{T \in U(u,v)} \sum_{i,j} T_{ij} c(x_i, y_j) \tag{9}$$

with $c(x_i, y_j)$ representing the distance between words x_i and y_j. $U(u, v)$ is the transport polytope

$$\left\{ T \mid T_{ij} \geq 0, \sum_i T_{ij} = u_i, \sum_j T_{ij} = v_j, \forall i, j \right\} \tag{10}$$

Following these definitions, u_i would be the relative frequency of word $x_i \in X$, and v_j for $y_j \in Y$. In this setting, words from one language can be assigned to different words with relative frequency T_{ij}.

Zhang et al. [2017a] then presented two different methods. The first one was Wasserstein GAN (WGAN), where a Generative Adversarial Network (GAN) minimised the Wasserstein distance between the two distributions without. At this stage, the transformation matrix W was not enforced to be orthogonal. The second method (EMDOT) optimized the Earth Mover Distance by iteratively solving the EMD program from Equation 9 and the orthogonal Procrustes problem (problem equivalent to minimise Equation 7). They found that combining both methods (first WGAN and then EMDOT) would give the best results.

Another interesting application of optimal transport algorithms can be found in Alvarez-Melis and Jaakkola [2018]. They used the same formulation as Zhang et al. [2017a] (the same probability distributions, the Cost Function 9 and the Transport Polytope 10) but regularized the cost function by adding an entropy penalization, which allows to use the efficient Sinkhorn-Knopp algorithm [Knight, 2008]. They observed that such a formulation does not take into account the effects of a global transformation matrix W, and proposed a method that worked in pairs of points (distances) rather than the points themselves.

The Gromov-Wasserstein distance compares metric spaces rather than the sample points. Intuitively, we can think of it as a distance of distances, and as a result operates on pairs of points, turning the problem into a quadratic one. For instance, for point pairs (x_i, y_j) and $(x_{i'}, x_{j'})$, we have an associated distance $L(c(x_i, y_j), c'(x_{i'}, y_{j'}))$. We reformulate Problem 9:

$$GW(P_1, P_2, C, C') = \min_{T \in U(u,v)} \sum_{i,j,i',j'} T_{ij} T_{i'j'} L(c(x_i, y_j), c'(x_{i'}, y_{j'})) \tag{11}$$

where we generally assume that the distance functions c and c' are the same. This new cost function looks harder to optimize as now it is non-linear with a four-order tensor L. However, there are efficient first-order optimization algorithms [Peyré and Cuturi, 2020] that consist in iteratively solving an optimal transport problem. After converging to a good transport polytope T, Alvarez-Melis and Jaakkola [2018] solve the orthogonal Procrustes problem to obtain W.

2.2 Methods based on Generative Adversarial Networks

In the previous subsection we have seen methods that frame the problem of realigning word embeddings as an optimal transport problem. This is motivated by the fact that we are mapping points together, which has a cost (distance/similarity) for each pair. Hence, two languages that are correctly aligned would have relatively small distances between embeddings, but it is not clear that by optimising these cost functions we are obtaining the best possible mappings.

There have been numerous approaches that have used Generative Adversarial Networks (GANs). GANs are Neural Networks that consist of two parts. These parts can be interpreted as two competitive players that are engaging in a sort of a zero-sum game, where the outcome is the

training loss of each player [Goodfellow et al., 2014].

Arjovsky et al. [2017] presented Wasserstein-GAN (WGAN), which would be later used by Zhang et al. [2017a]. This network was proven to optimise the Earth Mover Distance 9 (which is the discrete version of Wasserstein distance). We can see their approach in Figure 2.

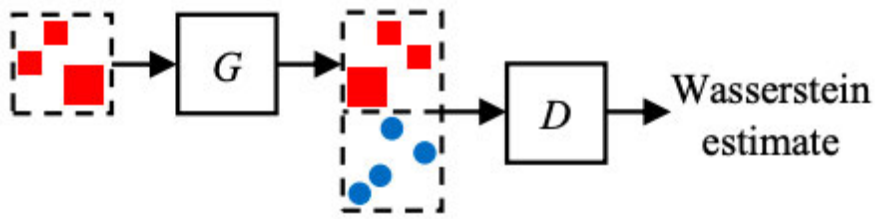

Figure 2: Scheme of a WGAN. Source: Zhang et al. [2017a]

In this adversarial scheme, the generator network (G) induces a transformation W on the source embeddings. Then the critic network (D) estimates the Wasserstein distance between the two clouds of points. Then, this estimation is sent back to G, so that it learns how to minimise it. In fact, this approach does not differ much from some of the methods we have seen in the previous subsection, as we are actually trying to minimise the Wasserstein distance, although D is just learning an approximation.

Zhang et al. [2017b] presented other ways in which GANs could be useful. Instead of using the critic for estimating the distance between points, they trained a discriminator network, which is given a point and has to guess whether that point belongs to the first or the second language. The generator is trained to fool the discriminator, producing W transformations on the first language that make it *look like* the second one. This line of work, although promising, has not achieved state of the art results.

Lample et al. [2018] presented MUSE: an adversarial approach similar to Zhang et al. [2017b]'s, but which set a milestone in the field. MUSE was either on par or it even outperformed state of the art methods for BLI, including all of them that used supervision. This was the first time that unsupervised methods improved the ones that included cross-lingual annotated data. Besides, MUSE standardised a dataset and evaluation measures with several languages available, which is still used for the comparison of BLI models.

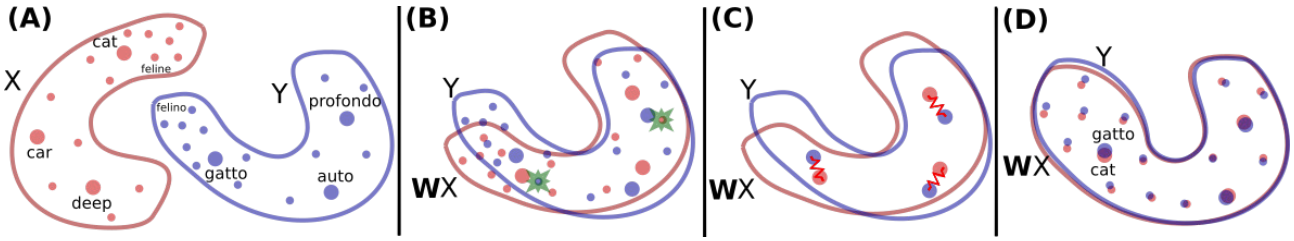

Figure 3: The four steps of MUSE. Source: Lample et al. [2018]

Figure 3 is a scheme on how MUSE works. We start by having a language pair (in this case

English and Italian) and their corresponding embeddings. In the first step (B) the rotation matrix W is learned through the adversarial game: the discriminator is given a couple of points (the two green stars from the image) and guesses whether they both come from the same distribution; weights for GAN are then updated. Then (step C) a dictionary is created from the pairs of points that are the closest to each other. The rotation matrix W is refined using this dictionary, in a supervised method called Procrustes (inspired by the orthogonal Procrustes method). Finally (D), words are mapped with each other using CSLS.

A typical way to match points from one set to the other would be to assign them to their closest neighbor from the other set (Nearest Neighbors). However, this approach has *the hubness problem* for word embeddings [Dinu and Baroni, 2014]: there are some words (hubs) that are close to too many points of the other set, making them over represented in the dictionary. Lample et al. [2018] introduced Cross-Domain Similarity Local Scaling (CSLS) in order to penalize hubs. If a (rotated) source word Wx_i has K neighbors $N(Wx_i)$, we define its mean similarity:

$$r_T(Wx_i) = \frac{1}{K} \sum_{y_j \in N(Wx_i)} \cos(Wx_i, y_j) \tag{12}$$

where $\cos(Wx_i, y_j)$ is the cosine similarity of words y_j and Wx_i. We define $r_S(y_j)$ analogously. Then, we can define CSLS:

$$CSLS(y_j, Wx_i) = 2\cos(Wx_i, y_j) - r_T(Wx_i) - r_S(y_j) \tag{13}$$

if either y_j or Wx_i are a hub, then they are penalised: CSLS enforces maps between isolated words. This mapping technique has been successful and was used in most recent approaches.

2.3 Other methods

There have been methods that applied other techniques. In general, researchers are inspired in some way by the loss functions presented in Subsection 2.1, but they might try to solve these problems after approximating the loss function.

Hoshen and Wolf [2018] presented Iterative Closest Point (ICP). This method is inspired by a well-known approach that aligns three-dimensional points. This consists on iteratively learning a transformation and enforcing its orthogonality with a regularisation term. They report that using small batches of 5000 words and PCA projection of just 50 components gives a good initialization.

Grave et al. [2019] used the Wasserstein-Procrustes loss function:

$$\underset{W \in O(d), P \in \pi(N)}{\arg\min} \|WX - YP\|_2^2 \tag{14}$$

where the permutation matrix P is part of the optimization scheme. They approximate Objective 14 by $\arg\min \left\|X^T PY\right\|_2^2$ in what is known as the Gold-Rangarajan relaxation. This new problem is solved with the Frank-Wolfe algorithm [Frank and Wolfe, 1956]. The Wasserstein-Procrustes loss function has also inspired some recent work [Alaux et al., 2019, Ramírez et al., 2020].

Finally, Ormazabal et al. [2020] is the current state of the art. Their method is somewhat different to all the previous approaches: instead of rotating the embedding space, they fix the

embeddings of the first language and for the second one their method learns new word embeddings that are already aligned with the first one. They use some weak supervision (they assume words that are written the same mean the same) and then they apply an extension of skip-gram that leverages translated context words as anchor points. However, their study was limited to a handful of European languages, which are generally thought to be isomorphic.

3 Discussion

We have seen many different approaches that rotate the embedding spaces for the unsupervised BLI task. The nature of the problem in itself is surprising: we are learning how to translate by reading text from different sources. The fact that in order to translate we are just looking at the word embeddings is astonishing.

However, this approach has serious limitations. Figure 4 shows the performance of one of the latest methods for unsupervised BLI [Zhang et al., 2019]. We can see that languages that there are languages that are reasonably translated to English (e.g. Spanish, French or Portuguese) but this is not the case of languages such as Korean or Thai. Vulić et al. [2020] exposed that it is difficult to achieve higher accuracies without accounting for the fact that these languages are not isomorphic to English. This non-isomorphism is thought to happen because of limited monolingual resources for these languages rather than having to do anything with the language structure or vocabulary.

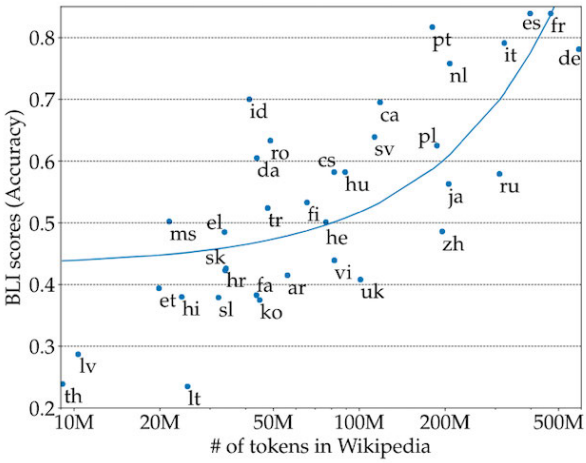

Figure 4: Performance of Zhang et al. [2019]'s model mapping from English to a target language. The x axis accounts for the amount of monolingual resources. Source: Vulić et al. [2020]

Another issue that has recently gained attention is the fact that cross-lingual word embeddings that perform well for BLI do not necessarily work well for other NLP tasks. Glavaš et al. [2019] evaluated BLI models in three downstream tasks: information retrieval, document classification and natural language inference. They reported that the performance of each model was dependent on the task, and that a higher BLI score was not correlated with better results overall.

There are some other works that criticise using BLI as a measure on how good cross-lingual embeddings are. Zhang et al. [2020], while working on supervised BLI, reported that their model was sometimes underfit. After forcing overfitting —which could lower the BLI test loss—

they discovered that the new model had a better performance in other downstream tasks. This suggests that just comparing the BLI accuracies might be insufficient.

4 Summary & Conclusion

The task of Bilingual Lexicon Induction has recently turned its attention into cross-lingual word embeddings. There are numerous unsupervised methods that, by rotating word embeddings, achieve a good performance in the BLI task. The first models for unsupervised cross-lingual word embeddings for BLI often deviated from optimal transport problems. The reason for this is that it is easy to frame the problem of realigning word embeddings as a distance optimization. These methods, although promising, could not beat supervision.

The appearance of adversarial networks, and in particular Lample et al. [2018], revealed that unsupervised methods were really competitive and could even improve supervised methods. There have been other unsupervised models that have been inspired by other methods [Hoshen and Wolf, 2018, Grave et al., 2019, Ormazabal et al., 2020].

This field has some important limitations, that ought to be considered. The hypothesis of having absolutely no bilingual data is a bit extreme, as there are in general numerous dictionaries available for most pairs of languages in the world. In particular, research generally translates English, which is a language with enormous resources. In my opinion, it would be interesting if there was a line of work that dealt with low-resource languages.

There are good reasons to think that the mentioned methods would struggle with such languages. Vulić et al. [2020] exposed that BLI methods are dependent on monolingual resources and the quality of word embeddings. The assumption that the transformation matrix W is orthogonal and that the two word embedding spaces are isomorphic could be wrong for certain pairs of languages. Methods that have relaxed this assumption have obtained promising results [Patra et al., 2019] and there should be more work in this direction. Furthermore, it is worth realigning word embeddings in settings other than BLI. It has been seen that a high BLI score does not necessarily mean a better NLP model overall [Glavaš et al., 2019].

Finally, despite the aforementioned limitations, the fact that we can realign word embeddings and obtain good translations is remarkable. This field has exploded recently and has seen a huge improvement to BLI models.

References

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations (Workshop Track)*, ICLR '13, Scottsdale, AZ, USA, 2013a.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1532–1543, Doha, Qatar, 2014.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information, 2017.

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

Ann Irvine and Chris Callison-Burch. A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics*, 43(2):273–310, June 2016. doi: 10.1162/COLI_a_00284. URL `https://www.aclweb.org/anthology/J17-2001`.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL '17, pages 451–462, Vancouver, BC, Canada, 2017.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations*, ICLR '18, Vancouver, BC, Canada, 2018.

Yedid Hoshen and Lior Wolf. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 469–478, Brussels, Belgium, 2018.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 7375–7388, Seattle, WA, USA, 2020.

Jiaming Luo, Frederik Hartmann, Enrico Santus, Yuan Cao, and Regina Barzilay. Deciphering undersegmented ancient scripts using phonetic prior, 2020.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 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, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1070. URL `https://www.aclweb.org/anthology/P19-1070`.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630, May 2019. ISSN 1076-9757. doi: 10.1613/jair.1.11640. URL `https://doi.org/10.1613/jair.1.11640`.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P08-1088`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of the 2017 International Conference on Learning Representations*, ICLR '17, Toulon, France, 2017.

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

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '15, pages 1006–1011, Denver, CO, USA, 2015.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '16, pages 1307–1317, San Diego, CA, USA, 2016.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP '16, pages 2289–2294, Austin, TX, USA, 2016.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1072. URL https://www.aclweb.org/anthology/P18-1072.

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

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 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, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1018. URL https://www.aclweb.org/anthology/P19-1018.

N. Tomizawa. On some techniques useful for solution of transportation network problems. *Networks*, 1(2):173–194, 1971. doi: https://doi.org/10.1002/net.3230010206. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/net.3230010206.

Sebastian Ruder, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard. A discriminative latent-variable model for bilingual lexicon induction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 458–468, Brussels, Belgium, 2018.

R. Jonker and A. Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, November 1987. ISSN 0010-485X. doi: 10.1007/BF02278710. URL https://doi.org/10.1007/BF02278710.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31: 1–10, 1966.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 1934–1945, Copenhagen, Denmark, 2017a.

David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1214. URL https://www.aclweb.org/anthology/D18-1214.

Philip A. Knight. The sinkhorn–knopp algorithm: Convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, jan 2008. doi: 10.1137/060659624. URL https://doi.org/10.1137%2F060659624.

Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada, July 2017b. Association for Computational Linguistics. doi: 10.18653/v1/P17-1179. URL https://www.aclweb.org/anthology/P17-1179.

Georgiana Dinu and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. 12 2014.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with Wasserstein Procrustes. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, AISTATS '2019, pages 1880–1890, Naha, Japan, 2019.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956. doi: https://doi.org/10.1002/nav.3800030109. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800030109`.

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. Unsupervised hyperalignment for multilingual word embeddings, 2019.

Guillem Ramírez, Rumen Dangovski, Preslav Nakov, and Marin Soljačić. On a novel application of wasserstein-procrustes for unsupervised cross-lingual learning, 2020.

Aitor Ormazabal, Mikel Artetxe, Aitor Soroa, Gorka Labaka, and Eneko Agirre. Beyond offline mapping: Learning cross lingual word embeddings through context anchoring, 2020.

Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. Are girls neko or shōjo? cross-lingual alignment of non-isomorphic embeddings with iterative normalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3180–3189, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1307. URL `https://www.aclweb.org/anthology/P19-1307`.

Mozhi Zhang, Yoshinari Fujinuma, Michael J. Paul, and Jordan Boyd-Graber. Why overfitting isn't always bad: Retrofitting cross-lingual word embeddings to dictionaries, 2020.