Unsupervised Multilingual Alignment using Wasserstein Barycenter

Xin Lian[1,3] , **Kshitij Jain**[2] , **Jakub Truszkowski**[2] , **Pascal Poupart**[1,2,3] and **Yaoliang Yu**[1,3]

[1]University of Waterloo, Waterloo, Canada
[2]Borealis AI, Waterloo, Canada
[3]Vector Institute, Toronto, Canada
{x9lian, k22jain, yaoliang.yu, ppoupart}@uwaterloo.ca, {jakub.truszkowski}@borealisai.com

Abstract

We study unsupervised multilingual alignment, the problem of finding word-to-word translations between multiple languages without using any parallel data. One popular strategy is to reduce multilingual alignment to the much simplified bilingual setting, by picking one of the input languages as the pivot language that we transit through. However, it is well-known that transiting through a poorly chosen pivot language (such as English) may severely degrade the translation quality, since the assumed transitive relations among all pairs of languages may not be enforced in the training process. Instead of going through a rather arbitrarily chosen pivot language, we propose to use the Wasserstein barycenter as a more informative "mean" language: it encapsulates information from all languages and minimizes all pairwise transportation costs. We evaluate our method on standard benchmarks and demonstrate state-of-the-art performances.

1 Introduction

Many natural language processing tasks, such as part-of-speech tagging, machine translation and speech recognition, rely on learning a distributed representation of words. Recent developments in computational linguistics and neural language modeling have shown that word embeddings can capture both semantic and syntactic information. This led to the development of the zero-shot learning paradigm as a way to address the manual annotation bottleneck in domains where other vector-based representations must be associated with word labels. This is a fundamental step to make natural language processing more accessible. A key input for machine translation tasks consists of embedding vectors for each word. Mikolov *et al.* [2013b] were the first to release their pre-trained model and gave a distributed representation of words. After that, more software for training and using word embeddings emerged.

The rise of continuous word embedding representations has revived research on the bilingual lexicon alignment problem [Rapp, 1995; Fung, 1995], where the initial goal was to learn a small dictionary of a few hundred words by leveraging statistical similarities between two languages. Mikolov *et al.*

[2013a] formulated bilingual word embedding alignment as a quadratic optimization problem that learns an explicit *linear* mapping between word embeddings, which enables us to even infer meanings of out-of-dictionary words [Zhang *et al.*, 2016; Dinu *et al.*, 2015; Mikolov *et al.*, 2013a]. Xing *et al.* [2015] showed that restricting the linear mapping to be orthogonal further improves the result. These pioneering works required some parallel data to perform the alignment. Later on, [Smith *et al.*, 2017; Artetxe *et al.*, 2017; Artetxe *et al.*, 2018a] reduced the need of supervision by exploiting common words or digits in different languages, and more recently, unsupervised methods that rely solely on monolingual data have become quite popular [Gouws *et al.*, 2015; Zhang *et al.*, 2017b; Zhang *et al.*, 2017a; Lample *et al.*, 2018; Artetxe *et al.*, 2018b; Dou *et al.*, 2018; Hoshen and Wolf, 2018; Grave *et al.*, 2019].

Encouraged by the success on bilingual alignment, the more ambitious task that aims at *simultaneously and unsupervisedly* aligning multiple languages has drawn a lot of attention recently.

A naive approach that performs all pairwise bilingual alignment *separately* would not work well, since it fails to exploit all language information, especially when there are low resource ones. A second approach is to align all languages to a pivot language, such as English [Smith *et al.*, 2017], allowing us to exploit recent progresses on bilingual alignment while still using information from all languages.

More recently, [Chen and Cardie, 2018; Taitelbaum *et al.*, 2019b; Taitelbaum *et al.*, 2019a; Alaux *et al.*, 2019; Wada *et al.*, 2019] proposed to map all languages into the same language space and train all language pairs simultaneously. Please refer to the related work section for more details.

In this work, we first show that the existing work on unsupervised multilingual alignment (such as [Alaux *et al.*, 2019]) amounts to *simultaneously* learning an arithmetic "mean" language from all languages and aligning all languages to the common mean language, instead of using a rather arbitrarily pre-determined input language (such as English). Then, we argue for using the (learned) Wasserstein barycenter as the pivot language as opposed to the previous arithmetic barycenter, which, unlike the Wasserstein barycenter, fails to preserve distributional properties in word embeddings. Our approach exploits available information from all languages to enforce coherence among language spaces by enabling accurate com-

positions between language mappings. We conduct extensive experiments on standard publicly available benchmark datasets and demonstrate competitive performance against current state-of-the-art alternatives.

2 Multilingual Lexicon Alignment

In this section we set up the notations and define our main problem: the multilingual lexicon alignment problem.

Given m languages $\mathcal{L}_1, \ldots, \mathcal{L}_m$, each represented by a vocabulary $\mathcal{V}_i$ consisting of n_i respective words. Following Mikolov *et al.* [2013a], we assume a monolingual word embedding $\mathbf{X}_i = [\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,n_i}]^\top \in \mathbb{R}^{n_i \times d_i}$ for each language $\mathcal{L}_i$ has been trained *independently* on its own data. We are interested in finding *all* pairwise mappings $\mathsf{T}_{i \to k} : \mathbb{R}^{d_i} \to \mathbb{R}^{d_k}$ that translate a word $\mathbf{x}_{i,j_i}$ in language $\mathcal{L}_i$ to a corresponding word $\mathbf{x}_{k,j_k} = \mathsf{T}_{i \to k}(\mathbf{x}_{i,j_i})$ in language $\mathcal{L}_k$. In the following, for the ease of notation, we assume w.l.o.g. that $n_i \equiv n$ and $d_i \equiv d$. Note that we do not have access to any parallel data, i.e., we are in the much more challenging unsupervised learning regime.

Our work is largely inspired by that of Alaux *et al.* [2019], which we review below first. Along the way we point out some crucial observations that motivated our further development. Alaux *et al.* [2019] employ the following joint alignment approach that minimizes the total sum of mis-alignment costs between every pair of languages:

$$\min_{Q_i \in \mathfrak{O}_d, P_{ik} \in \mathcal{P}_n} \sum_{i=1}^{m} \sum_{k=1, k \neq i}^{m} \|\mathbf{X}_i Q_i - P_{ik} \mathbf{X}_k Q_k\|^2, \quad (1)$$

where $Q_i \in \mathfrak{O}_d$ is a $d \times d$ orthogonal matrix and $P_{ik} \in \mathcal{P}_n$ is an $n \times n$ permutation matrix[1]. Since Q_i is orthogonal, this approach ensures transitivity among word embeddings: Q_i maps the i-th word embedding space $\mathbf{X}_i$ into a common space $\mathbf{X}$, and conversely $Q_i^{-1} = Q_i^\top$ maps $\mathbf{X}$ back to $\mathbf{X}_i$. Thus, $Q_i Q_k^\top$ maps $\mathbf{X}_i$ to $\mathbf{X}_k$, and if we transit through an intermediate word embedding space $\mathbf{X}_t$, we still have the desired transitive property $Q_i Q_t^\top \cdot Q_t Q_k^\top = Q_i Q_k^\top$.

The permutation matrix P_{ik} serves as an "inferred" correspondence between words in language $\mathcal{L}_i$ and language $\mathcal{L}_k$. Naturally, we would again expect some form of transitivity in these pairwise correspondences, i.e., $P_{ik} \cdot P_{kt} \approx P_{it}$, which, however, is not enforced in (1). A simple way to fix this is to decouple P_{ik} into the product $P_i^\top P_k$, in the same way as how we dealt with Q_i. This leads to the following variant:

$$\underset{Q_i \in \mathfrak{O}_d, P_i \in \mathcal{P}_n}{\operatorname{argmin}} \sum_{i=1}^{m} \sum_{k=1}^{m} \|P_i \mathbf{X}_i Q_i - P_k \mathbf{X}_k Q_k\|^2 \quad (2)$$

$$= \underset{Q_i \in \mathfrak{O}_d, P_i \in \mathcal{P}_n}{\operatorname{argmin}} \sum_{i=1}^{m} \|P_i \mathbf{X}_i Q_i - \frac{1}{m} \sum_{k=1}^{m} P_k \mathbf{X}_k Q_k\|^2 \quad (3)$$

$$= \underset{Q_i \in \mathfrak{O}_d, P_i \in \mathcal{P}_n}{\operatorname{argmin}} \min_{\bar{\mathbf{X}} \in \mathbb{R}^{n \times d}} \sum_{i=1}^{m} \|P_i \mathbf{X}_i Q_i - \bar{\mathbf{X}}\|^2 \quad (4)$$

[1] Alaux *et al.* [2019] also introduced weights $\alpha_{ik} > 0$ to encode the relative importance of the language pair (i, k).
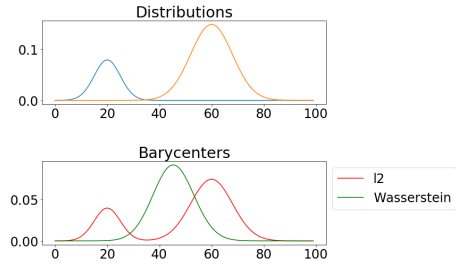

Figure 1: Comparing the Wasserstein barycenter and arithmetic mean (bottom panel) for two input distributions (top panel).

where Eq. 3 follows from the definition of variance and $\bar{\mathbf{X}}$ in Eq. 4 admits the closed-form solution:

$$\bar{\mathbf{X}} = \frac{1}{m} \sum_{k=1}^{m} P_k \mathbf{X}_k Q_k. \quad (5)$$

Thus, had we known the *arithmetic* "mean" language $\bar{\mathbf{X}}$ beforehand, the joint alignment approach of Alaux *et al.* [2019] would reduce to a separate alignment of each language $\mathbf{X}_i$ to the "mean" language $\bar{\mathbf{X}}$ that serves as the pivot. An efficient optimization strategy would then consist of alternating between separate alignment (i.e., computing Q_i and P_i) and computing the pivot language (i.e., (5)).

We now point out two problems in the above formulation. First, a permutation assignment is a 1-1 correspondence that completely ignores polysemy in natural languages, that is, a word in language $\mathcal{L}_i$ can correspond to multiple words in language $\mathcal{L}_k$. To address this, we propose to relax the permutation P_i into a coupling matrix that allows splitting a word into different words. Second, the pivot language in (5), being a simple arithmetic average, may be statistically very different from any of the m given languages, see Figure 1 and below. Besides, intuitively it is perhaps more reasonable to allow the pivot language to have a larger dictionary so that it can capture all linguistic regularities in all m languages. To address this, we propose to use the Wasserstein barycenter as the pivot language.

The advantage of using Wasserstein barycenter instead of the arithmetic average is that the Wasserstein metric gives a natural geometry for probability measures supported on a geometric space. In Figure 1, we demonstrate the difference between Wasserstein Barycenter and arithmetic average of two input distributions.

It is intuitively clear that the Wasserstein barycenter preserves the geometry of the input distributions.

3 Our Approach

We take a probabilistic approach, treating each language $\mathcal{L}_i$ as a probability distribution over its word embeddings:

$$\pi_i = \sum_{j=1}^{n} p_{ij} \delta_{\mathbf{x}_j^i} \quad (6)$$

where p_{ij} is the probability of occurrence of the j-th word $\mathbf{x}_j^i$ in language $\mathcal{L}_i$ (often approximated by the relative frequency of word $\mathbf{x}_j^i$ in its training documents), and $\delta_{\mathbf{x}_j^i}$ is the

unit mass at $\mathbf{x}_j^i$. We project word embeddings into a common space through the orthogonal matrix $Q_i \in \mathfrak{O}_d$. Taking a word $\mathbf{x}^i$ from each language $\mathcal{L}_i$, we associate a cost $c(Q_1\mathbf{x}^1, \ldots, Q_m\mathbf{x}^m) \in \mathbb{R}_+$ for bundling these words in our joint translation. To allow polysemy, we find a joint distribution π with fixed marginals π_i so that the average cost

$$\int c(Q_1\mathbf{x}^1, \ldots, Q_m\mathbf{x}^m) \, d\pi(\mathbf{x}^1, \ldots, \mathbf{x}^m) \qquad (7)$$

is minimized. If we fix Q_i, then the above problem is known as multi-marginal optimal transport [Gangbo and Świkech, 1998].

To simplify the computation, we take the pairwise approach of Alaux *et al.* [2019], where we set the joint cost c as the total sum of all pairwise costs:

$$c(\mathbf{x}^1, \ldots, \mathbf{x}^m) = \sum_{i,k} \|\mathbf{x}^i - \mathbf{x}^j\|^2. \qquad (8)$$

Interestingly, with this choice, we can significantly simplify the numerical computation of the multi-marginal optimal transport.

We recall the definition of Wasserstein barycenter ν of m given probability distributions $\pi_1, \ldots, \pi_m$:

$$\nu = \arg\min_\mu \sum_{i=1}^m \lambda_i \cdot \mathsf{W}_2^2(\pi_i, \mu), \qquad (9)$$

where $\lambda \geq 0$ are the weights, and the (squared) Wasserstein distance W_2^2 is given as:

$$\mathsf{W}_2^2(\pi_i, \mu) = \min_{\Pi_i \in \Gamma(\pi_i, \mu)} \int \|\mathbf{x} - \mathbf{y}\|^2 d\Pi_i(\mathbf{x}, \mathbf{y}). \qquad (10)$$

The notation $\Gamma(\pi_i, \mu)$ denotes all joint probability distributions (i.e. couplings) Π_i with (fixed) marginal distributions π_i and μ. As proven by Agueh and Carlier [2011], with the pairwise distance (8), the multi-marginal problem in (7) and the barycenter problem in (9) are formally equivalent. Hence, from now on we will focus on the latter since efficient computational algorithms for it exist. We use the push-forward notation $(Q_i)_\# \pi_i$ to denote the distribution of $Q_i\mathbf{x}^i$ when $\mathbf{x}^i$ follows the distribution π_i. Thus, we can write our approach succinctly as:

$$\min_\mu \min_{Q_i \in \mathfrak{O}_d} \sum_{i=1}^m \lambda_i \cdot \mathsf{W}_2^2[(Q_i)_\# \pi_i, \mu], \qquad (11)$$

where the barycenter μ serves as the pivot language in some common word embedding space. Unlike the arithmetic average in (5), the Wasserstein barycenter can have a much larger support (dictionary size) than the m given language distributions.

We can again apply the alternating minimization strategy to solve (11): fixing all orthogonal matrices Q_i, we find the Wasserstein barycenter using an existing algorithm of [Cuturi and Doucet, 2014] or [Claici *et al.*, 2018]; fixing the Wasserstein barycenter μ, we solve each orthogonal matrix Q_i separately:

$$\min_{Q_i \in \mathfrak{O}_d} \min_{\Pi_i \in \Gamma(\pi_i, \mu)} \int \|Q_i\mathbf{x} - \mathbf{y}\|^2 d\Pi_i(\mathbf{x}, \mathbf{y}). \qquad (12)$$

Algorithm 1: Barycenter Alignment

Input: Language distribution $L_i = (X_i, p_i)_{i=1}^m, p$
Output: Translation for L_k and L_m
for $i = 1, \ldots, m$ **do**
$\quad \mathbf{X}_i \leftarrow \mathbf{X}_i - \mathsf{mean}(\mathbf{X}_i)$
$\quad \{C_i\} \leftarrow cosine\_dist(\mathbf{X}_{i,j}, \mathbf{X}_{i,k}) \forall j, k$
$\quad \Pi_i \leftarrow \mathsf{GW}(C_i, C_1, p_i, p_1)$
$\quad U\Sigma V^\top \leftarrow \mathsf{SVD}(\mathbf{X}_i^\top \Pi_i \mathbf{X}_1)$
$\quad Q_i \leftarrow UV^\top$
$\quad \mathbf{X}_i \leftarrow \mathbf{X}_i Q_i$
while *not converged* **do**
$\quad \nu \leftarrow \mathsf{WB}(\pi_1, \cdots, \pi_m; \lambda_1, \cdots, \lambda_m)$
$\quad$ **for** $i = 1, \ldots, m$ **do**
$\quad\quad \Pi_i \leftarrow \mathsf{OT}(\pi_i, \nu)$
$\quad\quad U\Sigma V^\top \leftarrow \mathsf{SVD}(\mathbf{X}_i^\top \Pi_i \mathbf{Y})$
$\quad\quad Q_i \leftarrow UV^\top$
$\quad\quad \mathbf{X}_i \leftarrow \mathbf{X}_i Q_i$;
return $(\Pi_1, \ldots, \Pi_m, Q_1, \ldots, Q_m)$

For fixed coupling $\Pi_i \in \mathbb{R}^{n \times s}$, where s is the dictionary size for the barycenter μ, the integral can be simplified as:

$$\sum_{jl} (\Pi_i)_{jl} \|Q_i\mathbf{x}_j^i - \mathbf{y}_l\|^2 \equiv -\langle \mathbf{X}_i^\top \Pi_i \mathbf{Y}, Q_i \rangle. \qquad (13)$$

Thus, using the well-known theorem of Schönemann [1966], Q_i is given by the closed-form solution $U_i V_i^\top$, where $U_i \Sigma_i V_i^\top = \mathbf{X}_i^\top \Pi_i \mathbf{Y}$ is the singular value decomposition. Our approach is presented in Algorithm 1.

4 Experiments

We evaluate our algorithm on two standard publicly available datasets: MUSE [Lample *et al.*, 2018] and XLING [Glavas *et al.*, 2019]. The MUSE benchmark is a high-quality dictionary containing up to 100k pairs of words and has now become a standard benchmark for cross-lingual alignment tasks [Lample *et al.*, 2018]. On this dataset, we conducted an experiment with 6 European languages: English, French, Spanish, Italian, Portuguese, and German. The MUSE dataset contains a direct translation for any pair of languages in this set. We also conducted an experiment with the XLING dataset with a more diverse set of languages: Croatian (HR), English (EN), Finnish (FI), French (FR), German (DE), Italian (IT), Russian (RU), and Turkish (TR). In this set of languages, we have languages coming from three different Indo-European branches, as well as two non-Indo-European languages (FI from Uralic and TR from the Turkic family) [Glavas *et al.*, 2019].

4.1 Implementation Details

To speed up the computation, we took a similar approach as Alaux *et al.* [2019] and initialized space alignment matrices with the Gromov-Wasserstein approach [Alvarez-Melis and Jaakkola, 2018] applied to the first 5k vectors (Alaux *et al.* [2019] used the first 2k vectors) and with regularization parameter ϵ of $5e^{-5}$. After the initialization, we use the space alignment matrices to map all languages into the language

space of the first language. Multiplying all language embedding vectors with the corresponding space alignment matrix, we realign all languages into a common language space. In the common space, we compute the Wasserstein barycenter of all projected language distributions. The support locations for the barycenter are initialized with random samples from a standard normal distribution.

The next step is to compute the optimal transport plans from the barycenter distribution to all language distributions. After obtaining optimal transport plans T_i from the barycenter to every language $\mathcal{L}_i$, we can imply translations from $\mathcal{L}_i$ to $\mathcal{L}_j$ from the coupling $T_i T_j^\top$. The coupling is not necessarily a permutation matrix, and indicates the probability with which a word corresponds to another. Method and code for computing accuracies of bilingual translation pairs are borrowed from Alvarez-Melis and Jaakkola [2018].

4.2 Baselines

We compare the results of our method on MUSE with the following methods: 1) Procrustes Matching with RSLS as similarity function to imply translation pairs [Lample *et al.*, 2018]; 2) the state-of-the-art bilingual alignment method, Gromov-Wasserstein alignment (GW) [Alvarez-Melis and Jaakkola, 2018]; 3) the state-of-the-art multilingual alignment method (UMH) [Alaux *et al.*, 2019]; 4) bilingual alignment with multilingual auxiliary information (MPPA) [Taitelbaum *et al.*, 2019b]; and 5) unsupervised multilingual word embeddings trained with multilingual adversarial training [Chen and Cardie, 2018].

We compare the results of our method on XLING dataset with Ranking-Based Optimization (RCSLS) [Joulin *et al.*, 2018], solution to the Procrustes Problem (PROC) [Artetxe *et al.*, 2018b; Lample *et al.*, 2018; Glavas *et al.*, 2019], Gromov-Wasserstein alignment (GW) [Alvarez-Melis and Jaakkola, 2018], and VECMAP [Artetxe *et al.*, 2018b]. RCLS and PROC are supervised methods, while GW and VECMAP are both unsupervised methods.

The translation accuracies for Gromov-Wasserstein are computed using the source code released by [Alvarez-Melis and Jaakkola, 2018]. For the multilingual alignment method (UMH) [Alaux *et al.*, 2019], and the two multilingual adversarial methods [Chen and Cardie, 2018], [Taitelbaum *et al.*, 2019b], we directly compare our accuracies to previous methods as reported in [Glavas *et al.*, 2019].

4.3 Results

Table 2 depicts precision@1 results for all bilingual tasks on the MUSE benchmark [Lample *et al.*, 2018]. For most language pairs, our method Barycenter Alignment (BA) outperforms all current unsupervised methods. Our barycenter approach infers a "potential universal language" from input languages. Transiting through that universal language, we infer translation for all pairs of languages. From the experimental results in Table 2, we can see that our approach is clearly at an advantage and it benefits from using the information from all languages. Our method achieves statistically significant improvement for 22 out of 30 language pairs ($p \leq 0.05$, McNemar's test, one-sided).

Table 3 shows mean average precision (MAP) for 10 bilingual tasks on the XLING dataset [Glavas *et al.*, 2019].

In Table 1, we show several German to English translations and compare the results to Gromov-Wasserstein direct bilingual alignment. Our method is capable of incorporating both the semantic and syntactic information of one word. For example, the top ten predicted English translations for the German word *München*, are "Cambridge, Oxford, Munich, London, Birmingham, Bristol, Edinburgh, Dublin, Hampshire, Baltimore". In this case, we hit the English translation Munich. What's more important in this example is that all predicted English words are the name of some city. Therefore, our method is capable of implying *München* is a city name. Another example is the German word *sollte*, which means "should" in English. The top five words predicted for *sollte* are syntactically correct - *would*, *could*, *will*, *should* and *might* are all modal verbs. The last three examples show polysemous words, and in all these cases our method performs better than the Gromov-Wasserstein. For German word *erschienen*, our algorithm predicts all three words *released*, *appeared*, and *published* in the top ten translations as compared to Gromov-Wasserstein which only predicts *published* .

4.4 Ablation Study

In this section, we show the impact of some of our design choices and hyperparameters. One of the parameters is the number of support locations. In theory, the optimal barycenter distribution could have as many support locations as the sum of the total number of support locations for all input distributions. In Figure 2, we show the impact on translation performance when we have a different number of support locations. Let n_j be the number of words we have in language L_j. We picked the three most representative cases: the average number of words $= \sum_{j=1}^m n_j/m$, twice the average number of words $= 2\sum_{j=1}^m n_j/m$, and the total number of words $= \sum_{j=1}^m n_j$. As we increase the number of support locations for the barycenter distribution, we can see in Figure 2 that the performance for language translation improves. However, when we increase the number of support locations for the barycenter, the algorithm becomes costly. Therefore, in an effort to balance accuracy and computational complexity, we decided to use 10000 support locations (twice the average number of words).

We also conducted a set of experiments to determine whether the inclusion of distant languages increases bilingual translation accuracy. Excluding two non-Indo-European languages Finnish and Turkish, we calculated the barycenter of Croatian (HR), English (EN), French (FR), German (DE), and Italian (IT). Figure 3 contains results for common bilingual pairs. The red bar show the bilingual translation accuracy when translating through the barycenter for all languages including Finnish and Turkish, whereas blue bar indicate the accuracy of translations that use the barycenter of the five Indo-European languages.

5 Related Work

We briefly describe related work on supervised and unsupervised techniques for bilingual and multilingual alignment.

German	English	GW Prediction	BA Prediction
München	Munich	London, Dublin, Oxford, Birmingham, Wellington Glasgow, Edinburgh, Cambridge, Toronto, Hamilton	Cambridge, Oxford, **Munich**, London, Birmingham Rristol, Edinburgh, Dublin, Hampshire, Baltimore
sollte	should	would, could, might, will, needs, supposed, put, willing, wanted, meant	would, could, will, supposed, might, meant, needs, expected, able, **should**
erschienen	released appeared published	**published**, editions, publication, edition printed, volumes, compilation publications, releases, titled	**published**, editions, volumes, publication **released**, titled, printed **appeared**, edition, compilation
aufgenommen	admitted recorded taken, included	**recorded**, **taken**, recording, selected roll, placed, performing eligible, motion, assessed	**recorded**, **taken**, recording, **admitted** selected, sample, **included** track, featured, mixed
viel	lots, lot much	**much**, **lot**, little, more, less bit, too, plenty, than, better	**much**, **lot**, little, less, too more, than, bit, far, **lots**

Table 1: German-to-English translation prediction comparing results by 1) using GW alignment to imply direct bilingual mapping and 2) using Barycenter Alignment method described in Algorithm 1. We show top-10 translations of both methods. Last three examples show the polysemous words and their corresponding translations.
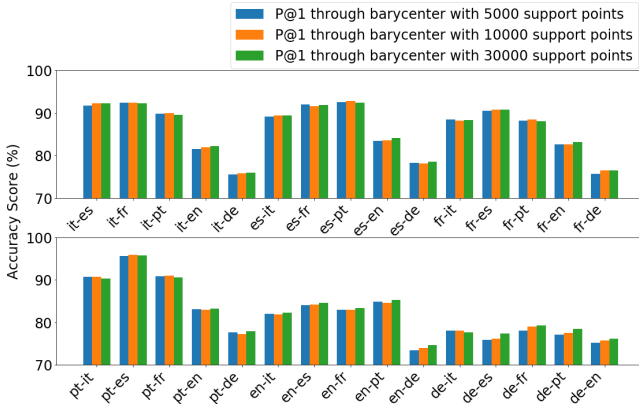

Figure 2: Accuracies for language pairs using different numbers of support locations for the barycenter. In our experimental setup, we have 5000 words in each language.
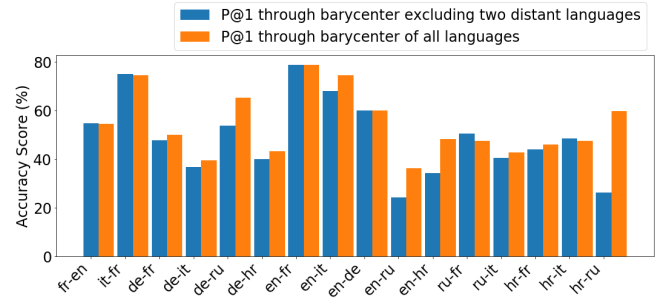

Figure 3: This graph shows the accuracy of bilingual translation pairs. The red bar indicate translation accuracy using the barycenter of all languages (HR, EN, FI, FR, DE, RU, IT, TR), while the blue bar correspond to the barycenter of (HR, EN, FR, DE, IT, RU).

5.1 Supervised Bilingual Alignment

Mikolov *et al.* [2013a] formulated the problem of aligning word embeddings as a quadratic optimization problem to find an explicit linear map Q between the word embeddings $\mathbf{X}_1$ and $\mathbf{X}_2$ of two languages.

$$\min_Q \ ||\mathbf{X}_1 Q - P\mathbf{X}_2||_2^2 \qquad (14)$$

This setting is supervised since the assignment matrix P that maps words of one language to another is known. Later, [Xing *et al.*, 2015] showed that the results can be improved by restricting the linear mapping Q to be orthogonal. This corresponds to Orthogonal Procrustes [Schönemann, 1966].

5.2 Unsupervised Bilingual Alignment

In the unsupervised setting, the assignment matrix P between words is unknown, and we resort to the joint optimization:

$$\min_Q \min_P \ ||\mathbf{X}_1 Q - P\mathbf{X}_2||_2^2. \qquad (15)$$

As a result, the optimization problem becomes non-convex and therefore more challenging. The problem can be relaxed into a (convex) semidefinite program.This method provides high accuracy at the expense of high computation complexity.

Therefore, it is not suitable for large scale problems. Another way to solve (15) is to use Block Coordinate Relaxation, where we iteratively optimize each variable with other variables fixed. When Q is fixed, optimizing P can be done with the Hungarian algorithm in $O(n^3)$ time (which is prohibitive since n is the number of words). Cuturi and Doucet [2014] developed an efficient approximation (complexity $O(n^2)$) achieved by adding a negative entropy regularizer.Observing that both P and Q preserve the intra-language distances, Alvarez-Melis and Jaakkola [2018] cast the unsupervised bilingual alignment problem as a Gromov-Wasserstein optimal transport problem, and give a solution with minimum hyper-parameter to tune.

5.3 Multilingual Alignment

In multilingual alignment, we seek to align multiple languages together while taking advantage of inter-dependencies to ensure consistency among them. A common approach consists of mapping each language to a common space X_0 by minimizing some loss function l:

$$\min_{Q_i \in \mathfrak{O}_d, P_i \in \mathcal{P}_n} \sum_i l(X_i Q_i, P_i X_0) \qquad (16)$$

The common space may be a pivot language such as English [Smith *et al.*, 2017; Lample *et al.*, 2018; Joulin *et al.*, 2018]. Nakashole and Flauger [2017] and Alaux *et al.* [2019] showed that constraining coherent word alignments between triplets of nearby languages improves the quality of induced

	it-es	it-fr	it-pt	it-en	it-de	es-it	es-fr	es-pt	es-en	es-de
GW	**92.63**	91.78	89.47	80.38	74.03	89.35	91.78	92.82	81.52	75.03
GW^o	-	-	-	75.2	-	-	-	-	80.4	-
PA	87.3	87.1	81.0	76.9	67.5	83.5	85.8	87.3	82.9	68.3
MAT+MPPA	87.5	87.7	81.2	77.7	67.1	83.7	85.9	86.8	83.5	66.5
MAT+MPSR	88.2	88.1	82.3	77.4	69.5	84.5	86.9	87.8	83.7	69.0
UMH	87.0	86.7	80.4	79.9	67.5	83.3	85.1	86.3	**85.3**	68.7
BA	92.32	**92.54***	**90.14**	**81.84***	**75.65***	89.38	92.19	92.85	83.5*	**78.25***

	fr-it	fr-es	fr-pt	fr-en	fr-de	pt-it	pt-es	pt-fr	pt-en	pt-de
GW	88.0	90.3	87.44	82.2	74.18	90.62	**96.19**	89.9	81.14	74.83
GW^o	-	-	-	82.1	-	-	-	-	-	-
PA	83.2	82.6	78.1	82.4	69.5	81.1	91.5	84.3	80.3	63.7
MAT+MPPA	83.1	83.6	78.7	82.2	69.0	82.6	92.2	84.6	80.2	63.7
MAT+MPSR	83.5	83.9	79.3	81.8	71.2	82.6	92.7	86.3	79.9	65.7
UMH	82.5	82.7	77.5	83.1	69.8	81.1	91.7	83.6	82.1	64.4
BA	**88.38**	**90.77***	**88.22***	**83.23***	**76.63***	**91.08**	96.04	**91.04***	**82.91***	**76.99***

	en-it	en-es	en-fr	en-pt	en-de	de-it	de-es	de-fr	de-pt	de-en	**Average**
GW	80.84	82.35	81.67	83.03	71.73	75.41	72.18	77.14	74.38	72.85	82.84
GW^o	78.9	81.7	81.3	-	71.9	-	-	-	-	72.8	78.04
PA	77.3	81.4	81.1	79.9	73.5	69.5	67.7	73.3	59.1	72.4	77.98
MAT+MPPA	78.5	82.2	82.7	81.3	74.5	70.1	68.0	75.2	61.1	72.9	78.47
MAT+MPSR	78.8	82.5	82.4	81.5	74.8	72.0	69.6	76.7	63.2	72.9	79.29
UMH	78.9	82.5	82.7	82.0	**75.1**	68.7	67.2	73.5	59.0	75.5	78.46
BA	**81.45***	**84.26***	**82.94***	**84.65***	74.08*	**78.09***	**75.93***	**78.93***	**77.18***	**75.85***	**84.24**

Table 2: Pairs of languages in multilingual alignment problem results for English, German, French, Spanish, Italian, and Portuguese. All reported results are precision@1 percentage. The method achieving the highest precision for each bilingual pair is highlighted in bold. Methods we are comparing to in the table are: Procrustes Matching with CSLS metric to infer translation pairs (PA) [Lample *et al.*, 2018]; Gromov-Wasserstein alignment (GW) [Alvarez-Melis and Jaakkola, 2018] (reproduced by us using their source code); GW^o refers to the results reported by Alvarez-Melis and Jaakkola [2018] in the paper; bilingual alignment with multilingual auxiliary information (MPPA) [Taitelbaum *et al.*, 2019b]; Multilingual pseudo-supervised refinement method [Chen and Cardie, 2018]; multilingual alignment method (UMH) [Alaux *et al.*, 2019]. Asterisks denote significant differences between BA and GW (McNemar's test, one-sided), the only methods for which predictions were available.

	en-de	it-fr	hr-ru	en-hr	de-fi	tr-fr	ru-it	fi-hr	tr-hr	tr-ru
PROC (1k)	0.458	0.615	0.269	0.225	0.264	0.215	0.360	0.187	0.148	0.168
PROC (5k)	0.544	0.669	0.372	0.336	0.359	0.338	0.474	0.294	0.259	0.290
PROC-B	0.521	0.665	0.348	0.296	0.354	0.305	0.466	0.263	0.210	0.230
RCSLS (1k)	0.501	0.637	0.291	0.267	0.288	0.247	0.383	0.214	0.170	0.191
RCSLS (5k)	0.580	0.682	0.404	0.375	0.395	0.375	0.491	0.321	0.285	0.324
VECMAP	0.521	0.667	0.376	0.268	0.302	0.341	0.463	0.280	0.223	0.200
GW	0.667	0.751	**0.683**	0.123	0.454	0.485	0.508	**0.634**	**0.482**	0.295
BA	**0.683**	**0.799**	0.667	**0.646**	**0.508**	**0.513**	**0.512**	0.601	0.481	**0.355**

Table 3: Mean average precision (MAP) accuracies of several current methods on XLING dataset.

bilingual lexicons. Chen and Cardie [2018] extended the work of [Lample *et al.*, 2018] to the multilingual case using adversarial algorithms. Taitelbaum *et al.* extended Procrustes Matching to the multi-Pairwise case [Taitelbaum *et al.*, 2019b], and also designed an improved representation of the source word using auxiliary languages [Taitelbaum *et al.*, 2019a].

6 Conclusion

In this paper, we discussed previous attempts to solve the multilingual alignment problem, compared similarity between the approaches and pointed out a problem with existing formulations. Then we proposed a new method using the Wasserstein barycenter as a pivot for the multilingual alignment problem. At the core of our algorithm lies a new inference method based on an optimal transport plan to predict the similarity between words. Our barycenter can be interpreted as a virtual universal language, capturing information from all languages. The algorithm we proposed improves the accuracy of pairwise translations compared to the current state-of-the-art method..

References

[Agueh and Carlier, 2011] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[Alaux *et al.*, 2019] Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. Unsupervised Hyper-alignment for Multilingual Word Embeddings. In *ICLR*, 2019.

[Alvarez-Melis and Jaakkola, 2018] David Alvarez-Melis and Tommi S Jaakkola. Gromov-Wasserstein Alignment of Word Embedding Spaces. In *ACL*, 2018.

[Artetxe *et al.*, 2017] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *ACL*, 2017.

[Artetxe *et al.*, 2018a] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-Step Framework of Linear Transformations. In *AAAI*, 2018.

[Artetxe *et al.*, 2018b] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*, 2018.

[Chen and Cardie, 2018] Xilun Chen and Claire Cardie. Unsupervised Multilingual Word Embeddings. *EMNLP*, 2018.

[Claici *et al.*, 2018] Sebastian Claici, Edward Chien, and Justin Solomon. Stochastic Wasserstein Barycenters. In *ICML*, 2018.

[Cuturi and Doucet, 2014] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. *ICML*, 2014.

[Dinu *et al.*, 2015] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. In *ICLR workshop*, 2015.

[Dou *et al.*, 2018] Zi-Yi Dou, Zhi-Hao Zhou, and Shujian Huang. Unsupervised Bilingual Lexicon Induction via Latent Variable Models. In *EMNLP*, 2018.

[Fung, 1995] Pascale Fung. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *Third Workshop on Very Large Corpora*, 1995.

[Gangbo and Świkech, 1998] Wilfrid Gangbo and Andrzej Świkech. Optimal maps for the multidimensional Monge-Kantorovich problem. *Communications on Pure and Applied Mathematics*, 51(1):23–45, 1998.

[Glavas *et al.*, 2019] Goran Glavas, Robert Litschko, Sebastian Ruder, and Ivan Vulic. How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconception. In *ACL*, 2019.

[Gouws *et al.*, 2015] Stephan Gouws, Yoshua Bengio, and Greg Corrado. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In *ICML*, 2015.

[Grave *et al.*, 2019] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised Alignment of Embeddings with Wasserstein Procrustes. In *AISTATS*, 2019.

[Hoshen and Wolf, 2018] Yedid Hoshen and Lior Wolf. Non-Adversarial Unsupervised Word Translation. In *EMNLP*, 2018.

[Joulin *et al.*, 2018] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *EMNLP*, 2018.

[Lample *et al.*, 2018] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *ICLR*, 2018.

[Mikolov *et al.*, 2013a] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting Similarities among Languages for Machine Translation. *CoRR*, abs/1309.4168, 2013.

[Mikolov *et al.*, 2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, 2013.

[Nakashole and Flauger, 2017] Ndapandula Nakashole and Raphael Flauger. Knowledge Distillation for Bilingual Dictionary Induction. In *EMNLP*, 2017.

[Rapp, 1995] Reinhard Rapp. Identifying Word Translations in Non-parallel Texts. In *ACL*, 1995.

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

[Smith *et al.*, 2017] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *ICLR*, 2017.

[Taitelbaum *et al.*, 2019a] Hagai Taitelbaum, Gal Chechik, and Jacob Goldberger. Multilingual word translation using auxiliary languages. In *EMNLP*, 2019.

[Taitelbaum *et al.*, 2019b] Hagai Taitelbaum, Gal Chechik, and Jacob Goldberger. A multi-pairwise extension of procrustes analysis for multilingual word translation. In *EMNLP*, 2019.

[Wada *et al.*, 2019] Takashi Wada, Tomoharu Iwata, and Yuji Matsumoto. Unsupervised Multilingual Word Embedding with Limited Resources using Neural Language Models. In *ACL*, 2019.

[Xing *et al.*, 2015] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL*, 2015.

[Zhang *et al.*, 2016] Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. Ten pairs to tag-multilingual pos tagging via coarse mapping between embeddings. In *ACL*, 2016.

[Zhang *et al.*, 2017a] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Earth Mover's Distance Minimization for Unsupervised Bilingual Lexicon Induction. In *EMNLP*, 2017.

[Zhang *et al.*, 2017b] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial Training for Unsupervised Bilingual Lexicon Induction. In *ACL*, 2017.