UNSUPERVISED WORD TRANSLATION PAIRING USING REFINEMENT BASED POINT SET REGISTRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-lingual alignment of word embeddings play an important role in knowledge transfer across languages, for improving machine translation and other multi-lingual applications. Current unsupervised approaches rely on similarities in geometric structure of word embedding spaces across languages, to learn structure-preserving linear transformations using adversarial networks and refinement strategies. However, such techniques, in practice, tend to suffer from instability and convergence issues, requiring tedious fine-tuning for precise parameter setting. This paper proposes *BioSpere*, a novel framework for unsupervised mapping of bi-lingual word embeddings onto a shared vector space, by combining *adversarial initialization* and *refinement procedure* with *point set registration* algorithm used in image processing. We show that our framework alleviates the shortcomings of existing methodologies, and is relatively invariant to variable adversarial learning performance, depicting robustness in terms of parameter choices and training losses. Experimental evaluation on parallel dictionary induction task demonstrates state-of-the-art results for our framework on diverse language pairs.

1 INTRODUCTION AND BACKGROUND

With the success of *distributed word representation*, like Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017), in capturing rich semantic meaning, the use of these embeddings has permeated a wide range of Natural Language Processing (NLP) tasks such as text classification, document clustering, text summarization and question answering (Klementiev et al., 2012) to name a few. Unsupervised learning of such continuous high dimensional vector representation for words rely on the *distributional hypothesis* (Harris, 1954).

Motivation. As a natural generalization, methods for obtaining multi-lingual word embeddings across diverse languages have recently gained significant attention in the NLP research community (Wang et al., 2020). Learning *cross-lingual word embeddings* (CLWE) entails mapping the vocabularies of different languages onto a single vector space for capturing syntactic and semantic similarity of words across languages boundaries (Upadhyay et al., 2016). Thus, CLWE provides an effective approach for knowledge transfer across languages for several downstream linguistics tasks such as machine translation (Artetxe et al., 2018a; Lample et al., 2018a;b), POS tagging (Zhang et al., 2016), dependency parsing (Ahmad et al., 2019), named entity recognition (Xie et al., 2018; Chen et al., 2019), entity linking (Tsai & Roth, 2016), language inference (Conneau et al., 2018b) and low-resource language understanding (Xiao & Guo, 2014). In fact, word alignment across languages also finds interesting applications in the study of cultural connotations (Kozlowski et al., 2019) and spatio-linguistic commonalities (Zwarts, 2017; Yun & Choi, 2018; Pederson et al., 1998).

Linguistic Correlation. Monolingual representation spaces learnt independently for different languages tend to exhibit similarity in terms of *geometric properties and orientations* (Mikolov & Sutskever, 2013). For example, the vector distribution of numbers and animals in English show a similar geometric constellation formation as their Spanish counterparts. Further, the frequency of words across languages have been shown to follow the *Zipf's distribution* [1], with nearly 70% most frequent word overlap (Aldarmaki et al., 2018) and 60% synonym overlap (Dinu et al., 2015) across language pairs. Existing techniques for extracting cross-lingual word correspondences rely on above inter-dependencies to efficiently learn transformations across the monolingual embedding spaces.

[1] observed on 10 million words from Wikipages across 30 languages as shown in `en.wikipedia.org/wiki/Zipf's_law`

State-of-the-art & Challenges. Early approaches for directly obtaining multi-lingual word embeddings relied on the availability of large parallel corpora (Gouws et al., 2015) or document-aligned comparable corpora (Mogadala & Rettinger, 2016; Vulić & Moens, 2016). However, such methods are not scalable as annotations are expensive and large parallel datasets, especially for low-resource languages, are scarce in practice. To address the above challenges, linear transformations between two monolingual embedding space using small manually created bi-lingual dictionaries were proposed (Mikolov & Sutskever, 2013; Artetxe et al., 2016). Words having similar surface forms across languages were used to induce seed dictionaries and other augmented refinement strategies were explored in the semi-supervised approaches of Artetxe et al. (2017); Zhou et al. (2019); Doval et al. (2018). Subsequently, improvements in orthogonality and optimization constraints were explored for generalization beyond bi-lingual settings for supervised cross-lingual alignment and joint training methods (Joulin et al., 2018; Jawanpuria et al., 2019; Alaux et al., 2019; Wang et al., 2020).

Unsupervised framework for bi-lingual word alignment was first proposed by Barone (2016); Zhang et al. (2017a;b) using *adversarial training*. The use of post-mapping refinements were shown to produce high quality results in the MUSE framework (Conneau et al., 2018a) across diverse languages, and was used for machine translation system in (Lample et al., 2018a;b). Parallel dictionary construction using *CSLS* (Conneau et al., 2018a) (adopted in this paper) or inverted softmax (Smith et al., 2017) was shown to tackle the "hubness problem" (Radovanović et al., 2010) caused due to highly dense vector space regions (called *hubs*), which adversely affects reliable retrieval of bi-lingual word translation pairs. However, the performance of adversarial learning techniques have been shown to suffer from instability, convergence issues, and dependence of precise parameter settings. Further, Søgaard et al. (2018) found the above unsupervised approaches to fail for morphologically rich languages. Hence, optimization formulations using Gromov-Wasserstein, Sinkhorn distance, and Iterative Closest Point were explored (Grave et al., 2019; Alvarez-Melis & Jaakkola, 2018; Xu et al., 2018; Hoshen & Wolf, 2018). Recently, *adversarial auto-encoders* using *cyclic loss optimization* in latent space supplemented with refinements (Mohiuddin & Joty, 2019; 2020) has achieved state-of-the-art results for bi-lingual word embedding alignment on diverse languages.

Proposed Approach. In this paper, we propose *BioSpere* (Bi-Lingual Word Translation via Point Set Registration), a novel framework for *fully unsupervised bi-lingual word correspondence induction*. Given two independently learnt monolingual word embedding space, *BioSpere* uses a combination of adversarial training, refinement procedure, and point set registration approach to efficiently extract word translations. Specifically, the input vector spaces are initially aligned using *CycleGAN*, a Generative Adversarial Network (GAN) trained using *cycle-consistency loss* optimization criteria, as word translation pairs are *symmetric*, i.e., if word w_x is a translation of w_y, then w_y is also a translation of w_x. The cyclic loss criteria has been shown to be better in capturing bi-directional distributional similarities (Xu et al., 2018) and in training adversarial networks in (Mohiuddin & Joty, 2020) (auto-encoders with a latent space of the embeddings). The word alignments obtained from CycleGAN are then refined via *symmetric re-whitening* or spherical transformation (Artetxe et al., 2018b) to remove correlations among the different components of the language embeddings. It is interesting to note that extracting *word correspondences is akin to point set registration* (Zhu et al., 2019) in image processing. To this end, *BioSpere* finally utilized the *Coherent Point Drift* (CPD) algorithm (Myronenko & Song, 2010) to compute an affine transformation between the aligned and refined vector spaces. Our choice of CPD hinges on two key insights: **(i)** CPD inherently works on the concept of *Gaussian Mixture Model* (GMM), which has been shown to tackle the *"hubness" problem* (Zhou et al., 2019); and **(ii)** CPD being an unsupervised approach might reduce error propagation from the adversarial or refinement steps, as opposed to the supervised Procrustes refinement (Schönemann, 1966) (extensively used in the literature) that requires an intermediate synthetic (possibly erroneous) dictionary creation from the adversarial training stage. Extensive empirical results on diverse languages (reported in Section 3) demonstrate that the proposed *BioSpere* framework outperforms existing approaches in terms of accuracy for parallel dictionary creation. We further show that *BioSpere* can robustly handle adversarial convergences issues, sub-optimal parameter settings, as well as morphologically rich and low-resource languages.

Contributions. In a nutshell, the key contributions of this paper can be described as follows:

- *BioSpere*, an *unsupervised* framework for learning bi-lingual word translations from two independent monolingual embedding spaces – thus aligning the vocabularies to a common vector representation for capturing semantic similarities between words across languages;

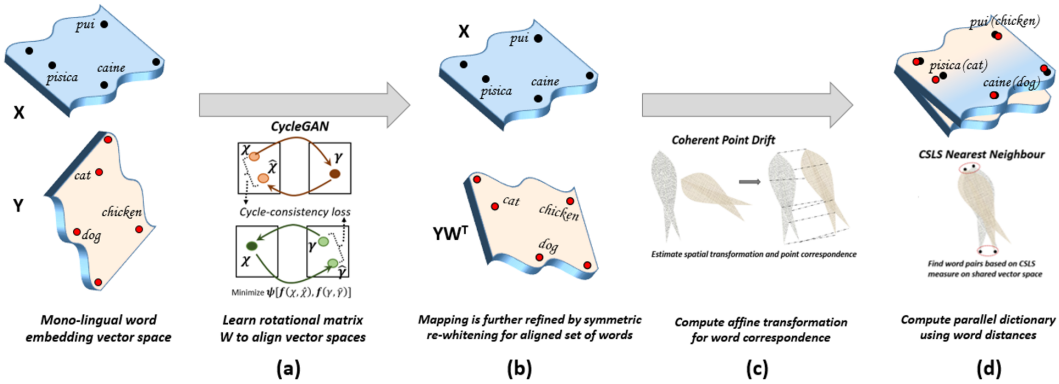

Figure 1: Toy illustration (on *en-ro* language pair) of the different modules of *BioSpere* – (a) *Align*, (b) *Correspond*, (c) *Transform*, and (d) *Generate* – for unsupervised parallel dictionary construction.

- A novel combination of adversarial training, refinement procedure, and point set registration algorithm – coupling the advantages of *cycle-consistence loss* and *Gaussian Mixture Model* – to alleviate the challenges for word embedding space alignment;
- Unsupervised stopping criterion incorporating *cycle-loss consistency* measure, with better correlation with mapping quality, for selection of adversarial training model parameters;
- Experimental evaluation on diverse language pairs showcasing *enhanced accuracy* (nearly at par with supervised approaches) compared to existing techniques, for parallel dictionary construction task, even for small vocabulary sizes; and,
- *Robustness* study of *BioSpere* framework in efficiently handling hubness problem, dependencies on adversarial learning convergence and precise parameter choice, as well as morphologically rich or low-resourced languages.

2 FRAMEWORK

We assume the existence of two sets $X = \{x_n\}_{n=1}^{N}$ and $Y = \{y_m\}_{m=1}^{M}$ of word embeddings trained independently on monolingual data from a source and a target language, respectively. The aim of our *BioSpere* framework is to map each word in the source language to its translation in the target language, in a manner that does not require any cross-lingual supervision. Equivalently, we wish to align the two embedding sets in such a way that words that are semantically similar across languages are close to each other.

To achieve this, we hinge on 4 modules, namely *Align, Correspond, Transform and Generate* (**ACTG**) [2] A pictorial depiction of the overview of the functionality of the different modules is presented in Figure 1. We now look at each module individually.

2.1 ALIGN

Our first module estimates an initial mapping using a domain-adversarial approach (Ganin et al., 2016). Let $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$ be the data distributions. We learn two linear mappings $F : X \rightarrow Y$ and $G : Y \rightarrow X$, that we refer to as forward and backward *generators*, respectively. We then train a model D_Y to discriminate between synthetic target embeddings $Y_{syn} = FX = \{Fx_n\}_{n=1}^{N}$, and real ones Y. Similarly, we train D_X to discriminate between synthetic source embeddings $X_{syn} = GY = \{Gy_m\}_{m=1^M}$ and X. Note the notation overloading: we have used F and G to refer both to the parametric linear operators, as well as to the matrices of their parameters. We continue this way for simplicity, unless the context makes the reference ambiguous.

This results in a two-player game, where the discriminators aim to distinguish real and synthetic embeddings, while the generators aim at making their image as close to their codomain as possible, prevent discriminators from making accurate predictions.

We resemble this game in our training objective, which includes two categories of terms. The *adversarial loss*, formulated for matching the distribution of the synthetic embeddings to the real

[2]Inspired by the 4 bases (Adenine, Cytosine, Thymine & Guanine) in DNA, the building block of life.

distribution. For the forward generator $F : X \rightarrow Y$, and its corresponding discriminator D_Y, our adversarial loss is:

$$\mathcal{L}_{adv}(F, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D_Y(F(x))] \quad (1)$$

We use a similar adversarial loss $\mathcal{L}_{adv}(G, D_X, Y, X)$ for the backward generator $G : Y \rightarrow X$ and its corresponding discriminator D_X.

The second objective category is in line with the work of Mohiuddin & Joty (2020). Similar to them, we note that an adversarial generator could map the same set of source embeddings to any random permutation of target embeddings, as long as the synthetic distribution matches the target distribution. To account for this possibility, we argue that the learned generators should not contradict each other, but should be cycle-consistent. That is, given a source embedding x, the forward translation cycle should attempt to produce an output that coincides with x, i.e. $G(F(x)) \approx x$. Analogously for the backward translation cycle, $G(F(y)) \approx y$. We capture this endeavour with the addition of a *cyclic loss* to our objective:

$$L_{cyc}(F, G) = \mathbb{E}_{x \sim_{data}(x)} \|G(F(x))\|_2 + \|F(G(y))\|_2 \quad (2)$$

Following Conneau et al. (2018a), we make sure F and G remain roughly orthogonal during training by alternating model parameter update with $F \leftarrow (1 + \beta)F - \beta(FF^T)F$, proceeding analogously for G. Intuitively, this preserves the monolingual quality of our embeddings by preserving their dot product and l_2 distances.

The output of this module are the two sets $X_A = F(X)$ and $Y_A = G(Y)$ of aligned embeddings (the images of the learned transformations).

2.2 CORRESPOND

Our vanilla CPD results, despite better than previous adversarial networks, are not au par with supervised work. To address this, we perform a set of refinement steps. In Correspond, the first refinement module, we perform symmetric re-weighting, successfully applied in previous work for word embedding alignment refinement(Artetxe et al., 2018a; 2016; 2017; Mohiuddin & Joty, 2020). This requires a seed parallel dictionary. We induce such a dictionary by considering mutual nearest neighbours across the the original and mapped embeddings in both directions. That is, given mappings $f : X \rightarrow Y$ and $g : Y \rightarrow X$, the similarity between x_n and y_m is $\sigma_{nm} = \delta(f(x_n), y_m) + \delta(x_n, g(y_m))$, where δ is a metric in both X and Y. Our metric of choice is cross-domain similarity local scaling (CSLS) (Conneau et al., 2018a), shown by (Conneau et al., 2018a) to effectively address the hubbness problem, stereotypical especially when working in high-dimensional spaces. Using the bidirectional nature of our adversarial network when computing the similarity has not been done in previous, work to our knowledge, and we found it to considerably improve word translation performance. During dictionary induction, we only consider the 25K most frequent words from the source and target languages.

In the first step of this module we length-normalise and mean-center X and Y, then apply a linear transformation with corresponding whitening matrices $W_x = (X^T X)^{-1/2}$ and $W_y = (Y^T Y)^{-1/2}$, i.e. $X_w = XW_x$ and $Y_w = YW_y$. This makes the embedding dimensions ucorrelated among themselves.

Next, let X^d and Y^d be two matrices that reflect our seed dictionary, with X_i^d being the embedding of a source word Y_i^d being the embedding of its translation. We perform an orthogonal transformation with symmetric re-weighting. Specifically, we compute $X_o = X_w U S^{1/2}$ and $Y_o = Y_w V S^{1/2}$ where U, S, and V come from the singular value decomposition $USV^T = (X_w^d)^T Y_w^d$. This transposes the source and target embeddings into a common vector space.

In a final step, we perform de-whitening, to restore the original covariance in the embedding dimension distributions. That is, this module outputs $X_C = X_o U^T (X^T X)^{1/2} U$ and $Y_C = Y_o V^T (Y^T Y)^{1/2} V$.

2.3 TRANSFORM

In this module we perform a further refinement of the transformed embeddings X_C and Y_C using affine Coherent Point Drift (CPD), a probabilistic framework suggested by Myronenko & Song

(2010) to perform point set registration, particularly in computer vision applications. The main idea is to consider the task of aligning the two embedding sets as a density estimation problem, where one set represents Gaussian mixture model (GMM) centroids, and the other the data points. With the two sets aligned, word translations can be obtained using the maximum of the GMM posterior probability, given a source embedding. Specifically, we consider the embeddings in Y_C as GMM centroids and the ones in X_C as data points, generated by the GMM. The GMM density has has the form:

$$p(x) = \sum_{m=1}^{M+1} p(m)p(x|m) \tag{3}$$

where $p(x|m) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(\frac{\|x-y_m\|_2}{2\sigma^2}\right)$ and $x \in X_C$, $y_m \in Y_C$. We also add a uniform distribution $p(x|M+1) = 1/N$ to account for outliers, resulting in a Uniform-Gaussian mixture model. Following the authors, we use equal isotropic covariances σ^2, and equal membership probabilities $P(m) = 1/M$ for all GMM components. We estimate the GMM centroid locations θ by minimising the negative log-likelihood function:

$$L(\theta, \sigma^2) = -\sum_{n=1}^{N} \log \sum_{m=1}^{M} P(m)p(x|m). \tag{4}$$

We use the Expectation Maximization (EM) algorithm (Dempster et al., 1977). to find the parameters θ and σ^2. We direct the interested reader to a more detailed description of CPD provided by its original authors(Myronenko & Song, 2010).

We use the affine version of CPD, which provides a tuple (R, t, s), where R is a rotation matrix, t is a translation vector, and s is a scaling constant. The transformed source embedding set is computed as $X_T = (RX_C^T * s + t)^T$. We run CPD twice for each language pair, once in each directions, providing us with X_T and Y_T.

2.4 GENERATE

We iterate between Correspond and Transform modules until an model selection criterion degrades for two consecutive iterations. The criterion is specified in Section 2.5. Equipped with the final X_T and Y_T, we compute the final estimated parallel dictionary using the same procedure as in Section 2.2. We compare this with ground truth parallel dictionaries to compute word translation accuracy.

2.5 UNSUPERVISED MODEL SELECTION

Being in an unsupervised setting, we cannot use a validation set to direct us in choosing the best performing setting of our framework. We follow approaches suggested in previous work to address this issue, that we adapt to our framework. We follow Conneau et al. (2018a) in considering the closeness of the source and target mapped embedding spaces. Specifically, we consider the 25K most frequent source words, use CSLS to generate a translation for each of them, and compute the average cosine similarity between these pairs. In our scenario, we consider similarity in both the source and target spaces, as specified in Section 2.2, criterion that we found to be better linked to word translation accuracy, compared to the unidirectional setting used in previous work (Conneau et al., 2018a; Mohiuddin & Joty, 2020).

3 EMPIRICAL EVALUATION

In this section, we evaluate the performance of the proposed *BioSpere* framework in mapping the input word embeddings onto a shared vector space, such that semantically similar words across languages are close to each other (in terms of distance) in the common space. We benchmark the accuracy of *BioSpere* against several existing approaches on the tasks of *bi-lingual dictionary induction* and *sentence translation retrieval* across a diverse set of languages.

3.1 EXPERIMENTAL SETUP

Dataset. Our experimental setup closely follows that of Conneau et al. (2018a), extensively used in the literature. As input vocabulary, we use the FastText monolingual vector embeddings (with a dimensionality of 300) (Bojanowski et al., 2017) of the top $200K$ most frequent words in each language. We consider *seven* different language pairs including morphologically rich and low-resourced languages. Specifically, we use English (en), German (de), French (fr), Spanish (es), Russian (ru), Hebrew (he), Finnish (fi), and Romanian (ro) – a diverse mix of *isolating, fusional and agglutinative language* with *dependent and mixed marking* as reported in Søgaard et al. (2018).

Evaluation. We report the *Precision@1* (P@1) accuracy scores based on the CSLS criteria (Conneau et al., 2018a) for our empirical evaluations. In the *word translation task*, we use the gold dictionary with 1,500 source test words (and full 200K target vocabulary) for different language pairs (obtained from `github.com/facebookresearch/MUSE`). We also perform the above evaluations with a smaller input vocabulary, to simulate scenarios of limited domain-specific resources.

Baselines. The performance of *BioSpere* is compared with the following *unsupervised* approaches:
(1) MUSE (Conneau et al., 2018a) – GAN (Goodfellow et al., 2014) trained for extracting a synthetic parallel dictionary to learn transformations via Procrustes refinement (Schönemann, 1966) [3];
(2) Adv-Auto (Mohiuddin & Joty, 2020) – Current state-of-the-art using adversarial auto-encoder to create synthetic dictionary, which is refined by symmetric re-whitening & Procrustes strategies [4];
(3) VecMap (Artetxe et al., 2018a) – Robust self-learning iterative algorithms exploiting structural similarities between embedding spaces for alignment [5];
(4) SinkHorn (Xu et al., 2018): GAN trained using a combination of cyclic consistency loss and Sinkhorn distance (Cuturi, 2013) as objective function;
(5) Non-Adv (Hoshen & Wolf, 2018) – Proposes an alternative approach using dimensionality reduction with Iterative Closest Point (Besl & McKay, 1992) algorithm to find word correspondences;
(6) Was-Proc (Grave et al., 2019) – A bi-stochastic matrix is computed using the assignment problem by jointly optimizing Wasserstein distance (Mémoli, 2011) and Procrustes transformation;
(7) GW-Proc (Alvarez-Melis & Jaakkola, 2018) – Word translation is formulated as an optimal flow problem across different domains using Gromov-Wasserstein distance (Mémoli, 2011); and
(8) UMH (Alaux et al., 2019) – Uses correlation between multiple languages for jointly learning embedding alignment using constraint optimization.

For completeness, we also report the accuracies achieved by state-of-the-art *supervised* approaches:
(1) RCSLS (Joulin et al., 2018): State-of-the-art supervised method for training a learning architecture based on optimizing the CSLS criteria (Conneau et al., 2018a);
(2) GeoMM (Jawanpuria et al., 2019): Language specific geometric rotations are learnt, and subsequently a network architecture is trained to align the languages; and
(3) DeMa-BME (Zhou et al., 2019): Provides a weakly-supervised approach for learning a Gaussian Mixture Model by characterizing the probability density between embeddings spaces.

Despite obtaining state of the art results, we emphasize that achieving the best possible accuracy was not our focus. Rather, we aimed to build a framework robust to adversarial instability and data noise. Most parameters were set to fixed values. As such, following Conneau et al. (2018a), we only fed the adversarial discriminator with the 50K most frequent words; the discriminator had an input dropout layer with rate 0.1. Production deployments may consider further parameter tuning. In our experiments, we only tuned the weight assigned to the cyclic loss between 5 and 10, and ran the framework under different random seeds, always picking the best model using the unsupervised criterion.

3.2 RESULTS AND DISCUSSION

Word Translation. Similar to machine translation, this task involves the retrieval of the translation of a given source word for a target language (from the target vocabulary). Observe, *polysemy* of words and *hubness* in embedding space provide a significant challenge in this setting. We evaluate the approaches using a similar setting and the ground-truth dictionaries from Conneau et al. (2018a).

[3] Code available at `github.com/facebookresearch/MUSE`

[4] Code from `ntunlpsg.github.io/project/unsup-word-translation` is updated as in Mohiuddin & Joty (2020)

[5] Code obtained from `github.com/artetxem/vecmap`

Table 1: CSLS@1 results on well-resourced languages for the dataset of Conneau et al. (2018a).

Algorithm	en-es		en-de		en-fr		en-ru	
	→	←	→	←	→	←	→	←
Supervised Approaches								
Non-Adv (Hoshen & Wolf, 2018)	81.4	82.9	73.5	72.4	81.1	82.4	51.7	63.7
DeMa-BME (Zhou et al., 2019)	82.8	85.4	77.2	75.1	83.2	83.5	49.2	63.6
GeoMM (Jawanpuria et al., 2019)	81.4	85.5	74.7	76.7	82.1	84.1	51.3	*67.6*
RCSLS (Joulin et al., 2018)	*84.1*	*86.3*	*79.1*	*76.3*	*83.3*	*84.1*	*57.9*	67.2
Unsupervised Approaches								
SinkHorn (Xu et al., 2018)**	79.5	77.8	69.3	67.0	77.9	75.5	-	-
Non-Adv (Hoshen & Wolf, 2018)	82.1	84.1	74.7	73.0	82.3	82.9	47.5	61.8
Was-Proc (Grave et al., 2019)	82.8	84.1	75.4	73.3	82.6	82.9	43.7	59.1
GW-Proc (Alvarez-Melis & Jaakkola, 2018)	81.7	80.4	71.9	72.8	81.3	78.9	45.1	43.7
MUSE (Conneau et al., 2018a)	81.7	83.3	74.0	72.2	82.3	82.1	44.0	59.1
VecMap (Artetxe et al., 2018a)††	82.3	84.7	75.1	74.3	82.3	*83.6*	*49.2*	*65.6*
UMH (Alaux et al., 2019)	82.5	84.9	74.8	73.7	**82.9**	83.3	45.3	62.8
Adv-Auto (Mohiuddin & Joty, 2020)	*83.0*	**85.2**	**76.2**	*74.7*	82.3	83.5	47.6	-
BioSpere	**83.1**	*85.0*	*75.7*	**75.2**	*82.4*	**83.8**	**49.5**	**66.1**

** Uses cosine similarity instead of CSLS and failed to reasonably converge for *en-ru* as reported in Zhou et al. (2019)
†† Results taken from Zhou et al. (2019)

Table 2: CSLS@1 results on morphologically rich and low-resource languages for Conneau et al. (2018a) data.

Algorithm	en-fi		en-he		en-ro	
	→	←	→	←	→	←
MUSE	43.7	53.7	36.9	-	57.8	66.0
VecMap	**49.9**	63.1	44.6	57.5	64.2	71.8
Adv-Auto	49.8	65.7	46.1	58.6	61.8	71.9
BioSpere	49.7	**67.3**	**46.3**	**59.1**	**65.4**	**74.3**

Table 3: CSLS@1 results for *limited vocabulary* word translation on Conneau et al. (2018a) data.

Algorithm	en-de		en-fi		en-ro	
	→	←	→	←	→	←
MUSE	71.0	77.5	-	71.7	72.7	75.5
VecMap	72.5	78.4	62.4	76.7	77.2	78.9
Adv-Auto	-	-	-	-	-	-
BioSpere	**72.8**	**79.4**	**60.0**	73.3	76.3	**80.7**

From Table 1, we observe that our *BioSpere* framework provides state-of-the-art translation results in nearly all of the language pairs. In fact, for certain language pairs like *fr→en*, the performance of *BioSpere* is almost at par with existing supervised methods (83.8 compared to 84.1 by RCSLS).

However, the challenges in word translation are compounded for *morphologically rich* and *low-resources languages* due to high vocabulary variation and limited accuracy of word embeddings respectively. To this end, we explore the performance of the competing algorithms on Finnish, Hebrew and Romanian – generally identified as "difficult" languages in the literature (Søgaard et al., 2018). From Table 2 it can be seen that *BioSpere* significantly outperforms the existing approaches with an accuracy improvement of 1.5% on average across the languages.

Limited Vocabulary. An interesting application for cross-lingual word embedding alignment is translation tasks in *domain-specific context*. For example, an organization expanding its scale of operations to geographically distributed markets and consumers. This would necessitate the efficient expansion of supporting languages for existing documents like manuals, FAQs, etc. as well as for customer services like Chatbots (Qiu et al., 2017; Massaro et al., 2018). Observe, that in such cases, the domain-specific vocabulary is relatively small, depending on the organization's range of business range and limited training resources. We simulate such application scenario in this setting, and observe the performances of the algorithms in face of with limited vocabulary.

The input mono-lingual word embeddings are limited to the 10K most frequent words (instead of 200K most frequent words) in each of the languages, which can potentially severely impact the training stages of existing techniques. However, we initially study the *word coverage* of Wiki articles with varying vocabulary sizes. Figure 2 depicts the percentage of word coverage with varying frequent word vocabulary sizes using the plain texts of Wikipedia articles from 2018 (Rosa, 2018). It can be observed, that all the languages depict similar characteristics, with a plateau around 50-75K vocabulary size. Recall, that our empirical setting is based on training the architectures on the $50K$ most frequent word embeddings. However, with around $10K$ vocabulary size, the coverage is in general not overtly bad (around 5-10% lower), but provides valuable insights as to the robustness to domain-specific or niche applications, and is hence used in this *low vocabulary* setting. From Table 3, we see that *BioSpere* performs better than existing methods on most of the language pairs.

Ablation Study. Finally, to understand the effects of the different modules in *BioSpere* on the overall performance, we perform ablation study by incrementally adding and removing the separate components. Table 4 tabulates the obtained results on different language pairs (including morphologically

Table 4: Ablation and Robustness Study: Effect of the different modules on the overall word translation performance of *BioSpere*.

Algorithm	en-de		en-fi		en-ro	
	→	←	→	←	→	←
MUSE GAN	70.1	66.4	22.3	24.1	34.5	49.6
CycleGAN	71.2	70.7	28.7	48.7	43.5	48.7
CycleGAN	71.2	70.7	28.7	48.7	43.5	48.7
CycleGAN + Sym-Wh.	75.5	74.9	47.9	66.1	63.8	72.5
BioSpere	**75.7**	**75.2**	**49.7**	**67.3**	**65.4**	**74.3**
Bad-GAN	57.7	66.7	27.0	31.1	37.9	46.8
***BioSpere* with Bad-GAN**	75.1	75.3	50.9	66.3	64.3	73.7

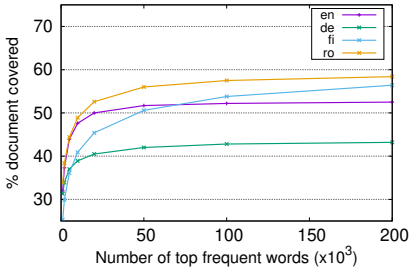

Figure 2: Distribution of document coverage with truncated vocabulary set.

rich and low-resourced). We observe, that the adversarial network, CycleGAN, using the cycle-loss consistency optimization criteria, in general performs better than MUSE GAN, the traditional GAN framework of Conneau et al. (2018a). In terms of refinement performed in the *Correspond* module of *BioSpere*, we compared the performance of symmetric re-whitening (used in this work) with the orthogonal Procrustes strategy. Both the refinement processes are seen to be comparable in performance, however since Procrustes, by definition, is a supervised approach, errors from the adversarial training in the *Align* module might be propagated, degrading the efficacy of the entire framework. Finally, addition of the Coherent Point Drift point-set registration in the *Transform* module (i.e., the complete *BioSpere* pipeline) is seen to further improve the results over the refinement strategy.

One important criticism for the performance adversarial training based alignment techniques is their dependence on precise parameter settings to tackle convergence instability (reported previously in our empirical results). Hence, we study the *robustness* of *BioSpere* to such issues, by intentionally selecting a sub-optimal CycleGAN model (from the training epochs) as the final output from the adversarial based *Align* module, denoted as *Bad-GAN* in Table 4. We observe *BioSpere* to robustly handle such situations, and provide a final accuracy score that is comparable to that achieved with a properly trained adversarial model selected based on our *cyclic unsupervised criteria*. Specifically, for *en* $\rightarrow$ *de* and *fi* $\rightarrow$ *en* languages, the performance of Bad-GAN is around 15% worse than the properly selected CycleGAN model, however, the final accuracy of *BioSpere* for word translation is seen to differ by only 1% (Table 4) – depicting robustness to noisy training.

In summary, the above empirical evaluations showcase that our framework, *BioSpere*, provides better unsupervised cross-lingual alignment of embedding spaces, by not only outperforming existing techniques in terms of translation accuracy even on morphologically rich and low-resource languages, but also demonstrating robustness in gracefully handling potential adversarial training loss.

4 Conclusion

We introduced *BioSpere*, an unsupervised cross-lingual alignment framework for word embedding. We use adversarial training with a cycle-consistency loss to induce a seed bidirectional mapping, that we subsequently refine and generate word correspondences using *point set registration method*. Extensive experiments on multiple languages for parallel dictionary creation not only demonstrate state-of-the-art results for our framework, but also depict robustness to variable adversarial performance, a considerable limitation of past work.

REFERENCES

W. U. Ahmad, Z. Zhang, X. Ma, E. Hovy, K. Chang, and N. Peng. On difficulties of Cross-lingual Transfer with Order Differences: A Case Study on Dependency Parsing. In *NAACL*, pp. 2440–2452, 2019.

J. Alaux, E. Grave, M. Cuturi, and A. Joulin. Unsupervised Hyperalignment for Multilingual Word Embeddings. In *ICLR*, pp. 1–11, 2019.

H. Aldarmaki, M. Mohan, and M. Diab. Unsupervised Word Mapping Using Structural Similarities in Monolingual Embeddings. *Transactions of the Association for Computational Linguistics*, 6: 185–196, 2018.

D. Alvarez-Melis and T. Jaakkola. Gromov-Wasserstein Alignment of Word Embedding Spaces. In *EMNLP*, pp. 1881–1890, 2018.

M. Artetxe, G. Labaka, and E. Agirre. Learning Principled Bilingual Mappings of Word Embeddings while Preserving Monolingual Invariance. In *EMNLP*, pp. 2289–2294, 2016.

M. Artetxe, G. Labaka, and E. Agirre. Learning Bilingual Word Embeddings with (almost) no Bilingual Data. In *ACL*, pp. 451–462, 2017.

M. Artetxe, G. Labaka, and E. Agirre. A Robust Self-learning Method for Fully Unsupervised Cross-lingual Mappings of Word Embeddings. In *ACL*, pp. 789–798, 2018a.

M. Artetxe, G. Labaka, and E. Agirre. Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-step Framework of Linear Transformations. In *AAAI*, pp. 5012–5019, 2018b.

A. V. M. Barone. Towards Cross-lingual Distributed Representations without Parallel Text Trained with Adversarial Autoencoders. In *Workshop on Representation Learning for NLP*, pp. 121–126, 2016.

P. J. Besl and N. D. McKay. A Method for Registration of 3-D Shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

X. Chen, A. H. Awadallah, H. Hassan, W. Wang, and C. Cardie. Multi-source Cross-lingual Model Transfer: Learning what to Share. In *ACL*, pp. 3098–3112, 2019.

A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. Word Translation Without Parallel Data. In *ICLR*, pp. 1–14, 2018a.

A. Conneau, R. Rinott, G. Lample, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov. XNLI: Evaluating Cross-lingual Sentence Representations. In *EMNLP*, pp. 2475–2485, 2018b.

M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *NIPS*, pp. 2292–2300, 2013.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

A. Dinu, L. P. Dinu, and A. S. Uban. Cross-lingual Synonymy Overlap. In *RANLP*, pp. 147–152, 2015.

Y. Doval, J. Camacho-Collados, L. Espinosa-Anke, and S. Schockaert. Improving Cross-Lingual Word Embeddings by Meeting in the Middle. In *EMNLP*, pp. 294–304, 2018.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, January 2016. ISSN 1532-4435.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *NIPS*, pp. 2672–2680, 2014.

S. Gouws, Y. Bengio, and G. Corrado. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In *ICML*, pp. 748–756, 2015.

E. Grave, A. Joulin, and Q. Berthet. Unsupervised Alignment of Embeddings with Wasserstein Procrustes. In *AISTATS*, pp. 1880–1890, 2019.

Z. S. Harris. Distributional Structure. *Word*, 10(2-3):146–162, 1954.

Y. Hoshen and L. Wolf. Non-Adversarial Unsupervised Word Translation. In *EMNLP*, pp. 469–478, 2018.

P. Jawanpuria, A. Balgovind, A. Kunchukuttan, and B. Mishra. Learning Multilingual Word Embeddings in Latent Metric Space: A Geometric Approach. *Transactions of the Association for Computational Linguistics*, 7:107–120, 2019.

A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *EMNLP*, pp. 2979–2984, 2018.

A. Klementiev, I. Titov, and B. Bhattarai. Inducing Cross-lingual Distributed Representations of Words. In *COLING*, pp. 1459–1474, 2012.

A. C. Kozlowski, M. Taddy, and J. A. Evans. The Geometry of Culture: Analyzing Meaning through Word Embeddings. *American Sociological Review*, 84(5):905–949, 2019.

G. Lample, A. Conneau, L. Denoyer, and M. Ranzato. Unsupervised Machine Translation using Monolingual Corpora only. In *ICLR*, pp. 1–14, 2018a.

G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato. Phrase-based & Neural Unsupervised Machine Translation. In *EMNLP*, pp. 5039–5049, 2018b.

A. Massaro, V. Maritati, and A. Galiano. Automated Self-learning Chatbot Initially Build as a FAQs Database Information Retrieval System. *Informatica (Slovenia)*, 42(4):515–525, 2018.

F. Mémoli. Gromov–Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, 11:417–487, 2011.

Q. V. Mikolov, T. Le and I. Sutskever. Exploiting Similarities among Languages for Machine Translation, 2013. arXiv:1309.4168.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, pp. 3111–3119, 2013.

A. Mogadala and A. Rettinger. Bilingual Word Embeddings from Parallel and Non-Parallel Corpora for Cross-language Text Classification. In *NAACL-HLT*, pp. 692–702, 2016.

T. Mohiuddin and S. Joty. Revisiting Adversarial Autoencoder for Unsupervised Word Translation with Cycle Consistency and Improved Training. In *NAACL-HLT*, pp. 3857–3867, 2019.

T. Mohiuddin and S. Joty. Unsupervised Word Translation with Adversarial Autoencoder. *Computational Linguistics*, 46(2):257–288, 2020.

A. Myronenko and X. Song. Point Set Registration: Coherent Point Drift. *Transactions on Pattern Analysis and Machine Intelligence*, 32:2262–2275, 2010.

E. Pederson, E. Danziger, D. Wilkins, S. Levinson, S. Kita, and G. Senft. Semantic Typology and Spatial Conceptualization. *Language*, 74(3):557–589, 1998.

J. Pennington, R. Socher, and C. D. Manning. GloVe: Global Vectors for Word Representation. In *EMNLP*, pp. 1532–1543, 2014.

M. Qiu, F. Li, S. Wang, X. Gao, Y. Chen, W. Zhao, H. Chen, J. Huang, and W. Chu. AliMe Chat: A Sequence to Sequence and Rerank based Chatbot Engine. In *ACL*, pp. 498–503, 2017.

M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

R. Rosa. Plaintext Wikipedia Dump 2018, 2018. URL `http://hdl.handle.net/11234/1-2735`. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Charles University.

P. H. Schönemann. A Generalized Solution of the Orthogonal Procrustes Problem. *Psychometrika*, 31(1):1–10, 1966.

S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla. Offline Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax. In *ICLR*, pp. 1–10, 2017.

A. Søgaard, S. Ruder, and I. Vulić. On the Limitations of Unsupervised Bilingual Dictionary Induction. In *ACL*, pp. 778–788, 2018.

C. Tsai and D. Roth. Cross-Lingual Wikification using Multilingual Embeddings. In *NAACL-HLT*, pp. 589–598, 2016.

S. Upadhyay, M. Faruqui, C. Dyer, and D. Roth. Cross-lingual Models of Word Embeddings: An Empirical Comparison, 2016. arXiv:1604.00425.

I. Vulić and M. Moens. Bilingual Distributed Word Representations from Document Aligned Comparable Data. *Journal of Artificial Intelligence Research*, 55(1):953–994, 2016.

Z. Wang, J. Xie, R. Xu, Y. Yang, G. Neubig, and J. Carbonell. Cross-lingual Alignment vs Joint Training: A Comparative Study and A Simple Unified Framework. In *ICLR*, pp. 1–15, 2020.

M. Xiao and Y. Guo. Distributed Word Representation Learning for Cross-lingual Dependency Parsing. In *CoNLL*, pp. 119–129, 2014.

J. Xie, Z. Yang, G. Neubig, N. A. Smith, and J. Carbonell. Neural Cross-lingual Named Entity Recognition with Minimal Resources. In *EMNLP*, pp. 369–379, 2018.

R. Xu, Y. Yang, N. Otani, and Y. Wu. Unsupervised Cross-lingual Transfer of Word Embedding Spaces. In *EMNLP*, pp. 2465–2474, 2018.

H. Yun and S. Choi. Spatial Semantics, Cognition, and Their Interaction: A Comparative Study of Spatial Categorization in English and Korean. *Cognitive Science*, 42(6):1736–1776, 2018.

M. Zhang, Y. Liu, H. Luan, and M. Sun. Adversarial Training for Unsupervised Bilingual Lexicon Induction. In *ACL*, pp. 1959–1970, 2017a.

M. Zhang, Y. Liu, H. Luan, and M. Sun. Earth Mover's Distance Minimization for Unsupervised Bilingual Lexicon Induction. In *EMNLP*, pp. 1934–1945, 2017b.

Y. Zhang, D. Gaddy, R. Barzilay, and T. Jaakkola. Ten Pairs to Tag – Multilingual POS Tagging via Coarse Mapping between Embeddings. In *NAACL-HLT*, pp. 1307–1317, 2016.

C. Zhou, X. Ma, D. Wang, and G. Neubig. Density Matching for Bilingual Word Embedding. In *NAACL-HLT*, pp. 1588–1598, 2019.

H. Zhu, B. Guo, K. Zou, Y. Li, K. V. Yuen, L. Mihaylova, and H. Leung. A Review of Point Set Registration: From Pairwise Registration to Groupwise Registration. *Sensors*, 19(1191):1–20, 2019.

J. Zwarts. Spatial Semantics: Modeling the meaning of Prepositions. *Language and Linguistics Compass*, 11(5), 2017.