Multi-Stage Framework with Refinement based Point Set Registration for Unsupervised Bi-Lingual Word Alignment

Anonymous ACL submission

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 learning structure-preserving linear transformations using adversarial networks and refinement strategies. However, such techniques, tend to suffer from instability and convergence issues, requiring tedious fine-tuning of parameter setting. This paper proposes *BioSpere*, a novel multi-stage framework for unsupervised mapping of bi-lingual word embeddings onto a shared vector space, by combining *adversarial initialization*, *refinement procedure* and *point set registration* algorithm. We show that our framework alleviates the above shortcomings, and is robust against variable adversarial learning performance and parameter choices. Experiments for parallel dictionary induction, sentence translation and word similarity demonstrate state-of-the-art results for *BioSpere* 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 such embeddings has permeated a range of Natural Language Processing (NLP) tasks such as text classification, document clustering, summarization and question answering (Klementiev et al., 2012). Unsupervised learning of such continuous high dimensional vector representation for words rely on *distributional hypothesis* (Harris, 1954).

Motivation. As a natural generalization, learning *cross-lingual word embeddings* (CLWE) entails mapping 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 (Tsai and Roth, 2016; Xie et al., 2018; Chen et al., 2019), and low-resource language understanding (Xiao and Guo, 2014; Conneau et al., 2018b). Word alignment across languages also finds applications in the study of cultural connotations (Kozlowski et al., 2019) and spatio-linguistic commonalities (Zwarts, 2017; Yun and 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 and Sutskever, 2013) [1]. The frequency of words across languages have also been shown to follow the *Zipf's distribution* [2], with an overlap of nearly 70% for the most frequent words (Aldarmaki et al., 2018) and 60% for synonyms (Dinu et al., 2015) across language pairs. Existing techniques for extracting cross-lingual word correspondences rely on above inter-dependencies to learn transformations across monolingual embedding spaces.

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 and Rettinger, 2016; Vulić and Moens, 2016). However, such methods are not scalable as annotations are expensive and large parallel datasets, especially for low-resource languages, are scarce. To address the above challenges, linear transformations between two monolingual embedding space using

[1] For example, the embedding vector distribution of numbers and animals in English show a similar geometric structural formation as their Spanish counterparts.

[2] observed on 10 million words from Wikipages on 30 languages (`en.wikipedia.org/wiki/Zipf's_law`)

small manually created bi-lingual dictionaries were proposed (Mikolov and Sutskever, 2013; Artetxe et al., 2016). These approaches tend to learn a transformation $T : X \rightarrow Y$ between the language embeddings of X and Y. This can mathematically be represented as an optimization problem solving $min_T||X - T(Y)||^2_F$, where $|| \cdot ||_F$ is the Frobenius norm. This formulation when constrained to orthonormal matrices solutions only, results in the closed-form *orthogonal Procrustes* (Schönemann, 1966) refinement strategy. Words having similar surface forms across languages were used to induce seed dictionaries and other augmented refinement strategies were explored in semi-supervised approaches (Artetxe et al., 2017; Zhou et al., 2019; Doval et al., 2018). Rigid transformation based point set registration was also studied in Cao and Zhao (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), with feedback-based learning (Yuan et al., 2020).

Unsupervised framework for bi-lingual word alignment was first proposed using *adversarial training* (Barone, 2016; Zhang et al., 2017a,b) . 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 systems (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 bi-lingual word translation. 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 and Jaakkola, 2018; Xu et al., 2018; Hoshen and Wolf, 2018). A survey of different methods can be found in Hartmann et al. (2019). *Adversarial auto-encoders* using *cyclic loss optimization* in latent space with stacked refinements (Mohiuddin and Joty, 2019, 2020) achieved improved results for bi-lingual embedding alignment on diverse languages.

Contributions. This paper proposes *BioSpere* (Bi-Lingual Word Translation via Point Set Registration and Refinement), a novel approach for *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 to align the vocabularies to a common vector representation. Our key contributions are as follows:

- *BioSpere*, an *unsupervised multi-stage* framework for learning bi-lingual word translations from independent monolingual embedding spaces, capturing cross-lingual word semantic similarities;
- A novel multi-stage framework coupling *cycle-consistence loss* and *Gaussian Mixture Model* for improved cross-lingual embedding alignment;
- Unsupervised criterion using *cycle-loss consistency* for adversarial training parameter choice;
- Experiments on diverse language pairs for *enhanced state-of-the-art accuracy* (comparable to supervised methods), for parallel dictionary creation, translation retrieval and word similarity;
- *Robustness* study of *BioSpere* framework in efficiently handling hubness problem, and adversarial learning convergence issues.

We next describe the detailed working of the different modules in the *BioSpere* framework.

2 *BioSpere* Framework

Consider, two monolingual word embedding spaces $X = \{x_n\}_{n=1}^N$ and $Y = \{y_m\}_{m=1}^M$, trained independently on monolingual data, to be provided as the source and target language representations, respectively. *BioSpere* aims to map each word in the source language to its translation in the target language, without the need for any cross-lingual supervision or pre-processing (Zhang et al., 2019). Equivalently, it aligns the language embeddings, such that semantically similar words are close to each other in the common vector space.

To achieve this, the working of *BioSpere* hinges on 4 modules, namely *Align, Correspond, Transform* and *Generate*. Fig. 1 provides an overview of the different modules, which we discuss next.

2.1 Align Module

The *Align* module uses an adversarial training approach (Ganin et al., 2016) to estimate an ini-

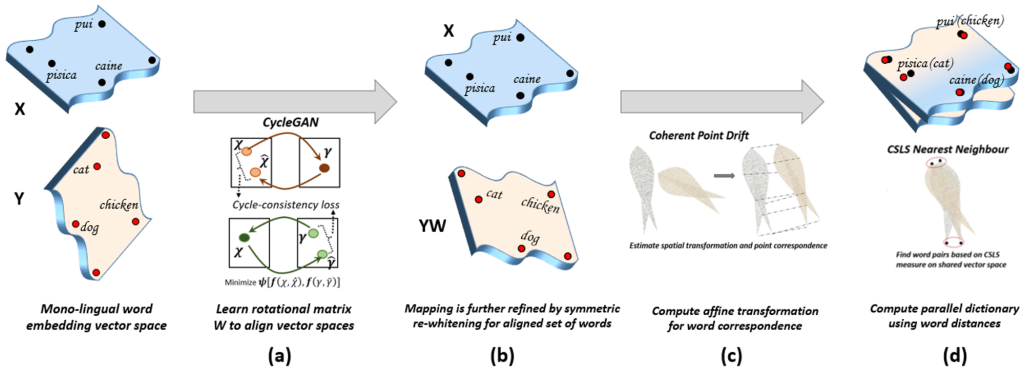
X
pui
pisica
caine
Y
cat
chicken
dog
Mono-lingual word embedding vector space
CycleGAN
χ
$\hat{\chi}$
γ
Cycle-consistency loss
χ
γ
$\hat{\gamma}$
Minimize $\psi[f(\chi,\hat{\chi}), f(\gamma,\hat{\gamma})]$
Learn rotational matrix W to align vector spaces
(a)
X
pui
pisica
caine
YW
cat
chicken
dog
Mapping is further refined by symmetric re-whitening for aligned set of words
(b)
Coherent Point Drift
Estimate spatial transformation and point correspondence
Compute affine transformation for word correspondence
(c)
pui (chicken)
pisica (cat)
caine (dog)
CSLS Nearest Neighbour
Find word pairs based on CSLS measure on shared vector space
Compute parallel dictionary using word distances
(d)

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.

tial mapping between the words across the languages, by learning an rotational transformation between the input embeddings spaces. Assuming $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$ to be the input data distributions, we learn two linear mappings $F : X \rightarrow Y$ and $G : Y \rightarrow X$, referred to as *forward* and *backward generators*, respectively. A generative adversarial network is then used to train a model D_Y (*discriminator*) to discriminate between generated synthetic target embeddings $Y_{syn} = FX = \{F(x_n)\}_{n=1}^{N}$, and the original embeddings Y. Similarly, we train another discriminator, D_X, in the opposite direction to discriminate between synthetic source embeddings $X_{syn} = GY = \{G(y_m)\}_{m=1}^{M}$ and the original X. The *discriminators* aim to distinguish between the real and synthetic embeddings, while the *generators* attempt to produce outputs that prevent the discriminators from making accurate predictions.

We resemble this in our training objective as two factors. First, the *adversarial loss* is formulated for matching the distribution of the synthetic embeddings to the real distribution. Thus, for the forward generator $F : X \rightarrow Y$, and its corresponding discriminator model D_Y, the adversarial loss is:

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

A similar loss $\mathcal{L}_{adv}(G, D_X, Y, X)$ is used for backward generator $G : Y \rightarrow X$ and discriminator D_X.

The second objective used is reported by Mohiuddin and Joty (2020) – 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, the backward translation cycle should ensure $F(G(y)) \approx y$. Since word translations are symmetric in general, this criterion is captured by a *cyclic-loss consistency* measure in:

$$\begin{aligned}L_{cyc}(F, G) = & \mathbb{E}_{x\sim_{data}(x)} \left\|G(F(x))\right\|_2 + \\ & \mathbb{E}_{y\sim_{data}(y)} \left\|F(G(y))\right\|_2 \qquad (2)\end{aligned}$$

Following Conneau et al. (2018a), we make sure F and G remain roughly orthogonal during training by alternating parameter update with $F \leftarrow (1 + \beta)F - \beta(FF^T)F$ (and analogously for G). Intuitively, this preserves the monolingual quality (dot product and L_2 distances) of embeddings.

Specifically, the above formulation corresponds to *CycleGAN* (Zhu et al., 2017), a generative adversarial network architecture, which we adopt in the *Align* module of *BioSpere*. This provides an initial aligned embedding space, obtained as two word vector sets, $X_A = F(X)$ and $Y_A = G(Y)$, as embeddings from the learned transformations.

2.2 Correspond Module

The above word alignments obtained based on cyclic loss, despite being better than other adversarial network based approaches, are not at par with state-of-the-art results and might suffer from convergence instability. To address this issue, the *Correspond* module performs a refinement step based on *symmetric re-weighting*, shown to be effective in word embedding alignment (Artetxe et al., 2018a, 2016, 2017; Mohiuddin and Joty, 2020).

To this end, a synthetic seed parallel dictionary, $\mathcal{D}$, is induced by considering the mutual nearest neighbour relation (in both directions) across the aligned embeddings (X_A and Y_A) obtained from the *Align* module. That is, given mappings $F : X \rightarrow Y$ and $G : Y \rightarrow X$, the similarity between words x_n and y_m is computed as:

$$\sigma_{nm} = \delta(F(x_n), y_m) + \delta(x_n, G(y_m)) \qquad (3)$$

where δ is a distance measure in both X_A and Y_A. As in Conneau et al. (2018a), we adopt the *cross-domain similarity local scaling* (CSLS) measure,

which addresses the "hubness" problem faced especially when working in high-dimensional spaces. Similar to the our adversarial network, σ_{nm} uses bi-directional similarity computation. In our experiments, the dictionary induction was performed on the 25K most frequent words (out of 200K words) from source and target languages. *Symmetric re-weighting* refinement is next done using 3 steps:
(i) Whitening: This makes the embedding dimensions uncorrelated with unit variance by applying *spherical transformation*. We use Mahalanobis or ZCA whitening, where original embeddings X and Y are length-normalized and mean-centered, followed by a linear transformation via whitening matrices $W_x = (X^T X)^{-1/2}$ and $W_y = (Y^T Y)^{-1/2}$, to obtain $X_w = XW_x$ and $Y_w = YW_y$.
(ii) Orthogonal Transformation: This provides an intermediate transformation of the whitened vector embeddings onto a common space. Initially, U, Σ, and V^T are obtained via singular value decomposition of $(X_w^{\mathcal{D}})^T Y_w^{\mathcal{D}}$, where $X_w^{\mathcal{D}}$ and $Y_w^{\mathcal{D}}$ are whitened embeddings of words of above seed dictionary $\mathcal{D}$. The orthogonal transformation is computed as $X_o = X_w U \Sigma^{1/2}$ and $Y_o = Y_w V \Sigma^{1/2}$.
(iii) De-Whitening: The final de-whitening step restores the original variance in the embedding dimensions in the above orthogonally transformed vector space. That is, the *Correspond* module outputs a refined vector embedding space as $X_C = X_o U^T (X^T X)^{1/2} U$ and $Y_C = Y_o V^T (Y^T Y)^{1/2} V$.

2.3 Transform Module

The *Transform* module performs a further refinement on the transformed embeddings X_C and Y_C (using the concept of *point set registration*). Specifically, we uses the *Coherent Point Drift* (CPD) algorithm (Myronenko and Song, 2010), an unsupervised probabilistic framework which assigns *point-to-point correspondence* between two sets of points, akin to finding word translation pairs in our setting. The idea here is to consider the task of aligning the two embedding spaces as a density estimation problem based on the *Gaussian Mixture Model* (GMM). This considers word embeddings of one language as GMM centroids, and the other embedding space to represent data points. The centroids are then fitted to data points by maximizing the likelihood, and at optimum point correspondences are obtained using GMM posterior probabilities.

Thus, we consider the target embeddings Y_C as the centroids and the source embedding space X_C as data points, to have been generated by the GMM probability density function. The centroid locations are estimated by Expectation Maximization (EM) algorithm (Dempster et al., 1977). We direct interested readers to the details of CPD algorithm provided by Myronenko and Song (2010).

The use of CPD provides the following advantages. The inherent use of GMM by CPD enables *BioSpere* to efficiently tackle the "hubness" problem (shown in Zhou et al. (2019)) and improve robustness. Further, CPD imposes the *Motion Coherence Theory* (MCT) (Yuille and Grzywacz, 1988) to force the GMM centroids to move coherently as a group, which preserves the underlying topological structure of the data. This would maintain the local geometric structures within the languages after alignment, benefiting downstream applications.

In *BioSpere* we use *affine transformation* for CPD, providing a higher degree of transformational freedom compared to rigid procedures of (Cao and Zhao, 2018) and Procrustes. The *Transform* module computes the 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 space is computed as $X_T = (RX_C^T * s + t)^T$. Similar to the re-weighting process, mutual nearest neighbours among the 25K most frequent words in the source and target languages (X_C and Y_C) were provided to CPD for computing correspondences. We run CPD twice for each language pair, once in each directions, generating the transformed source and target language embeddings X_T and Y_T.

2.4 Generate Module

The *Generate* module iterates between the above correspond and transform steps until convergence is reached. Equipped with the final aligned X_T and Y_T embedding spaces, the resultant parallel dictionary is computed using the bi-directional CSLS measure, similar to the construction of the intermediate dictionary in the *Correspond* module (using Eq. 3 of Sec. 2.2). For convergence of the iterative symmetric re-weighting refinement and CPD, we adopt the criteria as in Artetxe et al. (2018b); Mohiuddin and Joty (2020). The generated word pairs are compared with ground-truth parallel dictionaries to compute the accuracy of *BioSpere*.

In the next section, we show that the proposed *multi-stage framework*, *BioSpere* outperforms existing approaches in parallel dictionary creation, sentence translation retrieval, and word similarity tasks – robustly handling adversarial convergences issues and sub-optimal parameter settings.

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*, *sentence translation retrieval*, and *word similarity* across a diverse set of languages.

3.1 Experimental Setup

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

Evaluation. We report the *Precision@1* (P@1) accuracy scores based on 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 (from `github.com/facebookresearch/MUSE`). While for *sentence translation retrieval*, we consider the Europarl corpus with 2,000 source sentence queries and 200K target sentences for each of the language pairs. For the cosine based *word similarity* task, we use SemEval 2017 data (Camacho-Collados et al., 2017) and report the Pearson's correlation.

Baselines. The performance of *BioSpere* is compared against the following *unsupervised* methods:
(1) MUSE (Conneau et al., 2018a) – Uses GAN (Goodfellow et al., 2014) to learn transformations with Procrustes (Schönemann, 1966) [3];
(2) Adv-Auto (Mohiuddin and Joty, 2020) – State-of-the-art using adversarial auto-encoder to create synthetic dictionary, refined by symmetric re-weighting & Procrustes strategies [4];
(3) VecMap (Artetxe et al., 2018a) – Self-learning iterative algorithms exploiting structural similarities between embedding spaces for alignment [5];
(4) SinkHorn (Xu et al., 2018): GAN trained on cyclic loss and Sinkhorn distance (Cuturi, 2013);
(5) Non-Adv (Hoshen and Wolf, 2018) – Uses dimensionality reduction with Iterative Closest Point (Besl and McKay, 1992) algorithm;
(6) Was-Proc (Grave et al., 2019) – Computes bi-stochastic matrix for assignment by jointly optimizing Wasserstein dist. (Mémoli, 2011) & Procrustes;
(7) GW-Proc (Alvarez-Melis and Jaakkola, 2018) – Formulates optimal flow across domains using Gromov-Wasserstein distance (Mémoli, 2011); and
(8) UMH (Alaux et al., 2019) – Uses language correlation for learning via constraint optimization.

We also report the *supervised* approaches:
(1) RCSLS (Joulin et al., 2018): Optimizes CSLS criteria for learning (Conneau et al., 2018a);
(2) GeoMM (Jawanpuria et al., 2019): Language specific geometric rotations are learnt to align; and
(3) DeMa-BME (Zhou et al., 2019): Weakly-supervised approach for learning Gaussian Mixture Model between embeddings spaces.

3.1.1 Unsupervised Model Selection

Choosing the best performing model setting for adversarial training and convergence for iterative refinement (in Sec. 2.4) poses a challenge in an unsupervised setting, as we cannot use a validation set to direct our choices. To address this issue, we follow Conneau et al. (2018a) and use *CSLS* measure (denoted as DMC) to quantify the closeness of source and target mapped embedding spaces.

However, in line with our forward-backward or cyclic-consistency theme, we extend CSLS to measure the similarity in both the source and target spaces, as in Sec. 2.2. Specifically, we consider the 25K most frequent source words to generate a translation for them, and compute the average *bi-directional* cosine similarity between the pairs, to decide on model convergence. This revised criterion (termed as DualDMC) was found to be better correlated with word translation accuracy, than the unidirectional setting (DMC) used previously (Conneau et al., 2018a; Mohiuddin and Joty, 2020) – and closely correlated with CSLS@1 (on a validation set containing ground-truth word translations).

Parameter Setting. Despite obtaining state-of-the-art results, we emphasize that achieving the best possible accuracy (by extensive parameter

[3] Code from `github.com/facebookresearch/MUSE`

[4] `ntunlpsg.github.io/project/unsup-word-translation`

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

search) is not the focus of this work. Rather, we aim to build a framework robust to adversarial instability and parameter settings. Most parameters were set to fixed values, or selected via two successive degradation of the unsupervised DualDMC criteria of the previous section. Following Conneau et al. (2018a), we fed the adversarial discriminator with the 50K most frequent words, and the discriminator had an input dropout layer with a rate of 0.1. 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, picking the best model using unsupervised DualDMC.

Algorithm	**en-es**		**en-de**		**en-fr**		**en-ru**	
	→	←	→	←	→	←	→	←
				Supervised Approaches				
Non-Adv	81.4	82.9	73.5	72.4	81.1	82.4	51.7	63.7
DeMa-BME	82.8	85.4	77.2	75.1	83.2	83.5	49.2	63.6
GeoMM	81.4	85.5	74.7	76.7	82.1	84.1	51.3	*67.6*
RCSLS	*84.1*	*86.3*	*79.1*	*76.3*	*83.3*	*84.1*	*57.9*	67.2
				Unsupervised Approaches				
SinkHorn[**]	79.5	77.8	69.3	67.0	77.9	75.5	-	-
Non-Adv	82.1	84.1	74.7	73.0	82.3	82.9	47.5	61.8
Was-Proc	82.8	84.1	75.4	73.3	82.6	82.9	43.7	59.1
GW-Proc	81.7	80.4	71.9	72.8	81.3	78.9	45.1	43.7
MUSE	81.7	83.3	74.0	72.2	82.3	82.1	44.0	59.1
VecMap[††]	82.3	84.7	75.1	74.3	82.3	*83.6*	*49.2*	**65.6**
UMH	82.5	84.9	74.8	73.7	*82.9*	83.3	45.3	62.8
Adv-Auto	*83.0*	*85.2*	**76.2**	*74.7*	82.3	83.5	48.4	*64.5*
BioSpere	**83.3**	**85.4**	*75.8*	**75.8**	**83.4**	**84.1**	**49.5**	64.0

'-' denotes failure of the training network to converge reasonably

[**] Uses cosine similarity instead of CSLS, and results as reported in Zhou et al. (2019)

[††] Results taken from Zhou et al. (2019)

Table 1: CSLS@1 *word translation* results on the dataset of Conneau et al. (2018a).

Algorithm	**en-fi**		**en-he**		**en-ro**	
	→	←	→	←	→	←
MUSE	43.7	53.7	38.0	-	58.0	66.0
VecMap	**49.9**	63.5	44.6	57.7	64.2	71.8
Adv-Auto	49.8	**65.5**	46.1	58.6	62.6	71.9
BioSpere	**49.9**	**65.5**	**46.6**	**59.1**	**65.4**	**74.3**

Table 2: CSLS@1 *word translation* results on morphologically rich "difficult" languages.

3.2 Experimental Results

Word 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. Parallel dictionary construction forms a major application of word embedding alignment, and we compare all baselines using the ground-truth dictionaries of Conneau et al. (2018a).

From Table 1, we observe that *BioSpere* provides *state-of-the-art translation results* in nearly all of the four language pairs (for both directions). It should be noted that we achieve better results across the languages even when compared to supervised methods like Non-Adv and DeMa-BME. In fact, for certain language pairs like *en → es, en → fr, and fr → en*, the performance of *BioSpere* is almost at par with state-of-the-art supervised method of RCSLS. Since, unsupervised MUSE, VecMap, and Adv-Auto were seen to consistently perform well across the languages, they are selected as competing baselines for the remaining experiments.

The challenges in word translation are compounded for *morphologically rich languages* due to high vocabulary variation. To this end, we explore the performance of the competing algorithms on Finnish, Hebrew and Romanian, considered as "difficult" languages (Søgaard et al., 2018). From Table 2, we see that *BioSpere* is efficient even in such settings – outperforming existing approaches with an accuracy improvement across the languages.

Semantic Word Similarity. This task evaluates the quality of alignment of cross-lingual word embedding space by evaluating how the cosine similarity between words in different languages correlates with human-annotated word similarity scores (based on a well-defined similarity scale).

Table 3(a) shows that *BioSpere* achieves a better Pearson's correlation to human-annotated word similarity scores across the languages (except Italian). This depicts that our framework generates better alignment of different language embedding spaces – providing better understanding of semantic similarity between words across languages.

Sentence Translation Retrieval. We explore a higher level abstraction of multi-lingual word embedding space alignment, and study sentence translation retrieval on Europarl corpus. Similar to Conneau et al. (2018a), a sentence is represented as a bag-of-words, and the idf-weighted average of word embeddings are considered as sentence encoding. For each source sentence, the closest sentence (based on embedding space distance) from the target language is considered as its translation.

Table 3(b) depicts that *BioSpere* provides better sentence translation retrieval accuracy, outperforming the competing algorithms across language pairs with upto 1.5% P@1 score improvements – providing better cross-lingual alignment.

Language Models. Multi-lingual contextualized language models (CLM) like mBERT (Devlin et al., 2019) and XLM (Lample and Conneau, 2019) capture semantic meaning of words and provide "dynamic" token embeddings based on the context. Although, CLMs generate aligned multi-lingual

Algorithm	en-de		en-es		en-it	
	→	←	→	←	→	←
MUSE	0.708	0.713	0.712	0.711	0.710	0.712
VecMap	0.719	0.719	0.721	0.721	**0.746**	**0.746**
Adv-Auto	-	0.720	0.724	0.718	0.722	0.721
BioSpere	**0.726**	**0.725**	**0.730**	**0.728**	0.722	0.723

(a)

Algorithm	en-es		en-fr		en-fi	
	→	←	→	←	→	←
MUSE	75.1	73.9	69.1	69.9	64.2	64.0
VecMap	74.7	74.4	69.6	69.3	64.4	64.1
Adv-Auto	75.0	75.7	68.0	**71.0**	64.1	64.5
BioSpere	**76.7**	**76.3**	**70.2**	70.9	**65.1**	**65.9**

(b)

Table 3: Performance of competing approaches on (a) Pearson's Correlation score for *word similarity* task on SemEval 2017 dataset, and (b) Precision@1 results for *sentence translation retrieval* on Europarl data.

Algorithm	en-de	en-es	en-fi	en-ro
mBERT-CLS	70.0	80.2	40.8	65.0
***BioSpere*-WMD**	**90.2**	**93.2**	**79.1**	**94.9**

Table 4: *Sentence translation retrieval* P@1 result of *BioSpere* & multi-lingual language model on Europarl.

Algorithm	en-de		en-fi		en-ro	
	→	←	→	←	→	←
MUSE GAN	59.8	60.5	22.3	24.1	34.5	49.6
CycleGAN	69.8	69.6	27.7	48.3	44.4	52.5
CycleGAN + Procrustes	73.8	73.3	46.2	62.0	59.5	67.2
CycleGAN + SR	75.5	74.7	46.9	64.9	63.5	71.6
CycleGAN + rigid CPD	74.5	74.2	45.9	62.3	60.5	67.3
CycleGAN + affine CPD	75.2	74.7	**50.2**	**65.7**	**65.5**	72.5
BioSpere	**75.8**	**75.8**	49.9	65.5	65.4	**74.3**
Bad-GAN	70.5	62.9	25.1	36.3	42.1	51.4
Bad-GAN + Procrustes	74.5	73.3	46.7	61.7	59.5	68.9
Bad-GAN + SR	**75.9**	73.8	45.7	61.7	63.1	72.3
Bad-GAN + affine CPD	75.3	74.7	**51.7**	**65.7**	63.1	72.6
***BioSpere* with Bad-GAN**	**75.9**	**75.9**	**51.7**	65.4	**64.0**	**73.1**

Table 5: Ablation and Robustness study of *BioSpere* on *word translation* with (Conneau et al., 2018a) dataset.

contextual word embeddings (Pires et al., 2019; Wu and Dredze, 2019), parallel dictionary construction in this context becomes challenging. However, to evaluate the effect of cross-lingual word embedding alignment quality on downstream tasks, we perform sentence translation retrieval on Europarl with 2K sentence pairs. In this setting, for a source sentence, the closest target sentence is considered as translation using Word Mover's Distance (Kusner et al., 2015) on word embeddings obtained from *BioSpere*, while for mBERT we use sentence embedding similarity based on the CLS token.

Table 4 depicts that *BioSpere* achieves a large margin of improvement in translation retrieval compared to the multi-lingual language models – thus providing enhanced accuracy in capturing word semantic similarity across languages.

3.2.1 Ablation Study

To understand the impact of different modules on the performance of *BioSpere*, we perform ablation by incrementally evaluating the components.

Varying Components. Table 5 tabulates the results for different variants of our proposed framework on different language pairs. We observe, that the adversarial network, CycleGAN, using the cycle-loss consistency criteria, performs better than MUSE GAN, the framework of Conneau et al. (2018a). In terms of refinement performed in the *Correspond* module of *BioSpere*, we compared the performance of symmetric re-weighting (SR) with Procrustes. Both of them are seen to perform nearly similar, however, SR is seen to perform slightly better for morphologically rich languages, and is thus adopted in *BioSpere*. As discussed previously, we empirically observe that the higher degrees of translational freedom provided by *affine CPD* performances better than rigid CPD (used in Cao and Zhao (2018)). Note that CycleGAN + affine CPD achieves the best accuracy (with *BioSpere* performing nearly the same) for certain language pairs. We next discuss the advantages of symmetric re-weighting within our framework.

Adversarial Convergence. One important criticism for adversarial based alignment techniques is training convergence instability. Hence, we study the *robustness* of *BioSpere* to such issues, by intentionally selecting a sub-optimal CycleGAN model from the *Align* module, denoted as *Bad-GAN* in Table 5. We observe that symmetric re-weighting (SR) refinement is able to recover from such convergence issues (better than Procrustes) – providing an accuracy score comparable to that achieved by a properly trained adversarial model (selected using *DualDMC*). Specifically, for *fi* → *en* language pair, the performance of Bad-GAN is around 12% worse than the best CycleGAN model. However, the final accuracy of *BioSpere* for word translation differs by only 1% (in Table 5) even with the Bad-GAN initialization. Note, extensive parameter search for the best trained model was not performed.

Intuitively, the interactions across the different components in *BioSpere* are as follows: The adversarial module provides an initial embedding space alignment, but might be prone to convergence issues. The refinement stage then provides robustness against such training losses. However, the refinement process being a supervised approach by definition, errors in intermediate synthetic dictionary construction might propagate, degrading the efficacy. The final point correspondence CPD step, being unsupervised, is agnostic to such errors and provides enhanced cross-lingual embedding space

Algorithm	**en-es**		**en-fr**		**en-fi**		**en-ro**	
	→	←	→	←	→	←	→	←
MUSE	80.9	82.3	83.3	82.0	-	58.3	68.0	77.0
VecMap	82.2	85.7	84.7	85.4	62.4	76.7	77.2	79.9
Adv-Auto	**82.9**	85.7	84.5	85.4	-	78.3	-	79.9
BioSpere	82.8	**86.2**	**85.2**	**85.8**	**63.5**	**85.0**	**79.1**	**80.1**

Table 6: CSLS@1 results for *limited vocabulary* word translation on Conneau et al. (2018a) data.

alignment. The overall *BioSpere* framework (CycleGAN + SR + affine CPD) thus performs the best and robustly across all the different languages.

Limited Vocabulary. We now study the effect of smaller vocabulary size on the alignment accuracy of *BioSpere*. Observe, in scenarios with domain-specificity and for low-resource languages, the vocabulary space might be relatively small, which can potentially impact the training performance of existing learning techniques. Here, we limit the input monolingual word embeddings to only 10K most frequent words (instead of 200K).

From Table 6, we see that *BioSpere* outperforms the competing methods across the different language pairs. In fact, competing algorithms fail to converge (marked as '-') in certain scenarios – which can be attributed to limited training data for learning. Thus, we see that *BioSpere* provides stability and scalability in computing efficient embedding alignment across various input sizes.

In summary, the above empirical evaluations showcase that the proposed *BioSpere* framework provides better cross-lingual alignment of embedding spaces, by not only outperforming existing techniques (even supervised methods in certain cases) in translation accuracy even on morphologically rich languages, but also demonstrating robustness in handling potential training losses.

4 Related Background

Generative Adversarial Networks (GANs) couples the training of machine learning architecture between a *generative* and a *discriminative* network that work in tandem for "indirect" training in an unsupervised manner (Goodfellow et al., 2014). GANs have been shown to achieve impressive results in the domain image processing (Zhu et al., 2017), representation learning (Radford et al., 2016) and reinforcement learning (Ho and Ermon, 2016). The task of supervised image-to-image translation involves learning the transformation from an input image to an output image (Long et al., 2015). Unsupervised image-to-image translation approach, Co-GAN (Liu and Tuzel, 2016) was proposed based on weight sharing scheme. Removal of dependencies on task-specific similarity functions and low-dimensionality in this aspect was proposed by Zhu et al. (2017), and was shown in visual tracking by enforcing forward-backward consistency (Kalal et al., 2010). Improving translations via "back translation and reconciliation" is used by human translators (Brislin, 1970). We thus adopt the unsupervised CycleGAN (Zhu et al., 2017) adversarial training based on cycle-consistency loss.

Point Set Registration algorithms aim to compute the transformation for aligning two input point sets. Rigid transformation involving rotation, translation and reflection, were used in Iterative Closest Point (ICP) algorithm (Besl and McKay, 1992) and other variants (Rusinkiewicz and Levoy, 2001) for probabilistic alignment. Spectral methods (Scott and Longuet-Higgins, 1991) and closed-form solution for rigid probabilistic registration in multi-dimensional cases was presented in Myronenko and Song (2010). In addition to the rotation, translation and reflection, *affine* transformation also considers scaling, homothety, similarity and shear – providing more degrees of freedom for better point set registration (Ho et al., 2007). Non-rigid transformations are based on Gaussian Mixture model and filters (Hinton et al., 1992; Gao and Tedrake, 2019), Bayesian modelling (Hirose, 2020) or Thin Plate Spline (TPS) parameterization (Bookstein, 1989). Recent developments use convolutional neural networks (Huang et al., 2017) and other learning frameworks (Yew and Lee, 2018). An extensive literature survey can be found in Tam et al. (2013). We adopt Coherent Point Drift (CPD) (Myronenko and Song, 2010) combining Gaussian Mixture Model and Motion Coherence Theory.

5 Conclusion

This paper proposed *BioSpere*, a *multi-stage unsupervised cross-lingual word embedding alignment framework* – based on the novel coupling of *generative adversarial training*, *refinement procedure* and *point set registration*. We show that the bi-directional cycle-loss based training and convergence criteria with the inherent GMM formulation provides enhanced input vector spaces alignment. Extensive experiments on multiple languages for parallel dictionary creation, sentence translation retrieval, and word similarity not only demonstrate improved results, but also depict robustness to hubness and inconsistent adversarial performance.

References

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

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

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

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

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

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

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

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

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

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

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

F. L. Bookstein. 1989. Principal Warps: Thin-Plate Splines and the Decomposition of Deformations. *Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585.

R. W. Brislin. 1970. Back-translation for Cross-cultural Research. *Journal of Cross-Cultural Psychology*, 1(3):185–216.

J. Camacho-Collados, M. T. Pilehvar, N. Collier, and R. Navigli. 2017. Semeval-2017 Task 2: Multilingual and cross-lingual semantic word similarity. In *SemEval*.

H. Cao and T. Zhao. 2018. Point Set Registration for Unsupervised Bilingual Lexicon Induction. In *IJCAI*, pages 3991–3997.

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

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

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

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

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, pages 4171–4186.

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

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

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(1):2096–2030.

W. Gao and R. Tedrake. 2019. FilterReg: Robust and Efficient Probabilistic Point-Set Registration Using Gaussian Filter and Twist Parameterization. In *CVPR*, pages 11087–11096.

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

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

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

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

Mareike Hartmann, Yova Kementchedjhieva, and Anders Sø gaard. 2019. Comparing Unsupervised Word Translation Methods Step by Step. In *NeurIPS*, pages 6033–6043.

G. E. Hinton, C. K. I. Williams, and M. D. Revow. 1992. Adaptive Elastic Models for Hand-printed Character Recognition. In *NIPS*, pages 512–519.

O. Hirose. 2020. A Bayesian Formulation of Coherent Point Drift. *Transactions on Pattern Analysis and Machine Intelligence*.

J. Ho and S. Ermon. 2016. Generative Adversarial Imitation Learning. In *NIPS*, pages 4565–4573.

J. Ho, M. H. Yang, A. Rangarajan, and B. Vemuri. 2007. A New Affine Registration Algorithm for Matching 2D Point Sets. In *WACV*, pages 25–25.

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

H. Huang, E. Kalogerakis, S. Chaudhuri, D. Ceylan, V. G. Kim, and E. Yumer. 2017. Learning Local Shape Descriptors from Part Correspondences with Multiview Convolutional Networks. *Transactions on Graphics*, 37(1).

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

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

Z. Kalal, K. Mikolajczyk, and J. Matas. 2010. Forward-Backward Error: Automatic Detection of Tracking Failures. In *ICPR*, pages 2756–2759.

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

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

M. Kusner, Y. Sun, N. Kolkin, and K Weinberger. 2015. From Word Embeddings To Document Distances. In *ICML*, pages 957–966.

G. Lample and A. Conneau. 2019. Cross-Lingual Language Model Pretraining. In *NeurIPS*, pages 7059–7069.

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

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

M. Y. Liu and O. Tuzel. 2016. Coupled Generative Adversarial Networks. In *NIPS*, pages 469–477.

J. Long, E. Shelhamer, and T. Darrell. 2015. Fully Convolutional Networks for Semantic Segmentation. In *CVPR*, pages 3431–3440.

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

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

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

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

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

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

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

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

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

T. Pires, E. Schlinger, and D. Garrette. 2019. How Multilingual is Multilingual BERT? In *ACL*, pages 4996–5001.

A. Radford, L. Metz, and S. Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *ICLR*, pages 1–16.

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

S. Rusinkiewicz and M. Levoy. 2001. Efficient Variants of the ICP Algorithm. In *3DIM*, pages 145–152.

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

G. L. Scott and C. Longuet-Higgins. 1991. An Algorithm for Associating the Features of Two Images. *Royal Society London: Biological Sciences*, 244(1309):21–26.

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

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

G. K. L. Tam, Z. Cheng, Y. Lai, F. C. Langbein, Y. Liu, D. Marshall, R. R. Martin, X. Sun, and P. L. Rosin. 2013. Registration of 3D Point Clouds and Meshes: A Survey from Rigid to Nonrigid. *Transactions on Visualization and Computer Graphics*, 19(7):1199–1217.

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

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

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

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

S. Wu and M. Dredze. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *EMNLP-IJCNLP*, pages 833–844.

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

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

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

Z. J. Yew and G. H. Lee. 2018. 3DFeat-Net: Weakly Supervised Local 3D Features for Point Cloud Registration. In *ECCV*, pages 630–646.

M. Yuan, M. Zhang, B. Van Durme, L. Findlater, and J. Boyd-Graber. 2020. Interactive Refinement of Cross-Lingual Word Embeddings. In *EMNLP*, pages 5984–5996.

A. L. Yuille and N. M. Grzywacz. 1988. The motion coherence theory. In *ICCV*, pages 344–353.

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

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

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

Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. 2019. 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 (ACL)*, pages 3180–3189.

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

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

J. Zhu, T. Park, P. Isola, and A. A. Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *ICCV*, pages 2242–2251.

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