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

Silviu Vlad Oprea*
Amazon Alexa AI
Cambridge, UK
silviuvo@amazon.co.uk

Sourav Dutta
Huawei Research Centre
Dublin, Ireland
sourav.dutta2@huawei.com

Haytham Assem†
Amazon Alexa AI
Cambridge, UK
hithsala@amazon.co.uk

Abstract

Cross-lingual alignment of word embeddings are important in knowledge transfer across languages, for improving machine translation and other multi-lingual applications. Current unsupervised approaches relying on learning structure-preserving transformations, using adversarial networks and refinement strategies, suffer from instability and convergence issues. 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*. Experiments for parallel dictionary induction and word similarity demonstrate state-of-the-art unsupervised results for *BioSpere* on diverse languages – showcasing robustness against variable adversarial performance.

1 Introduction and Background

Distributed word representations like Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017) capture rich semantic meaning, and is used for a range of NLP tasks. *Cross-lingual word embeddings* (CLWE) entails mapping vocabularies of different languages onto a single vector space for capturing semantic similarity across languages (Upadhyay et al., 2016) – for machine translation (Artetxe et al., 2018a; Lample et al., 2018a,b), POS tagging (Zhang et al., 2016; Ahmad et al., 2019), & named entity recognition (Tsai and Roth, 2016; Xie et al., 2018; Chen et al., 2019).

Linguistic Correlation. This work is based on the observation that, monolingual representation spaces learnt independently for different languages tend to exhibit similarity in terms of *geometric properties and orientations* (Mikolov and Sutskever, 2013) [1]. Further, 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 obtaining multi-lingual word embeddings used parallel or comparable corpora (Gouws et al., 2015; Mogadala and Rettinger, 2016; Vulić and Moens, 2016). However, such methods are not scalable as parallel datasets, especially for low-resource languages, are scarce. Linear transformations between two monolingual embedding spaces (via optimization formulation (Schönemann, 1966)) using small manually created bi-lingual dictionaries were thus proposed (Mikolov and Sutskever, 2013; Artetxe et al., 2016). Words having similar surface forms across languages were used to induce seed dictionaries for semi-supervised approaches (Artetxe et al., 2017; Zhou et al., 2019; Doval et al., 2018). Rigid transformation based point set registration Cao and Zhao (2018), supervised cross-lingual alignment, joint training (Joulin et al., 2018; Jawanpuria et al., 2019; Alaux et al., 2019; Wang et al., 2020) with feedback-based learning (Yuan et al., 2020) were also studied. Unsupervised bi-lingual word alignment using *adversarial training* (Barone, 2016; Zhang et al., 2017a,b) were shown to produce good results in MUSE (Conneau et al., 2018). Inverted softmax (Smith et al., 2017) approach was shown to tackle the "hubness problem" (Radovanović et al., 2010) caused by dense vector space regions (called

[*] Work done during internship at Huawei Research, Ireland.

[†] Work done while the author was at Huawei Research, Ireland.

[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)

Proceedings of the 29th International Conference on Computational Linguistics, pages 1089–1097
October 12–17, 2022.

hubs), which adversely affected bi-lingual word translation. However, the performance of adversarial techniques were shown to suffer from convergence instability. Further, Søgaard et al. (2018) found the above approaches to fail for morphologically rich languages. Hence, optimization 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; Hartmann et al., 2019). *Adversarial auto-encoders* using *cyclic loss* and stacked refinements (Mohiuddin and Joty, 2019, 2020) recently achieved improved results.

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*. Our key contributions are as follows:

- *BioSpere*, an *unsupervised multi-stage* framework for learning bi-lingual word alignment, by using a combination of adversarial training, refinement procedure, and point set registration;
- Unsupervised criterion using *cycle-loss consistency* for adversarial model choice;
- Experiments on diverse language pairs showing *improved accuracy* on different tasks; and,
- *Robustness* to hubness and 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, to be provided as source and target language representations, respectively. *BioSpere* aims to map a word in the source language to its translation (or semantically similar word) in a target language, without cross-lingual supervision (Zhang et al., 2019). *BioSpere* consists of 4 modules – *Align, Correspond, Transform* and *Generate* (shown in Figure 1), as discussed next.

• **Align Module** – The *Align* module uses an adversarial training (Ganin et al., 2016) to estimate an initial 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 $F : X \to Y$ and $G : Y \to X$, referred to as *forward* and *backward generators*, respectively. A generative adversarial network is 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 *adversarial loss* formulates matching the distribution of synthetic embeddings to the real distribution. Thus, for forward generator $F : X \to Y$ and its discriminator model D_Y, the loss is: $\mathcal{L}_{adv}(F, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D_Y(F(x))]$ (refer Appendix B).

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

We also incorporate the objective used in Mohiuddin and Joty (2020) (considering word translations are symmetric in general) – 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., $G(F(x)) \approx x$; and vice-versa for the backward translation cycle. Thus, we have:

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

Following Conneau et al. (2018), to preserve dot product and L_2 distances from the monolingual space, we ensure 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). This corresponds to *CycleGAN* (Zhu et al., 2017), a generative adversarial network (used in our *Align* module), to provide an initial aligned embedding space, $X_A = F(X)$ and $Y_A = G(Y)$.

• **Correspond Module** – The above alignment obtained based on cyclic loss, might suffer from adversarial network convergence instability. To address this issue, the *Correspond* module performs a refinement step based on *symmetric re-weighting*, shown to be effective for alignment (Artetxe et al., 2018a, 2016, 2017; Mohiuddin and Joty, 2020).

A synthetic seed parallel dictionary, $\mathcal{D}$, is thus induced by computing the mutual nearest neighbour (in both directions) across the aligned embeddings (X_A and Y_A), as: $\sigma_{nm} = \delta(F(x_n), y_m) + \delta(x_n, G(y_m))$, where δ is a distance measure in both X_A and Y_A. As in Conneau et al. (2018), we adopt the *cross-domain similarity local scaling* (CSLS) measure, which addresses the "hubness" problem. Observe, σ_{nm} also uses bi-directional similarity computation. In our experiments, the dictionary induction

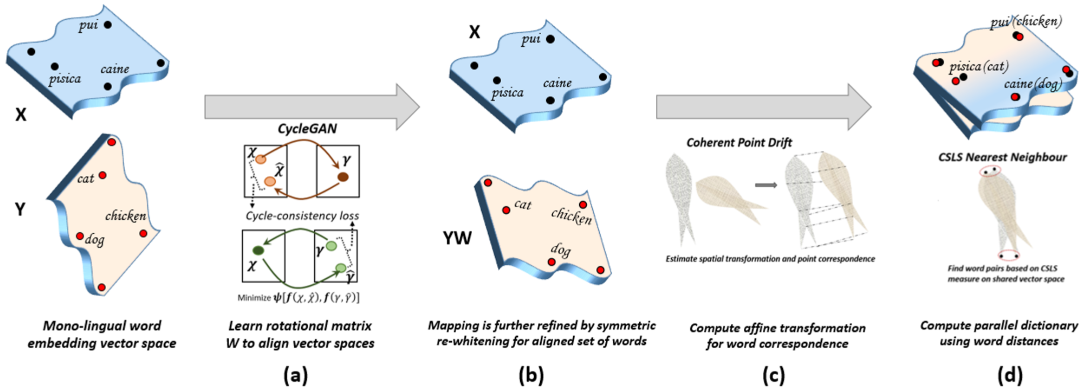
X
pui
pisica
caine
Y
cat
chicken
dog
Mono-lingual word embedding vector space
CycleGAN
χ
$\hat{\chi}$
γ
Cycle-consistency loss
χ
γ
$\hat{\gamma}$
Minimize $\psi[f(\chi,\hat{\chi}), f(\gamma,\hat{\gamma})]$
Learn rotational matrix W to align vector spaces
(a)
X
pui
pisica
caine
YW
cat
chicken
dog
Mapping 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.

is on 25K most frequent words (out of 200K words) from source and target languages. *Symmetric reweighting* is now performed via 3 steps:
(i) Whitening: makes the embedding dimensions uncorrelated with unit variance using *spherical transformation*. We use ZCA whitening, wherein the original embeddings X and Y are normalized and mean-centered, followed by a linear transformation via matrices $W_x = (X^TX)^{-1/2}$ and $W_y = (Y^TY)^{-1/2}$, to obtain $X_w = XW_x$ and $Y_w = YW_y$.
(ii) Orthogonal Transformation: provides a transformation of the whitened embeddings onto a common space. U, Σ, and V^T are obtained via singular value decomposition of $(X_w^{\mathcal{D}})^TY_w^{\mathcal{D}}$, where $X_w^{\mathcal{D}}$ and $Y_w^{\mathcal{D}}$ are whitened embeddings from the seed dictionary $\mathcal{D}$. The transformation is computed as $X_o = X_wU\Sigma^{1/2}$ and $Y_o = Y_wV\Sigma^{1/2}$.
(iii) De-Whitening: restores the original variance in the embedding dimensions of the transformed vectors – computes a refined vector embedding as: $X_C = X_oU^T(X^TX)^{1/2}U$ and $Y_C = Y_oV^T(Y^TY)^{1/2}V$.
• **Transform Module** – The *Transform* module performs a further refinement on the embeddings X_C and Y_C using the concept of *point set registration*. Specifically, we use 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. Here, the task of aligning two embedding spaces is performed using a density estimation problem based on the *Gaussian Mixture Model* (GMM). We direct interested readers to the details of CPD algorithm provided by Myronenko and Song (2010), and briefly in Appendix A.

The use of CPD provides the following advantages – (i) GMM enables *BioSpere* to tackle the "hubness" problem (shown in Zhou et al. (2019)), and (ii) CPD imposes *Motion Coherence Theory* (MCT) (Yuille and Grzywacz, 1988) to force the GMM centroids to move coherently as a group, preserving the underlying topological structure.

We use *affine CPD transformation*, providing a higher degree of freedom compared to the rigid procedure of (Cao and Zhao, 2018) and Procrustes, to compute the modified source embeddings as: $X_T = (RX_C^T * s + t)^T$, where R is a rotation matrix, t is a translation vector, and s is a scaling constant. 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.
• **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. . For convergence of the iterative symmetric re-weighting refinement and CPD, we adopt the criteria of 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*.

Overview. Intuitively, the interactions across the different components in *BioSpere* are as: 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,

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. (2018).

being unsupervised, is agnostic to such errors and provides enhanced cross-lingual embedding space alignment. The overall *BioSpere* framework (CycleGAN + SR + affine CPD) is seen to perform the best and robustly across different languages in our empirical evaluations. More details and evaluations can be found in (Oprea et al., 2020).

3 Empirical Evaluation

We evaluate *BioSpere* on mapping semantically similar words across languages, for *bi-lingual dictionary induction*, *word similarity* and *sentence translation retrieval* tasks across diverse languages.

Dataset. We follow the setup of Conneau et al. (2018), and use FastText monolingual vector embeddings (with 300 dimensions) (Bojanowski et al., 2017) for the top $200K$ most frequent words of each language as input vocabulary. We consider ***8*** different language pairs (including morphologically rich) – 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 languages* with *dependent and mixed marking* (Søgaard et al., 2018).

Evaluation. On *word translation* (dictionary induction), we use the gold dictionary with 1,500 source test words (and 200K target vocabulary) (`github.com/facebookresearch/MUSE`), while *sentence translation retrieval* uses Europarl corpus containing 2,000 source and 200K target sentences. We report *Precision@1* (P@1) based on CSLS criteria (Conneau et al., 2018). For *word similarity* on SemEval 2017 data (Camacho-Collados et al., 2017) we report the Pearson's correlation.

Baselines. The performance of *BioSpere* is compared against the following *unsupervised* methods:

(1) MUSE (Conneau et al., 2018) – 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., 2018);

(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.

Unsupervised Model Selection. For choosing the best performing model setting during adversarial training and convergence (a challenge in unsupervised setting), we follow Conneau et al. (2018) and use *CSLS* measure (denoted as DMC) to quantify the closeness of source and target mapped embedding spaces. However, adopting cyclic-consistency, we extend CSLS (termed as DualDMC) to measure the average *bi-directional cosine similarity* between source and target spaces (as in *Correspond* module), for model selection.

Parameter Setting. For a robust framework, we did not perform extensive parameter search, and most parameters were set to fixed values, or selected via two successive degradation of the unsu-

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

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

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

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**

(a)

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

(b)

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**

(c)

Table 2: Performance of competing approaches for (a) CSLS@1 on *word translation* for morphologically rich languages, (b) Pearson's Correlation score for *word similarity* task on SemEval 2017 dataset, and (c) Precision@1 results for *sentence translation retrieval* on Europarl data.

pervised DualDMC criteria. Following Conneau et al. (2018), 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.

3.1 Experimental Results

Word Translation – involves the retrieval of a source word translation to a target language for parallel dictionary construction. We use the ground-truth dictionaries of Conneau et al. (2018). From Table 1, we observe that *BioSpere* provides *better translation results* in nearly all of the four language pairs (across unsupervised methods). We achieve better results compared to even supervised methods like Non-Adv and DeMa-BME, and are comparable to RCSLS (e.g., *en → es and en → fr*). Since, MUSE, VecMap, and Adv-Auto consistently perform well, they are selected as competing baselines for the remaining experiments. We also explore the performance on "difficult" *morphologically rich languages* like Finnish, Hebrew and Romanian (Søgaard et al., 2018). Table 2(a) shows that *BioSpere* is efficient in such settings, outperforming existing approaches, across the languages.

Semantic Word Similarity – evaluates the quality of cross-lingual word alignment based on the correlation between cosine similarity between words in different languages and human-annotated word similarity scores. Table 2(b) shows that *BioSpere* achieves a better Pearson's correlation to human-annotated scores across languages (except *it*) – providing better alignment across languages.

Sentence Translation Retrieval – studies sentence translation retrieval on Europarl corpus. Similar to Conneau et al. (2018), a sentence is represented as a bag-of-words and the idf-weighted average of word embeddings is considered as its encoding. For each source sentence, the closest sentence from the target language is returned as its translation. Table 2(c) depicts that *BioSpere* provides better sentence translation retrieval accuracy with upto 1.5% P@1 score improvements.

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

Ablation Study – Table 3 tabulates the results for varying components of *BioSpere*. CycleGAN (using cycle-loss consistency) performs better than MUSE GAN, while the refinement procedures of symmetric re-weighting (SR) and Procrustes seem to perform similarly (SR being slightly better for morphologically rich languages). As discussed previously, we observe that higher degrees of translational freedom provided by *affine CPD* performs better than rigid CPD (of Cao and Zhao (2018)). To study the *robustness* of *BioSpere* to adversarial convergence issues, we intentionally select a sub-optimal CycleGAN model from the *Align* module, denoted as *Bad-GAN* in Table 3. We observe that SR refinement recovers from such convergence issues (better than Procrustes) – providing an accuracy comparable to a properly trained model (selected using *DualDMC*). Specifically, for *fi → en*, the performance of Bad-GAN is around 12% worse than the best CycleGAN model. However, the final accuracy of *BioSpere* differs by only 1% (in Table 3) even with Bad-GAN initialization.

4 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*. Experiments with diverse tasks on multiple languages demonstrate improved results over existing methods, and also depict robustness to hubness and 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. 2018. Word Translation Without Parallel Data. In *ICLR*, pages 1–14.

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.

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.

M. Hartmann, Y. Kementchedjhieva, and A. Søgaard. 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.

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.

S. Oprea, S. Dutta, and H. Assem. 2020. Unsupervised Word Translation Pairing using Refinement based Point Set Registration. arXiv:2011.13200.

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

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.

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.

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.

A Background of CPD

Point Set Registration algorithms aim to compute the correspondences 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 modeling (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.

***BioSpere* Transform Module.** The *Transform* module performs a refinement on the transformed embeddings X_C and Y_C (obtained from the *Correspond* module) 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).

B 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.