Bilingual Lexicon Induction with Semi-supervision in Non-Isometric Embedding Spaces

Barun Patra*, **Joel Ruben Antony Moniz***, **Sarthak Garg***,
Matthew R. Gormley, **Graham Neubig**
Carnegie Mellon University
{bpatra, jrmoniz, sarthakg, mgormley, gneubig}@cs.cmu.edu

Abstract

Recent work on bilingual lexicon induction (BLI) has frequently depended either on aligned bilingual lexicons or on distribution matching, often with an assumption about the isometry of the two spaces. We propose a technique to quantitatively estimate this assumption of the isometry between two embedding spaces and empirically show that this assumption weakens as the languages in question become increasingly etymologically distant. We then propose Bilingual Lexicon Induction with Semi-Supervision (BLISS) — a semi-supervised approach that relaxes the isometric assumption while leveraging both limited aligned bilingual lexicons and a larger set of unaligned word embeddings, as well as a novel hubness filtering technique. Our proposed method obtains state of the art results on 15 of 18 language pairs on the MUSE dataset, and does particularly well when the embedding spaces don't appear to be isometric. In addition, we also show that adding supervision stabilizes the learning procedure, and is effective even with minimal supervision.*

1 Introduction

Bilingual lexicon induction (BLI), the task of finding corresponding words in two languages from comparable corpora (Haghighi et al., 2008; Xing et al., 2015; Zhang et al., 2017a; Artetxe et al., 2017; Lample et al., 2018), finds use in numerous NLP tasks like POS tagging (Zhang et al., 2016), parsing (Xiao and Guo, 2014), document classification (Klementiev et al., 2012), and machine translation (Irvine and Callison-Burch, 2013; Qi et al., 2018).

Most work on BLI uses methods that learn a mapping between two word embedding spaces

(Ruder, 2017), which makes it possible to leverage pre-trained embeddings learned on large monolingual corpora. A commonly used method for BLI, which is also empirically effective, involves learning an orthogonal mapping between the two embedding spaces (Mikolov et al. (2013a), Xing et al. (2015), Artetxe et al. (2016), Smith et al. (2017)). However, learning an orthogonal mapping inherently assumes that the embedding spaces for the two languages are isometric (subsequently referred to as the orthogonality assumption). This is a particularly strong assumption that may not necessarily hold true, and consequently we can expect methods relying on this assumption to provide sub-optimal results. In this work, we examine this assumption, identify where it breaks down, and propose a method to alleviate this problem.

We first present a theoretically motivated approach based on the Gromov-Hausdroff (GH) distance to check the extent to which the orthogonality assumption holds (§2). We show that the constraint indeed does not hold, particularly for etymologically and typologically distant language pairs. Motivated by the above observation, we then propose a framework for **B**ilingual **L**exicon **I**nduction with **S**emi-**S**upervision (**BLISS**) (§3.2) Besides addressing the limitations of the orthogonality assumption, BLISS also addresses the shortcomings of purely supervised and purely unsupervised methods for BLI (§3.1). Our framework jointly optimizes for supervised embedding alignment, unsupervised distribution matching, and a weak orthogonality constraint in the form of a back-translation loss. Our results show that the different losses work in tandem to learn a better mapping than any one can on its own (§4.2). In particular, we show that two instantiations of the semi-supervised framework, corresponding to different supervised loss objectives, outperform their supervised and unsupervised counterparts on nu-

---

*Equal Contribution
*Code to replicate the experiments presented in this work can be found at https://github.com/joelmoniz/BLISS.

merous language pairs across two datasets. Our best model outperforms the state-of-the-art on 10 of 16 language pairs on the MUSE datasets.

Our analysis (§4.4) demonstrates that adding supervision to the learning objective, even in the form of a small seed dictionary, substantially improves the stability of the learning procedure. In particular, for cases where either the embedding spaces are far apart according to GH distance or the quality of the original embeddings is poor, our framework converges where the unsupervised baselines fail to. We also show that for the same amount of available supervised data, leveraging unsupervised learning allows us to obtain superior performance over baseline supervised, semi-supervised and unsupervised methods.

2 Isometry of Embedding Spaces

Both supervised and unsupervised BLI often rely on the assumption that the word embedding spaces are isometric to each other. Thus, they learn an orthogonal mapping matrix to map one space to another Xing et al. (2015).

This orthogonality assumption might not always hold, particularly for the cases when the language pairs in consideration are etymologically distant — Zhang et al. (2017b) and Søgaard et al. (2018) provide evidence of this by observing a higher Earth Mover's distance and eigenvector similarity metric respectively between etymologically distant languages. In this work, we propose a novel way of a-priori analyzing the validity of the orthogonality assumption using the Gromov Hausdorff (GH) distance to check how well two language embedding spaces can be aligned under an isometric transformation[†].

The Hausdorff distance between two metric spaces is a measure of the worst case or the diametric distance between the spaces. Intuitively, it measures the distance between the nearest neighbours that are the farthest apart. Concretely, given two metric spaces $\mathcal{X}$, and $\mathcal{Y}$ with a distance function $d(.,.)$, the Hausdorff distance is defined as:

$$\mathcal{H}(\mathcal{X}, \mathcal{Y}) = \max\{ \sup_{x \in \mathcal{X}} \inf_{y \in \mathcal{Y}} d(x, y), \\ \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} d(x, y) \}. \quad (1)$$

The Gromov-Hausdorff distance minimizes the Hausdorff distance over all isometric transforms

---

[†]Note that since we mean center the embeddings, the orthogonal transforms are equivalent to isometric transforms

between $\mathcal{X}$ and $\mathcal{Y}$, thereby providing a quantitative estimate of the isometry of two spaces

$$\mathcal{H}(\mathcal{X}, \mathcal{Y}) = \inf_{f,g} \mathcal{H}(f(\mathcal{X}), g(\mathcal{Y})), \quad (2)$$

where f, g belong to set of isometric transforms.

Computing the Gromov-Hausdorff distance involves solving hard combinatorial problems, which is intractable in general. Following Chazal et al. (2009), we approximate it by computing the Bottleneck distance between the two metric spaces (the details of which can be found in Appendix (§A.1)). As can be observed from Table 2, the GH distances are higher for etymologically distant language pairs.

3 Semi-supervised Framework

In this section, we motivate and define our semi-supervised framework for BLI. First we describe issues with purely supervised and unsupervised methods, and then lay the framework for tackling them along with orthogonality constraints.

3.1 Drawbacks of Purely Supervised and Unsupervised Methods

Most purely supervised methods for BLI just use words in an aligned bilingual dictionary and do not utilize the rich information present in the topology of the embeddings' space. Purely unsupervised methods, on the other hand, can suffer from poor performance if the distribution of the embedding spaces of the two languages are very different from each other. Moreover, unsupervised methods can successfully align clusters of words, but miss out on fine grained alignment within the clusters.

We explicitly show the aforementioned problem of purely unsupervised methods with the help of the toy dataset shown in 1a, and 1b. In this dataset, due to the density difference between the two large blue clusters, unsupervised matching is consistently able to align them properly, but has trouble aligning the smaller embedded green and red sub-clusters. The correct transformation of the source space is a clockwise 90° rotation followed by reflection along the x-axis. Unsupervised matching converges to this correct transformation only half of the time; in rest of the cases, it ignores the alignment of the sub-clusters and converges to a 90° counter-clockwise transformation as shown in 1c.

We also find evidence of this problem in the real datasets used in our experiments as shown in Ta-

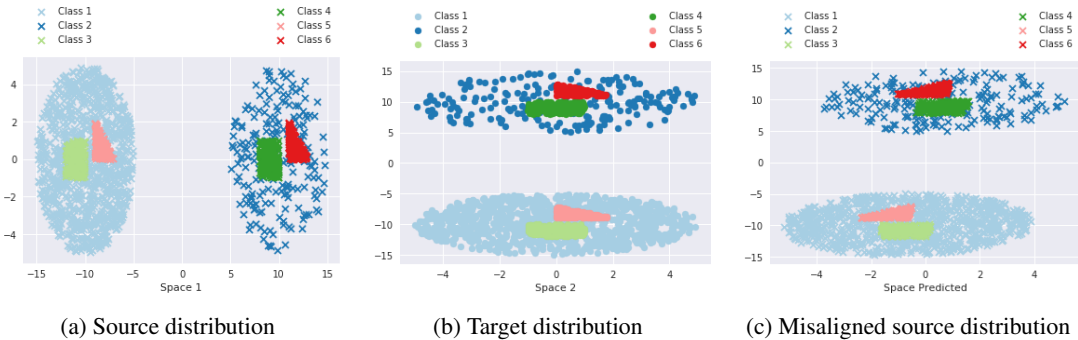| (a) Source distribution | (b) Target distribution | (c) Misaligned source distribution |

Figure 1: A toy dataset demonstrating the shortcomings of unsupervised distribution matching. Fig. a) and b) show two different distributions (source and target respectively) over six classes. Classes 1 and 2; classes 3 and 4; classes 5 and 6 were respectively drawn from a uniform distribution over a sphere, rectangle and triangle respectively. Fig. c) shows the misprojected source distribution obtained from unsupervised distribution matching which fails to align with the target distribution of Fig. b).

Source → Target	Incorrect Predicted
aunt → тётя	бабушка (Grandmother)
uruguay → уругвая	аргентины (Argentina)
regiments → полков	кавалерийские (Cavalry)
comedian → комик	актёр (Actor)

Table 1: Words for which semi-supervised method predicts correctly, but unsupervised method doesn't. The unsupervised method is able to guess the general family but fails to pinpoint exact match.

ble 1. It can be seen that the unsupervised method aligns clusters of similar words, but is poor at fine grained alignment. We hypothesize that this problem can be resolved by giving it some supervision in the form of matching anchor points inside these sub-clusters, which correctly aligns them. Analogously, for the task of BLI, generating a small supervised seed lexicon for providing the requisite supervision, is generally feasible for most language pairs, through bilingual speakers, existing dictionary resources, or Wikipedia language links.

3.2 A Semi-supervised Framework

In order to alleviate the problems with the orthogonality constraints, the purely unsupervised and supervised approaches, we propose a semi-supervised framework, described below.

Let $\mathcal{X} = \{x_1 \ldots x_n\}$ and $\mathcal{Y} = \{y_1 \ldots y_m\}$, $x_i, y_i \in \mathbb{R}^d$ be two sets of word embeddings from the source and target language respectively and let $\mathcal{S} = \{(x_1^s, y_1^s) \ldots (x_k^s, y_k^s)\}$ denote the bilingual aligned word embeddings.

For learning a linear mapping matrix W that maps $\mathcal{X}$ to $\mathcal{Y}$ we leverage unsupervised distribution matching, aligning known word pairs and a data-driven weak orthogonality constraint.

Unsupervised Distribution Matching: Given all word embeddings $\mathcal{X}$ and $\mathcal{Y}$, the unsupervised loss $\mathcal{L}_{W|D}$ aims to match the distribution of both embedding spaces. In particular, for our formulation, we use an adversarial distribution matching objective, similar to the work of Lample et al. (2018). Specifically, a mapping matrix W from the source to the target is learned to fool a discriminator D, which is trained to distinguish between the *mapped* source embeddings $W\mathcal{X} = \{Wx_1 \ldots Wx_n\}$ and $\mathcal{Y}$. We parameterize our discriminator with an MLP, and alternatively optimize the mapping matrix and the discriminator with the corresponding objectives:

$$\mathcal{L}_{D|W} = -\frac{1}{n} \sum_{x_i \in \mathcal{X}} \log(1 - D(Wx_i))$$
$$- \frac{1}{m} \sum_{x_i \in \mathcal{Y}} \log D(x_i) \quad (3)$$
$$\mathcal{L}_{W|D} = -\frac{1}{n} \sum_{x_i \in \mathcal{X}} \log D(Wx_i)$$

Aligning Known Word Pairs: Given aligned bilingual word embeddings $\mathcal{S}$, we aim to minimize a similarity function (f_s) which maximizes the similarity between the corresponding matched pairs of words. Specifically, the loss is defined as:

$$\mathcal{L}_{W|S} = -\frac{1}{|\mathcal{S}|} \sum_{(x_i^s, y_i^s) \in \mathcal{S}} f_s(Wx_i^s, y_i^s) \quad (4)$$

Weak Orthogonality Constraint: Given an embedding space $\mathcal{X}$, we define a consistency loss that maximizes a similarity function f_a between x and $W^T W x$, $x \in \mathcal{X}$. This cyclic consistency loss $\mathcal{L}_{\mathcal{W}|\mathcal{O}}$ encourages orthogonality of the W matrix based on the joint optimization:

$$\mathcal{L}_{W|O} = -\frac{1}{|\mathcal{X}|} \sum_{x_i \in \mathcal{X}} f_a(x_i, W^T W x_i) \quad (5)$$

The above loss term, used in conjunction with the supervised and unsupervised losses, allows the model to adjust the trade-off between orthogonality and accuracy based on the joint optimization. This is particularly helpful in the embedding spaces where the orthogonality constraint is violated (§4.4). Moreover, this data driven orthogonality constraint is more robust than an enforced hard constraint (A.3).

The final loss function for the mapping matrix is:

$$\mathcal{L} = \mathcal{L}_{W|D} + \mathcal{L}_{W|S} + \mathcal{L}_{W|O} \quad (6)$$

$\mathcal{L}_{W|D}$ enables the model to leverage the distributional information available from the two embedding spaces, thereby using all available monolingual data. On the other hand, $\mathcal{L}_{W|S}$ allows for the correct alignment of labeled pairs when available in the form of a small seed dictionary. Finally, $\mathcal{L}_{W|O}$ encourages orthogonality. One can think of $\mathcal{L}_{W|O}$ and $\mathcal{L}_{W|S}$ as working against each other when the spaces are not isometric. Jointly optimizing both helps the model to strike a balance between them in a data driven manner, encouraging orthogonality but still allowing for flexible mapping.

3.3 Nearest Neighbor Retrieval

For NN lookup, we use the CSLS distance defined by Lample et al. (2018). Let $\Gamma_{\mathcal{A}}(b)$ be the average cosine similarity between b and it's k-NN in $\mathcal{A}$. Then CSLS is defined as $\mathrm{CSLS}(x, y) = 2\cos(Wx, y) - \Gamma_{\mathcal{Y}}(Wx) - \Gamma_{W\mathcal{X}}(y)$.[*]

3.4 Iterative Procrustes Refinement and Hubness Mitigation

A common method of improving BLI is iteratively expanding the dictionary and refining the mapping matrix as a post-processing step (Artetxe et al., 2017; Lample et al., 2018). Given a learnt mapping matrix, Procrustes refinement first finds

---

[*] $W\mathcal{X}$ denotes the set $\{Wx : x \in \mathcal{X}\}$

the pair of points in the two languages that are very closely matched by the mapping matrix and constructs a bilingual dictionary from these pairs. These pair of points are found by considering the nearest neighbors (NN) of the projected source words in the target space. The mapping matrix is then refined by setting it to be the Procrustes solution of the dictionary obtained. Iterative Procrustes Refinement (also referred as Iterative Dictionary Expansion) applies the above step iteratively.

However, learning an orthogonal linear map in such a way leads to some words (known as hubs) to become nearest neighbors of a majority of other words (Radovanović et al., 2010; Dinu and Baroni, 2014). In order to estimate the hubness of a point, (Radovanović et al., 2010) first compute $N_x(k)$, the counts of all points y such that $x \in k-NN(y)$, normalized over all k. The skewness of the distribution over $N_x(k)$ is defined as the hubness of the point, with positive skew representing hubs and negative skew representing isolated points. An approximation to this would be $N_x(1)$, i.e the number of points for which x is the nearest neighbor.

We use a simple hubness filtering mechanism to filter out words in the target domain that are hubs, i.e., words in the target domain which have more than a threshold number of neighbors in the source domain are not considered in the iterative dictionary expansion. Empirically, this leads to a small boost in performance. In our models, we use iterative Procrustes refinement with hubness filtering at each refinement step.

4 Experiments and Results

In this section, we measure the GH distances between embedding spaces of various language pairs, and compute their correlation with several empirical measures of orthogonality. Next, we analyze the performance of the instantiations of our semi-supervised framework for two settings of supervised losses, and show that they outperform their supervised and unsupervised counterparts for a majority of the language pairs. Finally we analyze our performance with varying amounts of supervision and highlight the framework's training stability over unsupervised methods.

4.1 Empirical Evaluation of GH Distance

To evaluate the lower bound on the GH distance between the two embedding spaces, we select the

	ru-uk	en-fr	en-es	es-fr	en-uk	en-ru	en-sv	en-el	en-hi	en-ko	**|Corr|** (GH)	**|Corr|** (Λ)
GH	0.18	0.17	0.2	0.24	0.34	0.44	0.46	0.47	0.5	0.92	*	*
Λ	16.4	4.1	5.9	4.1	11.7	14.7	7.3	11.5	7.7	6.6	*	*
MUSE(U)	*	82.3	81.7	85.5	29.1	44.0	53.3	37.9	34.6	5.1	**0.87**	0.61
RCSLS	*	83.3	84.1	87.1	38.3	**57.9**	61.7	47.6	37.3	37.5	**0.74**	0.52
GeoMM	*	82.1	81.4	**87.8**	39.1	51.3	65.0	47.8	**39.8**	34.6	**0.76**	0.49
BLISS(R)	*	**83.9**	**84.3**	87.1	**40.7**	57.1	**65.1**	**48.5**	38.1	**39.9**	**0.73**	0.50
$\|\|I - W^T W\|\|^2$	0.03	0.01	0.03	0.02	59.8	54.3	71.6	72.6	106.3	98.46	**0.84**	0.75

Table 2: Correlation of GH and Eigenvector similarity with performance of BLI methods. Bold marks best metrics.

5000 most frequent words of the source and target language and compute the Bottleneck distance. These embeddings are mean centered, unit normed and the Euclidean distance is used as the distance metric.

Row 1 of Table 2 summarizes the GH distances obtained for different language pairs. We find that etymologically close languages such as en-fr and ru-uk have a very low GH distance and can possibly be aligned well using orthogonal transforms. In contrast, we find that etymologically distant language pairs such as en-ru and en-hi cannot be aligned well using orthogonal transforms.

To further corroborate this, similar to Søgaard et al. (2018) , we compute correlations of the GH distance with the accuracies of several methods for BLI. We find that the GH distance exhibits a strong negative correlation with these accuracies, implying that as the GH distance increases, it becomes increasingly difficult to align these language pairs. Søgaard et al. (2018) proposed the eigenvector similarity metric between embedding spaces for measuring similarity between the embedding spaces. We compute their metric over top n (100, 500, 1000, 5000 and 10000) embeddings (Column Λ in Table 2 shows correlation for the best setting of n) and show that the GH distance (Column GH) correlates better with the accuracies than eigenvector similarity. Furthermore, we also compute correlations against an empirical measure of the orthogonality of two embedding spaces by computing $\|\|I - W^T W\|\|^2$, where W is a mapping from one language to the other obtained from an unsupervised method (MUSE(U)). Note that an advantage of this metric is that it can be computed even when the supervised dictionaries are not available (ru-uk in Table 2). We obtain a strong correlation with this metric as well.

4.2 Benchmark Tasks: Setup

Baseline Methods

MUSE (U/S/IR): Lample et al. (2018) proposed two models: MUSE(U) and MUSE(S) for unsupervised and supervised BLI respectively. MUSE(U) uses a GAN based distribution matching followed by iterative Procrustes refinement. MUSE(S) learns an orthogonal map between the embedding spaces by minimizing the Euclidean distance between the supervised translation pairs. Note that for unit normed embedding spaces, this is equivalent to maximizing the cosine similarity between these pairs. MUSE(IR) is the semi-supervised extension of MUSE(S), which uses iterative refinement using the CSLS distance starting from the mapping learnt by MUSE(S). We also use our proposed hubness filtering technique during the iterative refinement process (MUSE(HR)) which leads to small performance improvements. We consequently use the hubness filtering technique in all our models.

RCSLS: Joulin et al. (2018) propose optimizing the CSLS distance[‡] directly for the supervised matching pairs. This leads to significant improvements over MUSE(S) and achieves state of the art results for a majority of the language pairs at the time of writing.

VecMap models: Artetxe et al. (2017) and Artetxe et al. (2018a) proposed two models, VecMap and VecMap^{++} which were based on Iterative Procrustes refinement starting from a small seed lexicon based on numeral matching.

We also compare against two well known methods GeoMM (Jawanpuria et al., 2018) and Vecmap $(U)^{++}$ (Artetxe et al., 2018b). These methods learn orthogonal mappings for both source and target spaces to a common embedding space, and

---

[‡]Since the CSLS distance requires computing the nearest neighbors over the whole embedding space, this can also be considered a semi-supervised method.

subsequently translate in the common space.

BLISS models

We instantiate two instances of our framework corresponding to the two supervised losses in the baseline methods mentioned above. BLISS(M) optimizes the cosine distance between supervised matching pairs as its supervised loss ($\mathcal{L}_{W|S}$), while BLISS(R) optimizes the CSLS distance between these matching pairs for its $\mathcal{L}_{W|S}$. We use the unsupervised CSLS metric as a stopping criterion during training. This metric, introduced by Lample et al. (2018), computes the average cosine similarity between matched source-target pairs using the CSLS distance for NN retrieval; and the authors showed that this correlates well with ground truth accuracy.

After learning the final mapping matrix, the translations of the words in the source language are mapped to the target space and their nearest neighbors according to the CSLS distance are chosen as the translations.

Datasets

We evaluate our models against baselines on two popularly used datasets: the MUSE dataset and the VecMap dataset. The MUSE dataset used by Lample et al. (2018) consists of embeddings trained by Bojanowski et al. (2017) on Wikipedia and bilingual dictionaries generated by internal translation tools used at Facebook. The VecMap dataset introduced by Dinu and Baroni (2014) consists of the CBOW embeddings trained on the WacKy crawling corpora. The bilingual dictionaries were obtained from the Europarl word alignments. We use the standard training and test splits available for both the datasets.

4.3 Benchmark Tasks: Results

In Tables 3 and 4, we group the instantiations of BLISS(M/R) with it's supervised counterparts. We use † to compare models within a group, and use **bold** do compare across different groups for a language pair.

As can be seen from Table 3, BLISS(M/R) outperform baseline methods within their groups for 9 of 10 language pairs. Moreover BLISS(R) gives the best accuracy across all baseline methods for 6 out of 10 language pairs.

We observe a similar trend for the VecMap datasets, where BLISS(M/R) outperforms its supervised and unsupervised counterparts (Table 4).

It can be seen that BLISS(M) and BLISS(R) outperform the MUSE baselines (MUSE(U), MUSE(R)) and RCSLS respectively.

We observe that GeoMM and VecMap(U)$^{++}$ outperform BLISS models on the VecMap datasets. A potential reason for this could be the slight disadvantage that BLISS suffers from because of translating in the target space, as opposed to in the common embedding space. This hypothesis is also supported by the results of Kementchedjhieva et al. (2018).

All the hyperparameters for the experiments can be found in the Appendix (§A.4)

4.4 Benefits of BLISS

Languages with high GH distance: As can be seen from Table 2, BLISS(R) substantially outperforms RCSLS on 6 of 9 language pairs, especially when the GH distance between the pairs is high (en-uk (2.4%), en-sv (3.4%), en-el (0.9%), en-hi(0.8%), en-ko (2.4%)). Results from Table 3 also underscores this point, wherein BLISS(R) performs at least at par with (and often better than) RCSLS on European languages, and performs significantly better on en-zh (2.8%) and zh-en (0.9%).

Performance with varying amount of supervision: Table 5 shows the performance of BLISS(R) as a function of the number of data points provided for supervision. As can be observed, the model performs reasonably well even for low amounts of supervision and outperforms the unsupervised baseline MUSE(U) and it's supervised counterpart RCSLS. Moreover, note that the difference is more prominent for the etymologically distant pair en↔zh. In this case the baseline models completely fail to train for 50 points, whereas BLISS(R) performs reasonably well.

Stability of Training: We also observe that providing even a little bit of supervision helps stabilize the training process, when compared to purely unsupervised distribution matching. We measure the stability during training using both the ground truth accuracy and the unsupervised CSLS metric. As can be seen from Figure 2, BLISS(M) is significantly more stable than MUSE(U), converging to better accuracy and CSLS values. Furthermore, for en↔zh, Vecmap(U)$^{++}$ fails to converge, while MUSE is somewhat unstable. However, BLISS does not suffer from this issue.

When the word vectors are not rich enough

Model	Type	Objective	Translation Space	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en
MUSE(U)	Unsup	GAN	target	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1	32.5	31.4
MUSE(S)	Sup	Cos	target	81.4	82.9	81.1	82.4	73.5	72.4	51.7	63.7	42.7†	36.7
MUSE(IR)	Semi	Cos + IR	target	81.9	83.5	82.1	82.4	74.3	72.7	51.7	63.7	42.7†	36.7
MUSE(HR)	Semi	Cos + IR	target	82.3†	83.3	82.5	83.2	75.7†	72.8	52.8	64.1†	42.7†	36.7
BLISS(M)	Semi	Cos + GAN	target	82.3†	84.3†	83.3†	83.9†	75.7†	73.8†	55.7†	63.7	41.1	41.4†
RCSLS	Semi	CSLS	target	84.1	**86.3†**	83.3	84.1	**79.1†**	76.3	**57.9†**	67.2	45.9	46.4
BLISS(R)	Semi	CSLS + GAN	target	**84.3†**	86.2	**83.9†**	**84.7†**	**79.1†**	76.6†	57.1	**67.7†**	48.7†	**47.3†**
GeoMM	Sup	Classification Loss	common	81.4	85.5	82.1	84.1	74.7	**76.7**	51.3	67.6	**49.1**	45.3
Vecmap(U)$^{++}$	Unsup	NN Based Dist matching + IR	common	82.2	84.5	82.5	83.6	75.2	74.2	48.5	65.1	0.0	0.0

Table 3: Performance comparison of BLISS on the MUSE dataset. Sup, Unsup and Semi refer to supervised, unsupervised and semi-supervised methods. Objective refers to the metric optimized. $\dagger$ marks the best in each category, while **bold** marks the best performance across all groups for a language pair.

Pairs	# seeds	Vec Map	Vec Map^{++}	MUSE (U)	MUSE (IR)	BLISS (M)	RCSLS	BLISS (R)	GeoMM	Vec Map(U)$^{++}$
en-it	all	39.7	45.3	45.8	45.3	45.9†	45.4	46.2†	48.3	**48.5**
	Num.	37.3	-	45.8†	0.7	44.3	0.3	44.6†	1.2	**48.5**
en-de	all	40.9	44.1	0.0	47.0	48.3†	47.3	48.1†	**48.9**	48.1
	Num.	39.6	-	0.0	39.9	47.2†	1.0	46.5†	2.3	**48.1**

Table 4: Performance of different models on the VecMap dataset. $\dagger$ marks the best in each category, while **bold** marks the best performance across different levels of supervision for a language pair.

# Datapoints	Model	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en
*	MUSE(U)	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1	32.5†	31.4†
*	Vecmap(U)$^{++}$	82.2†	84.5†	82.5†	83.6†	75.2†	**74.2†**	48.5†	65.1†	0.0	0.0
50	MUSE(IR)	0.3	82.7	0.5	1.6	31.9	72.7†	0.1	0.0	0.3	0.3
	GeoMM	0.3	1.9	0.3	1.0	0.3	0.3	0.0	0.6	0.0	0.0
	RCSLS	0.1	0.4	0.0	0.3	0.1	0.1	0.1	0.1	0.0	0.0
	BLISS (R)	82.1†	83.6†	**82.8†**	83.0†	75.1†	72.7†	39.3†	61.0†	32.6†	32.5†
500	MUSE(IR)	81.6	83.5†	82.1	82.0	73.1	72.7	40.3	62	34.5	32.2
	GeoMM	31.9	46.6	34.4	44.7	13.5	14.7	10.6	20.5	3.9	2.9
	RCSLS	22.9	44.9	22.4	43.5	9.9	10.2	7.9	19.6	6.6	7.1
	BLISS(R)	82.3†	83.4	82.3†	82.9†	74.7†	73.1†	41.6†	63.0†	36.3†	35.1†
5000	MUSE(IR)	81.9	82.8	82.2	82.1	75.2	72.4	50.4	63.7	39.2	36.3
	GeoMM	79.7	82.7	79.9	83.2	71.7	70.6	49.7	**65.5†**	**43.7†**	40.1
	RCSLS	80.9	82.9	80.4	82.5	72.5	70.9	51.3	63.8	42.5	41.9
	BLISS(R)	**82.4†**	**84.9†**	82.6†	**83.9†**	**75.7†**	72.5†	**52.1†**	65.2	42.5	**42.8†**

Table 5: Performance with different levels of supervision. $\dagger$ marks the best performance at a given level of supervision, while **bold** marks the best for a language pair.

(word2vec (Mikolov et al., 2013b) instead of fast-Text), the unsupervised method can completely fail to train. This can be observed for the case of en-de in Table 4. BLISS(M/R) does not face this problem: adding supervision, even in the form of 50 mapped words for the case of en-de, helps it to achieve reasonable performance.

5   Related Work

Mikolov et al. (2013a) first used anchor points to align two embedding spaces, leveraging the fact that these spaces exhibit similar structure across languages. Since then, several approaches have been proposed for learning bilingual dictionaries (Faruqui and Dyer, 2014; Zou et al., 2013; Xing
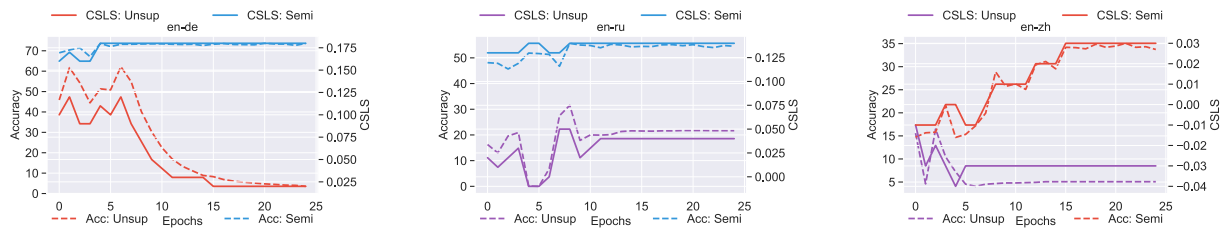
Figure 2: Training Stability of different language pairs (en-de), (en-ru), (en-zh)

et al., 2015). Xing et al. (2015) showed that adding an orthogonal constraint significantly improves performance, and admits a closed form solution. This was further corroborated by the work of Smith et al. (2017), who showed that in orthogonality was necessary for self-consistency. Artetxe et al. (2016) showed the equivalence between the different methods, and their subsequent work (Artetxe et al., 2018a) analyzed different techniques proposed in various works (like embedding centering, whitening etc.), and showed that leveraging a combination of different methods showed significant performance gains.

However, the validity of this orthogonality assumption has of late come into question: Zhang et al. (2017b) found that the Wasserstein distance between distant language pairs was considerably higher , while Søgaard et al. (2018) explored the orthogonality assumption using eigenvector similarity. We find our weak orthogonality constraint (along the lines of Zhang et al. (2017a)) when used in our semi-supervised framework to be more robust to this.

There has also recently been an increasing focus on generating these bilingual mappings without an aligned bilingual dictionary, i.e., in an unsupervised manner. Zhang et al. (2017a) and Lample et al. (2018) both use adversarial training for aligning two monolingual embedding spaces without any seed lexicon, while Zhang et al. (2017b) used a Wasserstein GAN to achieve this adversarial alignment, and use an earth-mover based fine-tuning approach; while Grave et al. (2018) formulate this as a joint estimation of an orthogonal matrix and a permutation matrix. However, we show that adding a little supervision, which is usually easy to obtain, improves performance.

Another vein of research (Jawanpuria et al., 2018; Artetxe et al., 2018b; Kementchedjhieva et al., 2018) has been to learn orthogonal map-

pings from both the source and the target embedding spaces into a common embedding space and doing the translations in the common embedding space. Artetxe et al. (2017) and Søgaard et al. (2018) motivate the utility of using both the supervised seed dictionaries and, to some extent, the structure of the monolingual embedding spaces. They use iterative Procrustes refinement starting with a small seed dictionary to learn a mapping; but doing may lead to sub-optimal performance for distant language pairs. However, these methods are close to our methods in spirit, and consequently form the baselines for our experiments.

Another avenue of research has been to try and modify the underlying embedding generation algorithms. Cao et al. (2016) modify the CBOW algorithm (Mikolov et al., 2013b) by augmenting the CBOW loss to match the first and second order moments from the source and target latent spaces, thereby ensuring the source and target embedding spaces follow the same distribution. Luong et al. (2015), in their work, use the aligned words to jointly learn the embedding spaces of both the source and target language, by trying to predict the context of a word in the other language, given an alignment. An issue with the proposed method is that it requires the retraining of embeddings, and cannot leverage a rich collection of precomputed vectors (like ones provided by Word2Vec (Mikolov et al., 2013b), Glove (Pennington et al., 2014) and FastText (Bojanowski et al., 2017)).

6 Conclusions

In this work, we analyze the validity of the orthogonality assumption and show that it breaks for distant language pairs. We motivate the task of semi-supervised BLI by showing the shortcomings of purely supervised and unsupervised approaches. We finally propose a semi-supervised framework which combines the advantages of supervised and

unsupervised approaches and uses a joint optimization loss to enforce a weak and flexible orthogonality constraint. We provide two instantiations of our framework, and show that both outperform their supervised and unsupervised counterparts. On analyzing the model errors, we find that a large fraction of them arise due to polysemy and antonymy (An interested reader can find the details in Appendix (§A.2).

We also find that translating in a common embedding space, as opposed to the target embedding space, obtains orthogonal gains for BLI, and plan on investigating this in the semi-supervised setting in future work.

Acknowledgements

References

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 789–798.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Hailong Cao, Tiejun Zhao, Shu Zhang, and Yao Meng. 2016. A distribution-based model to learn bilingual word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1818–1827.

Frédéric Chazal, David Cohen-Steiner, Leonidas J Guibas, Facundo Mémoli, and Steve Y Oudot. 2009. Gromov-hausdorff stable signatures for shapes using persistence. In *Computer Graphics Forum*, volume 28, pages 1393–1403. Wiley Online Library.

Georgiana Dinu and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. volume abs/1412.6568.

Herbert Edelsbrunner and Dmitriy Morozov. 2013. Persistent homology: theory and practice.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2018. Unsupervised alignment of embeddings with wasserstein procrustes. *arXiv preprint arXiv:1805.11222*.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio. Association for Computational Linguistics.

Ann Irvine and Chris Callison-Burch. 2013. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 262–270, Sofia, Bulgaria. Association for Computational Linguistics.

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2018. Learning multilingual word embeddings in latent metric space: a geometric approach. *arXiv preprint arXiv:1808.08773*.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing procrustes analysis for better bilingual dictionary induction. In *Proceedings of the 22nd Conference on*

Computational Natural Language Learning, pages 211–220.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. pages 1459–1474.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. volume 11, pages 2487–2531.

Sebastian Ruder. 2017. A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 778–788.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129.

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

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1959–1970.

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

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag-multilingual pos tagging via coarse mapping between embeddings. Association for Computational Linguistics.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.