An Alignment-Based Approach to Semi-Supervised Bilingual Lexicon Induction with Small Parallel Corpora

Kelly Marchisio kmarc@jhu.edu
Conghao Xiong cxiong5@jhu.edu
Philipp Koehn phi@jhu.edu
Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 21218, USA

Abstract

Aimed at generating a seed lexicon for use in downstream natural language tasks, unsupervised methods for bilingual lexicon induction have received much attention in the academic literature recently. While interesting, fully unsupervised settings are unrealistic; small amounts of bilingual data are usually available due to the existence of massively multilingual parallel corpora, or linguists can create small amounts of parallel data. In this work, we demonstrate an effective bootstrapping approach for semi-supervised bilingual lexicon induction that capitalizes upon the complementary strengths of two disparate methods for inducing bilingual lexicons. Whereas statistical methods are highly effective at inducing correct translation pairs for words frequently occurring in a parallel corpus, monolingual embedding spaces have the advantage of having been trained on large amounts of data, and therefore may induce accurate translations for words absent from the small corpus. By combining these relative strengths, our method achieves state-of-the-art results on 3 of 4 language pairs in the challenging VecMap test set using minimal amounts of parallel data and without the need for a translation dictionary. We release our implementation at `https://github.com/kellymarchisio/align-semisup-bli`.

1 Introduction

Unsupervised methods for machine translation (MT) and bilingual lexicon induction (BLI) have received considerable attention in recent years, showing impressive performance without bilingual data for supervision. While academically interesting, small amounts of supervised data can almost always help model performance.

The typical use case for unsupervised BLI is to provide initial synthetic training data for a traditional supervised setup where no parallel bitext exists, such as for MT or cross-lingual information retrieval. A starting lexicon is induced in an unsupervised manner, and then serves as initial training data to the supervised model. Practically, however, one struggles to identify a scenario where one would truly fail to have any parallel text whatesoever from which to gain some supervision. The Christian Bible, for instance, is translated into over 1600 world languages, providing multi-way parallel data for many of the world's languages that are typically considered "low-resource" (McCarthy et al., 2020). Human translators can also create a small translation corpus or seed dictionary. The practical necessity of fully unsupervised scenarios for BLI or MT therefore becomes hard to imagine.

Statistical translation/alignment models are very proficient at inducing bilingual lexicons from small amounts of parallel data. Particularly when words occur frequently in the corpus, statistical models easily recover the translation. At the same time, however, the number of seed translation pairs possible to extract is limited by the vocabulary of the parallel corpus.

We address a more realistic scenario: there is ample monolingual data and a small parallel corpus. We combine the strengths of statistical alignment and unsupervised mapping methods and achieve state-of-the-art results on 3 of 4 languages in the challenging VecMap dataset (Dinu et al., 2015; Artetxe et al., 2017, 2018a), trailing by only 0.1 in the 4th language pair.

2 Related Work

Automatic BLI has been a popular task in natural language processing for decades, beginning with statistical decipherment (e.g., Rapp, 1995; Fung, 1995; Koehn and Knight, 2000, 2002; Haghighi et al., 2008). With the advent of the ability to create large monolingual vector spaces from abundant monolingual text, the focus has shifted to finding an optimal linear transformation between such monolingual embedding spaces from which a seed lexicon can be extracted using nearest neighbors search. Practically, this often involves solving variations of the generalized Procrustes problem (e.g., Conneau et al., 2018; Artetxe et al., 2016, 2017; Patra et al., 2019; Artetxe et al., 2018b; Doval et al., 2018; Joulin et al., 2018; Jawanpuria et al., 2019; Alvarez-Melis and Jaakkola, 2018). Differing metrics and heuristics can be used to extract the seed lexicon once the mapping is found. Cross-domain similarity local scaling (CSLS) to mitigate the hubness is popular and effective (Conneau et al., 2018).

While the orthogonal variant of the Procrustes problem has a simple closed-form solution, one must know in advance the pairings of words one wants to be closest after the transformation (i.e., you already know the translations). To adapt to the unsupervised or semi-supervised scenario, such mapping-based BLI procedures must make a “guess” of some correct translation pairs. The solution can then iteratively refined through self-learning. The initial “guess” can come in the form of direct supervision using a bilingual training dictionary, or in an unsupervised manner, such as by identifying the nearest neighbors in a similarity matrix (e.g., Artetxe et al., 2018b) or via adversarial training (e.g., Conneau et al., 2018; Patra et al., 2019).

Like us, Shi et al. (2021) also use statistical alignment within a pipeline for BLI, but unlike our work, they do not use the induced alignments as seeds for monolingual embedding mapping.

3 Background

3.1 The Orthogonal Procrustes Problem

Let A and B be matrices in $\mathbb{R}^{m\times n}$. Let Q be a matrix in $\mathbb{R}^{n\times n}$. The goal of the orthogonal Procrustes problem is to find Q such that:

$$\arg\min_{QQ^T=I} \|AQ - B\|_F$$

The solution to the orthogonal Procrustes problem is $Q = VU^T$, where $U\Sigma V$ is the singular value decomposition of $B^T A$ (Schönemann, 1966).

3.2 IBM Model 2

IBM Model 2 (Brown et al., 1993) is designed to be a noisy channel model for MT, but it is a particularly useful statistical model for word alignment. We view the most likely alignment between a source sentence ***f*** and target sentence ***e*** as a hidden variable, modeled as the conditional probability

$$\arg\max_{a_1\ldots a_m} p\left(a_1\ldots a_m \mid f_1\ldots f_m, e_1\ldots e_l, m\right)$$

where m is the length of source sentence, l is the length of target sentence, $\{f_1...f_m\}$ and $\{e_1...e_l\}$ are the source words and target words respectively, and a_i is the alignment, indicating that f_i is aligned to e_{a_i}. To compute the alignment, we need two more definitions:

- $p(f|e)$: the lexical translation probabilities. e is a target word, and f is the source word. In addition to the whole vocabulary of target language, the target-side also includes a *NULL* token indicating that a source word aligned to none of the target words.

- $p(j \mid i, l, m)$: the alignment model. The probability of source position j being aligned to target position i.

The IBM models are trained via expectation-maximization. After training, alignments can be determined with:

$$a_i = \arg \max_{j \in \{0...l\}} (p(j \mid i, l, m) \times p(f_i \mid e_j))$$

4 Motivation

Different types of models have different strengths when it comes to determining translations of words. We discuss some contrasting strengths of inducing translations from statistical models versus monolingual embedding space mapping in this section as motivation for our method. We assert that to maximize accuracy, one should induce the translation of common words from statistical models and less frequent words from well-trained monolingual embedding spaces.

Statistical models succeed for common words, struggle for rare words.

In the IBM statistical translation models, word translation probabilities are typically initialized uniformly. In the IBM models, the probability $p(f|e)$ assigned to a given word pair in the translation table is iteratively refined according to the occurrence of f and e in the corpus. While this procedure can capture alignment and translation likelihoods of common words in a large bilingual corpus accurately, the probability can become inaccurate for rare words (not to mention those absent from the corpus). The risk of such inaccuracies of low-frequency words increases as corpus size shrinks.

There are 10,673 unique source tokens in the first 10,000 lowercased lines of the English-side of the Europarl v7 German-English corpus (Koehn, 2005), used later in this work. Of those, 4015 tokens occur just once. Only 5214 — less than half of the vocabulary — occur more than twice. Such a large percentage of rare words is explained by the well-known Zipf's law (Zipf, 1935, 1949; Mandelbrot, 1953, 1961), whereby the *kth* most common word tends to occur with a frequency approaching the below, where $\alpha \sim 1$ and $\beta \sim 2.7$ (Piantadosi, 2014).

$$freq(w) \propto \frac{1}{(rank(w) + \beta)^{\alpha}} \tag{1}$$

Embedding space mapping can take advantage of large amounts of monolingual data.

Just as statistical methods for word translation are more accurate for common words, inducing translations from monolingual word embeddings spaces for common words is also likely more accurate than for rare words, owing to the fact that the word embeddings for more common words are better trained than for rare words. The advantage that monolingual word embedding spaces have over traditional statistical MT methods, however, is that there is typically orders of magnitude more available monolingual text than there is translated parallel bitext for a given language pair. As such, a word that is rare in a bitext may occur frequently enough in a large monolingual corpus for its word embedding to be well-trained and useful.

More correct translation pairs → better embedding space mapping.

Empirically, more high-quality seed translation pairs improves the Procrustes mapping of monolingual embedding spaces for BLI. Our method is motivated by the desire to extract a large and accurate seed dictionary to solve Procrustes given only small amounts of parallel bitext from which to extract seeds.

Use the relative strengths of statistical vs. mapping methods to maximize performance.

Using 5000 seeds is common in the supervised BLI literature. In light of the fact that our 10,000-line Europarl bitext only has 5214 tokens that occur more than twice, we are hard-pressed to extract 5000 seed translations that we are confident are correct. We therefore use the relative strengths of IBM Model 2 and mapping-based methods for extracting a seed lexicon from monolingual embedding spaces to extract as many high-quality translation pairs as possible. Because of IBM Model 2's strength in identifying correct translations for high-frequency words, we trust its judgement for high-frequency words in the bitext. Monolingual embedding spaces, however, have the advantage of having a much larger vocabulary (the literature typically uses 200,000) and having been trained on much larger amounts of data. Thus we trust monolingual embedding mapping methods to identify the correct translations for any medium-frequency words, or high-frequency words that happened to not have been present in the parallel bitext given to IBM Model 2. We avoid the very lowest frequency words, but extract bilingual translation pairs for words seen frequently in the parallel corpus from IBM Model 2, and those seen less frequently (or not at all) from the embedding space mapping.

5 Method

5.1 Supervised statistical seed induction from bitext

We first run IBM Model 2 over a small parallel corpus. We rank the resulting word translation table by probability ("confidence"), and retain the top N translation pairs assigned the highest confidence. We discard pairs where either the source or target word occurred less than M times in the bitext, to avoid the problem of the statistical alignment model assigning erroneously high probabilities to rare words. We also discard pairs lower than a chosen confidence threshold.

5.2 Seed set expansion via embedding space mapping

Using the induced translations from the previous step as seeds, we map the monolingual embedding spaces using the public implementation of VecMap[1] in supervised mode (Artetxe et al., 2018a). In this method, word embeddings are length-normalized, mean-centered, and length-normalized again. A whitening transformation is performed, and then VecMap solves the orthogonal Procrustes problem over the known seeds, and the resulting spaces are reweighted and dewhitened. We extract a phrase table from the resulting mapped monolingual embedding spaces using Monoses[2](Artetxe et al., 2019). For a mapped source word e, let its k nearest neighbors in the mapped target embedding space be $N(x, k)$. Here, k=100. We calculate the translation probability for x and each of its k nearest neighbors using the softmax of the cosine similarity. Let $f \in N(x, k)$. Then,

$$p(f|e) = \frac{exp(\cos(e, f)/\tau)}{\sum\limits_{f' \in N(x,k)} exp(\cos(e, f')/\tau)}$$

See Artetxe et al. (2019) for further details.

[1] `https://github.com/artetxem/vecmap`

[2] `https://github.com/artetxem/monoses`

We extract the phrase table and rank the translations in descending order by forward translation probability. We again require the potential translation pairs to meet a minimum confidence threshold to be considered for use. We take the highest ranked translation per source word, therefore each source word is only used once.

5.3 Frequency-based seed selection with low-frequency agreement

We select our final seed set based on corpus frequency according to the motivation in Section 4. We retain the top K pairs from the embedding mapping method that are *disjoint* from the N word translations generated by IBM Model 2. In other words, if the source and target word in a potential translation occurred more than a pre-selected minimum number of times in the parallel bitext (M), we trust IBM Model 2 over VecMap. At the same time, we recognize the potential fault that the statistical alignment model could inaccurately guess a translation for a word it only sees once. To compensate for this weakness and allow for the creation of a larger seed dictionary on which to train our second round of VecMap, we turn to VecMap itself to induce the seeds of words rarely or never seen in the training corpus. In doing so, we can induce seed dictionaries larger than the vocabulary of the parallel bitext, but also with higher accuracy than if induced via VecMap alone in a self-learning fashion. Thus for words occurring infrequently (or never) in the parallel bitext, we trust VecMap over IBM Model 2. We merge the two potential seed dictionaries, only retaining low-frequency pairs induced by IBM Model 2 if VecMap can also confirm its desire for the potential pair to be retained.

5.4 Embedding space re-mapping with expanded seed set

Finally, the concatenated list of high-confidence translation pairs are used as seeds to again solve the Procrustes problem and re-map the monolingual embedding spaces. With the expanded joint seed set owing to the complementary strengths of IBM Model 2 and the previous embedding space mapping, this second round of embedding space mapping is expected to be more successful than would have been possible using only seeds from IBM Model 2, or only from self-learning.

6 Experimental Settings

Language	Corpus	# of words
English	WaCky, BNC, Wikipedia	2.8 B
Italian	itWac	1.6 B
German	SdeWaC	0.9 B
Spanish	News Crawl 2007-2012	386 M
Finnish	Common Crawl 2016	2.8 B

Table 1: Corpora used to train the word embeddings for each language in the VecMap dataset, with the number of words in billions (B) or millions (M).

6.1 Pretrained Word Embeddings

The pretrained embeddings from Dinu et al. (2015); Artetxe et al. (2017, 2018a) are 300-dimensional vectors of 200,000 words, trained with CBOW (Mikolov et al., 2013a). Table 1 details the parallel text used to train the embeddings. We conduct experiments on all four available language pairs (English-German, English-Spanish, English-Italian, English-Finnish).

6.2 Data

We use the popular and challenging VecMap data set, which is the original English-Italian data set of Dinu et al. (2015) with the subsequent extensions by Artetxe et al. (2017, 2018a). The dataset was obtained via alignment of the Europarl corpus (Koehn, 2005; Tiedemann, 2012). Test sets contain approximately 1500 source words and 2000 word pairs total. The source words are sampled evenly from frequency bins in the Europarl lexicon: one-fifth from each of frequency ranks [1000-5000], [5000-20,000], [20,000-50,000], [50,000-100,000], and [100,000-200,000]. This makes the test set considerably more challenging than the widely-used MUSE training and test sets (Conneau et al., 2018), where the test set consists of exactly source word frequencies 5,000-6,500 for each language pair. We create a development set for English-German and English-Finnish using the last 2,000 lines of the training seeds provided by Dinu et al. (2015); Artetxe et al. (2017, 2018a), which are disjoint from the test set.

We use Europarl v7 as our parallel bitext, which is a corpus of European Parliamentary proceedings available in 11 languages (Koehn, 2005). We normalize punctuation, tokenize, and clean the corpus to remove sentences with more than 100 tokens or with a source-to-target length ratio above 9. Each of these steps uses scripts from the Moses statistical MT system (Koehn et al., 2007). We then lowercase all bitext. For subsequent experiments varying the data size of the input corpus, we use the first N lines of the bitext, where N ranges from 500 to 50,000. We stop at 50,000 because our focus is on very small corpora. We use the NLTK[3] (Bird et al., 2009) implementation of IBM Model 2, and the public implementation of VecMap.

6.3 Hyperparameter Settings

For the IBM Model 2 step detailed in 5.1, we use N=3000, M=2, and minimum confidence threshold is set to 0.1. Final translations for the test set are retrieved by choosing the nearest neighbor in the target-side mapped space of the source word according to CSLS scaling, to mitigate the hubness problem. These settings are based on early experimentation with en-de using between 10k-100k lines of Europarl, where we observe that the subsequent VecMap stage needed about 3000 seeds extracted from 5,000 lines of Europarl to begin exceeding the unsupervised baseline performance. N=3000 and M=2 were chosen to encourage having 3000+ seeds from IBM2 for data conditions as low as 1k parallel lines. We then apply the chosen hyperparameters to all language pairs.

Seeds	en-de 1K	en-de 10K	en-fi 1K	en-fi 10K
0	38.1	64.1	14.0	44.8
200	48.7	65.0	18.9	46.8
500	55.8	65.5	26.0	47.0
1,000	58.8	65.7	29.8	48.3
3,000	61.2	**66.7**	33.5	48.8
5,000	60.5	**66.7**	33.7	**49.2**
10,000	**61.7**	66.6	**35.6**	48.2
15,000	61.2	65.9	**35.6**	**49.2**
20,000	61.1	65.7	**35.6**	48.3

Table 2: P@1 on en-de and en-fi development sets with increasing number of seeds induced from VecMap. Experiments are performed with models using 1K and 10K lines of parallel bitext input to IBM Model 2.

[3] https://www.nltk.org/

To determine the number of seeds that should be induced from VecMap, we performed experiments using the English-German and English-Finnish development sets. We train systems with N=3000 IBM seeds given 1,000 or 10,000 input sentences to IBM, and vary the amount of VecMap seeds that we extract from the resulting system to be concatenated with the IBM seeds to train the second round of VecMap. The results are presented in Table 2. Note that the vocabulary size is limited for 1,000 input sentences, the number of possible translation pairs is limited by vocabulary size and model confidence. This results in 1058 IBM-induced seeds for en-de and 791 for en-fi, for models using only 1,000 lines of parallel data. We examine all results, and select a number of seeds that appears to work well across all 4 conditions. We decide that this best seed set size is 10,000.

7 Results and Analysis

	en-it	en-de	en-fi	en-es
Unsupervised				
Conneau et al. (2018)* (avg.)	45.2	46.8	0.4	35.4
Artetxe et al. (2018b) (avg.)	48.1	48.2	32.6	37.3
Grave et al. (2019)	45.2	-	-	-
Mohiuddin and Joty (2020)	47.7	48.7	32.6	38.1
Alvarez-Melis and Jaakkola (2018)	49.2	46.5	18.3	37.6
Supervised / Semi-Supervised				
Smith et al. (2017)*	43.1	43.3	29.4	35.1
Patra et al. (2019) BLISS(M)	45.9	48.3	-	-
Patra et al. (2019) BLISS(R)	46.2	48.1	-	-
Mikolov et al. (2013b)*	34.9	35.0	25.9	27.7
Faruqui and Dyer (2014)*	38.4	37.1	27.6	26.8
Artetxe et al. (2016)*	39.3	41.9	30.6	31.4
Artetxe et al. (2017)	39.7	40.9	28.7	-
Artetxe et al. (2018a)	45.3	44.1	32.9	36.6
Jawanpuria et al. (2019) GeoMM	48.3	49.3	36.1	39.3
Mohiuddin et al. (2020)	46.7	47.7	34.1	37.8
Jawanpuria et al. (2019) GeoMMsemi	**50.0**	51.3	36.2	39.7
Ours, N=5,000	49.5	51.2	35.3	**40.0**
Ours, N=10,000	49.9	**51.7**	36.0	**40.1**
Ours, N=20,000	49.7	**51.4**	**36.8**	**40.1**
Ours, N=50,000	49.3	**51.4**	**37.1**	**39.9**

Table 3: Main results. P@1 BLI performance on the VecMap data set, compared with existing literature. *As reported in Artetxe et al. (2018b). "avg" are averaged over 10 runs. For our method, N is the number of sentences in the bitext given to IBM Model 2. Bold is best performance per language pair. We bold all of our models which outperform all previously published results.

Our main results compared with the existing literature are presented in Table 3. We achieve state-of-the-art results in the English-German, English-Finnish, and English-Spanish pairs. For English-Italian, we trail the state-of-the-art semi-supervised system of Jawanpuria et al. (2019) by only 0.1. However, Jawanpuria et al. (2019) use 80% of available training seeds from the VecMap test set (4000 seeds) while ours uses only 3000 seeds induced from a parallel bitext using IBM Model 2. For en-de and en-fi, our models trained on only 10,000 and 20,000 lines of bitext achieve state-of-the-art results, respectively. For en-es, even our model using only 5,000 parallel lines of bitext exceeds the performance of previous literature, achieving state-of-the-art performance.

7.1 Impact of Size of Input Corpus

	500	1000	5000	10000	20000
en-it	40.0	46.7	49.5	**49.9**	49.7
en-de	33.3	46.1	51.2	**51.7**	51.4
en-fi	8.4	24.4	35.3	36.0	**36.8**
en-es	32.7	37.6	40.0	**40.1**	40.1

Table 4: P@1 on VecMap test set varying the number of input parallel sentences. The number of induced seeds from IBM is 3,000 (or less, for lower data sizes with small vocabularies). 10,000 seeds are induced from VecMap. Top row is number of input sentences to IBM Model 2.

In Table 4, we examine the impact of the size of the input corpus to IBM Model 2 on downstream BLI performance. We feed between 500 and 20,000 parallel sentences from Europarl to the statistical translation model. In each experiment, we induce a maximum of 3,000 seeds from IBM Model 2.[4] In line with our intuition, performance generally increases as the size of the input corpus increases, and appears to plateau around 10,000 input sentences.

7.2 Ablation of frequency-based seed selection method

	en-de	en-fi
IBM only	64.1	44.8
VecMap Only	63.9	46.5
50% IBM + 50% VecMap	**64.8**	**47.9**

Table 5: P@1 on the development set of VecMap models trained with 3,000 seeds generated either from (1) IBM Model 2, (2) the previous run of VecMap, or (3) a combination of high-frequency translation pairs from IBM Model 2 and lower-frequency pairs from VecMap. IBM Model 2 was trained on 10,000 parallel sentences.

The size of the seed dictionary used for solving the Procrustes problem is a critically important parameter for success of mapping monolingual embedding spaces. Accordingly, a natural question to ask is whether our improved performance was due to the number of seeds induced alone, or our novel way of combining seeds extracted from both IBM and VecMap. To address this question, we use the en-de and en-fi models which used 10,000 lines of Europarl. In the first condition, we induce 3000 seeds from IBM Model 2 only, and train VecMap using these seeds. In the second condition, we extract 3000 from the first round training of VecMap, and feed only these into VecMap again for embedding space mapping retraining. In the third condition, we induce 1500 frequent words from IBM Model 2 and combine them with 1500 infrequent words induced from the phrase table generated from VecMap, according to our method for frequency-based seed selection with low-frequency agreement. We ensure that the resulting 3000 pair

[4] The number will be less for small vocabulary and if not enough potential translation pairs exceed the minimum confidence threshold.

seed set is split 50/50 between translation pairs induced from IBM Model 2 and those induced from VecMap. The results are presented in Table 5. We observe that when holding the number of induced seeds constant, best performance occurs using our combination method of keeping high-frequency translation pairs from IBM Model 2 and lower-frequency translation pairs from VecMap (according to the words' frequency in the 10,000 line parallel bitext).

Table 6 shows the relative importance of the two steps: induction from IBM 2 and inducing 10,000 additional seeds from VecMap to be fed back to VecMap for the final mapping. We use the first 3,000 seeds from the official VecMap training dictionaries as a baseline ("3k Artetxe Gold"), and show performance these gold seeds plus the additional 10,000 seeds induced from VecMap from the models trained using 10,000 lines of bitext (the models from row "Ours, N=10,000" of Table 3). For comparison, we show performance with the 3,000 pairs mined from IBM 2 only ("3k IBM2") from the same models, and report the development set performance of "Ours, N=10,000" under "3k IBM2 +10K VecMap". We observe that the secondary step of inducing 10,000 pairs from VecMap improves performance over the initial 3,000 seeds across all tested conditions, showing the magnitude of improvement between steps 1 (induction via IBM 2 or a given seed dictionary) and 2 (mining from word embedding space).

	3k Artetxe Gold	+10K VecMap	+3k IBM2	+10k VecMap
en-it	68.5	**70.3** *(+1.7)*	70.0	**70.3** *(+0.3)*
en-de	64.3	**65.3** *(+1.0)*	64.1	**66.6** *(+2.5)*
en-fi	48.9	**50.0** *(+1.1)*	44.8	**48.2** *(+3.4)*
en-es	66.0	**69.2** *(+3.3)*	66.4	**68.5** *(+2.0)*

Table 6: P@1 on the development set of models mapped with 3,000 seeds from the official VecMap Training Dictionary vs. 3,000 seeds induced from IBM2 with 10,000 lines of bitext, with or without an additional 10,000 pairs mined from the monolingual embedding spaces with VecMap.

8 Conclusion

Motivated by the strength of statistical translation and alignment models in inducing accurate word translation pairs from small amounts of data, the breadth of training data used to train monolingual word embedding spaces, we propose a motivated semi-supervised approach for bilingual lexicon induction that demonstrates state-of-the-art results on the challenging VecMap test sets. We capitalize upon the complementary strengths of statistical alignment and embedding space mapping methods for generating translation dictionaries, combining the methods for better downstream bilingual lexicon induction performance than either achieves alone. By taking this middle ground, we achieve state-of-the-art results with as little as 5,000 sentences - an amount readily available in thousands of language pairs. We release our implementation at `https://github.com/kellymarchisio/align-semisup-bli`.

References

Alvarez-Melis, D. and Jaakkola, T. (2018). Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium. Association for Computational Linguistics.

Artetxe, M., Labaka, G., and Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Con-*

ference on Empirical Methods in Natural Language Processing, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Artetxe, M., Labaka, G., and Agirre, E. (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.

Artetxe, M., Labaka, G., and Agirre, E. (2018a). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Artetxe, M., Labaka, G., and Agirre, E. (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)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Artetxe, M., Labaka, G., and Agirre, E. (2019). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2018). Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Dinu, G., Lazaridou, A., and Baroni, M. (2015). Improving zero-shot learning by mitigating the hubness problem.

Doval, Y., Camacho-Collados, J., Anke, L. E., and Schockaert, S. (2018). Improving cross-lingual word embeddings by meeting in the middle. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 294–304.

Faruqui, M. and Dyer, C. (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.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Third Workshop on Very Large Corpora*.

Grave, E., Joulin, A., and Berthet, Q. (2019). Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio. Association for Computational Linguistics.

Jawanpuria, P., Balgovind, A., Kunchukuttan, A., and Mishra, B. (2019). Learning multilingual word embeddings in latent metric space: a geometric approach. *Transactions of the Association for Computational Linguistics*, 7:107–120.

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., and Grave, É. (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.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P. and Knight, K. (2000). Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, page 711–715. AAAI Press.

Koehn, P. and Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 9–16, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Communication theory*, 84:486–502.

Mandelbrot, B. (1961). On the theory of word frequencies and on related markovian models of discourse. *Structure of language and its mathematical aspects*, 12:190–219.

McCarthy, A. D., Wicks, R., Lewis, D., Mueller, A., Wu, W., Adams, O., Nicolai, G., Post, M., and Yarowsky, D. (2020). The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mohiuddin, T., Bari, M. S., and Joty, S. (2020). Lnmap: Departures from isomorphic assumption in bilingual lexicon induction through non-linear mapping in latent space. *arXiv preprint arXiv:2004.13889*.

Mohiuddin, T. and Joty, S. (2020). Unsupervised word translation with adversarial autoencoder. *Computational Linguistics*, 46(2):257–288.

Patra, B., Moniz, J. R. A., Garg, S., Gormley, M. R., and Neubig, G. (2019). Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, page 320–322, USA. Association for Computational Linguistics.

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

Shi, H., Zettlemoyer, L., and Wang, S. I. (2021). Bilingual lexicon induction via unsupervised bitext construction and word alignment. *arXiv preprint arXiv:2101.00148*.

Smith, S. L., Turban, D. H. P., Hamblin, S., and Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Zipf, G. K. (1935). *The psycho-biology of language*. Houghton-Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.