Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Bilingual Lexicon Induction from Non-Parallel Data with Minimal Supervision

Meng Zhang,[1,2] **Haoruo Peng,**[3] **Yang Liu,**[1,2*] **Huanbo Luan,**[1] **Maosong Sun**[1,2]
[1]State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]Jiangsu Collaborative Innovation Center for Language Competence, Jiangsu, China
[3]University of Illinois, Urbana-Champaign
zmlarry@foxmail.com, hpeng7@illinois.edu, liuyang2011@tsinghua.edu.cn
luanhuanbo@gmail.com, sms@tsinghua.edu.cn

Abstract

Building bilingual lexica from non-parallel data is a long-standing natural language processing research problem that could benefit thousands of resource-scarce languages which lack parallel data. Recent advances of continuous word representations have opened up new possibilities for this task, e.g. by establishing cross-lingual mapping between word embeddings via a seed lexicon. The method is however unreliable when there are only a limited number of seeds, which is a reasonable setting for resource-scarce languages. We tackle the limitation by introducing a novel matching mechanism into bilingual word representation learning. It captures extra translation pairs exposed by the seeds to incrementally improve the bilingual word embeddings. In our experiments, we find the matching mechanism to substantially improve the quality of the bilingual vector space, which in turn allows us to induce better bilingual lexica with seeds as few as 10.

Introduction

Bilingual lexica provide valuable information for semantic equivalence of words across languages, and prove to be helpful for various cross-lingual tasks, including cross-lingual information retrieval (Levow, Oard, and Resnik 2005), statistical machine translation (Och and Ney 2003), and annotation projection for a variety of natural language processing tasks (Täckström et al. 2013, *inter alia*). Naturally, the size and quality of the bilingual lexica have a pivotal impact on these tasks (Levow, Oard, and Resnik 2005).

Although word alignment has proven effective for building bilingual lexica (Och and Ney 2003), it crucially relies on parallel data, and thus only applies to a limited number of domains between resource-rich languages. Therefore, researchers have focused their efforts on finding word translation pairs from non-parallel data, which is both more significant and more challenging (Koehn and Knight 2002; Fung and Cheung 2004; Haghighi et al. 2008). Most traditional approaches hinge on cross-lingual signals to link independent monolingual spaces: each word is associated with a vector that comprises monolingual statistics like PMI, and then the monolingual vector spaces are connected through bilingual signals, such as a seed lexicon or a bilingual topic model (Rapp 1999; Gaussier et al. 2004; Vulić, Smet, and Moens 2011; Vulić and Moens 2013a).

Recently, with the surge of continuous vector representation of words, commonly known as word embeddings, an interesting approach to bilingual lexicon induction is to replace the conventional statistics-based monolingual vector space with the new neural-network-inspired vector representation of words, which is supposed to carry semantic clues. This uncovers interesting findings. For example, an accurate linear transformation can be established between two monolingual embedding spaces by using a seed lexicon (Mikolov, Le, and Sutskever 2013).

However, this approach is limited by the number of seeds. As it relies on seed word pairs to provide supervision, large areas of the vector space may receive little guidance from seeds, especially when the seed lexicon is small in realistic settings. Attempts to translate words whose vectors reside in those areas are bound to fail miserably. Therefore, previous efforts typically require a rather large seed lexicon, with e.g. 5,000 entries (Mikolov, Le, and Sutskever 2013). Recent results indicate that at least a few hundred seeds are needed to achieve noticeable generalization (Vulić and Korhonen 2016).

Fortunately, as connection through seeds can reveal a few new translation pairs, we attempt to utilize them to provide further information about how the two monolingual vector spaces should be aligned, which should in turn expose more translation pairs. We can incrementally refine the quality of our bilingual word embeddings through this process, as it infuses information about proper alignment into extended areas of the vector space.

We encode our intuition into a novel *matching* term in the learning objective of bilingual word embeddings. Our matching term introduces latent variables that represent appropriate matching between words across languages. As every word in the vocabulary is considered for matching, the seed lexicon reaches its full potential because any newly discovered translation gets involved in refining the bilingual vector space. During training, our approach alternates between matching reliable translation pairs and adjusting the word vectors accordingly, which naturally reflects the intuition. In our experiments, we show that the matching mechanism substantially improves our system, compared to systems that only exploit seeds in superficial ways. This

*Corresponding author.

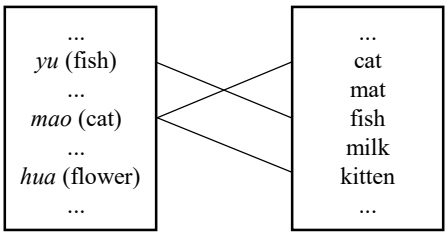

Figure 1: Illustration of the target-to-source matching for the Chinese-English language pair. Each target word in the corpus (right) is matched to a single source word (left). Target words shown without a link ("mat" and "milk") are implicitly matched to the special empty source word.

achievement in turn allows our method to successfully induce high-quality bilingual lexica with minimal supervision from only 10 seeds, which is particularly favorable for resource-scarce languages. The code of our system is available at http://nlp.csai.tsinghua.edu.cn/~zm/ Embedding-Matching.

Model

Like most approaches that learn bilingual word embeddings, our learning objective consists of monolingual and cross-lingual terms. However, unlike the usual cross-lingual term that draws signals from parallel sentences or a seed lexicon, ours additionally includes a matching term that attempts to improve bilingual word embeddings on its own. More formally, we maximize the following objective function

$$\mathcal{J}\left(W^{\mathrm{S}}, W^{\mathrm{T}}\right) = \mathcal{J}_{\mathrm{mono}} + \lambda_{\mathrm{s}}\mathcal{J}_{\mathrm{seed}} + \lambda_{\mathrm{m}}\mathcal{J}_{\mathrm{match}}, \quad (1)$$

where $W^{\mathrm{S}} \in \mathbb{R}^{D\times V^{\mathrm{S}}}$ and $W^{\mathrm{T}} \in \mathbb{R}^{D\times V^{\mathrm{T}}}$ are the model parameters, representing D-dimensional word embeddings of source and target languages, whose vocabulary sizes are V^{S} and V^{T}, respectively. Hyperparameters λ_{s} and λ_{m} control the relative weighting of the terms.

The monolingual term $\mathcal{J}_{\mathrm{mono}}$ is responsible for explaining regularities in corpora $\mathcal{C}^{\mathrm{S}}$ and $\mathcal{C}^{\mathrm{T}}$. Since the two corpora are non-parallel, $\mathcal{J}_{\mathrm{mono}}$ consists of two monolingual submodels that are independent of each other:

$$\mathcal{J}_{\mathrm{mono}} = \mathcal{J}^{\mathrm{S}}_{\mathrm{mono}}\left(W^{\mathrm{S}}\right) + \mathcal{J}^{\mathrm{T}}_{\mathrm{mono}}\left(W^{\mathrm{T}}\right). \quad (2)$$

As a common practice (Gouws, Bengio, and Corrado 2015), we choose the well established skip-gram model (Mikolov et al. 2013a) for our monolingual term.

The seed term $\mathcal{J}_{\mathrm{seed}}$ encourages embeddings of word translation pairs in a seed lexicon $\mathbf{d}$ to move near, which can be achieved via a L_2 regularizer:

$$\mathcal{J}_{\mathrm{seed}} = -\sum_{\langle s,t\rangle\in\mathbf{d}} \left\|W^{\mathrm{S}}_s - W^{\mathrm{T}}_t\right\|^2, \quad (3)$$

where $s \in \left\{1, ..., V^{\mathrm{S}}\right\}$ and W^{S}_s is the s-th column of W^{S} (i.e. the embedding of the s-th source word w^{S}_s), and notations are similar for the target side.

Our matching term is inspired by IBM model 1 (Brown et al. 1993). We begin by an exposition of target-to-source matching, while the reverse direction follows by symmetry. We assume each target word in the target corpus $\mathcal{C}^{\mathrm{T}}$ should be matched to a single source word or a special empty word (Figure 1), and multiple occurrences of the same target word should all be matched to the same source word. Thereby, we introduce a latent variable vector $\mathbf{m} \in \mathbb{N}^{V^{\mathrm{T}}}$. For each target word w^{T}_t in the target vocabulary, i.e. $t \in \left\{1, ..., V^{\mathrm{T}}\right\}$, $\mathbf{m}_t \in \left\{0, 1, ..., V^{\mathrm{S}}\right\}$ specifies which source word to link to, and $\mathbf{m}_t = 0$ indicates the empty word is linked. Then we can write out our target-to-source matching term:

$$\begin{aligned}\mathcal{J}^{\mathrm{T2S}}_{\mathrm{match}} &= \log P\left(\mathcal{C}^{\mathrm{T}} | \left\{w^{\mathrm{S}}_s\right\}_{s=1}^{V^{\mathrm{S}}}\right) &(4)\\ &= \log \sum_{\mathbf{m}} P\left(\mathcal{C}^{\mathrm{T}}, \mathbf{m} | \left\{w^{\mathrm{S}}_s\right\}_{s=1}^{V^{\mathrm{S}}}\right). &(5)\end{aligned}$$

We assume the matching process of each target word is independent of each other, and the matching probability should only depend on the matched word pair. Therefore, we have

$$\begin{aligned}&P\left(\mathcal{C}^{\mathrm{T}}, \mathbf{m} | \left\{w^{\mathrm{S}}_s\right\}_{s=1}^{V^{\mathrm{S}}}\right)\\ &= \prod_{w^{\mathrm{T}}\in\mathcal{C}^{\mathrm{T}}} P\left(w^{\mathrm{T}}, \mathbf{m} | \left\{w^{\mathrm{S}}_s\right\}_{s=1}^{V^{\mathrm{S}}}\right) &(6)\\ &= \prod_{t=1}^{V^{\mathrm{T}}} p\left(w^{\mathrm{T}}_t | w^{\mathrm{S}}_{\mathbf{m}_t}\right)^{c\left(w^{\mathrm{T}}_t\right)},\end{aligned}$$

where $c\left(w^{\mathrm{T}}_t\right)$ is the number of times the word w^{T}_t occurs in the corpus $\mathcal{C}^{\mathrm{T}}$. Finally, the parametrization of the matching probability has a number of possibilities. For example:

$$p\left(w^{\mathrm{T}}_t | w^{\mathrm{S}}_s\right) = \begin{cases}\epsilon & \text{if } s = 0\\ \frac{\exp\left(W^{\mathrm{T}}_t \cdot W^{\mathrm{S}}_s\right)}{\sum_{t'}\exp\left(W^{\mathrm{T}}_{t'} \cdot W^{\mathrm{S}}_s\right)} & \text{otherwise}\end{cases}, \quad (7)$$

where ϵ is a hyperparameter. We leave the discussion on a practical choice to a later section (Matching Probability).

Training

The optimization of monolingual and seed terms follows the established practice: we use negative sampling for our monolingual skip-gram model (Mikolov et al. 2013b), and computing the gradient of the seed term to apply stochastic gradient ascent is straightforward. However, due to the introduction of the latent variable, the optimization of the matching term poses a challenge. Using the EM technique is possible, but it requires significant computational cost. Therefore, we opt for the Viterbi EM algorithm. It alternates between a Viterbi E step and a subsequent M step.

Viterbi E Step

This step involves finding the most probable matching word pairs given the current parameters:

$$\begin{aligned}\widehat{\mathbf{m}} &= \arg\max_{\mathbf{m}}\left\{P\left(\mathcal{C}^{\mathrm{T}}, \mathbf{m} | \left\{w^{\mathrm{S}}_s\right\}_{s=1}^{V^{\mathrm{S}}}\right)\right\} &(8)\\ &= \arg\max_{\mathbf{m}}\left\{\prod_{t=1}^{V^{\mathrm{T}}} p\left(w^{\mathrm{T}}_t | w^{\mathrm{S}}_{\mathbf{m}_t}\right)^{c\left(w^{\mathrm{T}}_t\right)}\right\}. &(9)\end{aligned}$$

Due to independence, we can find matching for each word individually:

$$\widehat{\mathbf{m}}_t = \underset{s \in \{0,1,\ldots,V^{\mathrm{S}}\}}{\arg\max}\ p\left(w_t^{\mathrm{T}} | w_s^{\mathrm{S}}\right). \tag{10}$$

If the empty word is matched ($s = 0$), then the probability is simply ϵ. Otherwise, the matching is given by

$$\widetilde{\mathbf{m}}_t = \underset{s \in \{1,\ldots,V^{\mathrm{S}}\}}{\arg\max}\ p\left(w_t^{\mathrm{T}} | w_s^{\mathrm{S}}\right). \tag{11}$$

Therefore the Viterbi E step computes matching by

$$\widehat{\mathbf{m}}_t = \begin{cases} \widetilde{\mathbf{m}}_t & \text{if } p\left(w_t^{\mathrm{T}} | w_{\widetilde{\mathbf{m}}_t}^{\mathrm{S}}\right) > \epsilon \\ 0 & \text{otherwise} \end{cases}. \tag{12}$$

From this, we clearly see the role of ϵ: it serves as a threshold to keep out unreliable matching pairs.

M Step

The M step performs maximization as if the latent variable has been observed as given by the Viterbi E step. We treat the empty matching probability ϵ as a hyperparameter and hence hold it fixed during training. Therefore, the M step computes

$$\left(\widehat{W}^{\mathrm{S}}, \widehat{W}^{\mathrm{T}}\right) = \underset{W^{\mathrm{S}}, W^{\mathrm{T}}}{\arg\max}\, \mathcal{M}\left(W^{\mathrm{S}}, W^{\mathrm{T}}\right), \tag{13}$$

$$\mathcal{M} = \sum_{t=1}^{V^{\mathrm{T}}} \mathbb{I}\left[\widehat{\mathbf{m}}_t \neq 0\right] c\left(w_t^{\mathrm{T}}\right) \log p\left(w_t^{\mathrm{T}} | w_{\widehat{\mathbf{m}}_t}^{\mathrm{S}}\right). \tag{14}$$

Implementation

In this section, we explain the implementation details of our proposed system, which we omitted earlier.

Matching Probability

The parametrization of the matching probability given by Equation (7) is a softmax of inner products. The normalization factor poses significant computational cost and we choose to go unnormalized. For Viterbi E step, this simplifies to finding the maximum of inner products. Usually, the closely related cosine similarity better suits these scenarios, as will be used for retrieval during test time. Therefore, Equation (11) is replaced by

$$\widetilde{\mathbf{m}}_t = \underset{s \in \{1,\ldots,V^{\mathrm{S}}\}}{\arg\max}\ \cos\left(W_t^{\mathrm{T}}, W_s^{\mathrm{S}}\right), \tag{15}$$

and the overall Viterbi E step (12) is modified accordingly, where the hyperparameter ϵ now places a threshold on cosine similarity. For the M step, we treat the matched pairs as if they are correct translations (in the spirit of Viterbi EM), and use L_2 norm as well, making the M step (14) maximize

$$\mathcal{M} = -\sum_{t=1}^{V^{\mathrm{T}}} \mathbb{I}\left[\widehat{\mathbf{m}}_t \neq 0\right] c\left(w_t^{\mathrm{T}}\right) \left\| W_{\widehat{\mathbf{m}}_t}^{\mathrm{S}} - W_t^{\mathrm{T}} \right\|^2. \tag{16}$$

		# tokens (in million)	vocabulary size
zh-en	zh	21	3,349
	en	53	5,154
es-en	es	61	4,774
	en	95	6,637
it-en	it	73	8,490
	en	93	6,597
ja-zh	ja	38	6,043
	zh	16	2,814
tr-en	tr	6	7,482
	en	28	13,220

Table 1: Training set statistics. Language codes: zh = Chinese, en = English, es = Spanish, it = Italian, ja = Japanese, tr = Turkish.

Matching Directions

The original IBM model 1 is unidirectional and hence asymmetric. However, word translations are generally symmetric, meaning that translating back and forth would usually give the original word, especially if we assume the one-to-one translation constraint (Vulić and Moens 2013b; Gaussier et al. 2004). Therefore, it is natural to symmetrize our learning objective by including both matching directions:

$$\mathcal{J}_{\mathrm{match}} = \mathcal{J}_{\mathrm{match}}^{\mathrm{T2S}} + \mathcal{J}_{\mathrm{match}}^{\mathrm{S2T}}. \tag{17}$$

As we use online stochastic update for the M step (16), the counts $c\left(w_t^{\mathrm{T}}\right)$ and the corresponding $c\left(w_s^{\mathrm{S}}\right)$ may cause a weighting difference due to a disparity between the sizes of the two corpora. So we choose to use frequency instead, e.g. $c\left(w_t^{\mathrm{T}}\right) / \left|\mathcal{C}^{\mathrm{T}}\right|$ for the target side. This eliminates the need for separate matching weights, allowing a single λ_{m}.

In the same spirit as the one-to-one assumption, we restrict our search to words not covered by the seed lexicon during matching (as the seed term should take care of the rest), finalizing the Viterbi EM steps (15) (16) as

$$\widetilde{\mathbf{m}}_t = \underset{s \in \{1,\ldots,V^{\mathrm{S}}\} \wedge s \notin \mathbf{d}}{\arg\max}\ \cos\left(W_t^{\mathrm{T}}, W_s^{\mathrm{S}}\right), \tag{18}$$

$$\mathcal{M} = -\sum_{t=1}^{V^{\mathrm{T}}} \mathbb{I}\left[t \notin \mathbf{d}\right] \mathbb{I}\left[\widehat{\mathbf{m}}_t \neq 0\right] \frac{c\left(w_t^{\mathrm{T}}\right)}{\left|\mathcal{C}^{\mathrm{T}}\right|} \left\| W_{\widehat{\mathbf{m}}_t}^{\mathrm{S}} - W_t^{\mathrm{T}} \right\|^2. \tag{19}$$

Word Vectors and Context Vectors

We have presented our model with word vectors W^{S} and W^{T} as its parameters. In reality, each word is associated with a context vector as well (Mikolov et al. 2013b). While the usual representation of a word for evaluation is simply a word vector, some authors have suggested adding the context vector (Pennington, Socher, and Manning 2014; Levy, Goldberg, and Dagan 2015). Previously this means a simple post-processing step during evaluation, but in our setting we can bring the trick into training: the Viterbi EM steps (18) (19) consider context vectors when finding matchings and performing updates; the seed term (3) also encourages corresponding context vectors to be close.

Experimental Setup

Data

In our experiments, the tested systems induce bilingual lexica from Wikipedia comparable corpora[1] on five language pairs: Chinese-English, Spanish-English, Italian-English, Japanese-Chinese, and Turkish-English. Following (Vulić and Moens 2013a), we retain only nouns that occur at least 1,000 times in our corpora[2], except the resource-scarce Turkish-English pair, whose cutoff threshold is 100. For the Chinese side, we first use OpenCC[3] to normalize characters to be simplified, and then perform Chinese word segmentation and POS tagging with THULAC[4]. The preprocessing of the English side involves tokenization, POS tagging, lemmatization, and lowercasing, which we carry out with the NLTK toolkit[5] for the Chinese-English pair. For Spanish-English and Italian-English, we choose to use TreeTagger[6] for preprocessing, as in (Vulić and Moens 2013a). For the Japanese corpus, we use MeCab[7] for word segmentation and POS tagging. For Turkish, we utilize the preprocessing tools (tokenization and POS tagging) provided in LORELEI Language Packs (Strassel and Tracey 2016). The statistics of the preprocessed corpora is given in Table 1.

Ground Truth and Seed Word Translation Pairs

In order to carry out an objective evaluation, we need gold standard lexica for reference. For Chinese-English, we use Chinese-English Translation Lexicon Version 3.0[8] as the gold standard. For Spanish-English and Italian-English, we access Open Multilingual WordNet[9] through NLTK. For Japanese-Chinese, we use an in-house lexicon. For Turkish-English, we build a set of ground truth translation pairs in a way similar to (Vulić and Moens 2013a). First, we ask Google Translate to translate the source side vocabulary. Then the translations in the target language (English) are queried again in the reverse direction to translate back to the source language, and those that don't match with the original source words are discarded. This helps to ensure the quality of the ground truth translation, and embodies the one-to-one translation assumption. Finally, the target translations are lemmatized and lowercased, and discarded if they fall out of our target vocabulary. We reserve 10% of each gold standard lexicon for validation, and the remaining 90% for testing. As our task is bilingual lexicon induction, we retrieve the nearest neighbor in terms of cosine similarity in the target language space for each source word, and report accuracy as the evaluation metric, which is the percentage of source words with correctly retrieved translation.

Our approach only requires a minimal seed lexicon to work, and we choose to resort to Google Translate in the same way as we build the Turkish-English lexicon. We take the most frequent S word translation pairs as the seed lexicon. We vary S in an experiment to investigate its effect.

Baselines

We compare our approach to four baselines:

1. Statistics-based (`STAT`) (Gaussier et al. 2004).
2. Translation matrix (`TM`) (Mikolov, Le, and Sutskever 2013).
3. Isometric alignment (`IA`) (Zhang et al. 2016).
4. Bilingual Bag-of-Words without Word Alignments (`BilBOWA`) (Gouws, Bengio, and Corrado 2015).

The first baseline (`STAT`) is the traditional statistics-based approach, conventionally considered the standard approach to bilingual lexicon induction (Gaussier et al. 2004). We use a smoothed version of positive pointwise mutual information (PPMI) (Turney and Pantel 2010) as the monolingual association measure.

The following baselines are embedding-based. The second baseline (`TM`) is the pioneer of this type of approach mentioned in the introduction. It learns a translation matrix to establish a mapping between pretrained monolingual word embeddings. We use a publicly available implementation[10], and word2vec[11] to train monolingual word vectors.

The third baseline (`IA`) is an extension of `TM` that augments its learning objective with the isometric (orthonormal) constraint. The constraint was introduced to better cope with the limited seeds. Although Zhang et al. (2016) had subsequent steps for their POS tagging task, the `IA` technique could be used for bilingual lexicon induction as well. In our experiments, the same monolingual embeddings are provided for `TM` and `IA`.

The fourth baseline (`BilBOWA`) is a state-of-the-art bilingual distributed representation learner. It was designed to draw cross-lingual signals from parallel sentences, but this scenario does not fit the bilingual lexicon induction task. We nonetheless apply this method by treating each seed word translation pair as a parallel sentence pair.

Hyperparameters

Our system inherits hyperparameters from the monolingual skip-gram model, includes term weights λ_{s} and λ_{m}, and introduces matching threshold ϵ.

The monolingual hyperparameters are set as follows: embedding size D is 40; window size is 5; 5 negative samples; subsampling threshold is 10^{-5}; initial learning rate is 0.1; 10 training epochs. These hyperparameters do not impact performance much as long as they lie within a reasonable range. The same setting is used for the word2vec toolkit, and the statistics-based baseline uses a window size of 5 as well. The

[1] http://linguatools.org/tools/corpora/wikipedia-comparable-corpora

[2] Lower cutoff threshold makes the task more challenging with larger vocabularies, and we observe performance degradation for all systems.

[3] https://github.com/BYVoid/OpenCC

[4] http://thulac.thunlp.org

[5] http://www.nltk.org

[6] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

[7] http://taku910.github.io/mecab

[8] https://catalog.ldc.upenn.edu/LDC2002L27

[9] http://compling.hss.ntu.edu.sg/omw

[10] http://clic.cimec.unitn.it/~georgiana.dinu/down

[11] https://code.google.com/archive/p/word2vec

Method	zh-en	es-en	it-en	ja-zh	tr-en
STAT	10.63	20.69	19.22	13.88	5.66
TM	5.21	10.76	6.41	5.73	5.57
IA	17.39	32.36	23.66	16.64	8.64
BilBOWA	2.12	2.51	1.56	2.24	4.22
Ours	45.10	73.21	61.08	50.04	28.50

Table 2: Accuracies in percentage of the statistics-based (STAT), translation matrix (TM), isometric alignment (IA), BilBOWA baselines, and our system for inducing lexica of five language pairs with 50 seeds.

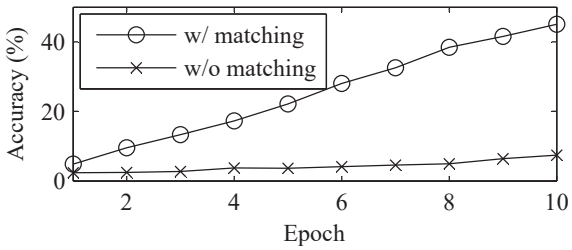

Figure 2: Accuracies after each epoch for our system with and without the matching term, with 50 seeds for Chinese-English. The matching mechanism provides substantial performance boost, and exhibits incremental refinement.

cross-lingual regularization weight of the BilBOWA baseline is tuned over $\{0.01, 0.1, 1\}$.

The seed term weight λ_{s} has limited impact as long as its value is not too low to tie up bilingual vector spaces, and we set it to 0.01. The matching threshold ϵ can also be set quite liberally as long as it is sufficiently low (in our experiments 0.5), otherwise few matching pairs could be found and the matching mechanism would be ineffective.

The matching term weight λ_{m} appears to be the most important hyperparameter, so we tune it on the validation set with values in $\{100, 1000, 10000\}$. Intuitively, too low a λ_{m} will render matching ineffective (equivalently disabling the matching term), while the other extreme will cause erroneous matched pairs to be reinforced too quickly with no hope to rectify. Our hypothesis is validated by the outcome on the validation set, and further backed up by observing that when the matching weight value is too large, matched words are mostly wrong with cosine similarities of almost 1 soon after training begins.

Results and Discussion

Overall Performance

Table 2 shows the performance of our system and the four baselines across the five language pairs. The seed lexicon size is 50.

Compared to our system, the baselines attain considerably lower performance for all the language pairs. The poor performance should be attributed to the harsh condition they have to face – 50 seed word translation pairs are after all too few for them to link vocabularies of two distinct languages. However, the success of our approach demonstrates that it is actually possible to connect two language spaces with such few seeds. Our approach manages to make full use of the limited information encoded in the seeds to generalize with little supervision, and obtain reasonable performance.

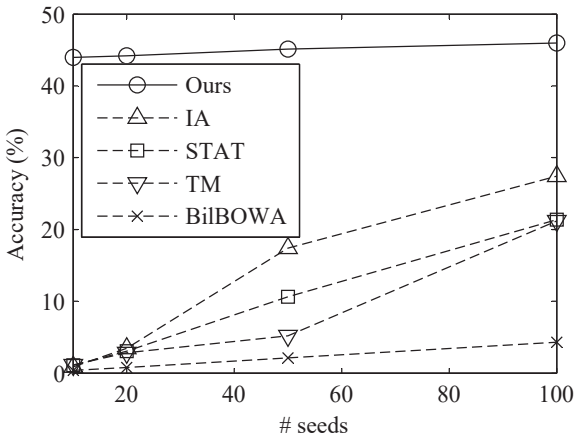

Figure 3: Accuracies of the tested systems for Chinese-English with varying seed lexicon size.

Across the four major language pairs, the performance on the closely related language pairs (Spanish-English and Italian-English) is generally higher than the relatively distant ones (Chinese-English and Japanese-Chinese).

Turkish-English is a resource-scarce language pair with limited parallel data. In fact, its non-parallel data in our experiments is considerably smaller (cf. Table 1). All systems suffer in this case, but ours still manages to significantly outperform the others.

Effect of the Matching Term

In order to investigate the effectiveness of the matching term, we run a version of our system with the matching term disabled, counting on the seed term to properly align the word vector spaces of the two languages. This can be considered another baseline, as a variant of the approach by (Shi et al. 2015) (discussed further in Related Work). We record the performance after each epoch to monitor the training process, shown in Figure 2. The plot is for Chinese-English; other language pairs exhibit similar trends.

From the figure we immediately see the important role the matching term plays in our lexicon inducer: enabling the matching term can result in dramatic accuracy gain. This finding, combined with the inadequate performance of the baselines, conforms to our conjecture about the limitation of insufficient seed lexicon exploitation. Looking closer at the curve trend, we observe that the matching mechanism incrementally increases accuracy as training proceeds, which echoes our intuition about how matching works. In contrast, the seed term alone is incompetent in discovering extra word translation pairs.

Effect of Seed Lexicon Size

In this section, we investigate how the number of seed word translation pairs may affect the performance of the

shui	*dongwu*	*jiezhi*
water	mammal	disguise
lake	reptile	curse
reservoir	**animal**	demon

Table 3: Three nearest English word translations to (non-seed) Chinese words, given by our system trained with 50 seeds. Bold typeface marks ground truth. For *jiezhi*, the ground truth is "ring" (finger ring).

bilingual lexicon inducers. We vary the seed lexicon size in $\{10, 20, 50, 100\}$. Figure 3 shows the accuracies of the tested systems for Chinese-English. We observe that our system always attains high performance, even when the seeds are as few as 10. It appears that new word translation pairs exposed through matching have well compensated for the small initial seed lexicon. We observe similar trends for the other language pairs except Turkish-English, for which our system needs 20 seeds. We conjecture this is due to the small non-parallel corpus. As for the baseline systems, a limited number of seeds considerably degrades their performance. Therefore, our system is particularly appealing in realistic resource-scarce scenarios for its minimal requirement for a seed lexicon, which can be expensive to obtain.

Qualitative Analysis

We analyze the behavior of our system through Chinese-English examples in Table 3. In particular, we look at what English words get matched to the Chinese word during training. Initially, *shui* (water) is matched to irrelevant words like "comeback" and "preacher". Soon more relevant words like "fluid" and "liquid" show up. In the second epoch, "water" is matched for the first time with cosine similarity 0.75. Then with gradually increasing similarity "water" sharply stands out with 0.99 at the end of training. This process demonstrates how the matching mechanism refines the bilingual word vector space. However, for *dongwu* (animal), our system does not sufficiently distinguish "animal" from semantically related words like "mammal" and "reptile", possibly due to the similar context of these words. For *jiezhi* (finger ring), the failure should be attributed to lexicalization difference between the two languages. In Chinese, *jiezhi* is a very specific word that occurs rarely in the corpus; its word vector may be unreliable even from a monolingual standpoint. In contrast, the English ground truth "ring" is a common polysemous word with tens of times more occurrences. This indicates that our system, like previous ones, is inadequate when the one-to-one translation assumption is violated. However, our work lays a foundation for introducing more sophisticated IBM models with fertilities (Brown et al. 1993) that can potentially address this issue. Integrating with multi-sense embeddings (Li and Jurafsky 2015) could be another solution. We leave the remedies to future work.

Related Work

There have been a number of papers that attempt to learn bilingual word embeddings (Upadhyay et al. 2016). However, most of them require parallel data as the cross-lingual signal (Zou et al. 2013; Chandar A P et al. 2014; Hermann and Blunsom 2014; Kočiský, Hermann, and Blunsom 2014; Gouws, Bengio, and Corrado 2015; Luong, Pham, and Manning 2015; Coulmance et al. 2015), which renders them unattractive for bilingual lexicon induction because word alignment can already find high-quality word translation pairs (Och and Ney 2003).

Here we discuss works that use a seed lexicon as the cross-lingual signal, as this is the appropriate setting for bilingual lexicon induction. These works can be roughly categorized into two types. The first type exploits the given lexicon to construct a pseudo-corpus from non-parallel data for bilingual representation learning (Gouws and Søgaard 2015; Wick, Kanani, and Pocock 2016; Ammar et al. 2016; Duong et al. 2016). However, it is conceivable that this idea would not work properly if the given lexicon is too small. The second type treats the seed lexicon as a source of supervision during the training of bilingual representation, either in a similar form to our seed term (Shi et al. 2015), or as an extension to the translation matrix approach (Dinu, Lazaridou, and Baroni 2015; Lazaridou, Dinu, and Baroni 2015; Faruqui and Dyer 2014; Lu et al. 2015; Ammar et al. 2016; Vulić and Korhonen 2016). Our experiments have revealed that, without our matching term, the seed term alone is insufficient to tie up the bilingual word vector space with a very small seed lexicon, nor is the translation matrix approach, which is consistent with (Vulić and Korhonen 2016).

Prior to the advent of embedding-based word translation, the idea of bootstrapping has been applied to bilingual lexicon induction (Vulić and Moens 2013b), and the intuition bears a resemblance to our work. However, all the other components of the system are quite different: they build on conventional statistics-based approach (Gaussier et al. 2004) and do not involve word embeddings, and the bootstrapping procedure calls for a number of heuristic design choices.

Recently, (Dong et al. 2015) proposed a model similar in spirit, as they also designed a matching term to iteratively improve learning. One difference is that their matching unit is phrase, and consequently their task is to mine parallel phrase pairs. Another crucial difference lies in the parametrization of the matching probability (in their case phrase translation probability). They use the standard IBM model 1 to define the phrase translation probability and their model does not involve continuous representation of words, which in turn leads to different optimization procedure. In this sense, their matching model is purely generative.

Conclusion

In this paper, we explore bilingual lexicon induction from non-parallel data with a small seed lexicon. We train bilingual word embeddings in a shared semantic space to achieve our goal. In addition to using the seed lexicon to properly embed words across languages, we maximize its potential of pairing words cross-lingually by introducing a matching term to our learning objective. We show the matching term dramatically enhances our system, and allows it to function well even when the seeds are as few as 10. This harsh condition baffles previous methods, but is highly desirable for truly resource-scarce scenarios.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61522204 and No. 61432013), the 863 Program (2015AA015407), and Fund from Online Education Research Center, Ministry of Education (No. 2016ZD102).

References

Ammar, W.; Mulcaire, G.; Tsvetkov, Y.; Lample, G.; Dyer, C.; and Smith, N. A. 2016. Massively Multilingual Word Embeddings. In *arXiv:1602.01925 [cs]*.

Brown, P.; Della Pietra, S.; Della Pietra, V.; and Mercer, R. 1993. The Mathematics for Statistical Machine Translation: Parameter Estimation. *CL*.

Chandar A P, S.; Lauly, S.; Larochelle, H.; Khapra, M.; Ravindran, B.; Raykar, V. C.; and Saha, A. 2014. An Autoencoder Approach to Learning Bilingual Word Representations. In *NIPS*.

Coulmance, J.; Marty, J.-M.; Wenzek, G.; and Benhalloum, A. 2015. Trans-gram, Fast Cross-lingual Word-embeddings. In *EMNLP*.

Dinu, G.; Lazaridou, A.; and Baroni, M. 2015. Improving Zero-Shot Learning by Mitigating the Hubness Problem. In *ICLR Workshop*.

Dong, M.; Liu, Y.; Luan, H.; Sun, M.; Izuha, T.; and Zhang, D. 2015. Iterative Learning of Parallel Lexicons and Phrases from Non-Parallel Corpora. In *IJCAI*.

Duong, L.; Kanayama, H.; Ma, T.; Bird, S.; and Cohn, T. 2016. Learning Crosslingual Word Embeddings without Bilingual Corpora. In *arXiv:1606.09403 [cs]*.

Faruqui, M., and Dyer, C. 2014. Improving Vector Space Word Representations Using Multilingual Correlation. In *EACL*.

Fung, P., and Cheung, P. 2004. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In *EMNLP*.

Gaussier, E.; Renders, J.; Matveeva, I.; Goutte, C.; and Dejean, H. 2004. A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. In *ACL*.

Gouws, S., and Søgaard, A. 2015. Simple task-specific bilingual word embeddings. In *NAACL-HLT*.

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

Haghighi, A.; Liang, P.; Berg-Kirkpatrick, T.; and Klein, D. 2008. Learning Bilingual Lexicons from Monolingual Corpora. In *ACL-HLT*.

Hermann, K. M., and Blunsom, P. 2014. Multilingual Distributed Representations without Word Alignment. In *ICLR*.

Koehn, P., and Knight, K. 2002. Learning a Translation Lexicon from Monolingual Corpora. In *ACL Workshop on Unsupervised Lexical Acquisition*.

Kočiský, T.; Hermann, K. M.; and Blunsom, P. 2014. Learning Bilingual Word Representations by Marginalizing Alignments. In *ACL*.

Lazaridou, A.; Dinu, G.; and Baroni, M. 2015. Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning. In *ACL-IJCNLP*.

Levow, G.-A.; Oard, D. W.; and Resnik, P. 2005. Dictionary-based techniques for cross-language information retrieval. *Information Processing & Management*.

Levy, O.; Goldberg, Y.; and Dagan, I. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *TACL*.

Li, J., and Jurafsky, D. 2015. Do Multi-Sense Embeddings Improve Natural Language Understanding? In *EMNLP*.

Lu, A.; Wang, W.; Bansal, M.; Gimpel, K.; and Livescu, K. 2015. Deep Multilingual Correlation for Improved Word Embeddings. In *NAACL-HLT*.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Bilingual Word Representations with Monolingual Quality in Mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient Estimation of Word Representations in Vector Space. In *ICLR Workshop*.

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

Mikolov, T.; Le, Q. V.; and Sutskever, I. 2013. Exploiting Similarities among Languages for Machine Translation. In *arXiv:1309.4168 [cs]*.

Och, F. J., and Ney, H. 2003. A Systematic Comparison of Various Statistical Alignment Models. *CL*.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*.

Rapp, R. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *ACL*.

Shi, T.; Liu, Z.; Liu, Y.; and Sun, M. 2015. Learning Cross-lingual Word Embeddings via Matrix Co-factorization. In *ACL-IJCNLP*.

Strassel, S., and Tracey, J. 2016. LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages. In *LREC*.

Turney, P. D., and Pantel, P. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *JAIR*.

Täckström, O.; Das, D.; Petrov, S.; McDonald, R.; and Nivre, J. 2013. Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging. *TACL*.

Upadhyay, S.; Faruqui, M.; Dyer, C.; and Roth, D. 2016. Cross-lingual Models of Word Embeddings: An Empirical Comparison. In *ACL*.

Vulić, I., and Korhonen, A. 2016. On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In *ACL*.

Vulić, I., and Moens, M.-F. 2013a. Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses. In *NAACL-HLT*.

Vulić, I., and Moens, M.-F. 2013b. A Study on Bootstrapping Bilingual Vector Spaces from Non-Parallel Data (and Nothing Else). In *EMNLP*.

Vulić, I.; Smet, W. D.; and Moens, M.-F. 2011. Identifying Word Translations from Comparable Corpora Using Latent Topic Models. In *ACL-HLT*.

Wick, M.; Kanani, P.; and Pocock, A. 2016. Minimally-Constrained Multilingual Embeddings via Artificial Code-Switching. In *AAAI*.

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

Zou, W. Y.; Socher, R.; Cer, D.; and Manning, C. D. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *EMNLP*.