Building Earth Mover's Distance on Bilingual Word Embeddings for Machine Translation

Meng Zhang[1] **Yang Liu**[1,2] **Huanbo Luan**[1] **Maosong Sun**[1,2] **Tatsuya Izuha**[3] **Jie Hao**[4]
[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]Toshiba Corporation Corporate Research & Development Center
[4]Toshiba (China) R&D Center
zmlarry@foxmail.com, {liuyang.china, luanhuanbo}@gmail.com, sms@tsinghua.edu.cn
tatsuya.izuha@toshiba.co.jp, haojie@toshiba.com.cn

Abstract

Following their monolingual counterparts, bilingual word embeddings are also on the rise. As a major application task, word translation has been relying on the nearest neighbor to connect embeddings cross-lingually. However, the nearest neighbor strategy suffers from its inherently local nature and fails to cope with variations in realistic bilingual word embeddings. Furthermore, it lacks a mechanism to deal with many-to-many mappings that often show up across languages. We introduce Earth Mover's Distance to this task by providing a natural formulation that translates words in a holistic fashion, addressing the limitations of the nearest neighbor. We further extend the formulation to a new task of identifying parallel sentences, which is useful for statistical machine translation systems, thereby expanding the application realm of bilingual word embeddings. We show encouraging performance on both tasks.

Introduction

Over the past few years, distributed representations of words, commonly referred to as word embeddings, have shown promise for a range of natural language processing tasks. They are welcomed by the community because they represent words by continuous vectors and overcome many limitations of the traditional discrete representations. Crucially, they allow capturing syntactic and semantic regularities of words by training neural networks on large-scale corpora. For example, semantically similar words are close to each other, and linear operations between word embeddings can capture interesting relationships between words (Mikolov et al. 2013a; 2013b; Mikolov, Yih, and Zweig 2013).

With the increasing popularity of monolingual word embeddings, their bilingual counterparts are also gaining attention. Ideally, bilingual word embeddings should capture cross-lingual regularities in addition to monolingual ones by, for example, placing monolingual synonyms and their translations in a cluster. A number of recent works have pursued this goal and produced high quality bilingual word embeddings (Zou et al. 2013; Mikolov, Le, and Sutskever 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; Vulić and Moens 2015; Soyer, Stenetorp, and Aizawa 2015).

Despite these efforts, existing bilingual word embeddings are still far from perfect due to the diversity and distinctiveness across natural languages. This is evidenced by the performance that leaves much to be desired on word translation, one major task that has been explored to test the quality of bilingual word embeddings.

However, the inadequate performance should also be attributed to the retrieval strategy for translation. Previously, a source word is translated by simply searching the target vocabulary for the nearest neighbors of the word vector (Mikolov, Le, and Sutskever 2013; Gouws, Bengio, and Corrado 2015; Vulić and Moens 2015). While this simple strategy would be perfectly effective if the bilingual word embeddings were ideal, it is not robust to variations in imperfect real settings due to its essentially local nature. As an illustrative example, Figure 1(a) shows a case where the nearest neighbor fails to translate both left-side words correctly because a right-side word is too close to them. To demonstrate the seriousness of this problem, we find in our English-Italian translation task that the nearest neighbor makes nearly three quarters of source words (1693 out of 2266) share target proposals, and only 71 of them are validated by the lexicon. Moreover, the nearest neighbor exhibits a counterintuitive behavior when we reverse the translation direction: Retrieving the nearest neighbor for each right-side word would result in correct translation in Figure 1(a). This asymmetry diverges from our general understanding of translation.

Conceptually, for the example in Figure 1, if we are able to minimize the total distance incurred by the translation paths subject to the one-to-one translation constraint, we will successfully find the correct translations, regardless of translation direction. We capture this intuition by introducing the Earth Mover's Distance (EMD). It is a well-studied optimization problem that aims to find the minimum distance required to move one set of points to another, which provides a suitable solution to our task at hand. In addition to the one-to-one translation discussed above, the EMD

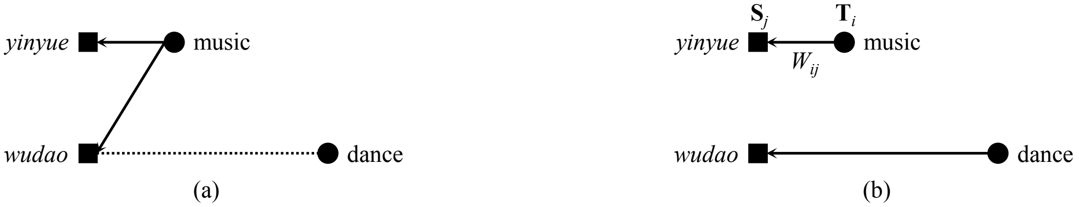

Figure 1: An illustration of bilingual word embeddings for translating from Chinese to English, with squares representing source side embeddings and circles target, and "*yinyue*"/"music" is a (romanized) Chinese-English translation pair, while "*wudao*"/"dance" is another. (a) The nearest neighbor incorrectly translates "*wudao*" to "music" because the dotted path is longer. (b) The Earth Mover's Distance correctly matches both word pairs. We assume each word carries the same weight in this illustration, as indicated by the matching sizes of squares and circles. In this case, the EMD automatically enforces the one-to-one translation constraint, and consequently moves earth from circles to squares as indicated by the arrows.

formulation naturally tackles multiple translations, which commonly occur cross-lingually but nearest neighbor cannot properly handle. Our evaluations show dramatic performance gains brought by the EMD, which alleviates the aforementioned problem by reducing the number of source words that share target proposals to 770, and 140 of them can be found in the lexicon. Furthermore, an analysis of the EMD-produced translations reveals its interesting behavior in dealing with multiple translations.

The idea of using the EMD to match word vectors in different vocabularies can be further extended for matching bilingual sentences. As its name suggests, the EMD provides a distance so that we can measure sentence pairs across languages. Drawing information from bilingual word embeddings, this sentence-level distance is expected to encode semantic closeness between a bilingual sentence pair. We verify its effectiveness by introducing a new task of finding parallel sentences out of a noisy bilingual corpus. With the help of the EMD, the debut of an embedding-based approach on this stage is encouragingly successful.

Related Work

Word translation is closely related to bilingual lexicon extraction, which is a basis for many cross-lingual applications. Although prior works also involve representing words with vectors (Haghighi et al. 2008; Vulić and Moens 2013), they differ substantially from the neural word embeddings we focus on, and from the methodology of our work. Similarly, previous works on bilingual corpus filtering (Khadivi and Ney 2005; Taghipour et al. 2010, *inter alia*) have never utilized neural word embeddings.

Mikolov, Le, and Sutskever (2013) pioneered the use of neural word embeddings for word translation by a translation matrix mapping between monolingual word embeddings. Gouws, Bengio, and Corrado (2015) followed up by providing standard bilingual word embeddings with improved results. Vulić and Moens (2015) explored training bilingual word embeddings with comparable corpora to induce bilingual lexica. As noted in our introduction, they all use nearest neighbors to translate words, which we attempt to improve upon.

Moving from word-level translation to sentence-level, we see a wealth of works that involve the use of word embeddings. However, most of them train a neural network model initialized by monolingual word embeddings, which are in turn trained separately on monolingual data, such as the work of (Zhang et al. 2014). The only work we are aware of that utilizes bilingual word embeddings is (Zou et al. 2013). In addition to training bilingual word embeddings on their own, they composed phrase embeddings by averaging word embeddings and computed distance between phrase pairs to serve as an additional feature for a phrase-based machine translation system. Our work uses the EMD instead of the simple averaging composition, and computes distance on the sentence level, though our idea can also be applied to the phrase level.

The Earth Mover's Distance is a well-studied transportation problem (Rubner, Tomasi, and Guibas 1998) and there exist fast specialized solvers (Pele and Werman 2009). The EMD has seen wide application in the computer vision literature (Rubner, Tomasi, and Guibas 1998; Ren, Yuan, and Zhang 2011). Recently, it has been successfully used to derive a document distance metric from monolingual word embeddings, which is then used for document categorization (Kusner et al. 2015). We apply the EMD to machine translation, a fundamental cross-lingual task, with the basis being bilingual word embeddings. Our novel formulation for word translation also differs substantially.

Earth Mover's Distance with Bilingual Word Embeddings

In this section, we first formulate the word translation task as an instance of the Earth Mover's Distance problem. As we get familiar with the framework, it can be naturally extended to derive a distance between a pair of bilingual sentences.

Word Translation

Our approach builds on bilingual word embeddings, which are fixed throughout. Each word in both languages is associated with a word vector in a shared D-dimensional semantic space, which can be looked up in matrices $\mathbf{S} \in \mathbb{R}^{D \times V_s}$ and $\mathbf{T} \in \mathbb{R}^{D \times V_t}$ for the source and target languages with vocabulary sizes V_s and V_t, respectively. In this D-dimensional space, the Euclidean distance between a target embedding $\mathbf{T}_i$ and a source embedding $\mathbf{S}_j$ naturally measures the cost of translation between the i-th target word and the j-th source word. We denote this cost as $C_{ij} = \|\mathbf{T}_i - \mathbf{S}_j\|$, where $i \in \{1, ..., V_t\}, j \in \{1, ..., V_s\}$.

We also associate each word with the number of times it appears in the parallel corpus. These frequencies are packed in vectors $\mathbf{s} \in \mathbb{R}^{V_s}$ and $\mathbf{t} \in \mathbb{R}^{V_t}$, where s_j represents the frequency for the j-th source word, and similarly for the target side. Importantly, since the frequencies are collected on a parallel corpus, we expect one-to-one translation pairs to show up roughly equal times.

We can imagine source side words as holes, and target side words as piles of earth, scattered in the D-dimensional space as specified by the bilingual word embeddings. Each word carries a frequency as its weight, representing the volume of a hole, or the amount of earth in a pile. Then our goal is to move the earth to fill up the holes with minimal cost. If we ensure there is sufficient earth, then every source word will get fully explained.

This formulation is helpful for both one-to-one and multiple translations. For one-to-one translations, the word frequencies on both sides are likely to match, which will naturally lead to correct translation as shown in Figure 1(b), because "music" would have used up all its earth to fill the "*yinyue*" hole, and hence would not interfere with the translation of "*wudao*". In the multiple-source-to-one-target case for example, the target word frequency will roughly equal the sum of frequencies of source equivalents, resulting in a distribution of earth to multiple holes.

We formalize the intuition as follows. Let $\mathbf{W} \in \mathbb{R}^{V_t \times V_s}$ be a (sparse) transportation matrix, with W_{ij} representing the amount of earth moved from the i-th pile to the j-th hole. The optimization problem becomes

$$\begin{aligned} &\min \sum_{i=1}^{V_t} \sum_{j=1}^{V_s} W_{ij} C_{ij} \\ &s.t.\ W_{ij} \geq 0 \\ &\quad \sum_{j=1}^{V_s} W_{ij} \leq t_i, i \in \{1, ..., V_t\} \\ &\quad \sum_{i=1}^{V_t} W_{ij} = s_j, j \in \{1, ..., V_s\} \end{aligned} \tag{1}$$

To ensure the problem is feasible, we require that

$$\sum_{j=1}^{V_s} s_j \leq \sum_{i=1}^{V_t} t_i. \tag{2}$$

This means there is at least as much earth to fill up all the holes, which will ensure all source words get translated for our task.

This optimization problem is a linear program, with specialized solvers available. Once the optimization problem is solved, the transportation matrix $\mathbf{W}$ will store the translation relations between words of the two languages, where a non-zero W_{ij} signifies a translation between the i-th target word and the j-th source word. Finally, the Earth Mover's Distance is defined as

$$\text{EMD}(\mathbf{s}, \mathbf{t}, \mathbf{C}) = \frac{\sum_{i=1}^{V_t} \sum_{j=1}^{V_s} W_{ij} C_{ij}}{\sum_{j=1}^{V_s} s_j}, \tag{3}$$

which is the objective value (total cost) normalized by the total volume of the holes. This will quantify the quality of the translation process based on bilingual word embeddings.

As a final note, the moving direction introduced by the metaphor of earth and holes is not essential to the optimization. In case the feasibility condition is violated, we can always exchange the roles of earth and holes, or equivalently swap the variable names $\mathbf{s}$ and $\mathbf{t}$, to arrive at a feasible program. This means the EMD optimization essentially gives a matching between V_s and V_t vocabularies. This invariance to translation direction makes our approach immune from the counterintuitive behavior of the nearest neighbor.

Bilingual Sentence Distance

The optimization problem (1) can be easily extended to other tasks by varying the notion of frequency. In this section, we would like to design a distance measure between a pair of bilingual sentences.

Naturally, we associate vectors $\mathbf{s} \in \mathbb{R}^{V_s}$ and $\mathbf{t} \in \mathbb{R}^{V_t}$ to a sentence pair, representing the source side and the target side respectively. A first idea would be using bag-of-words representations for the two sentences, and normalization is desired because we weigh both sentences equally, similar to the monolingual case in (Kusner et al. 2015). However, the commonest words, mostly function words, would carry overly high weights that overshadow the contribution of the more important content words. What is worse, function words often have no correspondence cross-lingually, which would result in huge piles of earth finding hardly any meaningful holes to move to. Kusner et al. simply removed stop words for their document categorization task. However, we argue that word removal is seldom an option under the translation scenario. Therefore, we propose to introduce the principled Inverse Document Frequency (IDF) to the weighting scheme to downplay the effect of common words.

Formally, for the j-th word in the source language, s_j now represents its term frequency in the source sentence, reweighted by its IDF and a normalization factor:

$$s_j = \frac{\text{TF-IDF}(j)}{\sum_k \text{TF-IDF}(k)}, j \in \{1, ..., V_s\}.$$

Things are similar for the target side.

Since a sentence is relatively short, $\mathbf{s}$ and $\mathbf{t}$ are usually very sparse. The sparsity allows a dramatic reduction in the number of constraints of the optimization problem (1), which will lead to fast solution. Furthermore, normalization causes the feasibility condition (2) to hold with equality, and the second set of constraints in (1) can be rewritten as equality constraints. Finally, the EMD (3) simplifies to the objective value in (1). This value gathers word distances specified by bilingual word embeddings to the sentence level through

Frequency bin	0-4K	0-1K	1-2K	2-3K	3-4K
M	4000	1000	1000	1000	1000
# OOL	1695	212	346	499	638
# OOC	39	9	4	7	19
N	2266	779	650	494	343
Coverage (%)	98.3	98.9	99.4	98.6	94.8

Table 1: English-Italian test set statistics drawn from the most frequent 4K English words and split over four bins. M is the number of words in the bin, and N is the number of test instances after ruling out out-of-lexicon (OOL) and out-of-coverage (OOC) words. Coverage shows the percentage of N over in-lexicon words.

the EMD optimization, and we hope it to encode the semantic distance between a sentence pair. We will use this sentence distance to filter out non-parallel sentence pairs in a bilingual corpus.

Experiments

Word Translation

We start out our experiments on English-Italian word translation before we move to the Chinese-English task.

Setup Our approach can be applied on top of any existing bilingual embeddings. We choose BilBOWA (Gouws, Bengio, and Corrado 2015) to train bilingual word embeddings, which holds the state of the art on this task. Our embeddings consist of 50 dimensions, and other hyperparameters are as recommended by the toolkit.

We take nearest neighbor (NN) as our primary baseline. Dinu, Lazaridou, and Baroni (2014) proposed a globally-corrected (GC) method to deal with the hubness problem NN faces. We implemented their approach and report results with zero and 5000 additional source word vectors, where more additional vectors are expected to enhance performance.

Out of computation efficiency concerns, we limit the source and target vocabularies to the most frequent 4000 and 10000 words, i.e. $V_s = 4000$ and $V_t = 10000$. We choose a relatively large V_t because we want to ensure the feasibility condition (2) is met so that all the V_s source words get translated. It is not common practice to limit the target vocabulary where NN retrieves from, but we actually observe a boost in NN performance, likely because it keeps NN from being distracted by rare garbage words. Indeed, it is fairly unlikely that a top 4000 frequent source word should translate to a target word that falls out of rank 10000.

Training and Testing Data We train BilBOWA on the English-Italian part of Europarl (Koehn 2005), a parallel corpus of 1.94M sentence pairs with 50.8M English words and 49.0M Italian words. A lexicon is needed to determine the correctness of word translation. We use an English-Italian lexicon[1] that consists of 4942 English entries, each having one or more Italian translations.

	K	R	A
NN	1	4.91	7.11
GC-0	1	11.05	15.98
GC-5000	1	11.44	16.55
EMD	1	19.50	28.20

Table 2: Overall recall and accuracy for English-Italian word translation when each approach only proposes one translation ($K = 1$).

Evaluation Metrics Our evaluation procedure is as follows. First, we take $M \leq V_s$ source words to look up the gold lexicon. If a word is not found, it is counted as out-of-lexicon (OOL). If all the translations of an in-lexicon word fall out of our V_t-sized target vocabulary, this source word is counted as out-of-coverage (OOC). Then we gather our test instances totaling $N = M - \#\text{OOL} - \#\text{OOC}$. We also report $\text{Coverage} = \frac{N}{M - \#\text{OOL}}$ to gauge our target vocabulary size. We summarize the test set information in Table 1, where we also split statistics over frequency bins. Finally, for each test instance $n \in \{1, ..., N\}$, we obtain gold translation set $S_G(n)$ and proposed translation set $S_P(n)$ to calculate evaluation metrics, including precision P, recall R, F_1 score, and accuracy A.

$$\begin{aligned}
P &= \frac{\sum_{n=1}^{N} |S_G(n) \cap S_P(n)|}{\sum_{n=1}^{N} |S_P(n)|} \\
R &= \frac{\sum_{n=1}^{N} |S_G(n) \cap S_P(n)|}{\sum_{n=1}^{N} |S_G(n)|} \\
F_1 &= \frac{2PR}{P + R} \\
A &= \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left[S_G(n) \cap S_P(n) \neq \emptyset\right]
\end{aligned}$$

We use K to denote the (average) number of translations a method outputs for a source word, i.e. $K \equiv \frac{1}{N} \sum_{n=1}^{N} |S_P(n)|$. For $K = 1$, accuracy will equal precision, and correlate with recall.

Previous works typically report accuracy only, but precision and recall are also important for $K > 1$, especially when we are interested in finding multiple translations of one source word.

Results We first report overall performance when $M = V_s$. For $K = 1$, we only report recall and accuracy, as shown in Table 2. Our approach substantially improves on NN by more than 21 accuracy points. GC also improves on NN, with more additional source vectors bringing slight benefit as expected, but still lags far behind our approach.

Next we allow more than one translation proposals per source word ($K > 1$). Unlike NN rigidly retrieving K target words for each test instance, our approach automatically determines the number of proposals. On average, each source word gets 3.8 translations, so we set $K = 3$ and 4 for NN to approach fairness. From Table 3, we observe increased accuracy over $K = 1$ as expected, and our approach still wins

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

# answers	1	2	3	4	5	6
# entries	1488	603	122	49	3	1
Average # proposals	3.3	4.4	5.4	5.3	3.0	11

Table 4: Average number of Italian proposals by the EMD for English entries having different number of answers in the lexicon.

	K	P	R	F_1	A
NN	3	4.34	9.00	5.86	12.80
NN	4	3.66	10.13	5.38	14.34
EMD	3.8	9.99	25.97	14.43	35.22

Table 3: Overall performance for English-Italian word translation when more than one translation is allowed ($K > 1$). Our approach outputs an average number of 3.8 proposals, so we list NN at $K = 3$ and 4.

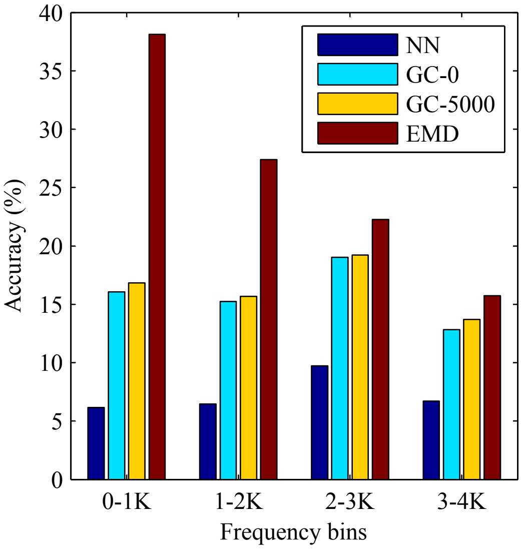

Figure 2: English-Italian accuracy over different frequency bins when each approach only proposes one translation ($K = 1$). The EMD wins by the largest gap for the commonest words (0-1K).

over NN by a large gap. Enlarging K naturally brings up recall at the cost of precision, but our approach manages to surpass both the recall of NN with $K = 4$ and the precision of NN with $K = 3$, even though this comparison offers NN an advantage.

To gain further insight into the behavior of the tested approaches, we split our test set into four frequency bins (Table 1) and report accuracy on them at $K = 1$ in Figure 2. Neither NN nor GC vary much across the four bins. It is clear that our approach outperforms them across all frequency bins, and the gain is more pronounced for the commoner words.

Finally, we are particularly interested in how the EMD deals with multiple translations. To analyze this trait, we make another split of our test set based on the number of answers provided by the lexicon, and then count the proposals for each split. As listed in Table 4, we observe a matching trend between the number of answers and the average number of proposals, unless entries having that many answers become too few to trust the statistics. It is surprising that the EMD manages to adaptively determine the length of its proposal list by mere word occurrence statistics and unsupervised embeddings.

	K	P	R	F_1	A
NN	1	26.14	11.47	15.94	26.14
GC-0	1	25.71	11.28	15.68	25.71
GC-5000	1	25.84	11.34	15.76	25.84
EMD	1	29.07	12.76	17.74	29.07
NN	2	17.85	15.66	16.68	34.53
NN	3	13.54	17.83	15.39	38.30
EMD	2.9	14.21	18.32	16.01	39.46

Table 5: Overall performance for Chinese-English word translation.

Chinese-English Translation Embeddings for this task are trained on a Chinese-English parallel corpus comprising 1.23M sentence pairs with 32.1M Chinese words and 35.4M English words. To evaluate the quality of translation, we use Chinese-English Translation Lexicon Version 3.0[2] as the gold standard. This lexicon includes 54170 Chinese headwords, each listed with at least one English translation. All the other settings are the same as the English-Italian case.

In this translation task, our approach returns an average of 2.9 translation proposals per word, so we list $K = 2$ and 3 for NN for reference. The results are summarized in Table 5. NN appears more competent for this language pair, while GC fails to improve on it. But our approach still manages to improve. For $K = 1$, we obtain an accuracy gain of roughly 3 points. For $K > 1$, our approach achieves higher recall even when compared to NN with $K = 3$ in adverse condition, highlighting its superiority in managing multiple translations.

Bilingual Corpus Filtering

This task aims to filter out non-parallel sentence pairs in a bilingual corpus because the noise can bring a detrimental effect to the translation system. We prepare a noisy corpus by introducing artificial noise into our parallel data. First, we remove duplicates from our Chinese-English corpus and split it into three parts: a portion of 100K sentence pairs to be made noisy, another 100K clean portion, and the remaining 842K data for training our bilingual embeddings.

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

Filtering scheme	A	Dev	MT 02	MT 03	MT 04	MT 05	MT 08	Average
None	N/A	26.69	29.85	28.56	29.44	27.18	20.77	26.93
Oracle	100	27.63	30.21	29.89	30.40	28.67	21.90	27.99
IBM 1	64.41	26.42	29.39	28.60	29.61	27.09	20.54	26.84
Ours	83.07	27.51	30.38	29.34	30.60	28.63	21.86	27.97
Ours w/o IDF	75.50	27.12	29.82	29.03	29.66	27.97	21.29	27.32

Table 6: Accuracy of filtering a bilingual corpus under noise level 0.2, and BLEU of the resulting system trained on filtered corpora. The "Average" column contains the average BLEU of test sets weighted by their sentence numbers. Row-wise, "None" means using the complete 200K noisy data for training the machine translation system, and "Oracle" means using the 100K clean portion for training, which represents an upper bound a filtering method could achieve. We also report a version of our approach without IDF weighting. Paired bootstrap resampling (Koehn 2004) reveals that the BLEU of "Ours" is significantly better than "None" and "IBM 1" on all sets, and "Ours w/o IDF" on all sets except MT 03, all with significance level $p < 0.01$.

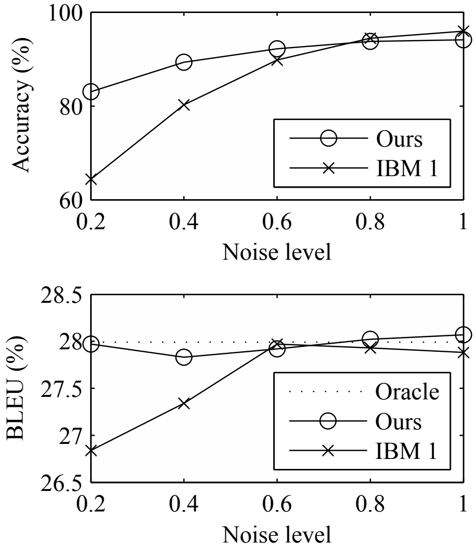

Figure 3: Accuracy and the resulting average BLEU of our approach and the baseline for filtering various levels of noise-corrupted corpus. Lower levels of noise more closely mimic reality but are harder to identify, for which our approach predominates over the baseline.

Then for each English sentence in the first portion, we randomly select $\lceil l \times n \rceil$ words, where l is the sentence length and $n \in [0, 1]$ is the noise level, and replace each of them with a random word in the vocabulary. Finally, the distorted sentences are mixed with the clean portion to obtain a 200K test set with half of them being noisy. A filter is asked to take out half of the test set with as many parallel sentences as possible. For this specific testing scenario, all the above evaluation metrics coincide.

Since our approach offers a measure of distance between a bilingual sentence pair, we use it to sort the test set in ascending order and take the first half to return. We choose as baseline a similar filter based on length-normalized log probabilities, which provides a score to sort in descending order (Khadivi and Ney 2005):

$$\text{Score}\left(f_1^J, e_1^I\right) = \frac{1}{J} \log p\left(f_1^J | e_1^I\right) + \frac{1}{I} \log p\left(e_1^I | f_1^J\right).$$

We use IBM model 1 (Brown et al. 1993) to estimate probabilities for both directions, which avoids the necessity of approximation in (Khadivi and Ney 2005).

In addition to reporting accuracy for identifying noise, we also test the quality of the filtered bilingual corpus by feeding it to the phrase-based translation system Moses (Koehn et al. 2007) as training data. We use NIST 2006 MT Chinese-English data set as the development set. Testing is performed on NIST 2002-2005, 2008 MT data sets, and evaluated with case-insensitive BLEU-4 score (Papineni et al. 2002).

We expect realistic corpora to contain relatively low level of noise, so we present our results under noise level 0.2 in Table 6. Although IBM model 1 identifies noise better than chance, it fails to improve on the unfiltered for machine translation, which means the reduction in noise does not counter the loss of parallel data. In contrast, our approach offers substantially better accuracy. As the effect of noise reduction begins to outweigh, our approach obtains better translation quality, surpassing the IBM baseline with 1.13 BLEU points on average, and nearing the oracle BLEU. We also test the version of our approach with IDF weighting disabled, retreating to a scheme similar to (Kusner et al. 2015). We observe a marked degradation in both accuracy and BLEU, and conclude that IDF weighting plays an indispensable role.

We also relax the difficulty of the task by varying the noise level. Intuitively, lower levels of noise constitute harder tasks, as noisy sentence pairs would be more similar to bitext and thus more difficult to tell apart. As charted in Figure 3, the accuracy trend matches our intuition, and with 0.6 or higher noise level, both methods achieve reasonably good performance. We also observe that accuracy over 80% basically leads to BLEU around oracle, and since our approach invariably attains that position regardless of noise level, its BLEU never falls far behind the oracle, while the baseline only catches up until the task becomes rather easy with sufficiently high level of noise.

Conclusion

Motivated by the shortcomings of using nearest neighbor for word translation with bilingual embeddings, we reformulate the task by introducing the Earth Mover's Distance. With the mere aid of occurrence statistics, it pairs up word vectors from distinct languages with substantially better performance, and naturally handles multiple translations. We further extend the vocabulary-level matching to sentence-level to arrive at a bilingual sentence distance metric that draws on information encoded in bilingual word embeddings. We apply the distance metric to a bilingual corpus filtering task, on which bilingual word embeddings come into play for the first time, and observe encouraging performance in terms of both filtering accuracy and resultant translation quality.

Acknowledgments

Yang Liu and Maosong Sun are supported by National Key Basic Research Program of China (973 Program 2014CB340500), the National Natural Science Foundation of China (No. 61522204 and No. 61432013), and Toshiba Corporation Corporate Research & Development Center. Huanbo Luan is supported by the National Natural Science Foundation of China (No. 61303075). This research is also supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme.

References

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

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-14*, 1853–1861.

Dinu, G.; Lazaridou, A.; and Baroni, M. 2014. Improving Zero-Shot Learning by Mitigating the Hubness Problem. In *arXiv:1412.6568 [cs]*.

Gouws, S.; Bengio, Y.; and Corrado, G. 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In *Proceedings of ICML-15*, 748–756.

Haghighi, A.; Liang, P.; Berg-Kirkpatrick, T.; and Klein, D. 2008. Learning Bilingual Lexicons from Monolingual Corpora. In *Proceedings of ACL-08: HLT*, 771–779.

Hermann, K. M., and Blunsom, P. 2014. Multilingual Distributed Representations without Word Alignment. In *Proceedings of ICLR-14*.

Khadivi, S., and Ney, H. 2005. Automatic Filtering of Bilingual Corpora for Statistical Machine Translation. In *Proceedings of NLDB-05*, 263–274.

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 ACL on Interactive Poster and Demonstration Sessions*.

Koehn, P. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP-04*.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT summit*, volume 5, 79–86. Citeseer.

Kočiský, T.; Hermann, K. M.; and Blunsom, P. 2014. Learning Bilingual Word Representations by Marginalizing Alignments. In *Proceedings of the 52nd Annual Meeting of the ACL*, volume 2, 224–229.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From Word Embeddings To Document Distances. In *Proceedings of ICML-15*, 957–966.

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*, 151–159.

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

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS-13*, 3111–3119.

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

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of NAACL-13: HLT*, 746–751.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the ACL*.

Pele, O., and Werman, M. 2009. Fast and Robust Earth Mover's Distances. In *Proceedings of ICCV-09*, 460–467.

Ren, Z.; Yuan, J.; and Zhang, Z. 2011. Robust Hand Gesture Recognition Based on Finger-earth Mover's Distance with a Commodity Depth Camera. In *Proceedings of MM-11*, 1093–1096.

Rubner, Y.; Tomasi, C.; and Guibas, L. 1998. A Metric for Distributions with Applications to Image Databases. In *Proceedings of ICCV-98*, 59–66.

Soyer, H.; Stenetorp, P.; and Aizawa, A. 2015. Leveraging Monolingual Data for Crosslingual Compositional Word Representations. In *Proceedings of ICLR-15*.

Taghipour, K.; Afhami, N.; Khadivi, S.; and Shiry, S. 2010. A Discriminative Approach to Filter out Noisy Sentence Pairs from Bilingual Corpora. In *IST-10*, 537–541.

Vulić, I., and Moens, M.-F. 2013. A Study on Bootstrapping Bilingual Vector Spaces from Non-Parallel Data (and Nothing Else). In *Proceedings of EMNLP-13*, 1613–1624.

Vulić, I., and Moens, M.-F. 2015. Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction. In *Proceedings of the 53rd Annual Meeting of the ACL and the 7th IJCNLP*, volume 2, 719–725.

Zhang, J.; Liu, S.; Li, M.; Zhou, M.; and Zong, C. 2014. Bilingually-constrained Phrase Embeddings for Machine Translation. In *Proceedings of the 52nd Annual Meeting of the ACL*, volume 1, 111–121.

Zou, W. Y.; Socher, R.; Cer, D.; and Manning, C. D. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *Proceedings of EMNLP-13*, 1393–1398.