# Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation

**Chao Xing**
CSLT, Tsinghua University
Beijing Jiaotong University
Beijing, P.R. China

**Dong Wang***
CSLT, RIIT, Tsinghua University
TNList, China
Beijing, P.R. China

**Chao Liu**
CSLT, RIIT, Tsinghua University
CS Department, Tsinghua University
Beijing, P.R. China

**Yiye Lin**
CSLT, RIIT, Tsinghua University
Beijing Institute of Technology
Beijing, P.R. China

## Abstract

Word embedding has been found to be highly powerful to translate words from one language to another by a simple linear transform. However, we found some inconsistence among the objective functions of the embedding and the transform learning, as well as the distance measurement. This paper proposes a solution which normalizes the word vectors on a hypersphere and constrains the linear transform as an orthogonal transform. The experimental results confirmed that the proposed solution can offer better performance on a word similarity task and an English-to-Spanish word translation task.

## 1 Introduction

Word embedding has been extensively studied in recent years (Bengio et al., 2003; Turian et al., 2010; Collobert et al., 2011; Huang et al., 2012). Following the idea that the meaning of a word can be determined by 'the company it keeps' (Baroni and Zamparelli, 2010), i.e., the words that it co-occurs with, word embedding projects discrete words to a low-dimensional and continuous vector space where co-occurred words are located close to each other. Compared to conventional discrete representations (e.g., the one-hot encoding), word embedding provides more robust representations for words, particulary for those that infrequently appear in the training data. More importantly, the embedding encodes syntactic and semantic content implicitly, so that relations among words can be simply computed as the distances among their embeddings, or word vectors. A well-known efficient word embedding approach was recently proposed by (Mikolov et al., 2013a), where two log-linear models (CBOW and skip-gram) are proposed to learn the neighboring relation of words in context. A following work proposed by the same authors introduces some modifications that largely improve the efficiency of model training (Mikolov et al., 2013c).

An interesting property of word vectors learned by the log-linear model is that the relations among relevant words seem linear and can be computed by simple vector addition and substraction (Mikolov et al., 2013d). For example, the following relation approximately holds in the word vector space: Paris - France + Rome = Italy. In (Mikolov et al., 2013b), the linear relation is extended to the bilingual scenario, where a linear transform is learned to project semantically identical words from one language to another. The authors reported a high accuracy on a bilingual word translation task.

Although promising, we argue that both the word embedding and the linear transform are ill-posed, due to the inconsistence among the objective function used to learn the word vectors (maximum likelihood based on inner product), the distance measurement for word vectors (cosine distance), and the objective function used to learn the linear transform (mean square error). This inconsistence may lead to



*Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 1006–1011,
Denver, Colorado, May 31 – June 5, 2015. 

suboptimal estimation for both word vectors and the bilingual transform, as we will see shortly.

This paper solves the inconsistence by normalizing the word vectors. Specifically, we enforce the word vectors to be in a unit length during the learning of the embedding. By this constraint, all the word vectors are located on a hypersphere and so the inner product falls back to the cosine distance. This hence solves the inconsistence between the embedding and the distance measurement. To respect the normalization constraint on word vectors, the linear transform in the bilingual projection has to be constrained as an orthogonal transform. Finally, the cosine distance is used when we train the orthogonal transform, in order to achieve full consistence.

## 2 Related work

This work largely follows the methodology and experimental settings of (Mikolov et al., 2013b), while we normalize the embedding and use an orthogonal transform to conduct bilingual translation.

Multilingual learning can be categorized into projection-based approaches and regularization-based approaches. In the projection-based approaches, the embedding is performed for each language individually with monolingual data, and then one or several projections are learned using multilingual data to represent the relation between languages. Our method in this paper and the linear projection method in (Mikolov et al., 2013b) both belong to this category. Another interesting work proposed by (Faruqui and Dyer, 2014) learns linear transforms that project word vectors of all languages to a common low-dimensional space, where the correlation of the multilingual word pairs is maximized with the canonical correlation analysis (CCA).

The regularization-based approaches involve the multilingual constraint in the objective function for learning the embedding. For example, (Zou et al., 2013) adds an extra term that reflects the distances of some pairs of semantically related words from different languages into the objective funtion. A similar approach is proposed in (Klementiev et al., 2012), which casts multilingual learning as a multitask learning and encodes the multilingual information in the interaction matrix.

All the above methods rely on a multilingual lexicon or a word/pharse alignment, usually from a machine translation (MT) system. (Blunsom et al., 2014) proposed a novel approach based on a joint optimization method for word alignments and the embedding. A simplified version of this approach is proposed in (Hermann and Blunsom, 2014), where a sentence is represented by the mean vector of the words involved. Multilingual learning is then reduced to maximizing the overall distance of the parallel sentences in the training corpus, with the distance computed upon the sentence vectors.

## 3 Normalized word vectors

Taking the skip-gram model, the goal is to predict the context words with a word in the central position. Mathematically, the training process maximizes the following likelihood function with a word sequence $w_1, w_2 ... w_N$:

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{-C\leq j\leq C, j\neq 0} logP(w_{i+j}|w_i) \qquad (1)$$

where $C$ is the length of the context in concern, and the prediction probability is given by:

$$P(w_{i+j}|w_i) = \frac{\exp(c_{w_{i+j}}^T c_{w_i})}{\sum_w \exp(c_w^T c_{w_i})} \qquad (2)$$

where $w$ is any word in the vocabulary, and $c_w$ denotes the vector of word $w$. Obviously, the word vectors learned by this way are not constrained and disperse in the entire $M$-dimensional space, where $M$ is the dimension of the word vectors. An inconsistence with this model is that the distance measurement in the training is the inner product $c_w^T c_{w'}$, however when word vectors are applied, e.g., to estimate word similarities, the metric is often the cosine distance $\frac{c_w^T c_{w'}}{||c_w|| ||c_{w'}||}$. A way to solve this consistence is to use the inner product in applications, however using the cosine distance is a convention in natural language processing (NLP) and this measure does

show better performance than the inner product in our experiments.

We therefore perform in an opposite way, i.e., enforcing the word vectors to be unit in length. Theoretically, this changes the learning of the embedding to an optimization problem with a quadratic constraint. Solving this problem by Lagrange multipliers is possible, but here we simply divide a vector by its $l$-2 norm whenever the vector is updated. This does not involve much code change and is efficient enough.[1]

The consequence of the normalization is that all the word vectors are located on a hypersphere, as illustrated in Figure 1. In addition, by the normalization, the inner product falls back to the cosine distance, hence solving the inconsistence between the embedding learning and the distance measurement.

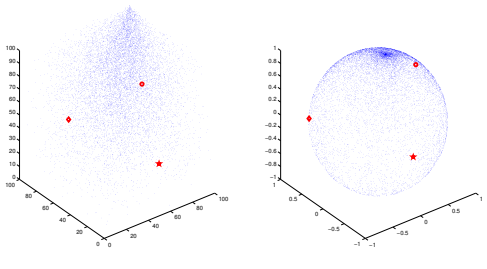

Figure 1: The distributions of unnormalized (left) and normalized (right) word vectors. The red circles/stars/diamonds represent three words that are embedded in the two vector spaces respectively.

## 4 Orthogonal transform

The bilingual word translation provided by (Mikolov et al., 2013b) learns a linear transform from the source language to the target language by the linear regression. The objective function is as follows:

$$\min_W \sum_i ||Wx_i - z_i||^2 \tag{3}$$

[1]For efficiency, this normalization can be conducted every $n$ mini-batches. The performance is expected to be not much impacted, given that $n$ is not too large.

where $W$ is the projection matrix to be learned, and $x_i$ and $z_i$ are word vectors in the source and target language respectively. The bilingual pair $(x_i, z_i)$ indicates that $x_i$ and $z_i$ are identical in semantic meaning. A high accuracy was reported on a word translation task, where a word projected to the vector space of the target language is expected to be as close as possible to its translation (Mikolov et al., 2013b). However, we note that the 'closeness' of words in the projection space is measured by the cosine distance, which is fundamentally different from the Euler distance in the objective function (3) and hence causes inconsistence.

We solve this problem by using the cosine distance in the transform learning, so the optimization task can be redefined as follows:

$$\max_W \sum_i (Wx_i)^T z_i. \tag{4}$$

Note that the word vectors in both the source and target vector spaces are normalized, so the inner product in (4) is equivalent to the cosine distance. A problem of this change, however, is that the projected vector $Wx_i$ has to be normalized, which is not guaranteed so far.

To solve the problem, we first consider the case where the dimensions of the source and target vector spaces are the same. In this case, the normalization constraint on word vectors can be satisfied by constraining $W$ as an orthogonal matrix, which turns the unconstrained problem (4) to a constrained optimization problem. A general solver such as SQP can be used to solve the problem. However, we seek a simple approximation in this work. Firstly, solve (4) by gradient descendant without considering any constraint. A simple calculation shows that the gradient is as follows:

$$\bigtriangledown_W = \sum_i x_i y_i^T, \tag{5}$$

and the update rule is simply given by:

$$W = W + \alpha \bigtriangledown_W \tag{6}$$

where $\alpha$ is the learning rate. After the update, $W$ is orthogonalized by solving the following constrained quadratic problem:

$$\min_{\bar{W}} ||W - \bar{W}|| \quad s.t. \ \bar{W}^T\bar{W} = I. \qquad (7)$$

One can show that this problem can be solved by taking the singular value decomposition (SVD) of $W$ and replacing the singular values to ones.

For the case where the dimensions of the source and target vector spaces are different, the normalization constraint upon the projected vectors is not easy to satisfy. We choose a pragmatic solution. First, we extend the low-dimensional vector space by padding a small tunable constant at the end of the word vectors so that the source and target vector spaces are in the same dimension. The vectors are then renormalized after the padding to respect the normalization constraint. Once this is done, the same gradient descendant and orthognalization approaches are ready to use to learn the orthogonal transform.

## 5 Experiment

We first present the data profile and configurations used to learn monolingual word vectors, and then examine the learning quality on the word similarity task. Finally, a comparative study is reported on the bilingual word translation task, with Mikolov's linear transform and the orthogonal transform proposed in this paper.

### 5.1 Monolingual word embedding

The monolingual word embedding is conducted with the data published by the EMNLP 2011 SMT workshop (WMT11)[2]. For an easy comparison, we largely follow Mikolov's settings in (Mikolov et al., 2013b) and set English and Spanish as the source and target language, respectively. The data preparation involves the following steps. Firstly, the text was tokenized by the standard scripts provided by WMT11[3], and then duplicated sentences were removed. The numerical expressions were tokenized as 'NUM', and special characters (such as !?,:) were removed.

The word2vector toolkit[4] was used to train the word embedding model. We chose the skip-gram model and the text window was set to 5. The training resulted in embedding of 169k English tokens and 116k Spanish tokens.

### 5.2 Monolingual word similarity

The first experiment examines the quality of the learned word vectors in English. We choose the word similarity task, which tests to what extent the word similarity computed based on word vectors agrees with human judgement. The WordSimilarity-353 Test Collection[5] provided by (Finkelstein et al., 2002) is used. The dataset involves 154 word pairs whose similarities are measured by 13 people and the mean values are used as the human judgement. In the experiment, the correlation between the cosine distances computed based on the word vectors and the humane-judged similarity is used to measure the quality of the embedding. The results are shown in Figure 2, where the dimension of the vector space varies from 300 to 1000. It can be observed that the normalized word vectors offer a high correlation with human judgement than the unnormalized counterparts.

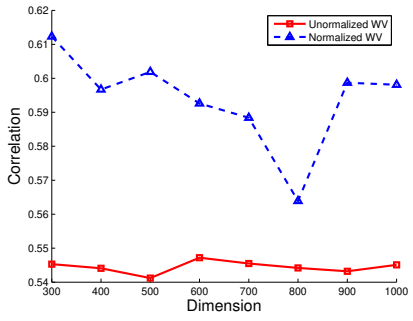


Figure 2: Results on the word similarity task with the normalized and unnormalized word vectors. A higher correlation indicates better quality.

[2]http://www.statmt.org/wmt11/

[3]http://www.statmt.org

[4]https://code.google.com/p/word2vec

[5]http://www.cs.technion.ac.il/ gabr/resources/data/wordsim353/

### 5.3 Bilingual word translation

The second experiment focuses on bilingual word translation. We select 6000 frequent words in English and employ the online Google's translation service to translate them to Spanish. The resulting 6000 English-Spanish word pairs are used to train and test the bilingual transform in the way of cross validation. Specifically, the 6000 pairs are randomly divided into 10 subsets, and at each time, 9 subsets are used for training and the rest 1 subset for testing. The average of the results of the 10 tests is reported as the final result. Note that not all the words translated by Google are in the vocabulary of the target language; the vocabulary coverage is 99.5% in our test.

#### 5.3.1 Results with linear transform

We first reproduce Mikolov's work with the linear transform. A number of dimension settings are experimented with and the results are reported in Table 1. The proportions that the correct translations are in the top 1 and top 5 candidate list are reported as P@1 and P@5 respectively. As can be seen, the best dimension setting is 800 for English and 200 for Spanish, and the corresponding P@1 and P@5 are 35.36% and 53.96%, respectively. These results are comparable with the results reported in (Mikolov et al., 2013b).

| D-EN | D-ES | P@1 | P@5 |
|---|---|---|---|
| 300 | 300 | 30.43% | 49.43% |
| 500 | 500 | 25.76% | 44.29% |
| 700 | 700 | 20.69% | 39.12% |
| 800 | 200 | 35.36% | 53.96% |

Table 1: Performance on word translation with unnormalized embedding and linear transform. 'D-EN' and 'D-ES' denote the dimensions of the English and Spanish vector spaces, respectively.

#### 5.3.2 Results with orthogonal transform

The results with the normalized word vectors and the orthogonal transform are reported in Table 2. It can be seen that the results with the orthogonal transform are consistently better than those reported in Table1 which are based on the linear transform. This confirms our conjecture that bilingual translation can be largely improved by the normalized embedding and the accompanied orthogonal transform.

| D-EN | D-ES | P@1 | P@5 |
|---|---|---|---|
| 300 | 300 | 38.99% | 59.16% |
| 500 | 500 | 39.91% | 59.82% |
| 700 | 700 | 41.04% | 59.38% |
| 800 | 200 | 40.06% | 60.02% |

Table 2: Performance on word translation with normalized embedding and orthogonal transform. 'D-EN' and 'D-ES' denote the dimensions of the English and Spanish vector spaces, respectively.

## 6 Conclusions

We proposed an orthogonal transform based on normalized word vectors for bilingual word translation. This approach solves the inherent inconsistence in the original approach based on unnormalized word vectors and a linear transform. The experimental results on a monolingual word similarity task and an English-to-Spanish word translation task show clear advantage of the proposal. This work, however, is still preliminary. It is unknown if the normalized embedding works on other tasks such as relation prediction, although we expect so. The solution to the orthogonal transform between vector spaces with mismatched dimensions is rather ad-hoc. Nevertheless, locating word vectors on a hypersphere opens a door to study the properties of the word embedding in a space that is yet less known to us.

## Acknowledgement

This work was conducted when CX & YYL were visiting students in CSLT, Tsinghua University. This research was supported by the National Science Foundation of China (NSFC) under the project No. 61371136, and the MESTDC PhD Foundation Project No. 20130002120011. It was also supported by Sinovoice and Huilan Ltd.

## References


Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

Phil Blunsom, Karl Moritz Hermann, Tomas Kocisky, et al. 2014. Learning bilingual word representations by marginalizing alignments. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 224–229.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. *Proceeding of EACL. Association for Computational Linguistics*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. In *Proceedings of The 2002 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 116–131.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual distributed representations without word alignment. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceeding of Annual Meeting of the Association for Computational Linguistics (ACL)*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING*. Citeseer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (ICLR)*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013d. Linguistic regularities in continuous space word representations. In *Proceedings of The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751. Citeseer.

Joseph Turian, Departement d'Informatique Et, Recherche Operationnelle (diro, Universite De Montreal, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semisupervised learning. In *Proceeding of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 384–394.

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