

# Learning Cross-lingual Word Embeddings via Matrix Co-factorization

Tianze Shi    Zhiyuan Liu    Yang Liu    Maosong Sun

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 100084, China

stz11@mails.tsinghua.edu.cn

{liuzy, liuyang2011, sms}@tsinghua.edu.cn

## Abstract

A joint-space model for cross-lingual distributed representations generalizes language-invariant semantic features. In this paper, we present a matrix co-factorization framework for learning cross-lingual word embeddings. We explicitly define monolingual training objectives in the form of matrix decomposition, and induce cross-lingual constraints for simultaneously factorizing monolingual matrices. The cross-lingual constraints can be derived from parallel corpora, with or without word alignments. Empirical results on a task of cross-lingual document classification show that our method is effective to encode cross-lingual knowledge as constraints for cross-lingual word embeddings.

## 1 Introduction

Word embeddings allow one to represent words in a continuous vector space, which characterizes the lexico-semantic relations among words. In many NLP tasks, they prove to be high-quality features, successful applications of which include language modelling (Bengio et al., 2003), sentiment analysis (Socher et al., 2011) and word sense discrimination (Huang et al., 2012).

Like words having synonyms in the same language, there are also word pairs across languages which share resembling semantic properties. Mikolov et al. (2013a) observed a strong similarity of the geometric arrangements of corresponding concepts between the vector spaces of different languages, and suggested that a cross-lingual mapping between the two vector spaces is technically plausible. In the meantime, the joint-space models for cross-lingual word embeddings are very desirable, as language-invariant semantic features can be generalized to make it easy

to transfer models across languages. This is especially important for those low-resource languages, where it allows one to develop accurate word representations of one language by exploiting the abundant textual resources in another language, e.g., English, which has a high resource density. Further, studies have shown that using cross-lingual correlation can improve the quality of word representations trained solely with monolingual corpora (Faruqui and Dyer, 2014).

Defining a cross-lingual learning objective is crucial at the core of the joint-space model. A sentence-aligned parallel bilingual corpus is a common starting point, where identical meanings are expressed in multiple languages. Hermann and Blunsom (2014) and Chandar A P et al. (2014) tried to calculate parallel sentence (or document) representations and to minimize the differences between the semantically equivalent pairs. These methods are useful in capturing semantic information carried by high-level units (such as phrases and beyond) and usually do not rely on word alignments. However, they suffer from reduced accuracy for representing rare tokens, whose semantic information may not be well generalized. In these cases, finer-grained information at lexical level, such as aligned word pairs, dictionaries, and word translation probabilities, is considered to be helpful.

Kočiský et al. (2014) integrated word aligning process and word embedding in machine translation models. This method makes full use of parallel corpora and produces high-quality word alignments. However, it is unable to exploit the richer monolingual corpora. On the other hand, Zou et al. (2013) and Faruqui and Dyer (2014) learnt word embeddings of different languages in separate spaces with monolingual corpora and projected the embeddings into a joint space, but they can only capture linear transformation. Of course one can manually design a non-linear projection function,

but the justification for the choice is not apparent.

In this paper, we address the above challenges with a framework of matrix co-factorization. We simultaneously learn word embeddings in multiple languages via matrix factorization, with induced constraints to assure cross-lingual semantic relations. It provides the flexibility of constructing learning objectives from separate monolingual and cross-lingual corpora. Intricate relations across languages, rather than simple linear projections, are automatically captured. Additionally, our method is efficient as it learns from global statistics. The cross-lingual constraints can be derived both with or without word alignments, given that there is a valid measure of cross-lingual co-occurrences or similarities.

We evaluate the bilingual embeddings induced with our method on the task of English-German cross-lingual document classification, where a classifier is trained with the embeddings in one language and tested on documents in another language. Empirical results demonstrate the validity of our model.

## 2 Framework

Without loss of generality, here we only consider bilingual embedding learning of the two languages  $l_1$  and  $l_2$ . Given monolingual corpora  $D^{l_i}$  and sentence-aligned parallel data  $D^{bi}$ , our task is to find a word embedding of dimensionality  $d$  for each word in the vocabularies  $V^{l_i}$ . These vector representations can be written in the form of matrices  $W^{l_i}$  of the size  $|V^{l_i}| \times d$  where each line corresponds to the embedding of a single word. Since we learn from the occurrences of words in their contexts, we also define vocabularies of contexts  $U^{l_i}$  and we learn context embedding matrices  $C^{l_i}$  of the size  $|U^{l_i}| \times d$  at the same time.<sup>1</sup>

These matrices are obtained by simultaneous matrix factorization of the monolingual word-context PMI (point-wise mutual information) matrices  $M^{l_i}$ . During monolingual factorization, we put a cross-lingual constraint (cost) on it, ensuring cross-lingual semantic relations. We formalize the global loss function as

$$L_{\text{total}} = \sum_{i \in \{1,2\}} \omega_i \cdot L_{\text{mono}}(W^{l_i}, C^{l_i}) + \omega_c \cdot L_{\text{cross}}(W^{l_1}, C^{l_1}, W^{l_2}, C^{l_2}), \quad (1)$$

where  $L_{\text{mono}}$  and  $L_{\text{cross}}$  are the monolingual and cross-lingual objectives respectively. They can be

<sup>1</sup>In this paper, we let  $U^{l_i} = V^{l_i}$ .

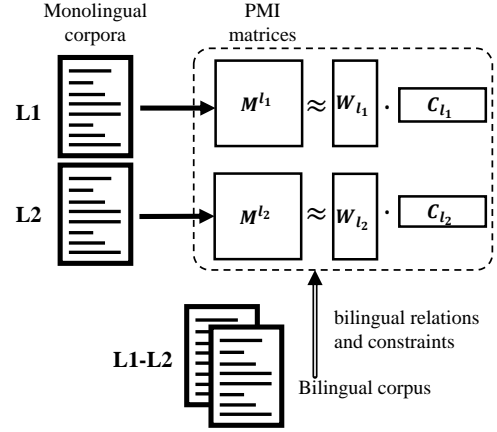


Figure 1: The framework of cross-lingual word embedding via matrix co-factorization.

derived from different corpora, i.e.  $D^{l_i}$  need not to be part of  $D^{bi}$ , which allows one to freely exploit the richer monolingual resources.  $\omega_i$  and  $\omega_c$  weigh the contribution of the different parts to the total objective. An overview of our algorithm is illustrated in Figure 1. In the following two sections, we explain the monolingual objective and choices of cross-lingual objective in detail.

## 3 Monolingual Objective

Our monolingual objective follows the GloVe model (Pennington et al., 2014), which learns from global word co-occurrence statistics. For a word-context pair  $(j, k)$  in language  $l_i$  which co-occurs more frequently than if they are independently distributed,  $j$  and  $k$  are assumed to be semantically correlated. Since PMI is a well-established measure of the association between two words (Levy and Goldberg, 2014), we try to minimize the difference between the dot product of the embeddings  $w_j^{l_i} \cdot c_k^{l_i}$  and their PMI value

$$M_{jk}^{l_i} = \frac{X_{jk}^{l_i} \cdot \sum_{j,k} X_{jk}^{l_i}}{\sum_j X_{jk}^{l_i} \cdot \sum_k X_{jk}^{l_i}}, \quad \text{where } X^{l_i} \text{ is the matrix of word-context co-occurrence counts. As Pennington et al. (2014), we add separate terms } b_{w_j}^{l_i}, b_{c_k}^{l_i} \text{ for each word and context to absorb the effect of any possible word-specific biases. We also add an additional matrix bias } b^{l_i} \text{ for the ease of sharing embeddings among matrices. The loss function is written as the sum of the weighted square error,}$$

$$L_{\text{mono}}^{l_i} = \sum_{j,k} f(X_{jk}^{l_i}) \left( w_j^{l_i} \cdot c_k^{l_i} + b_{w_j}^{l_i} + b_{c_k}^{l_i} + b^{l_i} - M_{jk}^{l_i} \right)^2, \quad (2)$$

where we choose the same weighting function as the GloVe model to place less confidence on those word-context pairs with rare occurrences,

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}. \quad (3)$$

Notice that we only have to optimize those  $X_{jk}^{l_i} \neq 0$ , which can be solved efficiently since the matrix of co-occurrence counts is usually sparse.

## 4 Cross-lingual Objectives

As the most important part in our model, the cross-lingual objective describes the cross-lingual word relations and sets constraints when we factorize monolingual co-occurrence matrices. It can be derived from either cross-lingual co-occurrences or similarities between cross-lingual word pairs.

### 4.1 Cross-lingual Contexts

The monolingual objective stems from the distributional hypothesis (Harris, 1954) and optimizes words in similar contexts into similar embeddings. It is natural to further extend this idea to define cross-lingual contexts, for which we have multiple choices.

For the definition of cross-lingual contexts, we have multiple choices. A straightforward option is to count all the word co-occurrences in aligned sentence pairs, which is equivalent to a uniform word alignment model adopted by Gouws et al. (2015). For the sentence-aligned bilingual corpus  $D^{\text{bi}} = \{(S^{l_1}, S^{l_2})\}$ , where each  $S^{l_i}$  is a monolingual sentence, we count the co-occurrences as

$$X_{jk}^{\text{bi}} = \sum_{(S^{l_1}, S^{l_2}) \in D^{\text{bi}}} \#(j, S^{l_1}) \times \#(k, S^{l_2}), \quad (4)$$

where  $X^{\text{bi}}$  is the matrix of cross-lingual co-occurrence counts, and  $\#(j, S)$  is a function counting the number of  $j$ 's in the sequence  $S$ . We then use a similar loss function as Equation 2, with the exception that we optimize for the dot products of  $w_j^{l_1} \cdot w_k^{l_2}$ . This method works without word alignments and we denote it as CLC-WA (Cross-lingual context without word alignments).

To define a cross-lingual context of  $j \in V^{l_1}$  that resembles the monolingual context of its translationally equivalent word  $k \in V^{l_2}$ , we also define CLC+WA (Cross-lingual context with word alignments). The idea is to count those words co-occurring with  $k$  as the context of  $j$ . Given the aligned pairs  $\{(a, b)\} = \text{Align}(S^{l_1}, S^{l_2})$ , where  $a$  and  $b$  are the indices of the aligned pair in the sentences, we count the cross-lingual co-occurrences with a window size of  $2d + 1$  as

$$X_{jk}^{\text{bi}} = \sum_{(S^{l_1}, S^{l_2}) \in D^{\text{bi}}} \sum_{\substack{(a, b) \in \text{Align}(S^{l_1}, S^{l_2}) \\ S_a^{l_1} = j}} \#(k, \langle S_{b-d}^{l_2}, \dots, S_{b+d}^{l_2} \rangle). \quad (5)$$

An example is shown in Figure 2. CLC+WA is expected to contain more precise information than CLC-WA, and we will compare the two definitions in the following experiments.

Once we have counted the co-occurrences, a naïve solution is to concatenate the bilingual vocabularies and perform matrix factorization as a whole. To induce additional flexibility, such as separate weighting, we divide the matrix into three parts. It is also more reasonable to calculate PMI values without mixing the monolingual and bilingual corpora.

### 4.2 Cross-lingual Similarities

An alternative way to set cross-lingual constraints is to minimize the distances between similar word pairs. We generalize the quantitative measure of cross-lingual similarity in a function  $\text{sim}(j, k)$ , where  $j \in V^{l_1}$  and  $k \in V^{l_2}$ . In this paper, we use the translation probabilities produced by a machine translation system. Minimizing the distances of related words in the two languages weighted by their similarities gives us the cross-lingual objective

$$L_{\text{cross}} = \sum_{j \in V^{l_1}, k \in V^{l_2}} \text{sim}(j, k) \cdot \text{distance}(w_j^{l_1}, w_k^{l_2}), \quad (6)$$

where  $w_j^{l_1}$  and  $w_k^{l_2}$  are the embeddings of  $j$  and  $k$  in  $l_1$  and  $l_2$  respectively. As we wish to position embeddings of semantically equivalent words as close as possible in the shared vector space, we choose the distance function to be the Euclidean distance, which rewrites Equation 6 into

$$L_{\text{cross}} = \sum_{j \in V^{l_1}, k \in V^{l_2}} \text{sim}(j, k) \cdot \|w_j^{l_1} - w_k^{l_2}\|^2. \quad (7)$$

Notice that similar to the monolingual objective, we may optimize for only those  $\text{sim}(j, k) \neq 0$ , which is efficient as the matrix of translation probabilities or dictionary is sparse. We call this method CLSim.

## 5 Experiments

The method introduced in this paper allows one to learn cross-lingual word embeddings in a shared vector space, where words with resembling semantic properties are similarly distributed in the space, regardless of their languages.

... we must do all we can, not just to ...  
 ... wir alles daran setzen müssen, nicht nur ...

Figure 2: An example of CLC+WA, where we show the cross-lingual context of the German word “müssen” in the dashed box.

Table 1: Accuracy for cross-lingual classification.

Model	en→de	de→en
Machine translation	68.1	67.4
Majority class	46.8	46.8
Klementiev et al.	77.6	71.1
BiCVM	83.7	71.4
BAE	91.8	74.2
BilBOWA	86.5	75.0
CLC-WA	91.3	77.2
CLC+WA	90.0	75.0
CLSim	<b>92.7</b>	<b>80.2</b>

To evaluate the quality of the relatedness between words in different languages, we induce the task of cross-lingual document classification for the English-German language pair, where a classifier is trained in one language and later used to classify documents in another. We exactly replicated the experiment settings of Klementiev et al. (2012). As they noted, the aim of this task is not to provide a state-of-the-art cross-lingual document classifier, where one should develop a task-specific solution to achieve higher accuracy, but rather to examine the validity of our joint semantic space model.

We then show some interesting cross-lingual relations observed in the semantic space through examples and visualization.

## 5.1 Data and Training

For optimizing the monolingual objectives, We used exactly the same subset of RCV1/RCV2 corpora (Lewis et al., 2004) as by Klementiev et al. (2012), which were sampled to balance the number of tokens between languages. Our preprocessing strategy followed Chandar A P et al. (2014), where we lowercased all words, removed punctuations and used the same vocabularies ( $|V^{\text{en}}| = 43,614$  and  $|V^{\text{de}}| = 50,110$ ). When counting word co-occurrences, we use a decreasing weighting function as Pennington et al. (2014), where  $d$ -word-apart word pairs contribute  $1/d$  to the total count. We used a symmetric window size of 10 words for all our experiments.

The cross-lingual constraints were derived using the English and German sections of the Europarl v7 parallel corpus (Koehn, 2005), which were similarly preprocessed. For CLC+WA and CLSim, we obtained word alignments and translation probabilities with SyMGIZA++ (Junczys-Dowmunt and Szał, 2012), which is slightly different from GIZA++ that it includes a symmetrization mechanism so that we exploit only bi-directional word alignments and translation probabilities. We did not use Europarl for monolingual training.

The documents for classification were randomly selected by Klementiev et al. (2012) from those in RCV1/RCV2 that are assigned to only one single topic among the four: CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), and MCAT (Markets). 1,000/5,000 documents in each language were used as a train/test set and we kept another 1,000 documents as a development set for hyperparameter tuning. Each document was represented as an idf-weighted average embedding of all its tokens, and a multi-class document classifier was trained for 10 epochs with an averaged perceptron algorithm as by Klementiev et al. (2012). A classifier trained with English documents is used to classify German documents and vice versa.

We trained our models using stochastic gradient descent. Non-negative elements (co-occurrences in monolingual objectives and CLC model, and word translation probabilities in CLSim model) are stochastically sampled and learned at a constant learning rate. We run 50 iterations for all of our experiments and the dimensionality of the embeddings is 40. We set  $x_{\text{max}}$  to be 100 for cross-lingual co-occurrences and 30 for monolingual ones, while  $\alpha$  is fixed to  $3/4$ . Other parameters are chosen according to the performance on the development set.

## 5.2 Results

We present the empirical results on the task of cross-lingual document classification in Table 1, where the performance of our models is compared with some baselines and previous work.

The baseline systems are *Majority class* where test documents are simply classified as the class with the most training samples, and *Machine translation* where a phrased-based machine translation system is used to translate test documents

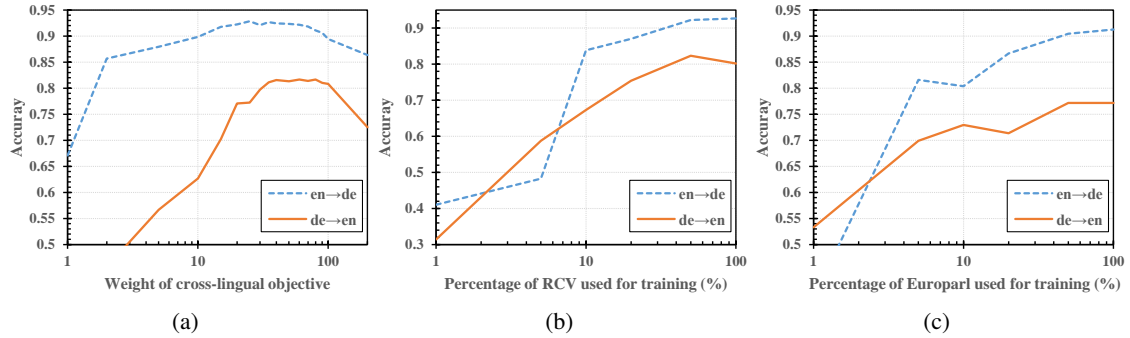


Figure 3: Cross-lingual document classification accuracy, with (a) varying weighting of cross-lingual objective (b) varying size of training monolingual corpora, and (c) varying size of training bilingual corpus.

Table 2: Example English words along with their nearest words in English (EN) and German (DE) measured by Euclidean distance.

Word	Lang	Nearest neighbors
january	EN	january, october, february
	DE	jänner, januar, april, märz
oil	EN	oil, crude, heating, fuel
	DE	oil, erdöl, rohöl, rohölpreise
microsoft	EN	microsoft, intel, ibm, exxon
	DE	microsoft, intel, ibm, exxon
selloff	EN	selloff, buoyed, rebound, slide
	DE	verlusten, kursaufschwung
nov	EN	nov, oct, aug, dec, jul, sep
	DE	dec, dez, sept, okt, jan, mar
ppi	EN	ppi, output, producer
	DE	industrieproduktion

into the same language as the training documents. This system was trained with the Europarl corpus, which is also the corpus we used to induce cross-lingual constraints.

We also summarize the classification accuracy reported in some previous work. The embeddings of each word is treated as a single task in *Multitask learning* (Klementiev et al., 2012) while cross-lingual relatedness information is encoded in an interaction matrix. *Bilingual compositional vector model (BiCVM)* (Hermann and Blunsom, 2014) learns to align the sentence embeddings of equivalent pairs. *Bilingual autoencoder for bags-of-words (BAE)* (Chandar A P et al., 2014) reconstructs the bag-of-words representations of equivalent sentence pairs. *BilBOWA* (Gouws et al., 2015) extends CBOW and skipgram models with a sampled cross-lingual loss between bag-of-words sentence vectors. BilBOWA is most related to CLC-WA that it also assumes a uniform word alignment model. A more recent work of Soyer et al. (2015) developed a compositional approach

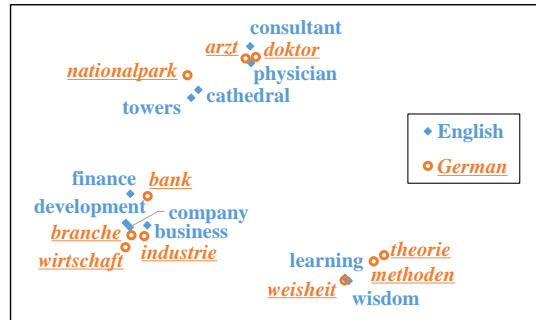


Figure 4: A visualization of the joint vector space.

and reported an accuracy of 90.8% (en→de) and 80.1% (de→en) when using full RCV and Europarl corpora.

Our method outperforms the previous work and we observe improvements when we exploit word translation probabilities (CLSim) over the model without word-level information (CLC-WA). The best result is achieved with CLSim. Furthermore, our model is considerably efficient and highly scalable, especially compared with Multitask learning and BAE. Their methods require days of training, while our model takes less than two hours on a single-core CPU, and even less than 10 minutes when we use 20 cores@2.00GHz. It is interesting to notice that CLC+WA, which makes use of word alignments in defining cross-lingual contexts, does not provide better performance than CLC-WA. We guess that sentence-level co-occurrence is more suitable for capturing sentence-level semantic relations in the task of document classification.

### 5.3 Analysis

As an important parameter of our method, the weight of cross-lingual objective versus monolingual ones determines the priority for our model to optimize, and the size of corpora used for training has an effect on the amount of informa-

tion available. Figure 3a demonstrates the effect of weighting between parts of the total objective on the accuracy of document classification, where we set monolingual weight  $\omega_i$  to 1 and adjust the cross-lingual weight  $\omega_c$ . Since monolingual and cross-lingual semantic relations both contribute to the classification accuracy, increasing  $\omega_c$  does not promise better results. The impact of the amount of monolingual data and bilingual constraints on the quality of the embeddings is shown in Figure 3b and Figure 3c, where training with more data generally improves the performance.

One drawback of our model is that it cannot learn multiple embeddings per word. At word level, homonymy and polysemy are common in natural languages, and they are often the sources of error in word embedding algorithms. When it comes to cross-lingual learning, the situation worsens as even translationally equivalent words in another language may have multiple meanings in other contexts. Bilingual resources have shown to be useful in word sense disambiguation (Guo et al., 2014), where learning multiple embeddings per word is required, but it is not trivial to develop such a technique for matrix factorization. Another shortcoming of our method is that it only learns at word level, where currently we cannot capture compositional semantics. We leave these for future work.

#### 5.4 Qualitative Examples and Visualization

We show some example English words and their nearest neighbours in the joint semantic vector space produced by our model in Table 2. For the words frequent in both monolingual and bilingual corpora, *january*, *oil* and *microsoft*, the nearest German neighbour are their direct translations or similar words. When we examine the words that appear only in monolingual corpora, but not in the bilingual corpus, such as *selloff*, *nov* (short for November) and *ppi* (producer price index), we also get semantically related words. Though these words do not occur as aligned pairs, their relations are automatically captured by the combined optimization of monolingual and cross-lingual objectives.

Figure 4 gives a visualization of some selected words using t-SNE (Van der Maaten and Hinton, 2008) where we observe the topical nature of word embeddings. Regardless of their source lan-

guages, words sharing a common topic, e.g. economy, are closely aligned with each other, revealing the semantic validity of the joint vector space.

## 6 Related Work

**Bilingual Representations** As introduced in the introduction section, joint-space models are useful in various NLP tasks, e.g., automatic bilingual lexicon extraction (Vulic and Moens, 2013). They are also useful to transfer models cross-lingually, where one projects feature representation in one vector space to another (Kozhevnikov and Titov, 2014). Similarly, bilingual word embeddings can also be projected to a shared vector space (Mikolov et al., 2013a; Faruqui and Dyer, 2014), but as pointed out by Kozhevnikov and Titov (2014) and Gouws et al. (2015), it is non-trivial to design corresponding projection procedures and whether a single transformation is adequate for the problem is under question. Some recent approaches, such as proposed by Klementiev et al. (2012), Chandar A P et al. (2014) and Gouws et al. (2015), jointly learn bilingual embeddings and avoid explicitly defining a projection process. They also allow one to simultaneously train with monolingual corpora. Our model is also a joint learning algorithm with an additional advantage that this unified framework can work with finest available parallel information, where one can encode word alignments into cross-lingual constraints for better performance, but the model also works without word alignments as long as there is a valid definition for cross-lingual co-occurrences or cross-lingual word similarities.

While we focus on word embeddings, cross-lingual document representations (Dumais et al., 1997; Platt et al., 2010) have been explored much earlier for the purpose of translingual information retrieval (Yang et al., 1998).

**Matrix Factorization** Matrix factorization has been successfully applied to learn word representations, which use several low-rank matrices to approximate the original matrix with extracted statistical information, usually word co-occurrence counts or PMI. Singular value decomposition (SVD) (Eckart and Young, 1936), SVD-based latent semantic analysis (LSA) (Landauer et al., 1998), latent semantic indexing (LSI) (Deerwester et al., 1990), and the more recently-proposed global vectors for word representation (GloVe) (Pennington et al., 2014) find their wide applications in

the area of NLP and information retrieval (Berry et al., 1995). Additionally, there is evidence that some neural-network-based models, such as Skip-gram (Mikolov et al., 2013b) which exhibits state-of-the-art performance, are also implicitly factorizing a PMI-based matrix (Levy and Goldberg, 2014). The strategy for matrix factorization in this paper, as Pennington et al. (2014), is in a stochastic fashion, which better handles unobserved data and allows one to weigh samples according to their importance and confidence.

**Matrix Co-factorization** Joint matrix factorization allows one to decompose matrices with some correlational constraints. Collective matrix factorization has been developed to handle pairwise relations (Singh and Gordon, 2008). Chang et al. (2013) generalized LSA to Multi-Relational LSA, which constructs a 3-way tensor to combine the multiple relations between words. While matrix factorization is widely used in recommender systems, matrix co-factorization helps to handle multiple aspects of the data and improves in predicting individual decisions (Hong et al., 2013). Multiple sources of information, such as content and linkage, can also be connected with matrix co-factorization to derive high-quality webpage representations (Zhu et al., 2007). The advantage of this approach is that it automatically finds optimal parameters to optimize both single matrix factorization and relational alignments, which avoids manually defining a projection matrix or transfer function. To the best of our knowledge, we are the first to introduce this technique to learn cross-lingual word embeddings.

## 7 Conclusions

In this paper, we introduced a framework of matrix co-factorization to learn cross-lingual word embeddings. It is capable of capturing the lexico-semantic similarities of different languages in a unified vector space, where the embeddings are jointly learnt instead of projected from separate vector spaces. The overall objective is divided into monolingual parts and a cross-lingual one, which enables one to use different weighting and learning strategies, and to develop models either with or without word alignments. Exploiting global context and similarity information instead of local ones, our proposed models are computationally efficient and effective.

With matrix co-factorization, it allows one to

integrate external information, such as syntactic contexts and morphology, which is not discussed in this paper. Its application in statistical machine translation and cross-lingual model transfer remains to be explored. Learning multiple embeddings per word and compositional embeddings with matrix factorization are also interesting future directions.

## Acknowledgments

This research is supported by the 973 Program (No. 2014CB340501) and the National Natural Science Foundation of China (NSFC No. 61133012, 61170196 & 61202140). We thank the anonymous reviewers for the valuable comments. We also thank Ivan Titov and Alexandre Klementiev for kindly offering their evaluation package, which allowed us to replicate their experiment settings exactly.

## References

- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *JMLR*, 3:1137–1155.
- Michael W Berry, Susan T Dumais, and Gavin W O’Brien. 1995. Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573–595.
- Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Proceedings of NIPS*, pages 1853–1861.
- Kai-Wei Chang, Wen-tau Yih, and Christopher Meek. 2013. Multi-relational latent semantic analysis. In *Proceedings of EMNLP*, pages 1602–1612.
- Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.
- Susan T Dumais, Todd A Letsche, Michael L Littman, and Thomas K Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, pages 15–21.
- Carl Eckart and Gale Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.
- Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*, pages 462–471.

- Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *ICML*, pages 748–756.
- Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning sense-specific word embeddings by exploiting bilingual resources. In *Proceedings of COLING*, pages 497–507.
- Zellig S Harris. 1954. Distributional structure. *Word*, 10(23):146–162.
- Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of ACL*, pages 58–68. ACL.
- Liangjie Hong, Aziz S Doumith, and Brian D Davison. 2013. Co-factorization machines: modeling user interests and predicting individual decisions in twitter. In *Proceedings of WSDM*, pages 557–566. ACM.
- Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL*, pages 873–882. ACL.
- Marcin Junczys-Dowmunt and Arkadiusz Szał. 2012. Symgiza++: symmetrized word alignment models for statistical machine translation. In *Security and Intelligent Information Systems*, pages 379–390. Springer.
- Alexandre Klementiev, Ivan Titov, and Binod Bhatnagar. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING. ICCL*.
- Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *Proceedings of ACL*, pages 224–229. ACL.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.
- Mikhail Kozhevnikov and Ivan Titov. 2014. Crosslingual model transfer using feature representation projection. In *Proceedings of ACL*, pages 579–585. ACL.
- Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.
- Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of NIPS*, pages 2177–2185.
- David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *JMLR*, 5:361–397.
- Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv:1309.4168*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. pages 1532–1543.
- John C Platt, Kristina Toutanova, and Wen-tau Yih. 2010. Translingual document representations from discriminative projections. In *Proceedings of EMNLP*, pages 251–261. ACL.
- Ajit P Singh and Geoffrey J Gordon. 2008. Relational learning via collective matrix factorization. In *Proceedings of SIGKDD*, pages 650–658. ACM.
- Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of EMNLP*, pages 151–161. ACL.
- Hubert Soyer, Pontus Stenetorp, and Akiko Aizawa. 2015. Leveraging monolingual data for crosslingual compositional word representations. In *Proceedings of ICLR*.
- Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *JMLR*, 9:2579–2605.
- Ivan Vulic and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of EMNLP*, pages 1613–1624.
- Yiming Yang, Jaime G Carbonell, Ralf D Brown, and Robert E Frederking. 1998. Translingual information retrieval: learning from bilingual corpora. *Artificial Intelligence*, 103(1):323–345.
- Shenghuo Zhu, Kai Yu, Yun Chi, and Yihong Gong. 2007. Combining content and link for classification using matrix factorization. In *Proceedings of SIGIR*, pages 487–494. ACM.
- Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of EMNLP*, pages 1393–1398.