Bilingual Word Representations with Monolingual Quality in Mind

Minh-Thang Luong Hieu Pham Christopher D. Manning
Computer Science Department, Stanford University, Stanford, CA, 94305
{lmlthang, hyhieu, manning}@stanford.edu

Abstract
Recent work in learning bilingual representations tend to tailor towards achieving good performance on bilingual tasks, most often the crosslingual document classification (CLDC) evaluation, but to the detriment of preserving clustering structures of word representations monolingually. In this work, we propose a joint model to learn word representations from scratch that utilizes both the context cooccurrence information through the monolingual component and the meaning equivalent signals from the bilingual constraint. Specifically, we extend the recently popular skipgram model to learn high quality bilingual representations efficiently. Our learned embeddings achieve a new state-of-the-art accuracy of 80.3 for the German to English CLDC task and a highly competitive performance of 90.7 for the other classification direction. At the same time, our models outperform best embeddings from past bilingual representation work by a large margin in the monolingual word similarity evaluation.¹

1 Introduction
Distributed word representations have been key to the recent success of many neural network models in tackling various NLP tasks such as tagging, chunking (Collobert et al., 2011), sentiment analysis (Maas et al., 2011; Socher et al., 2013b), and parsing (Socher et al., 2013a; Chen and Manning, 2014). So far, most of the focus has been spent on monolingual problems despite the existence of a wide variety of multilingual NLP tasks, which include not only machine translation (Brown et al., 1993), but also noun bracketing (Yarowsky and Ngai, 2001), entity clustering (Green et al., 2012), and bilingual NER (Wang et al., 2013). These multilingual applications have motivated recent work in training bilingual representations where similar-meaning words in two languages are embedded close together in the same high-dimensional space. However, most bilingual representation work tend to focus on learning embeddings that are tailored towards achieving good performance on a bilingual task, often the crosslingual document classification (CLDC) task, but to the detriment of preserving clustering structures of word representations monolingually.

In this work, we demonstrate that such a goal of learning representations of high quality both bilingually and monolingually is achievable through a joint learning approach. Specifically, our joint model utilizes both the context concurrence information present in the monolingual data and the meaning equivalent signals exhibited in the parallel data. The key for our approach to work is in designing a bilingual constraint consistent with monolingual components in our joint objective. To that end, we propose a novel bilingual skipgram model that extends the recently proposed skipgram approach (Mikolov et al., 2013a) to the bilingual context. Our model is efficient to train and achieves state-of-the-art performance in the CLDC task for the direction from German to English. At the same time, we demonstrate that our model well preserves the monolingual clustering structures in each language both quantitatively through the word similarity task and qualitatively through our detailed analysis.

2 Background

2.1 Monolingual Models
Existing approaches to distributed word representation learning divide into two categories: (a) neu-
eral probabilistic language models and (b) margin-based ranking models. The former specify either exactly or approximately distributions over all words \( w \) in the vocabulary given a context \( h \), and representatives of that approach include (Bengio et al., 2003; Morin, 2005; Mnih and Hinton, 2009; Mikolov et al., 2010; Mikolov et al., 2011). The later eschew the goal of training a language model and try to assign high scores for probable words \( w \) given contexts \( h \) and low scores for unlikely words \( \tilde{w} \) for the same contexts. Work in the later trend includes (Collobert and Weston, 2008; Huang et al., 2012; Luong et al., 2013).

Recently, Mikolov et al. (2013a) introduced the skipgram (SG) approach for learning solely word embeddings by reversing the prediction process, that is, to use the current word to infer its surrounding context, as opposed to using preceding contexts to predict subsequent words in traditional language model approaches. SG models greatly simplify the standard neural network-based architecture to only contain a linear projection input layer and an output softmax layer, i.e., there is no non-linear hidden layer. Despite its simplicity, SG models can achieve very good performances on various semantic tasks while having an advantage of fast training time.

We adapt SG models in our bilingual approach. Specifically, we follow Mikolov et al. (2013c) to use the negative sampling (NS) technique so as to avoid estimating the computationally expensive normalization terms in the standard softmax. Negative sampling is a simplified version of the noise contrastive estimation method (Gutmann and Hyvärinen, 2012), which attempts to differentiate data from noise by means of logistic regression. Specifically, in the SG-NS model, every word \( w \) has two distributed representations: the input vector \( x_w^{(i)} \) and the output one \( x_w^{(o)} \). For NS to work, one needs to define a scoring function to judge how likely a word \( w_n \) is likely to be a neighbor word of the current word \( w \). We use a simple scoring function (Mikolov et al., 2013c) as follows, \( \text{score}(w, w_n) = x_w^{(i)} \top x_{w_n}^{(o)} \). In our evaluation, we consider the embedding of a word as the sum of its input and output vectors.

### 2.2 Bilingual Models

Before delving further into comparing our models with those of others, let us first categorize different approaches to training bilingual word representa-

### 3 Our Approach

We hypothesize that by allowing the joint model to utilize both the cooccurrence context information within a language and the meaning-equivalent signals across languages, we can obtain better word vectors both monolingually and bilingualy. As such, we examine the following general joint objective similar to (Klementiev et al., 2012; Gouws...
et al., 2014):

\[ \alpha(Mono_1 + Mono_2) + \beta Bi \]  

(1)

In this formulation, each monolingual model, \( Mono_1 \) and \( Mono_2 \), aims to capture the clustering structure of each language, whereas the bilingual component, \( Bi \), is used to tie the two monolingual spaces together. The \( \alpha \) and \( \beta \) hyperparameters balance out the influence of the mono components over the bilingual one. When \( \alpha = 0 \), we arrive at the model proposed in (Hermann and Blunsom, 2014), whereas \( \alpha = 1 \) results in (Klementiev et al., 2012; Gouws et al., 2014) as well as our approach. Their models and ours, however, differ in terms of the choices of monolingual and bilingual components detailed next.

### 3.1 Model Choices

In terms of the monolingual component, any model listed in Section 2.1 can be a good candidate. Specifically, Klementiev et al. (2012) uses a neural probabilistic language model architecture, whereas Gouws et al. (2014) adapts the skipgram model trained with negative sampling.

When it turns to capturing bilingual constraints, these work generally use a different type of objectives for their bilingual models compared to the monolingual ones. For example, Klementiev et al. (2012) transforms the bilingual constraints into a multitask learning objective, whereas Gouws et al. (2014) minimizes the \( L_2 \)-loss between the bag-of-word vectors of parallel sentences.\(^2\)

In contrast to the existing approaches, we use the same type of models for both of our monolingual and bilingual constraints. Specifically, we adapt the skipgram model with negative sampling (SG-NS) to the bilingual context. Such a consistent choice of architectures results in a natural and effective way of building bilingual models from existing monolingual models (see §2.1).

In our case, we extend the \textit{word2vec} software\(^3\), an efficient implementation of the SG-NS, to build our fast code for bilingual representation learning. More importantly, we empirically show that our method is effective in learning representations both monolingually and bilingually as compared to existing approaches which use different architectures for monolingual and bilingual constraints.

### 3.2 Bilingual Skipgram Model (BiSkip)

The motivation behind our proposed \textit{bilingual skipgram} (BiSkip) model is to be able to predict words crosslingually rather than just monolingually as in the standard skipgram model. Imagine if we know that the word \textit{trade} is aligned to and has the same meaning as the German word \textit{Handels} as in Figure 1, we can simply substitute \textit{trade} and use \textit{Handels} to predict the surrounding words such as \textit{financial} and \textit{economic}.

Concretely, given an alignment link between a word \( w_1 \) in a language \( l_1 \) and a word \( w_2 \) in another language \( l_2 \), the BiSkip model uses the word \( w_1 \) to predict neighbors of the word \( w_2 \) and vice versa. That has the effect of training a single skipgram model with a joint vocabulary on parallel corpora in which we enrich the training examples with pairs of words coming from both sides instead of just from one language. Alternatively, one can also think of this BiSkip model as training four skipgram models jointly which predict words between the following pairs of languages: \( l_1 \rightarrow l_1 \), \( l_2 \rightarrow l_2 \), \( l_1 \rightarrow l_2 \), and \( l_2 \rightarrow l_1 \).

In our work, we experiment with two variants of our models: (a) \textit{BiSkip-UnsupAlign} where we utilize unsupervised alignment information learned by the Berkeley aligner (Liang et al., 2006) and (b) \textit{BiSkip-MonoAlign} where we simply assume monotonic alignments between words across languages. For the former, if a word is unaligned but at least one of its immediate neighbors is aligned, we will use either the only neighbor alignment or an average of the two neighbor alignments. For the latter, each source word at position \( i \) is aligned to the target word at position \( [i \ast T/S] \) where \( S \)
and $T$ are the source and target sentence lengths. These two variants are meant to attest how important unsupervised alignment information is in learning bilingual embeddings.

4 Experiments

4.1 Data

We train our joint models on the parallel Europarl v7 corpus between German (de) and English (en) (Koehn, 2005), which consists of 1.9M parallel sentences (49.7M English tokens and 52.0M German tokens). After lowercasing and tokenizing we map each digit into 0, i.e. 2013 becomes 0000. Other rare words occurring less than 5 times are mapped to $<$unk$>$. The resulting vocabularies are of size 40K for English and 95K for German.

4.2 Training

We use the following settings as described in (Mikolov et al., 2013c): stochastic gradient descent with a default learning rate of 0.025, negative sampling with 30 samples, skipgram with context window of size 5, and a subsampling rate\(^4\) of value $1e^{-4}$. All models are trained for 10 epochs and the learning rate is decayed to 0 once training is done. We set the hyperparameters in Eq. (1) to 1 for $\alpha$ and 4 for $\beta$ in our experiments.

4.3 Evaluation Tasks

We evaluate our models on two aspects: (a) monolingually with a word similarity task and (b) bilingually through a cross-lingual document classification setup.

4.3.1 Word Similarity

This task measures the semantic quality of the learned word vectors monolingually over various word similarity datasets which have been used in papers on word embedding learning lately. For each dataset, we report a Spearman’s rank correlation coefficient between similarity scores given by the learned word vectors and those rated by humans. For English, we utilize the following publicly available datasets: WordSim353 (353 pairs), $MC$ (30 pairs), $RG$ (65 pairs), SCWS (1762 pairs), and RW (2034 pairs). See (Luong et al., 2013) for more information about these datasets.

\(^4\)Smaller values mean frequent words are discarded more often, see (Mikolov et al., 2013c) and the word2vec code for more details.

To evaluate the semantic quality of German embeddings, we devise our own version of the WordSim353 German counterpart. Our procedure is as follows: we first used Google Translate to get German translations for the 437 distinct tokens in the English WordSim353. We then asked two German speakers to help us verify these translations, out of which, we fixed 23 translation pairs.

4.3.2 Cross-lingual Document Classification

To judge the bilingual aspect of our models, we follow (Klementiev et al., 2012) in using a cross-lingual document classification task: train with 1000 and test on 5000 RCV-labeled documents.\(^5\) In this setup, a multi-class classifier is trained using the averaged perceptron algorithm. The feature vector for each document is the averaged vector of words in the document weighted by their idf values. A classification model trained on one language is then applied directly to classify new documents in another language without retraining. This is an example of transfer learning of models from a resource-rich language into a resource-poor one. The premise for such a setup to work is because word vectors in these languages are embedded in the same space, so document feature vectors are constructed consistently across these two languages and trained weights can be reused.

5 Results

In this section, we present results of our joint models trained on the Europarl corpus. Our first focus is on the CLDC evaluation where we compare performances achieved by our BiSkip models over the best CLDC results from past work. Specifically, we utilize the best set of embeddings from each of the following bilingual work: (a) multitask learning model (Klementiev et al., 2012), (b) bilingual without alignment model (Gouws et al., 2014), (c) distributed word alignment model (Kočiský et al., 2014), (d) autoencoder model (Chandar A P et al., 2014), and (e) compositional model (Hermann and Blunsom, 2014).

The above models are compared against our two BiSkip models, one utilizing the unsupervised alignments ($UnsupAlign$) and one assuming monotonic alignments ($MonoAlign$); we trained both 40- and 128-dimensional vectors to be comparable with existing embeddings. Simultaneou
Table 1: **German (de) - English (en) bilingual embeddings** – results of various models in terms of both the monolingual (word similarity) and bilingual (cross-lingual document classification) tasks. Spearman’s rank correlation coefficients are reported for word similarity tasks, whereas accuracies on 1000 RCV-labeled documents are used for CLDC. We compare our BiSkip embeddings to the best ones from past work: multitask I-Matrix (Klementiev et al., 2012), bilingual without alignment BilBOWA (Gouws et al., 2014), distributed word alignment DWA (Kočiský et al., 2014), autoencoder BAE-cr (Chandar AP et al., 2014), and compositional CVM-Add (Hermann and Blunsom, 2014). Numbers in boldface highlight the best scores per metric. We italicize the second best results and mark _ for models where we do not have access to the trained embeddings.

Previously, we test if these learned bilingual embeddings still preserve the clustering properties monolingually in terms of their performance on the word similarity datasets.

At 40 dimensions, both our BiSkip embeddings outperform those produced by the model in (Klementiev et al., 2012) over all aspects. Our MonoAlign model also surpasses the CLDC performances of the BilBOWA model (Gouws et al., 2014). These two models we are comparing to are most similar to ours in terms of the joint objective, i.e. with two monolingual language models and a bilingual component.

The fact that the embeddings in (Klementiev et al., 2012) perform poorly on the monolingual aspects, i.e. the word similarity tasks, supports one of our early observations that it is important to design a bilingual component that is consistent with the monolingual documents (§3.1). Otherwise, the model will make a tradeoff between obtaining good performance for bilingual tasks over monolingual tasks as seems to be the case for the embeddings produced by the multitask learning model.

Our 40-dimensional embeddings also rival those trained by much more complex models than ours such as the autoencoder model BAE-cr (Chandar AP et al., 2014). It is worthwhile to mention that beside the Europarl corpus, the autoencoder model was also trained with the RCV documents on which the CLDC classifiers were built, which is an advantage over our model. Despite this, our MonoAlign representations outperform the embeddings in (Chandar AP et al., 2014) over all word similarity datasets and CLDC _de→en_.

**Larger dimensions** – When learning higher dimensional embeddings, which is an advantage of our joint models as it is very fast to train compared to other methods, the results across all metrics well correlate with the embedding sizes as we increase from 40, 128, 256, to 512. Our 256- and 512-dimensional embeddings trained with unsupervised alignments produce strong results, significantly better than all other models in terms of the word similarity datasets and achieve _state-of-the-art performance_ in terms of the CLDC _de→en_ with an accuracy of 80.3. For CLDC _en→de_, our model reaches a very high score of 90.7, close to the best published result of 91.8 produced by the autoencoder model.6

---

6The 256- and 512-dimensional MonoAlign models do
Table 2: Nearest neighbor words – shown are the top 10 nearest English (en) and German (de) words for each of the following words in the list \{January, Microsoft, Distinctive\} as measured by the Euclidean distances given a set of embeddings. We compare our learned vectors (BiSkip-UnsupAlign, \(d = 128\)) with those produced by the autoencoder model (Chandar A P et al., 2014). For the word distinctive, we provide Google Translate glosses for German words. The word telekommunikationsunternehmen is truncated into telekommun*.

Alignment effects – It is interesting to observe that the 40- and 80-dimensional MonoAlign models with a simple monotonic alignment assumption can rival the UnsupAlign models, which uses unsupervised alignments, in many metrics. Overall, all our models are superior to the DWA approach (Koˇcisk ´y et al., 2014) which learns distributed alignments and embeddings simultaneously.

Word similarity results – It is worthwhile to point out that this work does not aim to be best in terms of the word similarity metrics. Past work such as (Pennington et al., 2014; Faruqui and Dyer, 2014) among many others, have demonstrated that higher word similarity scores can be achieved by simply increasing the vocabulary coverage, training corpus size, and the embedding dimension. Rather, we show that our model can learn bilingual embeddings that are naturally better than those of existing approaches monolingually.

6 Analysis

Beside the previous quantitative evaluation, we examine our learned embeddings qualitatively in this section through the following methods: (a) nearest neighbor words and (b) embedding visualization.

6.1 Nearest Neighbor Words

For the former method, we follow (Chandar A P et al., 2014) to find, for each English word, a list of top 10 English and German words closest to it based on Euclidean distance in a learned bilingual space. Our list of words include \{January, Microsoft, Distinctive\}, in which the first two choices are made by the previous work. We compare our learned embeddings using the BiSkip-UnsupAlign model (\(d = 128\)) with those produced by the autoencoder model in (Chandar A P et al., 2014).

Examples in Table 2 demonstrate that our learned representations are superior in two aspects. Bilingually, our embeddings succeed in selecting the 1-best translations for all words in the list, whereas the other model fails to do so for the word distinctive. Monolingually, our embeddings possess a clearly better clustering structure. For example, all months are clustered together, around the word January, whereas that is not the case for the other embeddings with the occurrences of \{1999, month, year\} in the top 10 list. Our embeddings also find very relevant neighbor words for the word Microsoft such as \{IBM, Yahoo, etc.\}.

We also examine the BiSkip-MonoAlign model
6.2 Bilingual Embedding Visualization

In this section, we visualize embeddings of the top frequent German and English words from the WordSim353 datasets. The two-dimensional visualizations of word vectors are produced using the Barnes-Hut-SNE algorithm (van der Maaten, 2013). Figure 2 shows that most English-German words with similar meanings appear nearby, e.g., day-tag or article-artikel. Monolingually, we also see clusters of words such as {market, industry, production} (at the bottom) and {president, minister, government} (at the top).

7 Related Work

We have previously discussed in Section 2 models directly related to our work. In this section, we survey other approaches in learning monolingual and bilingual representations.

Current work in dimensionality reduction of word representations can be broadly grouped into three categories (Turian et al., 2010): (a) distributional representations learned from a co-occurrence matrix of words and contexts (documents, neighbor words, etc.) using techniques such as LSA (Dumais et al., 1988) or LDA (Blei et al., 2003), (b) clustering-based representations, e.g., Brown et al. (1992)’s hierarchical clustering algorithm which represents each word as a binary path through the cluster hierarchy, and (c) distributed representations, where each word is explicitly modeled by a dense real-valued vector and directly induced by predicting words from contexts or vice versa as detailed in Section 2.1.

Moving beyond monolingual representations, work in constructing bilingual vector-space models divides into two main streams: (a) those that make use of comparable corpora and (b) those that only require unaligned or monolingual text. The former includes various extensions to standard techniques such as bilingual latent semantic models (LSA) (Tam and Schultz, 2007; Ruiz and Federico, 2011) or bilingual/multilingual topic models (LDA) (Zhao and Xing, 2007; Ni et al., 2009; Mimno et al., 2009; Vulić et al., 2011). In this work, the general assumption is that aligned documents share identical topic distributions. The latter stream, which eschews the use of compara-
ble data, generally requires a small initial lexicon which is extracted either manually or automatically (e.g., cognates, string edit distances, etc.). Representatives of this strand include work that extends CCA (Haghighi et al., 2008; Boyd-Graber and Blei, 2009), mapping representations of words in different languages into the same space, as well as work that follows a bootstrapping style to iteratively enlarge the initial lexicon (Peirsman and Padó, 2010; Vulić and Moens, 2013).

8 Conclusion

This work proposes a novel approach that jointly learns bilingual representations from scratch by utilizing both the context concurrence information in the monolingual data and the meaning-equivalent signals in the parallel data. We advocate a new standard in training bilingual embeddings, that is, to be good in not only gluing representations bilingually but also preserving the clustering structures of words in each language.

We provide a key insight to train embeddings that meet the above two criteria, that is, to design a bilingual constraint that is consistent with the monolingual models in our joint objective. Our learned representations are superior to the best embeddings from past bilingual work in two tasks: (a) the crosslingual document classification one in which we achieve a new state-of-the-art performance for the direction from German to English, and (b) the word similarity evaluation where we outperform other embeddings by a large margin over all datasets. We also evaluate the learned vectors qualitatively by examining nearest neighbors of words and visualizing the representations.

Lastly, it would be interesting to extend our method to multiple languages as in (Hermann and Blunsom, 2014) and to be able to train on a large amount of monolingual data similar to (Gouws et al., 2014).

Acknowledgment

We gratefully acknowledge support from a gift from Bloomberg L.P. and from the Defense Advanced Research Projects Agency (DARPA) Broad Operational Language Translation (BOLT) program under contract HR0011-12-C-0015 through IBM. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA, or the US government. We also thank members of the Stanford NLP Group as well as the anonymous reviewers for their valuable comments and feedbacks.

References


Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In EACL.


Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *NAACL-HLT*.


Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.


Ivan Vušić and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *EMNLP*.


