Multilingual Training of Crosslingual Word Embeddings

Long Duong,[1] **Hiroshi Kanayama,**[2] **Tengfei Ma,**[3] **Steven Bird**[14] and **Trevor Cohn**[1]
[1]Department of Computing and Information Systems, University of Melbourne
[2]IBM Research – Tokyo
[3]IBM T.J. Watson Research Center
[4]International Computer Science Institute, University of California Berkeley

Abstract

Crosslingual word embeddings represent lexical items from different languages using the same vector space, enabling crosslingual transfer. Most prior work constructs embeddings for a pair of languages, with English on one side. We investigate methods for building high quality crosslingual word embeddings for many languages in a unified vector space. In this way, we can exploit and combine information from many languages. We report competitive performance on bilingual lexicon induction, monolingual similarity and crosslingual document classification tasks.

1 Introduction

Monolingual word embeddings have facilitated advances in many natural language processing tasks, such as natural language understanding (Collobert and Weston, 2008), sentiment analysis (Socher et al., 2013), and dependency parsing (Dyer et al., 2015). Crosslingual word embeddings represent words from several languages in the same low dimensional space. They are helpful for multilingual tasks such as machine translation (Brown et al., 1993) and bilingual named entity recognition (Wang et al., 2013). Crosslingual word embeddings can also be used in transfer learning, where the source model is trained on one language and applied directly to another language; this is suitable for the low-resource scenario (Yarowsky and Ngai, 2001; Duong et al., 2015b; Das and Petrov, 2011; Täckström et al., 2012).

Most prior work on building crosslingual word embeddings focuses on a pair of languages. English is usually on one side, thanks to the wealth of available English resources. However, it is highly desirable to have a crosslingual word embeddings for many languages so that different relations can be exploited.[1] For example, since Italian and Spanish are similar, they are excellent candidates for transfer learning. However, few parallel resources exist between Italian and Spanish for directly building bilingual word embeddings. Our multilingual word embeddings, on the other hand, map both Italian and Spanish to the same space without using any direct bilingual signal between them. In addition, multilingual word embeddings allow multiple source language transfer learning, producing a more general model and overcoming data sparseness (McDonald et al., 2011; Guo et al., 2016; Agić et al., 2016). Moreover, multilingual word embeddings are also crucial for multilingual applications such as multi-source machine translation (Zoph and Knight, 2016), and multi-source transfer dependency parsing (McDonald et al., 2011; Duong et al., 2015a).

We propose several algorithms to map bilingual word embeddings to the same vector space, either during training or during post-processing. We apply a linear transformation to map the English side of each pretrained crosslingual word embedding to the same space. We also extend Duong et al. (2016), which used a lexicon to learn bilingual word embeddings. We modify the objective function to jointly build multilingual word embeddings during training. Unlike most prior work which focuses on downstream applications, we measure the quality of our multilingual word embeddings in three ways: bilingual lexicon induction, monolingual word similarity, and crosslingual document classification tasks. Relative to a benchmark of

[1]From here on we refer to crosslingual word embeddings for a pair of languages and multiple languages as *bilingual word embeddings* and *multilingual word embeddings* respectively.

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 894–904,
Valencia, Spain, April 3-7, 2017.

training on each language pair separately and to various published multilingual word embeddings, we achieved high performance for all the tasks.

In this paper we make the following contributions: (a) novel algorithms for post hoc combination of multiple bilingual word embeddings, applicable to any pretrained bilingual model; (b) a method for jointly learning multilingual word embeddings, extending Duong et al. (2016), to jointly train over monolingual corpora in several languages; (c) achieving competitive results in bilingual, monolingual and crosslingual transfer settings.

2 Related work

Crosslingual word embeddings are typically based on co-occurrence statistics from parallel text (Luong et al., 2015; Gouws et al., 2015; Chandar A P et al., 2014; Klementiev et al., 2012; Kočiský et al., 2014; Huang et al., 2015). Other work uses more widely available resources such as comparable data (Vulić and Moens, 2015) and shared Wikipedia entries (Søgaard et al., 2015). However, those approaches rely on data from Wikipedia, and it is non-trivial to extend them to languages that are not covered by Wikipedia. Lexicons are another source of bilingual signal, with the advantage of high coverage. Multilingual lexical resources such as PanLex (Kamholz et al., 2014) and Wiktionary[2] cover thousands of languages, and have been used to construct high performance crosslingual word embeddings (Mikolov et al., 2013a; Xiao and Guo, 2014; Faruqui and Dyer, 2014).

Previous work mainly focuses on building word embeddings for a pair of languages, typically with English on one side, with the exception of Coulmance et al. (2015), Søgaard et al. (2015) and Ammar et al. (2016). Coulmance et al. (2015) extend the bilingual skipgram model from Luong et al. (2015), training jointly over many languages using the Europarl corpora. We also compare our models with an extension of Huang et al. (2015) adapted for multiple languages also using bilingual corpora. However, parallel data is an expensive resource and using parallel data seems to under-perform on the bilingual lexicon induction task (Vulić and Moens, 2015). While Coulmance et al. (2015) use English as the pivot language, Søgaard et al. (2015) learn multilingual word embeddings for many languages using Wikipedia entries which are the same for many languages. However, their approach is limited to languages covered in Wikipedia and seems to under-perform other methods. Ammar et al. (2016) propose two algorithms, MultiCluster and MultiCCA, for multilingual word embeddings using set of bilingual lexicons. MultiCluster first builds the graph where nodes are lexical items and edges are translations. Each cluster in this graph is an anchor point for building multilingual word embeddings. MultiCCA is an extension of Faruqui and Dyer (2014), performing canonical correlation analysis (CCA) for multiple languages using English as the pivot. A shortcoming of MultiCCA is that it ignores polysemous translations by retaining only one-to-one dictionary pairs (Gouws et al., 2015), disregarding much information. As a simple solution, we propose a simple post hoc method by mapping the English parts of each bilingual word embedding to each other. In this way, the mapping is always exact and one-to-one.

Duong et al. (2016) constructed bilingual word embeddings based on monolingual data and PanLex. In this way, their approach can be applied to more languages as PanLex covers more than a thousand languages. They solve the polysemy problem by integrating an EM algorithm for selecting a lexicon. Relative to many previous crosslingual word embeddings, their joint training algorithm achieved state-of-the-art performance for the bilingual lexicon induction task, performing significantly better on monolingual similarity and achieving a competitive result on cross lingual document classification. Here we also adopt their approach, and extend it to multilingual embeddings.

2.1 Base model for bilingual embeddings

We briefly describe the base model (Duong et al., 2016), an extension of the continuous bag-of-word (CBOW) model (Mikolov et al., 2013a) with negative sampling. The original objective function is

$$\sum_{i \in D} \left(\log \sigma(\mathbf{u}_{w_i}^\top \mathbf{h}_i) + \sum_{j=1}^{p} \log \sigma(-\mathbf{u}_{w_{ij}}^\top \mathbf{h}_i) \right), \quad (1)$$

where D is the training data, $\mathbf{h}_i = \frac{1}{2k} \sum_{j=-k;j\neq 0}^{k} \mathbf{v}_{w_{i+j}}$ is a vector encoding the context over a window of size k centred around position i, $\mathbf{V}$ and $\mathbf{U} \in \mathbb{R}^{|V_e| \times d}$ are learned matrices referred to as the context and centre word

[2] wiktionary.org

embeddings, where V_e is the vocabulary and p is the number of negative examples randomly drawn from a noise distribution, $w_{ij} \sim P_n(w)$.

Duong et al. (2016) extend the CBOW model for application to two languages, using monolingual text in both languages and a bilingual lexicon. Their approach augments CBOW by generating not only the middle word, but also its translation in the other language. This is done by first selecting a translation $\bar{w}_i$ from the lexicon for the middle word w_i, based on the cosine distance between the context h_i and the context embeddings $\mathbf{V}$ for each candidate foreign translation. In this way source monolingual training contexts must generate both source and target words, and similarly target monolingual training contexts also generate source and target words. Overall this results in compatible word embeddings across the two languages, and highly informative nearest neighbours across the two languages. This leads to the new objective function

$$\sum_{i \in D_s \cup D_t} \Bigg(\log \sigma(\mathbf{u}_{w_i}^\top \mathbf{h}_i) + \log \sigma(\mathbf{u}_{\bar{w}_i}^\top \mathbf{h}_i) + \sum_{j=1}^{p} \log \sigma(-\mathbf{u}_{w_{ij}}^\top \mathbf{h}_i) \Bigg) + \delta \sum_{w \in V_s \cup V_t} \|\mathbf{u}_w - \mathbf{v}_w\|_2^2 , \quad (2)$$

where D_s and D_t are source and target monolingual data, V_s and V_t are source and target vocabulary. Comparing with the CBOW objective function in Equation (1), this represents two additions: the translation cross entropy $\log \sigma(\mathbf{u}_{\bar{w}_i}^\top \mathbf{h}_i)$, and a regularisation term $\sum_{w \in V_s \cup V_t} \|\mathbf{u}_w - \mathbf{v}_w\|_2^2$ which penalises divergence between context and center word embedding vectors for each word type, which was shown to improve the embedding quality (Duong et al., 2016).

3 Post hoc Unification of Embeddings

Our goal is to learn multilingual word embeddings over more than two languages. One simple way to do this is to take several learned bilingual word embeddings which share a common target language (here, English), and map these into a shared space (Mikolov et al., 2013a; Faruqui and Dyer, 2014). In this section we propose post hoc methods, however in §4 we develop an integrated multilingual method using joint inference.

Formally, the input to the post hoc combination methods are a set of n pre-trained bilingual word embedding matrices, i.e., $C_i = \{(E_i, F_i)\}$ with $i \in \mathbf{F}$ is the set of foreign languages (not English), $E_i \in \mathbb{R}^{|V_{e_i}| \times d}$ are the English word embeddings and $F_i \in \mathbb{R}^{|V_{f_i}| \times d}$ are foreign language word embeddings for language i, with V_{e_i} and V_{f_i} being the English and foreign language vocabularies and d is the embedding dimension. These bilingual embeddings can be produced by any method, e.g., those discussed in §2.

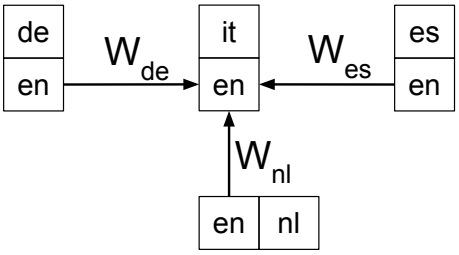

Figure 1: Examples of unifying four bilingual word embeddings between `en` and `it`, `de`, `es`, `nl` to the same space using post hoc linear transformation.

Linear Transformation. The simplest method is to learn a linear transformation which maps the English part of each bilingual word embedding into the same space (inspired by Mikolov et al. (2013a)), as illustrated in Figure 1. One language pair is chosen as the pivot, `en-it` in this example, and the English side of the other language pairs, `en-de`, `en-es`, `en-nl`, are mapped to closely match the English side of the pivot, `en-it`. This is achieved through learning linear transformation matrices for each language, $W_{\texttt{de}}, W_{\texttt{es}}$ and $W_{\texttt{nl}}$, respectively, where each $W_i \in \mathbb{R}^{d \times d}$ is learned to minimize the objective function $\|E_i \times W_i - E_{pivot}\|_2^2$ where E_{pivot} is the English embedding of the pivot pair, `en-it`.

Each foreign language f_i is then mapped to the same space using the learned matrix W_i, i.e., $F_i' = F_i \times W_i$. These projected foreign embeddings are then used in evaluation, along with the English side of the language pair with largest English vocabulary coverage, i.e., biggest $|V_{e_i}|$. Together these embeddings allow for querying of monolingual and cross-lingual word similarity, and multilingual transfer of trained models.

The advantage of this approach is that it is very fast and simple to train, since the objective function is strictly convex and has a closed form solution. Moreover, unlike Mikolov et al. (2013a) who learn the projection from a source to a target

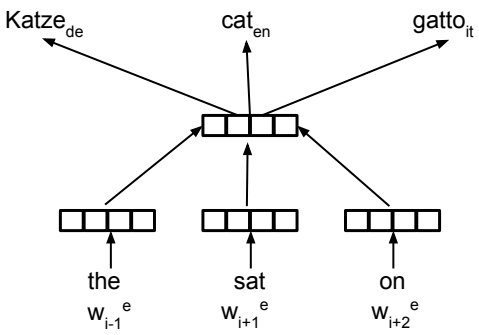

Figure 2: Examples of our multilingual joint training model without mapping for learning multilingual embeddings for three languages `en, it, de` using joint inference.

language, we learn the projection from English to English, thus do not require a lexicon, sidestepping the polysemy problem.[3]

4 Multilingual Joint Training

Instead of combining bilingual word embeddings in the post-processing step, it might be more beneficial to do it during training, so that languages can interact with each other more freely. We extend the method in §2.1 to jointly learn the multilingual word embeddings during training. The input to the model is the combined monolingual data for each language and the set of lexicons between any language pair.

We modify the base model (Duong et al., 2016) to accommodate more languages. For the first step, instead of just predicting the translation for a single target language, we predict the translation for all languages in the lexicon. That is, we compute $w_i^f = \text{argmax}_{w \in \text{dict}_e^f(w_i^e)} \cos(\mathbf{v}_w, \text{context})$, which is the best translation in language f of source word w_i^e in language e, given the bilingual lexicon dict_e^f and the context. For the second step, we jointly predict word w_i^e and all translations w_i^f in all foreign languages $f \in \mathbf{T}$ that we have dictionary dict_e^f as illustrated in Figure 2.

[3] A possible criticism of this approach is that a linear transformation is not powerful enough for the required mapping. We experimented with non-linear transformations but did not observe any improvements. Faruqui and Dyer (2014) extended Mikolov et al. (2013a) as they projected both source and target languages to the same space using canonical correlation analysis (CCA). We also adopted this approach for multilingual environment by applying multi-view CCA to map the English part of each pre-trained bilingual word embedding to the same space. However, we only observe minor improvements.

The English word *cat* might have several translations in German $\{$*Katze, Raupe, Typ*$\}$ and Italian $\{$*gatto, gatta*$\}$. In the first step, we select the closest translation given the context for each language, i.e. *Katze* and *gatto* for German and Italian respectively. In the second step, we jointly predict the English word *cat* together with selected translations *Katze* and *gatto* using the following modified objective function:

$$\mathcal{O} = \sum_{i \in D_{all}} \Bigg(\log \sigma(\mathbf{u}_{w_i^e}^\top \mathbf{h}_i) + \sum_{f \in \mathbf{T}} \log \sigma(\mathbf{u}_{w_i^f}^\top \mathbf{h}_i) + \sum_{j=1}^{p} \log \sigma(-\mathbf{u}_{w_{ij}}^\top \mathbf{h}_i) \Bigg) + \delta \sum_{w \in V_{all}} \|\mathbf{u}_w - \mathbf{v}_w\|_2^2 , \quad (3)$$

where D_{all} and V_{all} are the combined monolingual data and vocabulary for all languages. Each of the p negative samples, w_{ij}, are sampled from a unigram model over the combined vocabulary V_{all}.

Explicit mapping. As we keep adding more languages to the model, the hidden layer in our model – shared between all languages – might not be enough to accommodate all languages. However, we can combine the strength of the linear transformation proposed in §3 to our joint model as described in Equation (3). We explicitly learn the linear transformation jointly during training by adding the following regularization term to the objective function:

$$\mathcal{O}' = \mathcal{O} + \alpha \sum_{i \in D_e} \sum_{f \in \mathbf{F}} \|\mathbf{u}_{w_i^f} W_f - \mathbf{u}_{w_i^e}\|_2^2 , \quad (4)$$

where D_e is the English monolingual data (since we use English as the pivot language), $\mathbf{F}$ is the set of foreign languages (not English), $W_f \in \mathbb{R}^{d \times d}$ is the linear transformation matrix, and α controls the contribution of the regularization term and will be tuned in §6.[4] Thus, the set of learned parameters for the model are the word and context embeddings $\mathbf{U}, \mathbf{V}$ and $|\mathbf{F}|$ linear transformation matrices, $\{W_f\}_{f \in \mathbf{F}}$. After training is finished, we linearly transform the foreign language embeddings with the corresponding learned matrix W_f, such that all embeddings are in the same space.

5 Experiment Setup

Our experimental setup is based on that of Duong et al. (2016). We use the first 5 million sen-

[4] For an efficient implementation, we apply this constraint to only 10% of English monolingual data.

	Model	it-en		es-en		nl-en		nl-es		Average	
		rec_1	rec_5	rec_1	rec_5	rec_1	rec_5	rec_1	rec_5	rec_1	rec_5
Baselines	MultiCluster	35.6	64.3	34.9	62.5	-	-	-	-	-	-
	MultiCCA	63.4	77.3	58.5	72.7	-	-	-	-	-	-
	MultiSkip	57.6	68.5	49.3	58.9	-	-	-	-	-	-
	MultiTrans	72.1	83.1	71.5	82.2	-	-	-	-	-	-
Ours	Linear	78.5	88.2	69.3	81.8	74.9	87.0	66.3	79.7	72.2	84.2
	Joint	79.4	89.7	73.6	84.6	76.6	89.6	69.4	82.0	74.7	86.5
	+ Mapping	**81.6**	**90.5**	74.6	**87.4**	77.9	**91.4**	71.6	**83.5**	76.4	**88.2**
	BiWE	80.8	90.4	**74.7**	85.4	**79.1**	90.5	**71.7**	80.7	**76.6**	86.7

Table 1: Bilingual lexicon induction performance for four pairs. Bilingual word embeddings (BiWE) is the state-of-the-art result from Duong et al. (2016) where each pair is trained separately. Our proposed methods including linear transformation (Linear), joint prediction as in Equation (3) (Joint) and joint prediction with explicit mapping as in Equation (4) (+mapping). We report recall at 1 and 5 with respect to four baseline multilingual word embeddings. The best scores for are shown in bold.

tences from the tokenized monolingual data from the Wikipedia dump from Al-Rfou et al. (2013).[5] The dictionary is from PanLex which covers more than 1,000 language varieties. We build multilingual word embeddings for 5 languages (`en, it, es, nl, de`) jointly using the same parameters as Duong et al. (2016).[6] During training, for a fairer comparison, we only use lexicons between English and each target language. However, it is straightforward to incorporate a lexicon between any pair of languages into our model. The pretrained bilingual word embeddings for the postprocessing experiment in §3 are also from Duong et al. (2016).

In the following sections, we evaluate the performance of our multilingual word embeddings in comparison with bilingual word embeddings and previous published multilingual word embeddings (MultiCluster, MultiCCA, MultiSkip and MultiTrans) for three tasks: bilingual lexicon induction (§6), monolingual similarity (§7) and crosslingual document classification (§8). MultiCluster and MultiCCA are the models proposed from Ammar et al. (2016) trained on monolingual data using bilingual lexicons extracted from aligning Europarl corpus. MultiSkip is the reimplementation of the multilingual skipgram model from Coulmance et al. (2015). MultiTrans is the multilingual version of the translation invariance model from Huang et al. (2015). Both MultiSkip and MultiTrans are trained directly on parallel data from Europarl. All the previous work is trained with 512 dimensions on 12 languages acquired directly from Ammar et al. (2016).

6 Bilingual Lexicon Induction

In this section we evaluate our multilingual models on the bilingual lexicon induction (BLI) task, which tests the bilingual quality of the model. Given a word in the source language, the model must predict the translation in the target language. We report recall at 1 and 5 for the various models listed in Table 1. The evaluation data for `it-en`, `es-en`, and `nl-en` pairs was manually constructed (Vulić and Moens, 2015). We extend the evaluation for `nl-es` pair which do not involve English.[7]

The BiWE results for pairs involving English in Table 1 are from Duong et al. (2016), the current state of the art in this task. For the `nl-es` pair, we cannot build bilingual word embeddings, since we do not have a corresponding bilingual lexicon. Instead, we use English as the pivot language. To get the `nl-es` translation, we use two bilingual embeddings of `nl-en` and `es-en` from Duong et al. (2016). We get the best English translation for the Dutch word, and get the top 5 Spanish

[5] We will use the whole data if there are less than 5 million sentences.

[6] Default learning rate of 0.025, negative sampling with 25 samples, subsampling rate of value $1e^{-4}$, embedding dimension $d = 200$, window size 48, run for 15 epochs and $\delta = 0.01$ for combining word and context embeddings.

[7] We build 1,000 translation pairs for `nl-es` pair with the source word from Vulić and Moens (2015) and ground truth candidates from Google Translate but manually verified.

translations with respect to the English word. This simple trick performs surprisingly well, probably because bilingual word embeddings involving English such as nl-en and es-en from Duong et al. (2016) are very accurate.

For the linear transformation, we use the first pair it-en as the pivot and learn to project es-en, de-en, nl-en pairs to this space as illustrated in Figure 1. We use English part ($E'_{biggest}$) from transformed de-en pair as the English output. Despite simplicity, linear transformation performs surprisingly well.

Our joint model to predict all target languages simultaneously, as described in Equation (3), performs consistently better in contrast with linear transformation at all language pairs. The joint model with explicit mapping as described in Equation (4) can be understood as the combination of joint model and linear transformation. For this model, we need to tune α in Equation (4). We tested α with value in range $\{10^{-i}\}_{i=0}^{5}$ using es-en pair on BLI task. $\alpha = 0.1$ gives the best performance. To avoid over-fitting, we use the same value of α for all experiments and all other pairs. With this tuned value α, our joint model with mapping clearly outperforms other proposed methods on all pairs. More importantly, this result is substantially better than all the baselines across four language pairs and two evaluation metrics. Comparing with the state of the art (BiWE), our final model (joint + mapping) are more general and more widely applicable, however achieves relatively better result, especially for recall at 5.

7 Monolingual similarity

The multilingual word embeddings should preserve the monolingual property of the languages. We evaluate using the monolingual similarity task proposed in Luong et al. (2015). In this task, the model is asked to give the similarity score for a pair of words in the same language. This score is then measured against human judgment. Following Duong et al. (2016), we evaluate on three datasets, WordSim353 (WS-en), RareWord (RW-en), and the German version of WordSim353 (WS-de) (Finkelstein et al., 2001; Luong et al., 2013; Luong et al., 2015).

Table 2 shows the result of our multilingual word embeddings with respect to several baselines. The trend is similar to the bilingual lexicon induction task. Linear transformation performs surprisingly well. Our joint model achieves a similar result, with linear transformation (better on WS-de but worse on WS-en and RW-en). Our joint model with explicit mapping regains the drop and performs slightly better than linear transformation. More importantly, this model is substantially better than all baselines, except for MultiTrans on RW-en dataset. This can probably be explained by the low coverage of MultiTrans on this dataset. Our final model (Joint + Mapping) is also close to the best bilingual word embeddings (BiWE) performance reported by Duong et al. (2016).

	Model	WS-de	WS-en	RW-en
Baselines	MultiCluster	51.0 [98.3]	53.9 [100]	38.1 [57.6]
	MultiCCA	60.2 [99.7]	66.3 [100]	43.1 [71.1]
	MultiSkip	48.4 [96.6]	51.2 [99.7]	33.9 [55.4]
	MultiTrans	56.4 [92.6]	61.1 [97.2]	**51.1** [23.1]
Ours	Linear	67.5 [99.4]	**74.7** [100]	45.4 [75.5]
	Joint	68.5 [99.4]	74.6 [100]	43.8 [75.5]
	Joint + Mapping	**70.4** [99.4]	74.4 [100]	45.1 [75.5]
	BiWE	71.1 [99.4]	76.2 [100]	44.0 [75.5]

Table 2: Spearman's rank correlation for monolingual similarity measurement for various models on 3 datasets WS-de (353 pairs), WS-en (353 pairs) and RW-en (2034 pairs). We compare against 4 baseline multilingual word embeddings. BiWE is the result from Duong et al. (2016) where each pair is trained separately which serves as the reference for the best bilingual word embeddings. The best results for multilingual word embeddings are shown in bold. Numbers in square brackets are the coverage percentage.

8 Crosslingual Document Classification

In the previous sections, we have shown that our methods for building multilingual word embeddings, either in the post-processing step or during training, preserved high quality bilingual and monolingual relations. In this section, we demonstrate the usefulness of multi-language crosslingual word embeddings through the crosslingual document classification (CLDC) task.

This task exploits transfer learning, where the document classifier is trained on the source language and tested on the target language. The source language classifier is transferred to the target language using crosslingual word embeddings as the document is represented as the sum of bag-

		en→de	de→en	it→de	it→es	en→es	Avg
Baselines	MultiCluster	**92.9**	69.1	79.1	**81.0**	**63.1**	**77.0**
	MultiCCA	69.2	50.7	83.1	79.0	45.3	65.5
	MultiSkip	79.9	63.5	71.8	76.3	60.4	70.4
	MultiTrans	87.7	75.2	70.4	64.4	56.1	70.8
Ours	Linear	83.8	75.7	74.8	67.3	57.4	71.8
	Joint	86.2	75.7	82.3	70.7	56.0	74.2
	Joint + Mapping	89.5	**81.6**	**84.3**	74.1	53.9	76.7
Bilingual	Luong et al. (2015)	88.4	**80.3**	-	-	-	-
	Chandar A P et al. (2014)	**91.8**	74.2	-	-	-	-
	Duong et al. (2016)	86.3	76.8	-	-	**53.8**	-

Table 3: Crosslingual document classification accuracy for various model. Chandar A P et al. (2014) and Luong et al. (2015) achieved a state-of-the-art result for en→de and de→en respectively, served as the reference. The best results for bilingual and multilingual word embeddings are bold.

of-word embeddings weighted by *tf.idf*. This setting is useful for target low-resource languages where the annotated data is insufficient.

The train and test data are from multilingual RCV1/RCV2 corpus (Lewis et al., 2004) where each document is annotated with labels from 4 categories: CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social) and MCAT (Markets). We extend the evaluation from Klementiev et al. (2012) to cover more language pairs. We use the same data split for en→de and de→en pairs but additionally construct the train and test data for it→de, it→es and en→es. For each pair, we use 1,000 documents in the source language as the training data and 5,000 documents in the target language as the test data. The training data is randomly sampled, but the test data (for es) is evenly balanced among labels.

Table 3 shows the accuracy for the CLDC task for many pairs and models with respect to the baselines. For all bilingual models (Duong et al., 2016; Luong et al., 2015; Chandar A P et al., 2014), the bilingual word embeddings are constructed for each pair separately. In this way, they can only get the pairs involving English since there are many bilingual resources involving English on one side. For all our models, including Linear, Joint and Joint + Mapping, the embedding space is available for multiple languages; this is why we can exploit different relations, such as it→es. This is the motivation for the work reported in this paper. Suppose we want to build a document classifier for es but lack any annotations. It is common to build en-es crosslingual word embeddings for transfer learning, but this only achieves 53.8 % accuracy. Yet when we use it as the source, we get 81.0% accuracy. This is motivated by the fact that it and es are very similar.

The trend observed in Table 3 is consistent with previous observations. Linear transformation performs well. Joint training performs better especially for the it→de pair. The joint model with explicit mapping is generally our best model, even better than the base bilingual model from Duong et al. (2016). The de→en result improves on the existing state of the art reported in Luong et al. (2015). Our final model (Joint + Mapping) achieved competitive results compared with four strong baseline multilingual word embeddings, achieving best results for two out of five pairs. Moreover, the best scores for each language pairs are all from multilingual training, emphasizing the advantages over bilingual training.

9 Analysis

Mikolov et al. (2013b) showed that monolingual word embeddings capture some analogy relations such as $\vec{\text{Paris}} - \vec{\text{France}} + \vec{\text{Italy}} \approx \vec{\text{Rome}}$. It seems that in our multilingual embeddings, these relations still hold. Table 4 shows some examples of such relations where each word in the analogy query is in different languages.

All our baselines (MultiCluster, MultiCCA, MultiSkip, MultiTrans) are trained using different datasets. While MultiSkip and MultiTrans

chico$_{es}$ - bruder$_{de}$ + sorella$_{it}$ (boy - brother + sister)	ehemann$_{de}$ - padre$_{es}$ + madre$_{it}$ (husband - father + mother)	principe$_{it}$ - junge$_{de}$ + meisje$_{nl}$ (prince - boy + girl)
chica$_{es}$ (girl)	**echtgenote$_{nl}$** (wife)	**principessa$_{it}$** (princess)
ragazza$_{it}$ (girl)	**moglie$_{it}$** (wife)	**princess$_{en}$**
meisje$_{nl}$ (girl)	her$_{en}$	**princesa$_{es}$** (princess)
girl$_{en}$	marito$_{it}$ (husband)	príncipe$_{es}$ (prince)
mädchen$_{de}$ (girl)	haar$_{nl}$ (her)	**prinzessin$_{de}$** (princess)

Table 4: Top five closest words in our embeddings for multilingual word analogy. The transliteration is provided in parentheses. The correct output is bold.

	Tasks	MultiCluster	MultiCCA	Our model
Extrinsic	multilingual Dependency Parsing	61.0	58.7	**61.2**
	multilingual Document Classification	**92.1**	**92.1**	90.8
Intrinsic	monolingual word similarity	38.0	**43.0**	40.9
	multilingual word similarity	58.1	66.6	**69.8**
	word translation	43.7	35.7	**45.7**
	monolingual QVEC	10.3	10.7	**11.9**
	multilingual QVEC	**9.3**	8.7	8.6
	monolingual QVEC-CCA	62.4	**63.4**	46.4
	multilingual QVEC-CCA	**43.3**	41.5	31.0

Table 5: Performance of our model compared with MultiCluster and MultiCCA using extrinsic and intrinsic evaluation tasks on 12 languages proposed in Ammar et al. (2016), all models are trained on the same dataset. The best score for each task is bold.

are trained on parallel corpora, MultiCluster and MultiCCA use monolingual corpora and bilingual lexicons which are similar to our proposed methods. Therefore, for a strict comparison[8], we train our best model (Joint + Mapping) using the same monolingual data and set of bilingual lexicons on the same 12 languages with MultiCluster and MultiCCA. Table 5 shows the performance on intrinsic and extrinsic tasks proposed in Ammar et al. (2016). Multilingual dependency parsing and document classification are trained on a set of source languages and test on a target language in the transfer learning setting. Monolingual word similarity task is similar with our monolingual similarity task described in §7, multilingual word similarity is an extension of monolingual word similarity task but tested for pair of words in different languages. Monolingual QVEC, multilingual QVEC test the linguistic content of word embeddings in monolingal and multilingual setting. Monolingual QVEC-CCA and multilingual QVEC-CCA are the extended versions of monolingual QVEC and multilingual QVEC also proposed in Ammar et al. (2016). Table 5 shows that our model achieved competitive results, best at 4 out of 9 evaluation tasks.

[8]also with respect to the word coverage since MultiSkip and MultiTrans usually have much lower word coverage, biasing the intrinsic evaluations.

10 Conclusion

In this paper, we introduced several methods for building unified multilingual word embeddings. These represent an improvement because they exploit more relations and combine information from many languages. The input to our model is just a set of monolingual data and a set of bilingual lexicons between any language pairs. We induce the bilingual relationship for all language pairs while keeping high quality monolingual relations. Our multilingual joint training model with explicit mapping consistently achieves better performance compared with linear transformation. We achieve new state-of-the-art performance on bilingual lexicon induction task for recall at 5, similar excellent results with the state-of-the-art bilingual word embeddings on monolingual similarity task (Duong

et al., 2016). Moreover, our model is competitive at the crosslingual document classification task, achieving a new state of the art for de→en and it→de pair.

Acknowledgments

This work was conducted during Duong's internship at IBM Research Tokyo and partially supported by the University of Melbourne and National ICT Australia (NICTA). We are grateful for support from NSF Award 1464553 and the DARPA/I2O, Contract Nos. HR0011-15-C-0114 and HR0011-15-C-0115. We thank Yuta Tsuboi and Alvin Grissom II for helpful discussions, and Doris Hoogeveen for helping with the nl-es evaluation.

References

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR*, abs/1602.01925.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.*, 19(2):263–311, June.

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 Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1853–1861. Curran Associates, Inc.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA. ACM.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Transgram, fast cross-lingual word-embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1113, Lisbon, Portugal, September. Association for Computational Linguistics.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 600–609.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015a. Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 113–122, Beijing, China, July. Association for Computational Linguistics.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015b. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 845–850, Beijing, China. Association for Computational Linguistics.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning Crosslingual Word Embeddings without Bilingual Corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, Austin, Texas, USA, November. Association for Computational Linguistics.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343, Beijing, China, July. Association for Computational Linguistics.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden, April. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 406–414, New York, NY, USA. ACM.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In David Blei and Francis Bach, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 748–756. JMLR Workshop and Conference Proceedings.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2734–2740. AAAI Press.

Kejun Huang, Matt Gardner, Evangelos Papalexakis, Christos Faloutsos, Nikos Sidiropoulos, Tom Mitchell, Partha P. Talukdar, and Xiao Fu. 2015. Translation invariant word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1084–1088, Lisbon, Portugal, September. Association for Computational Linguistics.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–50, Reykjavik, Iceland. European Language Resources Association (ELRA).

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India, December. The COLING 2012 Organizing Committee.

Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 224–229, Baltimore, Maryland, June. Association for Computational Linguistics.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, December.

Thang Luong, Richard Socher, and Christopher D. Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, pages 104–113.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *NAACL Workshop on Vector Space Modeling for NLP*, Denver, United States.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 62–72.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. 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, Atlanta, Georgia. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual nlp. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1713–1722, Beijing, China, July. Association for Computational Linguistics.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 477–487. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 719–725, Beijing, China, July. Association for Computational Linguistics.

Mengqiu Wang, Wanxiang Che, and Christopher D. Manning. 2013. Joint word alignment and bilingual named entity recognition using dual decomposition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1082, Sofia, Bulgaria, August. Association for Computational Linguistics.

Min Xiao and Yuhong Guo, 2014. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, chapter Distributed Word Representation Learning for Cross-Lingual Dependency Parsing, pages 119–129. Association for Computational Linguistics.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8, Pittsburgh, Pennsylvania.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California, June. Association for Computational Linguistics.