Cross-lingual Models of Word Embeddings: An Empirical Comparison

Shyam Upadhyay[1] **Manaal Faruqui**[2] **Chris Dyer**[2] **Dan Roth**[1]
[1] Department of Computer Science, University of Illinois, Urbana-Champaign, IL, USA
[2] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
upadhya3@illinois.edu, mfaruqui@cs.cmu.edu
cdyer@cs.cmu.edu, danr@illinois.edu

Abstract

Despite interest in using cross-lingual knowledge to learn word embeddings for various tasks, a systematic comparison of the possible approaches is lacking in the literature. We perform an extensive evaluation of four popular approaches of inducing cross-lingual embeddings, each requiring a different form of supervision, on four typographically different language pairs. Our evaluation setup spans four different tasks, including intrinsic evaluation on mono-lingual and cross-lingual similarity, and extrinsic evaluation on downstream semantic and syntactic applications. We show that models which require expensive cross-lingual knowledge almost always perform better, but cheaply supervised models often prove competitive on certain tasks.

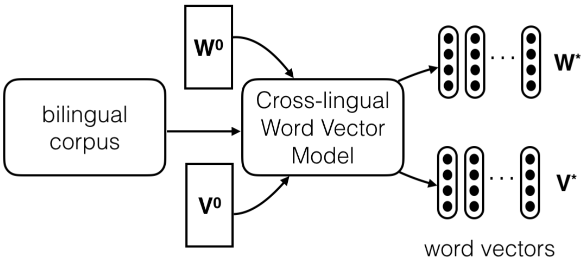
1 Introduction

Learning word vector representations using mono-lingual distributional information is now a ubiquitous technique in NLP. The quality of these word vectors can be significantly improved by incorporating cross-lingual distributional information (Klementiev et al., 2012; Zou et al., 2013; Vulić and Moens, 2013b; Mikolov et al., 2013b; Faruqui and Dyer, 2014; Hermann and Blunsom, 2014; Lauly et al., 2014, *inter alia*), with improvements observed both on mono-lingual (Faruqui and Dyer, 2014; Rastogi et al., 2015) and cross-lingual tasks (Guo et al., 2015; Søgaard et al., 2015; Guo et al., 2016).

Several models for inducing cross-lingual embeddings have been proposed, each requiring a different form of cross-lingual supervision – some can use document-level alignments (Vulić and Moens, 2015), others need alignments at the sentence (Hermann and Blunsom, 2014; Gouws et al., 2015) or word (Faruqui and Dyer, 2014; Gouws and Søgaard, 2015) level, while some require both sentence and word alignments (Luong et al., 2015). However, a systematic and extensive comparison of these models is missing from the literature, making it difficult to analyse which approach is suitable for a particular NLP task. In this paper, we fill this void by empirically comparing four cross-lingual word embedding models each of which require different form of alignment(s) as supervision, across several dimensions. To this end, we train these models on four different language pairs, and evaluate them on both mono-lingual and cross-lingual tasks. We will release our comparison setup and trained embeddings upon publication.

First, we show that different models can be viewed as instances of a more general framework for inducing bilingual word embeddings. Then, we evaluate these models on both extrinsic and intrinsic tasks. Our intrinsic evaluation assesses the quality of the vectors on mono-lingual (§4.2) and cross-lingual (§4.3) word similarity tasks, while our extrinsic evaluation spans semantic (cross-lingual document classification (§4.4)) and syntactic (cross-lingual dependency parsing (§4.5)) tasks.

Our experiments show that word vectors trained using expensive cross-lingual supervision (word alignments or sentence alignments) perform the best on semantic tasks. On the other hand, for syntactic tasks like cross-lingual dependency parsing, models requiring weaker form of cross-lingual supervision (such as context agnostic translation dictionary) are competitive to models requiring expensive supervision. We also analyze the embeddings qualitatively with the aim of revealing latent regularities in the vector space which explain why some models do better than others.



Algorithm 1 General Algorithm

1: Initialize $\mathbf{W} \leftarrow \mathbf{W}^0, \mathbf{V} \leftarrow \mathbf{V}^0$
2: $(\mathbf{W}^*, \mathbf{V}^*) \leftarrow \arg\min \alpha A(\mathbf{W}) + \beta B(\mathbf{V}) + C(\mathbf{W}, \mathbf{V})$

Figure 1: (**Above**) A general schema for induction of cross-lingual word vector representations, when provided with bilingual supervision for a language pair. The word vector model generates embeddings which incorporates distributional information *cross-lingually*. (**Below**) A general algorithm for inducing bilingual word embeddings, where $\alpha, \beta, \mathbf{W}^0, \mathbf{V}^0$ are parameters and A, B, C are suitably defined losses.

2 Bilingual Embeddings

A general schema for inducing bilingual embeddings is shown in Figure 1. Our comparison focuses on dense, fixed-length distributed embeddings which are obtained using some form of cross-lingual supervision.[1] We briefly describe the embedding induction procedure for each of the selected bilingual word vector models, with the aim to provide a unified algorithmic perspective for all methods, to facilitate better understanding and comparison. Our choice of models spans different forms of supervision required for inducing the embeddings, illustrated in Figure 2.

Notation. Let $W = \{w_1, w_2, \ldots, w_{|W|}\}$ be the vocabulary of a language l_1 with $|W|$ words, and $\mathbf{W} \in \mathbb{R}^{|W| \times l}$ be the corresponding word embeddings of length l. Let $V = \{v_1, v_2, \ldots, v_{|V|}\}$ be the vocabulary of another language l_2 with $|V|$ words, and $\mathbf{V} \in \mathbb{R}^{|V| \times m}$ the corresponding word embeddings of length m. We sometimes abuse notation and denote a word and its vector with boldface $\mathbf{v}$ (or $\mathbf{w}$).

2.1 Bilingual Skip-Gram Model (BiSkip)

Luong et al. (2015) proposed Bilingual Skip-Gram, a simple extension of the mono-lingual skip-gram model, which learns bilingual embeddings by using a parallel corpus along with word alignments (both sentence and word level alignments).

The learning objective is a simple extension of the skip-gram model, where the context of a word is expanded to include bilingual links obtained from word alignments, so that the model is trained to predict words cross-lingually. In particular, given a word alignment link from word $\mathbf{v} \in \mathbf{V}$ in language l_1 to $\mathbf{w} \in \mathbf{W}$ in language l_2, the model predicts the context words of $\mathbf{w}$ using $\mathbf{v}$ and vice-versa. Formally, the cross lingual part of the objective is,

$$D_{12}(\mathbf{W}, \mathbf{V}) = -\sum_{(\mathbf{v},\mathbf{w}) \in \mathbf{Q}} \sum_{\mathbf{w_c} \in \textsc{Nbr}_2(\mathbf{w})} \log P(\mathbf{w_c} \mid \mathbf{v}) \tag{1}$$

Where $\textsc{Nbr}_2(\mathbf{w})$ is the context of $\mathbf{w}$ in language l_2, $\mathbf{Q}$ is the set of word alignments, and $P(\mathbf{w_c} \mid \mathbf{v}) \propto \exp(\mathbf{w}_c^T \mathbf{v})$. Another similar term D_{21} models the objective for $\mathbf{v}$ and $\textsc{Nbr}_1(\mathbf{v})$. The objective can be cast into Algorithm 1 as,

$$C(\mathbf{W}, \mathbf{V}) = D_{12}(\mathbf{W}, \mathbf{V}) + D_{21}(\mathbf{W}, \mathbf{V}) \tag{2}$$

$$A(\mathbf{W}) = -\sum_{\mathbf{w} \in \mathbf{W}} \sum_{\mathbf{w_c} \in \textsc{Nbr}_2(\mathbf{w})} \log P(\mathbf{w_c} \mid \mathbf{w}) \tag{3}$$

$$B(\mathbf{V}) = -\sum_{\mathbf{v} \in \mathbf{V}} \sum_{\mathbf{v_c} \in \textsc{Nbr}_1(\mathbf{v})} \log P(\mathbf{v_c} \mid \mathbf{v}) \tag{4}$$

Where $A(\mathbf{W})$ and $B(\mathbf{V})$ are the familiar skip-gram formulation of the mono-lingual part of the objective.

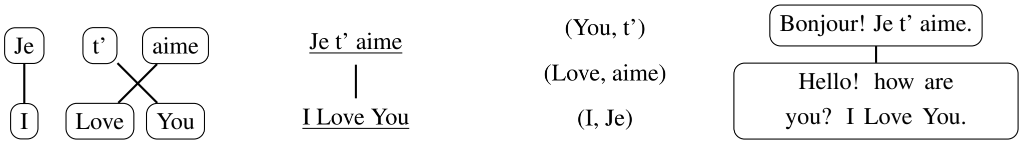
2.2 Bilingual Compositional Model (BiCVM)

Hermann and Blunsom (2014) present a method that learns bilingual word vectors from a sentence aligned corpus. Their model leverages the fact that aligned sentences have equivalent meaning, thus their sentence representations should be similar.

We denote two aligned sentences, $\vec{v} = \langle x_1, \ldots, \rangle$ and $\vec{w} = \langle y_1, \ldots \rangle$, where $x_i \in \mathbf{V}, y_i \in \mathbf{W}$, are vectors corresponding to the words in the sentences. Let functions $f : \vec{v} \to \mathbb{R}^n$ and $g : \vec{w} \to \mathbb{R}^n$, map sentences to their semantic representations in $\mathbb{R}^n$. BiCVM generates word vectors by minimizing the squared ℓ_2 norm between the sentence representations of aligned sentences. In order to prevent the degeneracy arising from directly minimizing the ℓ_2 norm, they use a noise-contrastive large-margin update, with randomly drawn sentence pairs $(\vec{v}, \vec{w}^n)$ as negative samples. The loss for the sentence pairs $(\vec{v}, \vec{w})$ and $(\vec{v}, \vec{w}^n)$ can be written as,

$$E(\vec{v}, \vec{w}, \vec{w}^n) = \max\left(\delta + \Delta E(\vec{v}, \vec{w}, \vec{w}^n), 0\right) \tag{5}$$

[1] We compare different cross-lingual word embeddings, which are not to be confused with a collection of mono-lingual word embeddings trained for different languages individually (Al-Rfou et al., 2013).



(a) **BiSkip** (b) **BiCVM** (c) **BiCCA** (d) **BiVCD**

Figure 2: Form of supervision required by the four models compared in this paper. From left to right, the cost of supervision required varies from expensive (BiSkip) to cheap (BiVCD). BiSkip requires a parallel corpus annotated with word alignments (Fig. 2a), BiCVM requires a sentence-aligned corpus (Fig. 2b), BiCCA only requires a bilingual lexicon (Fig. 2c) and BiVCD requires comparable documents (Fig. 2d).

where,

$$E(\vec{v},\vec{w}) = \|f(\vec{v}) - g(\vec{w})\|^2 \tag{6}$$

and,

$$\Delta E(\vec{v},\vec{w},\vec{w}^n) = E(\vec{v},\vec{w}) - E(\vec{v},\vec{w}^n) \tag{7}$$

This can be cast into Algorithm 1 by,

$$C(\mathbf{W},\mathbf{V}) = \sum_{\substack{\text{aligned } (\vec{v},\vec{w}) \\ \text{random } \vec{w}^n}} E(\vec{v},\vec{w},\vec{w}^n) \tag{8}$$

$$A(\mathbf{W}) = \|\mathbf{W}\|^2 \quad B(\mathbf{V}) = \|\mathbf{V}\|^2 \tag{9}$$

with $A(\mathbf{W})$ and $B(\mathbf{V})$ being regularizers, with $\alpha = \beta$.

2.3 Bilingual Correlation Based Embeddings (BiCCA)

The BiCCA model, proposed by Faruqui and Dyer (2014), showed that when (independently trained) mono-lingual vector matrices $\mathbf{W},\mathbf{V}$ are projected using CCA (Hotelling, 1936) to respect a translation lexicon, their performance improves on word similarity and word analogy tasks. They first construct $\mathbf{W}' \subseteq \mathbf{W}, \mathbf{V}' \subseteq \mathbf{V}$ such that $|\mathbf{W}'| = |\mathbf{V}'|$ and the corresponding words (w_i, v_i) in the matrices are translations of each other. The projection is then computed as:

$$\mathbf{P}_W, \mathbf{P}_V = \mathrm{CCA}(\mathbf{W}', \mathbf{V}') \tag{10}$$

$$\mathbf{W}^* = \mathbf{W}\mathbf{P}_W \quad \mathbf{V}^* = \mathbf{V}\mathbf{P}_V \tag{11}$$

where, $\mathbf{P}_V \in \mathbb{R}^{l\times d}, \mathbf{P}_W \in \mathbb{R}^{m\times d}$ are the projection matrices with $d = \min(l, m)$ and the $\mathbf{V}^* \in \mathbb{R}^{|V|\times d}, \mathbf{W}^* \in \mathbb{R}^{|W|\times d}$ are the word vectors that have been "enriched" using bilingual knowledge.

The BiCCA objective can be viewed[2] as the following instantiation of Algorithm 1

$$\mathbf{W}^0 = \mathbf{W}', \mathbf{V}^0 = \mathbf{V}' \tag{12}$$

$$C(\mathbf{W},\mathbf{V}) = \|\mathbf{W} - \mathbf{V}\|^2 + \gamma\left(\mathbf{V}^T\mathbf{W}\right) \tag{13}$$

$$A(\mathbf{W}) = \|\mathbf{W}\|^2 - 1 \quad B(\mathbf{V}) = \|\mathbf{V}\|^2 - 1 \tag{14}$$

where $\mathbf{W} = \mathbf{W}^0\mathbf{P}_W$ and $\mathbf{V} = \mathbf{V}^0\mathbf{P}_V$, where we set $\alpha = \beta = \gamma = \infty$ to set hard constraints.

2.4 Bilingual Vectors from Comparable Data (BiVCD)

Another approach of inducing bilingual word vectors, which we refer to as BiVCD, was proposed by Vulić and Moens (2015). Their approach is designed to use *comparable* corpus between the source and target language pair to induce cross-lingual vectors.

Let d_e and d_f denote a pair of comparable documents with length p and q respectively (assume $p > q$). BiVCD first merges these two comparable documents into a single pseudo-bilingual document using a deterministic strategy based on length ratio of two documents $R = \lfloor \frac{p}{q} \rfloor$. Every R^{th} word of the merged pseudo-bilingual document is picked sequentially from d_f. Finally, a skip-gram model is trained on the corpus of pseudo-bilingual documents, to generate vectors for all words in $\mathbf{W}^* \cup \mathbf{V}^*$. The vectors constituting $\mathbf{W}^*$ and $\mathbf{V}^*$ can then be easily identified.

Instantiation of BiVCD in the general algorithm is obvious: $C(\mathbf{W},\mathbf{V})$ assumes the familiar word2vec skip-gram objective over the pseudo-bilingual document,

$$C(\mathbf{W},\mathbf{V}) = -\sum_{\mathbf{s}\in\mathbf{W}\cup\mathbf{V}} \sum_{\mathbf{t}\in\textsc{Nbr}(\mathbf{s})} \log P(\mathbf{t} \mid \mathbf{s}) \tag{15}$$

where $\textsc{Nbr}(\mathbf{s})$ is defined by the pseudo-bilingual document and $P(\mathbf{t} \mid \mathbf{s}) \propto \exp(\mathbf{t}^T\mathbf{s})$.

Although BiVCD is designed to use comparable corpus, we provide it with parallel data in our experiments (to ensure comparability), and treat two aligned sentences as comparable.

[2]Canonical correlation versus distance, Section 6.5 (Hardoon et al., 2004)

l_1	l_2	#sent	#l_1-words	#l_2-words
en	de	1.9	53	51
	fr	2.0	55	61
	sv	1.7	46	42
	zh	2.0	58	50

Table 1: The size of parallel corpora (in millions) of different language pairs used for training multilingual word vectors.

3 Data

We train cross-lingual embeddings for 4 language pairs: English-German (en-de), English-French (en-fr), English-Swedish (en-sv) and English-Chinese (en-zh). For en-de and en-sv we use the Europarl v7 parallel corpus[3] (Koehn, 2005). For en-fr, we use Europarl combined with the news-commentary and UN-corpus dataset from WMT 2015.[4] For en-zh, we use the FBIS parallel corpus from the news domain (LDC2003E14). We use the Stanford Chinese Segmenter (Tseng et al., 2005) to preprocess the en-zh parallel corpus. Corpus statistics for all languages is shown in Table 1.

4 Evaluation

We measure the quality of induced cross-lingual word embeddings in terms of their performance, when used as features in the following tasks:

- Mono-lingual word similarity for English
- Cross-lingual dictionary induction
- Cross-lingual document classification
- Cross-lingual syntactic dependency parsing

The first two tasks intrinsically measure how much mono-lingual and cross-lingual similarity benefit from cross-lingual training. The last two tasks measure the ability of bilingually trained vectors to extrinsically facilitate model transfer across languages, for semantic and syntactic applications respectively. These tasks have been used in previous works (Klementiev et al., 2012; Luong et al., 2015; Vulić and Moens, 2013a; Guo et al., 2015) for evaluating bilingual embeddings, but no comparison exists which uses them in conjunction.

To ensure fair comparison, all models are trained with embeddings of size 200. We provide all models with parallel corpora, irrespective of their requirements. Whenever possible, we also report statistical significance of our results.

4.1 Parameter Selection

We follow the **BestAvg** parameter selection strategy from Lu et al. (2015). We selected the parameters for all models by tuning on a set of values (described below) and picking the parameter setting which did best on an average across all tasks.

BiSkip. All models were trained using a window size of 10 (tuned over $\{5, 10, 20\}$), and 30 negative samples (tuned over $\{10, 20, 30\}$). The cross-lingual weight was set to 4 (tuned over $\{1, 2, 4, 8\}$). The word alignments for training the model (available at `github.com/lmthang/bivec`) were generated using cdec (Dyer et al., 2010). The number of training iterations was set to 5 (no tuning).

BiCVM. We use the tool (available at `github.com/karlmoritz/bicvm`) released by Hermann and Blunsom (2014) to train all embeddings. We train an additive model (that is, $f(\vec{x}) = g(\vec{x}) = \sum_i x_i$) with hinge loss margin set to 200 (no tuning), batch size of 50 (tuned over $50, 100, 1000$) and noise parameter of 10 (tuned over $\{10, 20, 30\}$). All models are trained for 100 iterations (no tuning).

BiCCA. First, mono-lingual word vectors are trained using the skip-gram model[5] with negative sampling (Mikolov et al., 2013a) with window of size 5 (tuned over $\{5, 10, 20\}$). To generate a bilingual dictionary, word alignments are generated using cdec from the parallel corpus. Then, word pairs $(a, b), a \in l_1, b \in l_2$ are selected such that a is aligned to b the most number of times and vice versa. This way, we obtained dictionaries of approximately 36k, 35k, 30k and 28k word pairs for en-de, en-fr, en-sv and en-zh respectively.

The mono-lingual vectors are aligned using the above dictionaries with the tool (available at `github.com/mfaruqui/eacl14-_cca`) released by Faruqui and Dyer (2014) to generate the bilingual word embeddings. We use $k = 0.5$ as the number of canonical components (tuned over $\{0.2, 0.3, 0.5, 1.0\}$). Note that this results in a embedding of size 100 after performing CCA.

BiVCD. We use word2vec's skip gram model for training our embeddings, with a window size

[3] `www.statmt.org/europarl/v7/{de,_sv}-_en.tgz`

[4] `www.statmt.org/wmt15/translation-_task.html`

[5] `code.google.com/p/word2vec`

of 5 (tuned on $\{5, 10, 20, 30\}$) and negative sampling parameter set to 5 (tuned on $\{5, 10, 25\}$). Every pair of parallel sentences is treated as a pair of comparable documents, and merging is performed using the sentence length ratio strategy described earlier. We implemented the code for performing the merging as we could not find a tool provided by the authors.

4.2 Mono-lingual Evaluation

We first evaluate if the inclusion of cross-lingual knowledge using bilingual models improves the quality of English embeddings.

Word Similarity. Word similarity datasets contain word pairs which are assigned similarity ratings by humans. The task evaluates how well the notion of word similarity according to humans is emulated in the vector space. Evaluation is based on the Spearman's rank correlation coefficient (Myers and Well, 1995) between human rankings and rankings produced by computing cosine similarity between the vectors of two words.

We use SimLex dataset for English (Hill et al., 2014) which contains 999 pairs of English words, with a balanced set of noun, adjective and verb pairs. SimLex is claimed to capture word similarity exclusively instead of WordSim-353 (Finkelstein et al., 2001) which captures both word simiairty and relatedness. We declare significant improvement if $p < 0.1$ according to Steiger's method (Steiger, 1980) for calculating the statistical significant differences between two dependent correlation coefficients.

Table 2 shows performance of English embeddings induced by all the models by training on different language pairs on the SimLex word similarity task. Overall, across all language pairs, BiCVM is the best performing model in terms of spearman's correlation, but its improvement over BiSkip and BiVCD is often insignificant. It is notable that 2 of the 3 top performing models, BiCVM and BiVCD, need sentence aligned and document-aligned corpus only, which are easier to obtain than parallel data with word alignments required by BiSkip.

QVEC. Tsvetkov et al. (2015) proposed an intrinsic evaluation metric for estimating the quality of English word vectors. The score produced by QVEC measures how well a given set of word vectors is able to quantify linguistic properties of words, with higher being better. The metric is shown to have strong correlation with performance on downstream semantic applications. As it can be only used for English, we use it to evaluate the English vectors obtained using cross-lingual training of different models. Table 3 shows that on average across language pairs, BiSkip achieves the best score, followed by BiVCD and BiCCA. Interestingly, BiCVM which was the best model according to SimLex, ranks last overall according to QVEC. The fact that the best models according to QVEC and word similarities are different reinforces the observation of Tsvetkov et al. (2015) that performance on word similarity tasks alone does not reflect quantification of linguistic properties of words.

pair	BiSkip	BiCVM	BiCCA	BiVCD
en-de	0.34	**0.37**	0.30	0.32
en-fr	0.35	**0.39**	0.31	0.36
en-sv	0.32	**0.34**	0.27	0.32
en-zh	0.34	**0.39**	0.30	0.31
avg.	0.34	**0.37**	0.30	0.33

Table 2: Word similarity score measured in Spearman's correlation ratio for English on SimLex-999. The best score for each language pair is shown in **bold**. Scores which are significantly better (per Steiger's Method with $p < 0.1$) than the next lower score are underlined. For example, for en-zh, BiCVM is significantly better than BiSkip, which in turn is significantly better than BiVCD.

pair	BiSkip	BiCVM	BiCCA	BiVCD
en-de	**0.40**	0.31	0.33	0.37
en-fr	**0.40**	0.31	0.33	0.38
en-sv	**0.39**	0.31	0.32	0.37
en-zh	**0.40**	0.32	0.33	0.38
avg.	**0.40**	0.31	0.33	0.38

Table 3: Intrinsic evaluation of English word vectors measured in terms of QVEC score across models. Best scores for each language pair is shown in **bold**.

4.3 Cross-lingual Dictionary Induction

The task of cross-lingual dictionary induction (Vulić and Moens, 2013a; Gouws et al., 2015; Mikolov et al., 2013b) judges how good bilingual embeddings are at detecting word pairs that are semantically similar across languages. We follow the setup of Vulić and Moens (2013a), but instead of manually creating a gold bilingual dictionary, we derived our gold dictionaries using the Open Multilingual Wordnet data released by Bond and Foster (2013). The data includes synset alignments across 26 languages with over 90% accuracy. First, we prune out words from each synset whose frequency count is less than 1000 in the vocabulary of the training data from §3. Then, for

l_1	l_2	BiSkip	BiCVM	BiCCA	BiVCD
en	de	**79.7**	74.5	72.4	62.5
	fr	**78.9**	72.9	70.1	68.8
	sv	**77.1**	76.7	74.2	56.9
	zh	**69.4**	66.0	59.6	53.2
avg.		**76.3**	72.5	69.1	60.4

Table 4: Cross-lingual dictionary induction results (top-10 accuracy). The same trend was also observed across models when computing MRR (mean reciprocal rank).

each pair of aligned synsets $s_1 = \{k_1, k_2, \cdots\}$ $s_2 = \{g_1, g_2, \cdots\}$, we include all elements from the set $\{(k, g) \mid k \in s_1, g \in s_2\}$ into the gold dictionary, where k and g are the lemmas. Using this approach we generated dictionaries of sizes 1.5k, 1.4k, 1.0k and 1.6k pairs for en-fr, en-de, en-sv and en-zh respectively.[6]

We report top-10 accuracy, which is the fraction of the entries (e, f) in the gold dictionary, for which f belongs to the list of top-10 neighbours of the word vector of e, according to the induced bilingual embeddings. From the results (Table 4), it can be seen that for dictionary induction, the performance improves with the quality of supervision. As we move from cheaply supervised methods (eg. BiVCD) to more expensive supervision (eg. BiSkip), the accuracy improves. This suggests that for cross lingual similarity tasks, the more expensive the cross-lingual knowledge available, the better. Models using weak supervision like BiVCD perform poorly in comparison to models like BiSkip and BiCVM, with performance gaps upwards of 10 pts on an average.

4.4 Cross-lingual Document Classification

We follow the cross-lingual document classification (CLDC) setup of Klementiev et al. (2012), but extend it to cover all of our language pairs. We use the RCV2 Reuters multilingual corpus[7] for our experiments. In this task, for a language pair (l_1, l_2), a document classifier is trained using the document representations derived from word embeddings in language l_1, and then the trained model is tested on documents from language l_2 (and vice-versa). By using supervised training data in one language and evaluating without further supervision in another, CLDC assesses whether the learned multilingual representations are semantically coherent across multiple languages.

[6]We will release these dictionaries.

[7]`http://trec.nist.gov/data/reuters/reuters.html`

l_1	l_2	BiSkip	BiCVM	BiCCA	BiVCD
en	de	**85.2**	85.0	79.1	79.9
	fr	**77.7**	71.7	70.7	72.0
	sv	**72.3**	69.1	65.3	59.9
	zh	**75.5**	73.6	69.4	73.0
de	en	**74.9**	71.1	64.9	74.1
fr		**80.4**	73.7	75.5	77.6
sv		73.4	67.7	67.0	**78.2**
zh		**81.1**	76.4	77.3	80.9
avg.		**77.6**	73.5	71.2	74.5

Table 5: Cross-lingual document classification accuracy when trained on language l_1, and evaluated on language l_2. The best score for each language is shown in **bold**. Scores which are significantly better (per McNemar's Test with $p < 0.05$) than the next lower score are underlined. For example, for sv→en, BiVCD is significantly better than BiSkip, which in turn is significantly better than BiCVM.

All embeddings are learned on the data described in §3, and we only use the RCV2 data to learn document classification models. Following previous work, we compute document representation by taking the tf-idf[8] weighted average of vectors of the words present in it. A multi-class classifier is trained using an averaged perceptron (Freund and Schapire, 1999) for 10 iterations, using the document vectors of language l_1 as features[9]. Majority baselines for $en \rightarrow l_2$ and $l_1 \rightarrow en$ are 49.7% and 46.7% respectively, for all languages. Table 5 shows the performance of different models across different language pairs. We computed confidence values using the McNemar test (McNemar, 1947) and declare significant improvement if $p < 0.05$.

From Table 5, it is evident that in almost all the cases, BiSkip performs significantly better than the remaining models. For transferring semantic knowledge across languages via embeddings, sentence and word level alignment proves superior to sentence or word level alignment alone. This observation is consistent with the trend in cross-lingual dictionary induction, where too the most expensive form of supervision performed the best.

4.5 Cross-lingual Dependency Parsing

The use of cross lingual similarity measures for direct-transfer of dependency parsers was first shown in Täckström et al. (2012). The idea behind direct-transfer is to train a dependency parsing model using embeddings for language l_1 and

[8]tf-idf was computed using all documents for that language in RCV2

[9]We use the implementation of Klementiev et al. (2012)

then test the trained model on language l_2, replacing embeddings for language l_1 with those of l_2. The transfer relies on coherence of the embeddings across languages arising from the cross lingual training. For our experiments, we use the cross lingual transfer setup of Guo et al. (2015).[10] Their framework trains a transition-based dependency parser using nonlinear activation function, with the source-side embeddings as lexical features. These embeddings can be replaced by target-side embeddings at test time.

All models are trained for 5000 iterations with fixed word embeddings during training. Since our goal is to determine the utility of word embeddings in dependency parsing, we turn off other features that can capture distributional information like brown clusters, which were originally used in Guo et al. (2015). We use the universal dependency treebank (McDonald et al., 2013) version-2.0 for our evaluation. For Chinese, we use the treebank released as part of the CoNLL-X shared task (Buchholz and Marsi, 2006).

We first evaluate how useful the word embeddings are in cross-lingual model transfer of dependency parsers (Table 6). On an average, BiCCA does better than other models. BiSkip is a close second, with an average performance gap of less than 1 point. BiSkip outperforms BiCVM on German and French (over 2 point improvement), owing to word alignment information BiSkip's model uses during training. It is not surprising that English-Chinese transfer scores are low, due to the significant difference in syntactic structure of the two languages. Surprisingly, unlike the semantic tasks considered earlier, the models with expensive supervision requirements like BiSkip and BiCVM could not outperform a cheaply supervised BiCCA.

We also evaluate whether using bilingually trained vectors for learning dependency parsers is better than using mono-lingually trained vectors in Table 7. We compare against parsing models trained using mono-lingually trained word vectors (column marked Mono in Table 7). These vectors are the same mono-lingual used as input to the BiCCA model. All other settings remain the same. On an average across language pairs, improvement over the mono-lingual embeddings was obtained with the BiSkip and BiCCA models, while BiCVM and BiVCD consistently performed worse. A possible reason for this is that BiCVM and BiVCD operate on sentence level contexts to learn the embeddings, which only captures the semantic meaning of the sentences and ignores the internal syntactic structure. As a result, embedding trained using BiCVM and BiVCD are not informative for syntactic tasks. On the other hand, BiSkip and BiCCA both utilize the word alignment information to train their embeddings and thus do better in capturing some notion of syntax.

l_1	l_2	BiSkip	BiCVM	BiCCA	BiVCD
en	de	49.8	47.5	**51.3**	49.0
	fr	65.8	63.2	**65.9**	60.7
	sv	56.9	56.7	**59.4**	54.6
	zh	**6.4**	6.1	**6.4**	6.0
de	en	49.7	45.0	**50.3**	43.6
fr		53.3	50.6	**54.2**	49.5
sv		48.2	49.0	**49.9**	44.6
zh		**0.17**	0.12	**0.17**	0.15
avg.		41.3	39.8	**42.2**	38.5

Table 6: Labeled attachment score (LAS) for cross-lingual dependency parsing when trained on language l_1, and evaluated on language l_2. The best score for each language is shown in **bold**.

l	Mono	BiSkip	BiCVM	BiCCA	BiVCD
de	71.1	**72.0**	60.4	**71.4**	58.9
fr	78.9	**80.4**	73.7	**80.2**	69.5
sv	75.5	**78.2**	70.5	**79.0**	64.5
zh	73.8	73.1	65.8	71.7	67.0
avg.	74.8	**75.9**	67.6	**75.6**	66.8

Table 7: Labeled attachment score (LAS) for dependency parsing when trained and tested on language l. Mono refers to parser trained with mono-lingually induced embeddings. Scores in **bold** are better than the Mono scores for each language, showing improvement from cross-lingual training.

5 Qualitative Analysis

Figure 3 visualizes the word embeddings in a two-dimensional space using the t-SNE tool (Van der Maaten and Hinton, 2008). The plot shows projected embeddings of some randomly selected words from the most frequent words in the English segment of the parallel en-fr corpus, along with their French translations. We plot the BiCVM, BiSkip, and BiVCD vectors which are 200 dimensional.[11]

We can explain qualitatively why some models do better on the cross lingual similarity task using Figure 3. BiVCD embeddings moves similar pairs like $(children, enfants)$ and

[10] `github.com/jiangfeng1124/acl15-_clnndep`

[11] BiCCA could not be included as its 100 dimensional and is incompatible for the tool.

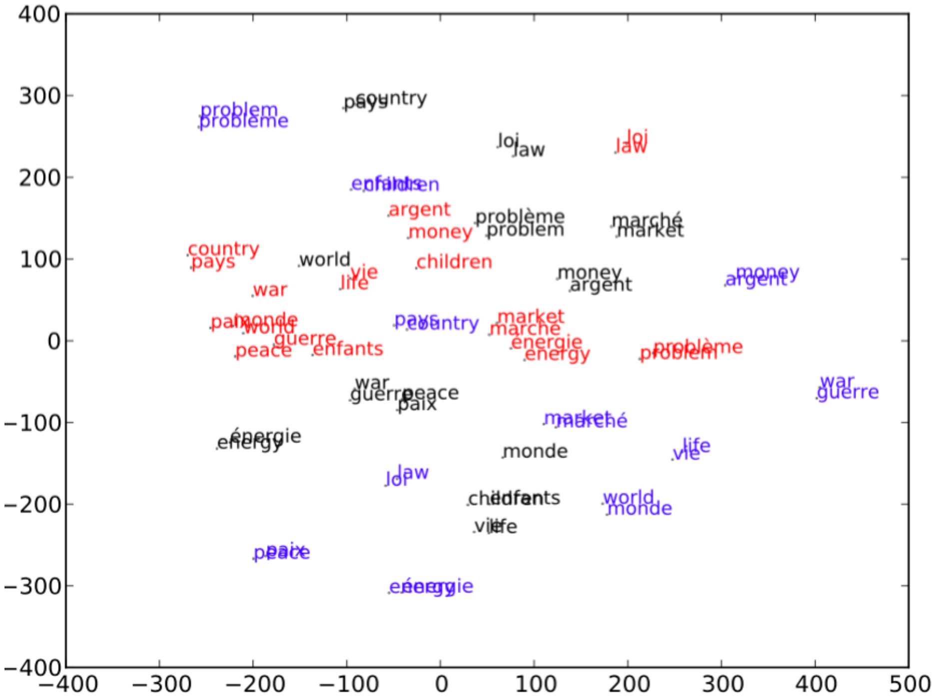


Figure 3: t-SNE visualisation of word embeddings of some frequent words present in en-fr corpus. BiCVM, BiSkip, and BiVCD vectors are shown in black, blue and red respectively.

$(war, guerre)$ apart, while BiCVM does the same with the pair $(world, monde)$, but overall is better than BiVCD. BiSkip does an almost perfect cross-lingual alignment, matching every word to its corresponding translation.

6 Discussion and Conclusion

We presented the first systematic comparative evaluation of cross-lingual embedding methods on several downstream NLP tasks, both intrinsic and extrinsic. We provided a unified representation for all approaches, showing them as instances of a general algorithm. Our choice of methods spans a diverse range of approaches, in that each requires a different form of alignment as supervision. It will be interesting to see how these embedding methods fare on other such tasks like frame semantic parsing (Johannsen et al., 2015), super-sense tagging (Gouws and Søgaard, 2015) etc.

The paper did not cover *all* approaches that generate cross-lingual word embeddings. Some methods do not have publicly available code (Coulmance et al., 2015; Zou et al., 2013); for others, like BilBOWA (Gouws et al., 2015), we identified problems in the available code, which caused it to consistently produced results that are inferior even to mono-lingually trained vectors.[12] However, the models that we included for comparison in our survey are representative of other cross-lingual models in terms of the form of cross-lingual supervision required by them. For example, BilBOWA (Gouws et al., 2015) and Bilingual Autoencoder (Lauly et al., 2014) are similar to BiCVM in this respect and Multi-view CCA (Rastogi et al., 2015) and deep CCA (Lu et al., 2015) can be viewed as extensions of BiCCA. Our choice of models was motivated to compare different forms of supervision, and therefore, adding these models, would not provide additional insight.

Our experiments reveal several interesting trends. When evaluating on intrinsic tasks such as mono-lingual word similarity, models (such as BiVCD) relying on cheaper forms of supervision perform almost on par (no statistically significant difference) with models requiring expensive supervision. On the other hand, for cross-lingual semantic tasks, like CLDC and dictionary induction, the model with the most informative supervision performs best overall. In contrast, for the syntactic task of dependency parsing, cheaply supervised models perform slightly better. It appears that for cross-lingual transfer of syntactic knowledge, word level alignment alone can suffice.

[12]We contacted the authors of the papers and were unable to resolve the issues in the toolkit.

References

[Al-Rfou et al.2013] Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proc. of CoNLL*.

[Bond and Foster2013] Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, August. Association for Computational Linguistics.

[Buchholz and Marsi2006] Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proc. of CoNLL*.

[Coulmance et al.2015] Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, 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.

[Dyer et al.2010] Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Hendra Setiawan, Ferhan Ture, Vladimir Eidelman, Phil Blunsom, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *In Proc. of ACL System Demonstrations*.

[Faruqui and Dyer2014] Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proc. of EACL*.

[Finkelstein et al.2001] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: the concept revisited. In *Proc. of WWW*.

[Freund and Schapire1999] Yoav Freund and Robert E Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296.

[Gouws and Søgaard2015] Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390, Denver, Colorado, May–June. Association for Computational Linguistics.

[Gouws et al.2015] Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proc. of ICML*.

[Guo et al.2015] Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proc. of ACL*.

[Guo et al.2016] Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *Proc. of AAAI*.

[Hardoon et al.2004] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664.

[Hermann and Blunsom2014] Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributional Semantics. In *Proc. of ACL*.

[Hill et al.2014] Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*.

[Hotelling1936] Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

[Johannsen et al.2015] Anders Johannsen, Héctor Martínez Alonso, and Anders Søgaard. 2015. Any-language frame-semantic parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2062–2066, Lisbon, Portugal, September. Association for Computational Linguistics.

[Klementiev et al.2012] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proc. of COLING*.

[Koehn2005] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. of MT summit*.

[Lauly et al.2014] Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Proc. of NIPS*.

[Lu et al.2015] Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of NAACL*.

[Luong et al.2015] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proc. of the Workshop on Vector Space Modeling for NLP*.

[McDonald et al.2013] Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proc. of ACL*.

[McNemar1947] Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

[Mikolov et al.2013a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

[Myers and Well1995] Jerome L. Myers and Arnold D. Well. 1995. *Research Design & Statistical Analysis*. Routledge.

[Rastogi et al.2015] Pushpendre Rastogi, Benjamin Van Durme, and Raman Arora. 2015. Multiview lsa: Representation learning via generalized cca. In *Proc. of NAACL*.

[Søgaard et al.2015] 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.

[Steiger1980] James H Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245.

[Täckström et al.2012] O. Täckström, R. McDonald, and J. Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *NAACL*.

[Tseng et al.2005] Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

[Tsvetkov et al.2015] Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proc. of EMNLP*.

[Van der Maaten and Hinton2008] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85.

[Vulić and Moens2013a] Ivan Vulić and Marie-Francine Moens. 2013a. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–116, Atlanta, Georgia, June. Association for Computational Linguistics.

[Vulić and Moens2013b] Ivan Vulić and Marie-Francine Moens. 2013b. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proc. of EMNLP*.

[Vulić and Moens2015] Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proc. of ACL*.

[Zou et al.2013] Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proc. of EMNLP*.