Learning Crosslingual Word Embeddings without Bilingual Corpora

Long Duong,[12] **Hiroshi Kanayama,**[3] **Tengfei Ma,**[3] **Steven Bird**[14] and **Trevor Cohn**[1]
[1]Department of Computing and Information Systems, University of Melbourne
[2]National ICT Australia, Victoria Research Laboratory
[3]IBM Research – Tokyo
[4]International Computer Science Institute, University of California Berkeley

Abstract

Crosslingual word embeddings represent lexical items from different languages in the same vector space, enabling transfer of NLP tools. However, previous attempts had expensive resource requirements, difficulty incorporating monolingual data or were unable to handle polysemy. We address these drawbacks in our method which takes advantage of a high coverage dictionary in an EM style training algorithm over monolingual corpora in two languages. Our model achieves state-of-the-art performance on bilingual lexicon induction task exceeding models using large bilingual corpora, and competitive results on the monolingual word similarity and cross-lingual document classification task.

1 Introduction

Monolingual word embeddings have had widespread success in many NLP tasks including sentiment analysis (Socher et al., 2013), dependency parsing (Dyer et al., 2015), machine translation (Bahdanau et al., 2014). Crosslingual word embeddings are a natural extension facilitating various crosslingual tasks, e.g. through transfer learning. A model built in a source resource-rich language can then applied to the target resource poor languages (Yarowsky and Ngai, 2001; Das and Petrov, 2011; Täckström et al., 2012; Duong et al., 2015). A key barrier for crosslingual transfer is lexical matching between the source and the target language. Crosslingual word embeddings are a natural remedy where both source and target language lexicon are presented as dense vectors in the same vector space (Klementiev et al., 2012).

Most previous work has focused on down-stream crosslingual applications such as document classification and dependency parsing. We argue that good crosslingual embeddings should preserve both monolingual and crosslingual quality which we will use as the main evaluation criterion through monolingual word similarity and bilingual lexicon induction tasks. Moreover, many prior work (Chandar A P et al., 2014; Kočiský et al., 2014) used bilingual or comparable corpus which is also expensive for many low-resource languages. Søgaard et al. (2015) impose a less onerous data condition in the form of linked Wikipedia entries across several languages, however this approach tends to underperform other methods. To capture the monolingual distributional properties of words it is crucial to train on large monolingual corpora (Luong et al., 2015). However, many previous approaches are not capable of scaling up either because of the complicated objective functions or the nature of the algorithm. Other methods use a dictionary as the bridge between languages (Mikolov et al., 2013a; Xiao and Guo, 2014), however they do not adequately handle translation ambiguity.

Our model uses a bilingual dictionary from Panlex (Kamholz et al., 2014) as the source of bilingual signal. Panlex covers more than a thousand languages and therefore our approach applies to many languages, including low-resource languages. Our method selects the translation based on the context in an Expectation-Maximization style training algorithm which explicitly handles polysemy through incorporating multiple dictionary translations (word sense and translation are closely linked (Resnik and Yarowsky, 1999)). In addition to the dictionary, our method only requires monolingual data. Our approach is an extension of the continuous bag-of-words (CBOW) model (Mikolov et al., 2013b) to inject multilingual training signal based on dictionary translations. We experiment with several variations of our model, whereby we predict only the translation or both word and its translation and consider different ways of using the different learned center-word versus context embeddings in application tasks. We also propose a regularisation method to combine the two embedding matrices during training. Together, these modifications substantially improve the performance across several tasks. Our final model achieves state-of-the-art performance on bilingual lexicon induction task, large improvement over word similarity task compared with previous published crosslingual word embeddings, and competitive result on cross-lingual document classification task. Notably, our embedding combining techniques are general, yielding improvements also for monolingual word embedding.

This paper makes the following contributions:

- Proposing a new crosslingual training method for learning vector embeddings, based only on monolingual corpora and a bilingual dictionary;
- Evaluating several methods for combining embeddings, which are shown to help in both crosslingual and monolingual evaluations; and
- Achieving consistent results which are competitive in monolingual, bilingual and crosslingual transfer settings.

2 Related work

There is a wealth of prior work on crosslingual word embeddings, which all exploit some kind of bilingual resource. This is often in the form of a parallel bilingual text, using word alignments as a bridge between tokens in the source and target languages, such that translations are assigned similar embedding vectors (Luong et al., 2015; Klementiev et al., 2012). These approaches are affected by errors from automatic word alignments, motivating other approaches which operate at the sentence level (Chandar A P et al., 2014; Hermann and Blunsom, 2014; Gouws et al., 2015) through learning compositional vector representations of sentences, in order that sentences and their translations representations closely match. The word embeddings learned this way capture translational equivalence, despite not using explicit word alignments. Nevertheless, these approaches demand large parallel corpora, which are not available for many language pairs.

Vulić and Moens (2015) use bilingual comparable text, sourced from Wikipedia. Their approach creates a psuedo-document by forming a bag-of-words from the lemmatized nouns in each comparable document concatenated over both languages. These pseudo-documents are then used for learning vector representations using `Word2Vec`. Their system, despite its simplicity, performed surprisingly well on a bilingual lexicon induction task (we compare our method with theirs on this task.) Their approach is compelling due to its lesser resource requirements, although comparable bilingual data is scarce for many languages. Related, Søgaard et al. (2015) exploit the comparable part of Wikipedia. They represent word using Wikipedia entries which are shared for many languages.

A bilingual dictionary is an alternative source of bilingual information. Gouws and Søgaard (2015) randomly replace the text in a monolingual corpus with a random translation, using this corpus for learning word embeddings. Their approach doesn't handle polysemy, as very few of the translations for each word will be valid in context. For this reason a high coverage or noisy dictionary with many translations might lead to poor outcomes. Mikolov et al. (2013a), Xiao and Guo (2014) and Faruqui and Dyer (2014) filter a bilingual dictionary for one-to-one translations, thus side-stepping the problem, however discarding much of the information in the dictionary. Our approach also uses a dictionary, however we use all the translations and explicitly disambiguate translations during training.

Another distinguishing feature on the above-cited research is the method for training embeddings. Mikolov et al. (2013a) and Faruqui and Dyer (2014) use a cascade style of training where the word embeddings in both source and target language are trained separately and then combined later using the dictionary. Most of the other works train multlingual models jointly, which appears to have better performance over cascade training (Gouws et al., 2015).

For this reason we also use a form of joint training in our work.

3 Word2Vec

Our model is an extension of the contextual bag of words (CBOW) model of Mikolov et al. (2013b), a method for learning vector representations of words based on their distributional contexts. Specifically, their model describes the probability of a token w_i at position i using logistic regression with a factored parameterisation,

$$p(w_i|w_{i\pm k\setminus i}) = \frac{\exp(\mathbf{u}_{w_i}^\top \mathbf{h}_i)}{\sum_{w\in W}\exp(\mathbf{u}_w^\top \mathbf{h}_i)}\,, \qquad (1)$$

where $\mathbf{h}_i = \frac{1}{2k}\sum_{j=-k;j\neq 0}^{k} \mathbf{v}_{w_{i+j}}$ is a vector encoding the context over a window of size k centred around position i, W is the vocabulary and the parameters $\mathbf{V}$ and $\mathbf{U} \in \mathbb{R}^{|W|\times d}$ are matrices referred to as the context and word embeddings. The model is trained to maximise the log-pseudo likelihood of a training corpus, however due to the high complexity of computing the denominator of equation (1), Mikolov et al. (2013b) propose negative sampling as an approximation, by instead learning to differentiate data from noise (negative examples). This gives rise to the following optimisation objective

$$\sum_{i\in D}\left(\log\sigma(\mathbf{u}_{w_i}^\top \mathbf{h}_i) + \sum_{j=1}^{p}\mathbb{E}_{w_j\sim P_n(w)}\log\sigma(-\mathbf{u}_{w_j}^\top \mathbf{h}_i)\right), \qquad (2)$$

where D is the training data and p is the number of negative examples randomly drawn from a noise distribution $P_n(w)$.

4 Our Approach

Our approach extends CBOW to model bilingual text, using two monolingual corpora and a bilingual dictionary. We believe this data condition to be less stringent than requiring parallel or comparable texts as the source of the bilingual signal. It is common for field linguists to construct a bilingual dictionary when studying a new language, as one of the first steps in the language documentation process. Translation dictionaries are a rich information source, capturing much of the lexical ambiguity in a language through translation. For example, the word *bank* in English might mean the *river bank* or *financial bank* which corresponds to two different translations *sponda* and *banca* in Italian. If we are able to learn to select good translations, then this implicitly resolves much of the semantic ambiguity in the language, and accordingly we seek to use this idea to learn better semantic vector representations of words.

Algorithm 1 EM algorithm for selecting translation during training, where $\theta = (\mathbf{U}, \mathbf{V})$ are the model parameters and η is the learning rate.

```
randomly initialize $\mathbf{V}$, $\mathbf{U}$
for $i <$ Iter do
    for $i \in D_e \cup D_f$ do
        $\mathbf{s} \leftarrow \mathbf{v}_{w_i} + \mathbf{h}_i$
        $\bar{w_i} = \text{argmax}_{w\in\text{dict}(w_i)} \cos(\mathbf{s}, \mathbf{v}_w)$
        $\theta \leftarrow \theta + \eta \frac{\partial \mathcal{O}(\bar{w_i}, w_i, \mathbf{h}_i)}{\partial \theta}$ {see (3) or (5)}
    end for
end for
```

4.1 Dictionary replacement

To learn bilingual relations, we use the context in one language to predict the translation of the centre word in another language. This is motivated by the fact that the context is an excellent means of disambiguating the translation for a word. Our method is closely related to Gouws and Søgaard (2015), however we only replace the middle word w_i with a translation $\bar{w_i}$ while keeping the context fixed. We replace each centre word with a translation on the fly during training, predicting instead $p(\bar{w_i}|w_{i\pm k\setminus i})$ but using the same formulation as equation (1) albeit with an augmented $\mathbf{U}$ matrix to cover word types in both languages.

The translation $\bar{w_i}$ is selected from the possible translations of w_i listed in the dictionary. The problem of selecting the correct translation from the many options is reminiscent of the problem faced in expectation maximisation (EM), in that cross-lingual word embeddings will allow for accurate translation, however to learn these embeddings we need to know the translations. We propose an EM-inspired algorithm, as shown in Algorithm 1, which operates over both monolingual corpora, D_e and D_f. The vector $\mathbf{s}$ is the semantic representation combining both the centre word, w_i, and the context,[1] which is used to choose the best translation into the other language from the bilingual dictionary $dict(w_i)$.[2] After selecting the translation, we use $\bar{w}_i$ together with the context vector $\mathbf{h}$ to make a stochastic gradient update of the CBOW log-likelihood.

4.2 Joint Training

Words and their translations should appear in very similar contexts. One way to enforce this is to jointly learn to predict both the word and its translation from its monolingual context. This gives rise to the following joint objective function,

$$\mathcal{O} = \sum_{i \in D_e \cup D_f} \left(\alpha \log \sigma(\mathbf{u}_{w_i}^\top \mathbf{h}_i) + (1-\alpha) \log \sigma(\mathbf{u}_{\bar{w}_i}^\top \mathbf{h}_i) + \sum_{j=1}^{p} \mathbb{E}_{w_j \sim P_n(w)} \log \sigma(-\mathbf{u}_{w_j}^\top \mathbf{h}_i) \right), \quad (3)$$

where α controls the contribution of the two terms. For our experiments, we set $\alpha = 0.5$. The negative examples are drawn from combined vocabulary unigram distribution calculated from combined data $D_e \cup D_f$.

4.3 Combining Embeddings

Many vector learning methods learn two embedding spaces $\mathbf{V}$ and $\mathbf{U}$. Usually only $\mathbf{V}$ is used in application. The use of $\mathbf{U}$, on the other hand, is understudied (Levy and Goldberg, 2014) with the exception of Pennington et al. (2014) who use a linear combination $\mathbf{U} + \mathbf{V}$, with minor improvement over $\mathbf{V}$ alone.

We argue that with our model, $\mathbf{V}$ is better at capturing the monolingual regularities and $\mathbf{U}$ is better at capturing bilingual signal. The intuition for this is as follows. Assuming that we are predicting the word *finance* and its Italian translation *finanze* from the context (*money, loan, bank, debt, credit*) as shown in figure 1. In $\mathbf{V}$ only the context word representations are updated and in $\mathbf{U}$ only the representations of *finance, finanze* and negative samples such as *tree* and *dog* are updated. CBOW learns good embeddings because each time it updates the parameters, the words in the contexts are pushed closer to each other in the $\mathbf{V}$ space. Similarly, the target word w_i and the translation $\bar{w}_i$ are also pushed closer in the $\mathbf{U}$ space. This is directly related to poitwise mutual information values of each pair of word and context explained in Levy and Goldberg (2014). Thus, $\mathbf{U}$ is bound to better at bilingual lexicon induction task and $\mathbf{V}$ is better at monolingual word similarity task.

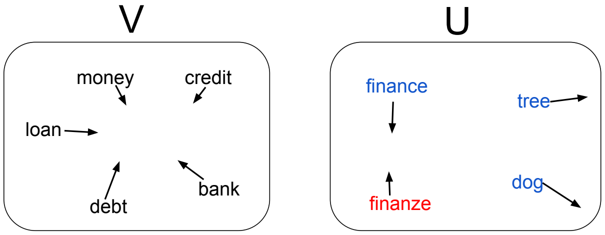


Figure 1: Example of $\mathbf{V}$ and $\mathbf{U}$ space during training.

The simple question is, how to combine both $\mathbf{V}$ and $\mathbf{U}$ to produce a better representation. We experiment with several ways to combine $\mathbf{V}$ and $\mathbf{U}$. First, we can follow Pennington et al. (2014) to *interpolate* $\mathbf{V}$ and $\mathbf{U}$ in the post-processing step. i.e.

$$\gamma \mathbf{V} + (1-\gamma)\mathbf{U} \quad (4)$$

where γ controls the contribution of each embedding space. Second, we can also *concatenate* $\mathbf{V}$ and $\mathbf{U}$ instead of interpolation such that $\mathbf{C} = [\mathbf{V} : \mathbf{U}]$ where $\mathbf{C} \in \mathbb{R}^{|W| \times 2d}$ and W is the combined vocabulary from $D_e \cup D_f$.

Moreover, we can also fuse $\mathbf{V}$ and $\mathbf{U}$ during training. For each word in the combined dictionary $V_e \cup V_f$, we encourage the model to learn similar representation in both $\mathbf{V}$ and $\mathbf{U}$ by adding a *regularization* term to the objective function in equation (3) during training.

$$\mathcal{O}' = \mathcal{O} + \delta \sum_{w \in V_e \cup V_f} \|\mathbf{u}_w - \mathbf{v}_w\|_2^2 \quad (5)$$

where δ controls to what degree we should bind two spaces together.[3]

5 Experimental Setup

Our experimental evaluation seeks to determine how well lexical distances in the learned embedding

[1] Using both embeddings gives a small improvement compared to just using context vector $\mathbf{h}$ alone.

[2] We also experimented with using expectations over translations, as per standard EM, with slight degredation in results.

[3] In the stochastic gradient update for a given word in context, we only compute the gradient of the regularisation term in (5) with respect to the words in the set of positive and negative examples.

spaces match with known lexical similarity judgements from bilingual and monolingual lexical resources. To this end, in §6 we test crosslingual distances using a bilingual lexicon induction task in which we evaluate the embeddings in terms of how well nearby pairs of words from two languages in the embedding space match with human judgements. Next, to evaluate the monolingual embeddings we evaluate word similarities in a single language against standard similarity datasets (§7). Lastly, to demonstrate the usefulness of our embeddings in a task-based setting, we evaluate on crosslingual document classification (§9).

Monolingual Data The monolingual data is taken from the pre-processed Wikipedia dump from Al-Rfou et al. (2013). The data is already cleaned and tokenized. We additionally lower-case all words. Normally monolingual word embeddings are trained on billions of words. However, obtaining that much monolingual data for a low-resource language is infeasible. Therefore, we only select the first 5 million sentences (around 100 million words) for each language.

Dictionary A bilingual dictionary is the only source of bilingual correspondence in our technique. We prefer a dictionary that covers many languages, such that our approach can be applied widely to many low-resource languages. We use Panlex, a dictionary which currently covers around 1300 language varieties with about 12 million expressions. The translations in PanLex come from various sources such as glossaries, dictionaries, automatic inference from other languages, etc. Accordingly, Panlex has high language coverage but often noisy translations.[4] Table 1 summarizes the sizes of monolingual corpora and dictionaries for each pair of language in our experiments.

	Source (M)	Target (M)	Dict (k)
en-es	120.1 (73.9%)	126.8 (74.4%)	712.0
en-it	120.1 (74.7%)	114.6 (67.4%)	560.1
en-nl	120.1 (69.1%)	80.2 (63.4%)	406.6
en-de	120.1 (77.8%)	90.8 (68.3%)	964.4
en-sr	120.1 (28.0%)	7.5 (17.5%)	35.1

Table 1: Number of tokens in millions for the source and target languages in each language pair. Also shown is the number of entries in the bilingual dictionary in thousands. The number in the parenthesis shows the token coverage in the dictionary on each monolingual corpus.

6 Bilingual Lexicon Induction

Given a word in a source language, the bilingual lexicon induction (BLI) task is to predict its translation in the target language. Vulić and Moens (2015) proposed this task to test crosslingual word embeddings. The difficulty of this is that it is evaluated using the recall of the top ranked word. The model must be very discriminative in order to score well.

We build the CLWE for 3 language pairs: it-en, es-en and nl-en, using similar parameters setting with Vulić and Moens (2015).[5] The remaining tunable parameters in our system are δ from Equation (5), and the choice of algorithm for combining embeddings. We use the regularization technique from §4.3 for combining context and word embeddings with $\delta = 0.01$, and word embeddings $\mathbf{U}$ are used as the output for all experiments (but see comparative experiments in §8.)

Qualitative evaluation We jointly train the model to predict both w_i and the translation $\bar{w_i}$, combine $\mathbf{V}$ and $\mathbf{U}$ during training for each language pair. Table 2 shows the top 10 closest words in both source and target languages according to cosine similarity. Note that the model correctly identifies the translation in en as the top candidate, and the top 10 words in both source and target languages are highly related. This qualitative evaluation initially demonstrates the ability of our CLWE to capture both the bilingual and monolingual relationship.

Quantitative evaluation Table 3 shows our results compared with prior work. We reimple-

[4] We also experimented with a crowd-sourced dictionary from Wiktionary. Our initial observation was that the translation quality was better but with a lower-coverage. For example, for en-it dictionary, Panlex and Wiktionary have a coverage of 42.1% and 16.8% respectively for the top 100k most frequent English words from Wikipedia. The average number of translations are 5.2 and 1.9 respectively. We observed similar trend using Panlex and Wiktionary dictionary in our model. However, using Panlex results in much better performance. We can run the model on the combined dictionary from both Panlex and Wiktionary but we leave it for future work.

[5] Default learning rate of 0.025, negative sampling with 25 samples, subsampling rate of value $1e^{-4}$, embedding dimension $d = 200$, window size $cs = 48$ and run for 15 epochs.

Model	es-en		it-en		nl-en		Average	
	rec$_1$	*rec*$_5$	*rec*$_1$	*rec*$_5$	*rec*$_1$	*rec*$_5$	*rec*$_1$	*rec*$_5$
Gouws and Søgaard (2015) + Panlex	37.6	63.6	26.6	56.3	49.8	76.0	38.0	65.3
Gouws and Søgaard (2015) + Wikt	61.6	78.9	62.6	81.1	65.6	79.7	63.3	79.9
BilBOWA: Gouws et al. (2015)	51.6	-	55.7	-	57.5	-	54.9	-
Vulić and Moens (2015)	68.9	-	68.3	-	39.2	-	58.8	-
Our model (random selection)	41.1	62.0	57.4	75.4	34.3	55.5	44.3	64.3
Our model (EM selection)	67.3	79.5	66.8	82.3	64.7	82.4	66.3	81.4
+ Joint model	68.0	80.5	70.5	83.3	68.8	84.0	69.1	82.6
+ combine embeddings ($\delta = 0.01$)	74.7	85.4	80.8	90.4	79.1	90.5	78.2	88.8
+ lemmatization	**74.9**	**86.0**	**81.3**	**91.3**	**79.8**	**91.3**	**78.7**	**89.5**

Table 3: Bilingual Lexicon Induction performance from es, it, nl to en. Gouws and Søgaard (2015) + Panlex/Wikt is our reimplementation using Panlex/Wiktionary dictionary. All our models use Panlex as the dictionary. We reported the recall at 1 and 5. The best performance is bold.

gravedad$_{es}$		*tassazione*$_{it}$	
es	en	it	en
gravitacional	**gravity***	tasse	**taxation***
gravitatoria	gravitation*	fiscale	taxes
aceleracin	acceleration	tassa	tax*
gravitacin	non-gravitational	imposte	levied
inercia	inertia	imposta	fiscal
gravity	centrifugal	fiscali	low-tax
msugra	free-falling	l'imposta	revenue
centrífuga	gravitational	tonnage	levy
curvatura	free-fall	tax	annates
masa	newton	accise	evasion

Table 2: Top 10 closest words in both source and target language corresponding to es word *gravedad* (left) and it word *tassazione* (right). They have 15 and 4 dictionary translations respectively. The en words in the dictionary translations are marked with (*). The correct translation is in bold.

ment Gouws and Søgaard (2015) using Panlex and Wiktionary dictionaries. The result with Panlex is substantially worse than with Wiktionary. This confirms our hypothesis in §2. That is the context might be corrupted if we just randomly replace the training data with the translation from noisy dictionary such as Panlex.

Our model when randomly picking the translation is similar to Gouws and Søgaard (2015), using the Panlex dictionary. The biggest difference is that they replace the training data (both context and middle word) while we fix the context and only replace the middle word. For a high coverage yet noisy dictionary such as Panlex, our approach gives better average score. Comparing our two most basic models (EM selection and random selection), it is clear that the model using EM to select the translation outperforms random selection by a significant margin.

Our joint model, as described in equation (3) which predicts both target word and the translation, further improves the performance, especially for nl-en. We use equation (5) to combine both context embeddings $\mathbf{V}$ and word embeddings $\mathbf{U}$ for all three language pairs. This modification during training substantially improves the performance. More importantly, all our improvements are consistent for all three language pairs and both evaluation metrics, showing the robustness of our models.

Our combined model out-performed previous approaches by a large margin. Vulić and Moens (2015) used bilingual comparable data, but this might be hard to obtain for some language pairs. Their performance on nl-en is poor because their comparable data between en and nl is small. Besides, they also use POS tagger and lemmatizer to filter only *Noun* and reduce the morphology complexity during training. These tools might not be available for many languages. For a fairer comparison to their work, we also use the same Treetagger (Schmid, 1995) to lemmatize the output of our combined model before evaluation. Table 3 (+lemmatization) shows some improvements but minor. It demonstrates that our model is already good at disambiguating morphology. For example, the top 2 translations for es word *lenguas* in en are *languages* and *language* which correctly prefer the plural translation.

7 Monolingual Word Similarity

Now we consider the efficacy of our CLWE on monolingual word similarity. We evaluate on En-

	Model	WS-de	WS-en	RW-en
Baselines	Klementiev et al. (2012)	23.8	13.2	7.3
	Chandar A P et al. (2014)	34.6	39.8	20.5
	Hermann and Blunsom (2014)	28.3	19.8	13.6
	Luong et al. (2015)	47.4	49.3	25.3
	Gouws and Søgaard (2015)	67.4	71.8	31.0
Mono	CBOW	62.2	70.3	42.7
	+ combine	65.8	74.1	43.1
	Yih and Qazvinian (2012)	-	81.0	-
	Shazeer et al. (2016)	-	74.8	48.3
Ours	Our joint-model	59.3	68.6	38.1
	+ combine	**71.1**	**76.2**	**44.0**

Table 4: Spearman's rank correlation for monolingual similarity measurement on 3 datasets WS-de (353 pairs), WS-en (353 pairs) and RW-en (2034 pairs). We compare against 5 baseline crosslingual word embeddings. The best CLWE performance is bold. For reference, we add the monolingual CBOW with and without embeddings combination, Yih and Qazvinian (2012) and Shazeer et al. (2016) which represents the monolingual state-of-the-art results for WS-en and RW-en.

glish monolingual similarity on WordSim353 (WS-en), RareWord (RW-en) and German version of WordSim353 (WS-de) (Finkelstein et al., 2001; Luong et al., 2013; Luong et al., 2015). Each of those datasets contain many tuples $(w_1, w_2, \mathtt{s})$ where s is a scalar denoting the semantic similarity between w_1 and w_2 given by human annotators. Good system should produce the score correlated with human judgement.

We train the model as described in §4, which is the *combine embeddings* setting from Table 3. Since the evaluation involves de and en word similarity, we train the CLWE for en−de pair. Table 4 shows the performance of our combined model compared with several baselines. Our combined model out-performed both Luong et al. (2015) and Gouws and Søgaard (2015)[6] which represent the best published crosslingual embeddings trained on bitext and monolingual data respectively.

We also compare our system with the monolingual CBOW model trained on the monolingual data for each language, using the same parameter settings from earlier (§6). Surprisingly, our combined model performs better than the monolingual CBOW baseline which makes our result close to the monolingual state-of-the-art on each different dataset. However, the best monolingual methods use much larger monolingual corpora (Shazeer et al., 2016), WordNet or the output of commercial search engines (Yih and Qazvinian, 2012).

Next we explain the gain of our combined model compared with the monolingual CBOW model. First, we compare the combined model with the joint-model with respect to monolingual CBOW model (Table 4). It shows that the improvement seems mostly come from combining $\mathbf{V}$ and $\mathbf{U}$. If we apply the combining algorithm to the monolingual CBOW model (CBOW + combine), we also observe an improvement. Clearly most of the improvement is from combining $\mathbf{V}$ and $\mathbf{U}$, however our $\mathbf{V}$ and $\mathbf{U}$ are more complementary as the gain is more marked. Other improvements can be explained by the observation that a dictionary can improve monolingual accuracy through linking synonyms (Faruqui and Dyer, 2014). For example, since *plane*, *airplane* and *aircraft* have the same Italian translation *aereo*, the model will encourage those words to be closer in the embedding space.

8 Model selection

Combining context embeddings and word embeddings results in an improvement in both monolingual similarity and bilingual lexicon induction. In §4.3, we introduce several combination methods including post-processing (interpolation and concatenation) and during training (regularization). In this section, we justify our parameter and model choices.

We use en−it pair for tuning purposes, considering the value of γ in equation 4. Figure 2 shows the performances using different values of γ. The two extremes where $\gamma = 0$ and $\gamma = 1$ corresponds to no interpolation where we just use $\mathbf{U}$ or $\mathbf{V}$ respectively. As γ increases, the performance on WS-en increases yet BLI decreases. These results confirm our hypothesis in §4.3 that $\mathbf{U}$ is better at capturing bilingual relations and $\mathbf{V}$ is better at capturing monolingual relations. As a compromise, we choose $\gamma = 0.5$ in our experiments. Similarly, we tune the regularization sensitivity δ in equation (5) which combines embeddings space during training. We test $\delta = 10^{-n}$ with $n = \{0, 1, 2, 3, 4\}$ and us-

[6] trained using the Panlex dictionary

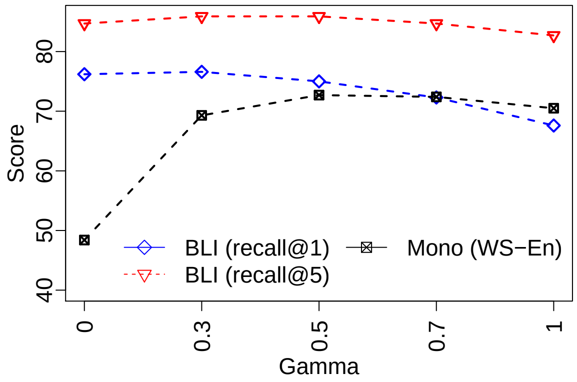


Figure 2: Performance of word embeddings interpolated using different values of γ evaluated using BLI (Recall@1, Recall@5) and English monolingual WordSim353 (WS-en).

	Model	BLI		Mono
		rec_1	rec_5	WS-en
Alone	Joint-model + $\mathbf{V}$	67.6	82.8	70.5
	Joint-model + $\mathbf{U}$	76.2	84.7	48.4
Combine	Interpolation $\left[\frac{\mathbf{V}+\mathbf{U}}{2}\right]$	75.0	85.9	72.7
	Concatenation	72.7	85.2	71.2
	Regularization + $\mathbf{V}$	80.3	89.8	45.9
	Regularization + $\mathbf{U}$	80.8	90.4	**74.8**
	Regularization + $\frac{\mathbf{V}+\mathbf{U}}{2}$	**80.9**	**91.1**	72.3

Table 5: Performance on en−it BLI and en monolingual similarity WordSim353 (WS-en) for various combining algorithms mentioned in §4.3 w.r.t just using $\mathbf{U}$ or $\mathbf{V}$ alone (after joint-training). We use $\gamma = 0.5$ for interpolation and $\delta = 0.01$ for regularization with the choice of $\mathbf{V}$, $\mathbf{U}$ or interpolation of both $\frac{\mathbf{V}+\mathbf{U}}{2}$ for the output. The best scores are bold.

ing $\mathbf{V}$, $\mathbf{U}$ or the interpolation of both $\frac{\mathbf{V}+\mathbf{U}}{2}$ as the learned embeddings, evaluated on the same BLI and WS-en. We select $\delta = 0.01$.

Table 5 shows the performance with and without using combining algorithms mentioned in §4.3. As the compromise between both monolingual and crosslingual tasks, we choose regularization + $\mathbf{U}$ as the combination algorithm. All in all, we apply the regularization algorithm for combining $\mathbf{V}$ and $\mathbf{U}$ with $\delta = 0.01$ and $\mathbf{U}$ as the output for all language pairs without further tuning.

9 Crosslingual Document Classification

In this section, we evaluate our CLWE on a downstream crosslingual document classification (CLDC) task. In this task, the document classifier is trained on a source language and then applied directly to classify a document in the target language. This is convenient for a target low-resource language where we do not have document annotations. The experimental setup is the same as Klementiev et al. (2012)[7] with the training and testing data sourced from Reuter RCV1/RCV2 corpus (Lewis et al., 2004).

Model	en → de	de → en
MT baseline	68.1	67.4
Klementiev et al. (2012)	77.6	71.1
Gouws et al. (2015)	86.5	75.0
Kočiský et al. (2014)	83.1	75.4
Chandar A P et al. (2014)	**91.8**	74.2
Hermann and Blunsom (2014)	86.4	74.7
Luong et al. (2015)	88.4	**80.3**
Our model	86.3	76.8

Table 6: CLDC performance for both en → de and de → en direction for many CLWE. The MT baseline uses phrase-based statistical machine translation to translate the source language to target language (Klementiev et al., 2012). The best scores are bold.

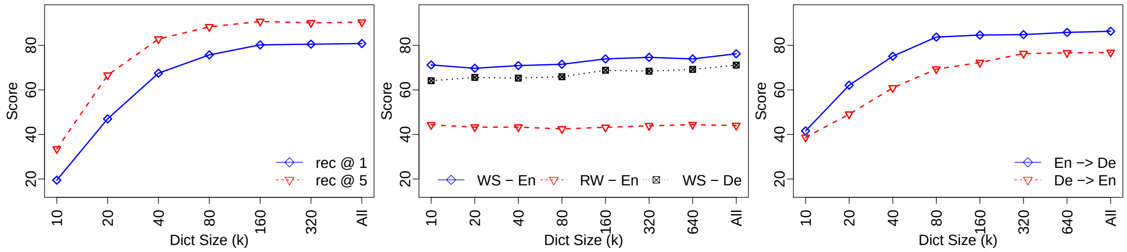
The documents are represented as the bag of word embeddings weighted by tf.idf. A multi-class classifier is trained using the average perceptron algorithm on 1000 documents in the source language and tested on 5000 documents in the target language. We use the CLWE, such that the document representation in the target language embeddings is in the same space with the source language.

We build the en-de CLWE using combined models as described in section §4. Following prior work, we also use monolingual data[8] from the RCV1/RCV2 corpus (Klementiev et al., 2012; Gouws et al., 2015; Chandar A P et al., 2014).

Table 6 shows the CLDC results for various CLWE. Despite its simplicity, our model achieves competitive performance. Note that aside from our model, all other models in Table 6 use a large bitext (Europarl) which may not exist for many low-resource languages, limiting their applicability.

[7] The data split and code are kindly provided by the authors.

[8] We randomly sample documents in RCV1 and RCV2 corpora and selected around 85k documents to form 400k monolingual sentences for both en and de. For each document, we perform basic pre-processing including: lower-casing, remove html tags and tokenization. These monolingual data are then concatenated with the monolingual data from Wikipedia to form the final training data.



(a) BLI (b) Mono Similarity (c) CLDC

Figure 3: Learning curve showing how task scores increase with increasing dictionary size; showing bilingual lexicon induction (BLI) task (left), monolingual similarity (center) and crosslingual document classification (right). BLI is trained on `en-it`, and monolingual similarity and CLDC are trained on `en-de`.

10 Low-resource languages

Our model exploits dictionaries, which are more widely available than parallel corpora. However the question remains as to how well this performs of a real low-resource language, rather than a simulated condition like above, whereupon the quality of the dictionary is likely to be worse. To test this, we evaluation on Serbian, a language with few annotated language resources. Table 1 shows the relative size of monolingual data and dictionary for `en-sr` compared with other language pairs. Both the Serbian monolingual data and the dictionary size is more than 10 times smaller than other language pairs. We build the `en-sr` CLWE using our best model (joint + combine) and evaluate on the bilingual word induction task using 939 gold translation pairs.[9] We achieved recall score of 35.8% and 45.5% at 1 and 5 respectively. Although worse than the earlier results, these numbers are still well above chance.

We can also simulate low-resource setting using our earlier datasets. For estimating the performance loss on all three tasks, we down sample the dictionary for `en-it` and `en-de` based on `en` word frequency. Figure 3 shows the performance with different dictionary sizes for all three tasks. The monolingual similarity performance is very similar across various sizes. For BLI and CLDC, dictionary size is more important, although performance levels off at around 80k dictionary pairs. We conclude that this size is sufficient for decent performance.

11 Conclusion

Previous CLWE methods often impose high resource requirements yet have low accuracy. We introduce a simple framework based on a large noisy dictionary. We model polysemy using EM translation selection during training to learn bilingual correspondences from monolingual corpora. Our algorithm allows to train on massive amount of monolingual data efficiently, representing monolingual and bilingual properties of language. This allows us to achieve state-of-the-art performance on bilingual lexicon induction task, competitive result on monolingual word similarity and crosslingual document classification task. Our combination techniques during training, especially using regularization, are highly effective and could be used to improve monolingual word embeddings.

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 No. HR0011-15-C-0114. We thank Yuta Tsuboi and Alvin Grissom II for helpful discussions, Jan Šnajder for helping with `sr-en` evaluation.

References

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.

[9] The `sr`→`en` translations are sourced from Google Translate by translating one word at a time, followed by manually verification, after which 61 translation pairs were ruled out as being bad or questionable.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

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.

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. 2015. 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.

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 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.

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.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 58–68, Baltimore, Maryland, June. 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.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as a factorization. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2177–2185.

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.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Nat. Lang. Eng.*, 5(2):113–133, June.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Noam Shazeer, Ryan Doherty, Colin Evans, and Chris Waterson. 2016. Swivel: Improving embeddings by noticing what's missing. *CoRR*, abs/1602.02215.

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.

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.

Wen-tau Yih and Vahed Qazvinian. 2012. Measuring word relatedness using heterogeneous vector space models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 616–620, Stroudsburg, PA, USA. Association for Computational Linguistics.