Improving Vector Space Word Representations Using Multilingual Correlation

Manaal Faruqui and **Chris Dyer**
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
{mfaruqui, cdyer}@cs.cmu.edu

Abstract

The distributional hypothesis of Harris (1954), according to which the meaning of words is evidenced by the contexts they occur in, has motivated several effective techniques for obtaining vector space semantic representations of words using unannotated text corpora. This paper argues that lexico-semantic content should additionally be invariant across languages and proposes a simple technique based on canonical correlation analysis (CCA) for incorporating multilingual evidence into vectors generated monolingually. We evaluate the resulting word representations on standard lexical semantic evaluation tasks and show that our method produces substantially better semantic representations than monolingual techniques.

1 Introduction

Data-driven learning of vector-space word embeddings that capture lexico-semantic properties is a technique of central importance in natural language processing. Using cooccurrence statistics from a large corpus of text (Deerwester et al., 1990; Turney and Pantel, 2010),[1] it is possible to construct high-quality semantic vectors — as judged by both correlations with human judgements of semantic relatedness (Turney, 2006; Agirre et al., 2009) and as features for downstream applications (Turian et al., 2010).

The observation that vectors representing cooccurrence tendencies would capture meaning is expected according to the **distributional hypothesis** (Harris, 1954), famously articulated by Firth (1957) as *You shall know a word by the company it keeps.* Although there is much evidence in favor of the distributional hypothesis, in this paper we argue for incorporating *translational* context when constructing vector space semantic models (VSMs). Simply put: knowing how words translate is a valuable source of lexico-semantic information and should lead to better VSMs.

Parallel corpora have long been recognized as valuable for lexical semantic applications, including identifying word senses (Diab, 2003; Resnik and Yarowsky, 1999) and paraphrase and synonymy relationships (Bannard and Callison-Burch, 2005). The latter work (which we build on) shows that if different words or phrases in one language often translate into a single word or phrase type in a second language, this is good evidence that they are synonymous. To illustrate: the English word forms *aeroplane*, *airplane*, and *plane* are observed to translate into the same Hindi word: वायुयान (*vaayuyaan*). Thus, even if we did not know the relationship between the English words, this translation fact is evidence that they all have the same meaning.

How can we exploit information like this when constructing VSMs? We propose a technique that first constructs independent VSMs in two languages and then projects them onto a common vector space such that translation pairs (as determined by automatic word alignments) should be maximally correlated (§2). We review latent semantic analysis (LSA), which serves as our monolingual VSM baseline (§3), and a suite of standard evaluation tasks that we use to measure the quality of the embeddings (§4). We then turn to experiments. We first show that our technique leads to substantial improvements over monolingual LSA (§5), and then examine how our technique fares with vectors learned using two different neural networks, one that models word sequences and a second that models bags-of-context

[1] Related approaches use the internal representations from neural network models of word sequences (Collobert and Weston, 2008) or continuous bags-of-context wordsels (Mikolov et al., 2013a) to arrive at vector representations that likewise capture cooccurence tendencies and meanings.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 462–471,
Gothenburg, Sweden, April 26-30 2014.

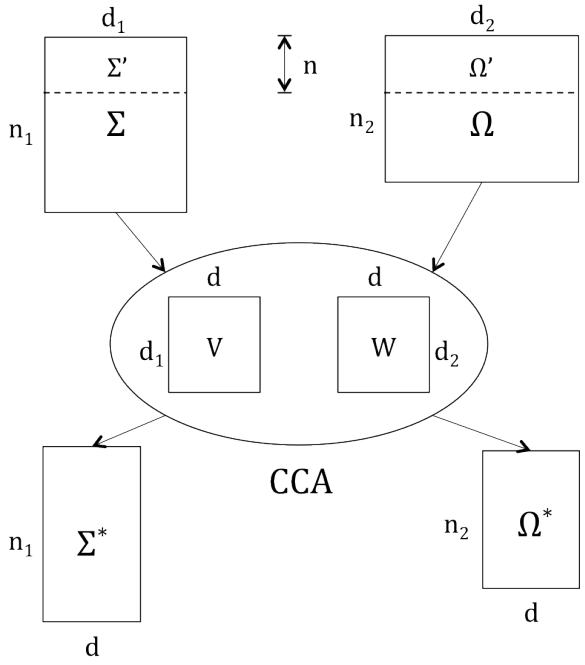

Figure 1: Cross-lingual word vector projection using CCA.

words. We observe substantial improvements over the sequential model using multilingual evidence but more mixed results relative to using the bags-of-contexts model (§6).

2 Multilingual Correlation with CCA

To gain information from the translation of a given word in other languages the most basic thing to do would be to just append the given word representation with the word representations of its translation in the other language. This has three drawbacks: first, it increases the number of dimensions in the vector; second, it can pull irrelevant information from the other language that doesn't generalize across languages and finally the given word might be out of vocabulary of the parallel corpus or dictionary.

To counter these problems we use CCA[2] which is a way of measuring the linear relationship between two multidimensional variables. It finds two projection vectors, one for each variable, that are optimal with respect to correlations. The dimensionality of these new projected vectors is equal to or less than the smaller dimensionality of the two variables.

Let $\mathbf{\Sigma} \in \mathbb{R}^{n_1 \times d_1}$ and $\mathbf{\Omega} \in \mathbb{R}^{n_2 \times d_2}$ be vector space embeddings of two different vocabularies where rows represent words. Since the two vocabularies are of different sizes (n_1 and n_2) and there might not exist translation for every word of $\mathbf{\Sigma}$ in $\mathbf{\Omega}$, let $\mathbf{\Sigma}' \subseteq \mathbf{\Sigma}$ where every word in $\mathbf{\Sigma}'$ is translated to one other word[3] in $\mathbf{\Omega}' \subseteq \mathbf{\Omega}$ and $\mathbf{\Sigma} \in \mathbb{R}^{n \times d_1}$ and $\mathbf{\Omega} \in \mathbb{R}^{n \times d_2}$.

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two corresponding vectors from $\mathbf{\Sigma}'$ and $\mathbf{\Omega}'$, and $\boldsymbol{v}$ and $\boldsymbol{w}$ be two projection directions. Then, the projected vectors are:

$$\boldsymbol{x}' = \boldsymbol{x}\boldsymbol{v} \quad \boldsymbol{y}' = \boldsymbol{y}\boldsymbol{w} \tag{1}$$

and the correlation between the projected vectors can be written as:

$$\rho(\boldsymbol{x}', \boldsymbol{y}') = \frac{E[\boldsymbol{x}'\boldsymbol{y}']}{\sqrt{E[\boldsymbol{x}'^2]E[\boldsymbol{y}'^2]}} \tag{2}$$

CCA maximizes ρ for the given set of vectors $\mathbf{\Sigma}'$ and $\mathbf{\Omega}'$ and outputs two projection vectors $\boldsymbol{v}$ and $\boldsymbol{w}$:

$$\begin{aligned} \boldsymbol{v}, \boldsymbol{w} &= \mathrm{CCA}(\boldsymbol{x}, \boldsymbol{y}) \\ &= \arg\max_{\boldsymbol{v},\boldsymbol{w}} \rho(\boldsymbol{x}\boldsymbol{v}, \boldsymbol{y}\boldsymbol{w}) \end{aligned} \tag{3}$$

Using these two projection vectors we can project the entire vocabulary of the two languages $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ using equation 1. Summarizing:

$$\boldsymbol{V}, \boldsymbol{W} = \mathrm{CCA}(\mathbf{\Sigma}', \mathbf{\Omega}') \tag{4}$$

$$\Sigma^* = \mathbf{\Sigma}\boldsymbol{V} \quad \mathbf{\Omega}^* = \mathbf{\Omega}\boldsymbol{W} \tag{5}$$

where, $\boldsymbol{V} \in \mathbb{R}^{d_1 \times d}$, $\boldsymbol{W} \in \mathbb{R}^{d_2 \times d}$ contain the projection vectors and $d = \min\{\mathrm{rank}(\boldsymbol{V}), \mathrm{rank}(\boldsymbol{W})\}$. Thus, the resulting vectors cannot be longer than the original vectors. Since $\boldsymbol{V}$ and $\boldsymbol{W}$ can be used to project the whole vocabulary, CCA also solves the problem of not having translations of a particular word in the dictionary. The schema of performing CCA on the monolingual word representations of two languages is shown in Figure 1.

Further Dimensionality Reduction: Since CCA gives us correlations and corresponding projection vectors across d dimensions which can be large, we perform experiments by taking projections of the original word vectors across only the top k correlated dimensions. This is trivial to implement as the projection vectors $\boldsymbol{V}$,

[2]We use the MATLAB module for CCA: `http://www.mathworks.com/help/stats/canoncorr.html`

[3]Further information on how these one-to-one translations are obtained in §5

$\boldsymbol{W}$ in equation 4 are already sorted in descending order of correlation. Therefore in,

$$\boldsymbol{\Sigma}_k^* = \boldsymbol{\Sigma}\boldsymbol{V}_k \quad \boldsymbol{\Omega}_k^* = \boldsymbol{\Omega}\boldsymbol{W}_k \tag{6}$$

$\boldsymbol{\Sigma}_k^*$ and $\boldsymbol{\Omega}_k^*$ are now word vector projections along the top k correlated dimensions, where, $\boldsymbol{V}_k$ and $\boldsymbol{W}_k$ are the column truncated matrices.

3 Latent Semantic Analysis

We perform latent semantic analysis (Deerwester et al., 1990) on a word-word co-occurrence matrix. We construct a word co-occurrence frequency matrix $\boldsymbol{F}$ for a given training corpus where each row $\boldsymbol{w}$, represents one word in the corpus and every column $\boldsymbol{c}$, is the context feature in which the word is observed. In our case, every column is a word which occurs in a given window length around the target word. For scalability reasons, we only select words with frequency greater than 10 as features. We also remove the top 100 most frequent words (mostly stop words) from the column features.

We then replace every entry in the sparse frequency matrix $\boldsymbol{F}$ by its pointwise mutual information (PMI) (Church and Hanks, 1990; Turney, 2001) resulting in $\boldsymbol{X}$. PMI is designed to give a high value to x_{ij} where there is a interesting relation between w_i and c_j, a small or negative value of x_{ij} indicates that the occurrence of w_i in c_j is uninformative. Finally, we factorize the matrix $\boldsymbol{X}$ using singular value decomposition (SVD). SVD decomposes $\boldsymbol{X}$ into the product of three matrices:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Psi}\boldsymbol{V}^\top \tag{7}$$

where, $\boldsymbol{U}$ and $\boldsymbol{V}$ are in column orthonormal form and $\boldsymbol{\Psi}$ is a diagonal matrix of singular values (Golub and Van Loan, 1996). We obtain a reduced dimensional representation of words from size $|\boldsymbol{V}|$ to k:

$$\boldsymbol{A} = \boldsymbol{U}_k\boldsymbol{\Psi}_k \tag{8}$$

where k can be controlled to trade off between reconstruction error and number of parameters, $\boldsymbol{\Psi}_k$ is the diagonal matrix containing the top k singular values, $\boldsymbol{U}_k$ is the matrix produced by selecting the corresponding columns from $\boldsymbol{U}$ and $\boldsymbol{A}$ represents the new matrix containing word vector representations in the reduced dimensional space.

4 Word Representation Evaluation

We evaluate the quality of our word vector representations on a number of tasks that test how well they capture both semantic and syntactic aspects of the representations.

4.1 Word Similarity

We evaluate our word representations on four different benchmarks that have been widely used to measure word similarity. The first one is the **WS-353** dataset (Finkelstein et al., 2001) containing 353 pairs of English words that have been assigned similarity ratings by humans. This data was further divided into two fragments by Agirre et al. (2009) who claimed that *similarity* (**WS-SIM**) and *relatedness* (**WS-REL**) are two different kinds of relations and should be dealt with separately. We present results on the whole set and on the individual fragments as well.

The second and third benchmarks are the **RG-65** (Rubenstein and Goodenough, 1965) and the **MC-30** (Miller and Charles, 1991) datasets that contain 65 and 30 pairs of nouns respectively and have been given similarity rankings by humans. These differ from **WS-353** in that it contains only nouns whereas the former contains all kinds of words. The fourth benchmark is the **MTurk-287** (Radinsky et al., 2011) dataset that constitutes of 287 pairs of words and is different from the above two benchmarks in that it has been constructed by crowdsourcing the human similarity ratings using Amazon Mechanical Turk.

We calculate similarity between a given pair of words by the *cosine* similarity between their corresponding vector representation. We then report Spearman's rank correlation coefficient (Myers and Well, 1995) between the rankings produced by our model against the human rankings.

4.2 Semantic Relations (SEM-REL)

Mikolov et al. (2013a) present a new semantic relation dataset composed of analogous word pairs. It contains pairs of tuples of word relations that follow a common semantic relation. For example, in *England : London :: France : Paris*, the two given pairs of words follow the country-capital relation. There are three other such kinds of relations: country-currency, man-woman, city-in-state and overall 8869 such pairs of words[4].

The task here is to find a word *d* that best fits the following relationship: *a : b :: c : d* given *a, b* and *c*. We use the vector offset method described

[4] 107 pairs were out of vocabulary for our vectors and were ignored.

in Mikolov et al. (2013a) that computes the vector $\boldsymbol{y} = \boldsymbol{x}_a - \boldsymbol{x}_b + \boldsymbol{x}_c$ where, $\boldsymbol{x}_a, \boldsymbol{x}_b$ and $\boldsymbol{x}_c$ are word vectors of a, b and c respectively and returns the vector $\boldsymbol{x}_w$ from the whole vocabulary which has the highest cosine similarity to $\boldsymbol{y}$:

$$\boldsymbol{x}_w = \arg\max_{\boldsymbol{x}_w} \frac{\boldsymbol{x}_w \cdot \boldsymbol{y}}{|\boldsymbol{x}_w| \cdot |\boldsymbol{y}|}$$

It is worth noting that this is a non-trivial $|V|$-way classification task where V is the size of the vocabulary.

4.3 Syntactic Relations (SYN-REL)

This dataset contains word pairs that are different syntactic forms of a given word and was prepared by Mikolov et al. (2013a). For example, in *walking* and *walked*, the second word is the past tense of the first word. There are nine such different kinds of relations: adjective-adverb, opposites, comaparative, superlative, present-participle, nation-nationality, past tense, plural nouns and plural verbs. Overall there are 10675 such syntactic pairs of word tuples. The task here again is identifying a word *d* that best fits the following relationship: *a : b :: c : d* and we solve it using the method described in §4.2.

5 Experiments

5.1 Data

For English, German and Spanish we used the WMT-2011[5] monolingual news corpora and for French we combined the WMT-2011 and 2012[6] monolingual news corpora so that we have around 300 million tokens for each language to train the word vectors.

For CCA, a one-to-one correspondence between the two sets of vectors is required. Obviously, the vocabulary of two languages are of different sizes and hence to obtain one-to-one mapping, for every English word we choose a word from the other language to which it has been aligned the maximum number of times[7] in a parallel corpus. We got these word alignment counts using cdec (Dyer et al., 2010) from the parallel news commentary corpora (WMT 2006-10) combined with the Europarl corpus for English-{German, French, Spanish}.

[5] http://www.statmt.org/wmt11/

[6] http://www.statmt.org/wmt12/

[7] We also tried weighted average of vectors across all aligned words and did not observe any significant difference in results.

5.2 Methodology

We construct LSA word vectors of length 640[8] for English, German, French and Spanish. We project the English word vectors using CCA by pairing them with German, French and Spanish vectors. For every language pair we take the top k correlated dimensions (cf. equation 6), where $k \in 10\%, 20\%, \ldots 100\%$ and tune the performance on **WS-353** task. We then select the k that gives us the best average performance across language pairs, which is $k = 80\%$, and evaluate the corresponding vectors on all other benchmarks. This prevents us from over-fitting k for every individual task.

5.3 Results

Table 1 shows the Spearman's correlation ratio obtained by using word vectors to compute the similarity between two given words and compare the ranked list against human rankings. The first row in the table shows the baseline scores obtained by using only the monolingual English vectors whereas the other rows correspond to the multilingual cases. The last row shows the average performance of the three language pairs. For all the tasks we get at least an absolute gain of 20 points over the baseline. These results are highly assuring of our hypothesis that multilingual context can help in improving the semantic similarity between similar words as described in the example in §1. Results across language pairs remain almost the same and the differences are most of the times statistically insignificant.

Table 1 also shows the accuracy obtained on predicting different kinds of relations between word pairs. For the **SEM-REL** task the average improvement in accuracy is an absolute 30 points over the baseline which is highly statistically significant ($p < 0.01$) according to the McNemar's test (Dietterich, 1998). The same holds true for the **SYN-REL** task where we get an average improvement of absolute 8 points over the baseline across the language pairs. Such an improvement in scores across these relation prediction tasks further enforces our claim that cross-lingual context can be exploited using the method described in §2 and it does help in encoding the meaning of a word better in a word vector than monolingual information alone.

[8] See section 5.5 for further discussion on vector length.

Lang	Dim	WS-353	WS-SIM	WS-REL	RG-65	MC-30	MTurk-287	SEM-REL	SYN-REL
En	640	46.7	56.2	36.5	50.7	42.3	51.2	14.5	36.8
De-En	512	68.0	**74.4**	64.6	**75.5**	**81.9**	53.6	43.9	**45.5**
Fr-En	512	**68.4**	73.3	**65.7**	73.5	81.3	**55.5**	43.9	44.3
Es-En	512	67.2	71.6	64.5	70.5	78.2	53.6	**44.2**	44.5
Average	–	56.6	64.5	51.0	62.0	65.5	60.8	44	44.7

Table 1: Spearman's correlation (left) and accuracy (right) on different tasks.

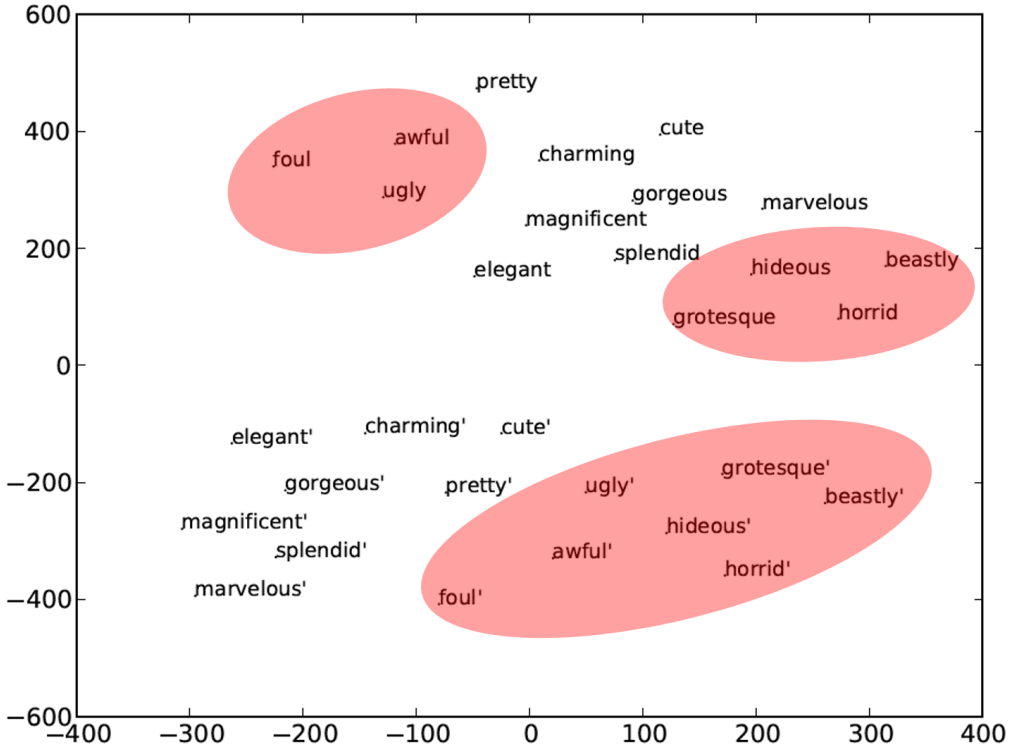

Figure 2: Monolingual (top) and multilingual (bottom; marked with apostrophe) word projections of the antonyms (shown in red) and synonyms of "beautiful".

5.4 Qualitative Example

To understand how multilingual evidence leads to better results in semantic evaluation tasks, we plot the word representations obtained in §3 of several synonyms and antonyms of the word "beautiful" by projecting both the transformed and untransformed vectors onto $\mathbb{R}^2$ using the t-SNE tool (van der Maaten and Hinton, 2008). The untransformed LSA vectors are in the upper part of Fig. 2, and the CCA-projected vectors are in the lower part. By comparing the two regions, we see that in the untransformed representations, the antonyms are in two clusters separated by the synonyms, whereas in the transformed representation, both the antonyms and synonyms are in their own cluster. Furthermore, the average intra-class distance between synonyms and antonyms is reduced.

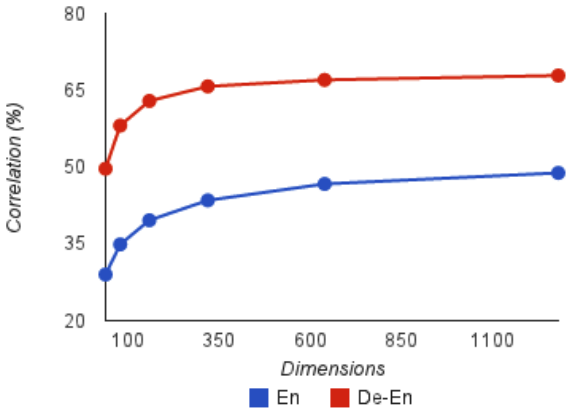

Figure 3: Performance of monolingual and multilingual vectors on **WS-353** for different vector lengths.

5.5 Variation in Vector Length

In order to demonstrate that the gains in performance by using multilingual correlation sustains

for different number of dimensions, we compared the performance of the monolingual and (German-English) multilingual vectors with $k = 80\%$ (cf. §5.2). It can be see in figure 3 that the performance improvement for multilingual vectors remains almost the same for different vector lengths strengthening the reliability of our approach.

6 Neural Network Word Representations

Other kinds of vectors shown to be useful in many NLP tasks are word embeddings obtained from neural networks. These word embeddings capture more complex information than just co-occurrence counts as explained in the next section. We test our multilingual projection method on two types of such vectors by keeping the experimental setting exactly the same as in §5.2.

6.1 RNN Vectors

The recurrent neural network language model maximizes the log-likelihood of the training corpus. The architecture (Mikolov et al., 2013b) consists of an input layer, a hidden layer with recurrent connections to itself, an output layer and the corresponding weight matrices. The input vector $\boldsymbol{w}(t)$ represents input word at time t encoded using 1-of-N encoding and the output layer $\boldsymbol{y}(t)$ produces a probability distribution over words in the vocabulary V. The hidden layer maintains a representation of the sentence history in $\boldsymbol{s}(t)$. The values in the hidden and output layer are computed as follows:

$$\boldsymbol{s}(t) = f(\boldsymbol{U}\boldsymbol{w}(t) + \boldsymbol{W}\boldsymbol{s}(t-1)) \quad (9)$$

$$\boldsymbol{y}(t) = g(\boldsymbol{V}\boldsymbol{s}(t)) \quad (10)$$

where, f and g are the logistic and softmax functions respectively. $\boldsymbol{U}$ and $\boldsymbol{V}$ are weight matrices and the word representations are found in the columns of $\boldsymbol{U}$. The model is trained using back-propagation. Training such a purely lexical model will induce representations with syntactic and semantic properties. We use the RNNLM toolkit[9] to induce these word representations.

6.2 Skip Gram Vectors

In the RNN model (§6.1) most of the complexity is caused by the non-linear hidden layer. This is avoided in the new model proposed in Mikolov et al. (2013a) where they remove the non-linear hidden layer and there is a single projection layer for the input word. Precisely, each current word is used as an input to a log-linear classifier with continuous projection layer and words within a certain range before and after the word are predicted. These vectors are called the skip-gram (SG) vectors. We used the tool[10] for obtaining these word vectors with default settings.

6.3 Results

We compare the best results obtained by using different types of monolingual word representations across all language pairs. For brevity we do not show the results individually for all language pairs as they follow the same pattern when compared to the baseline for every vector type. We train word vectors of length 80 because it was computationally intractable to train the neural embeddings for higher dimensions. For multilingual vectors, we obtain $k = 60\%$ (cf. §5.2).

Table 2 shows the correlation ratio and the accuracies for the respective evaluation tasks. For the RNN vectors the performance improves upon inclusion of multilingual context for almost all tasks except for **SYN-REL** where the loss is statistically significant ($p < 0.01$). For **MC-30** and **SEM-REL** the small drop in performance is not statistically significant. Interestingly, the performance gain/loss for the SG vectors in most of the cases is not statistically significant, which means that inclusion of multilingual context is not very helpful. In fact, for **SYN-REL** the loss is statistically significant ($p < 0.05$) which is similar to the performance of RNN case. Overall, the best results are obtained by the SG vectors in six out of eight evaluation tasks whereas SVD vectors give the best performance in two tasks: **RG-65**, **MC-30**. This is an encouraging result as SVD vectors are the easiest and fastest to obtain as compared to the other two vector types.

To further understand why multilingual context is highly effective for SVD vectors and to a large extent for RNN vectors as well, we plot (Figure 4) the correlation ratio obtained by varying the length of word representations by using equation 6 for the three different vector types on two word similarity tasks: **WS-353** and **RG-65**.

SVD vectors improve performance upon the increase of the number of dimensions and tend to

[9] http://www.fit.vutbr.cz/~imikolov/rnnlm/

[10] https://code.google.com/p/word2vec/

Vectors	Dim	Lang	WS-353	WS-SIM	WS-REL	RG-65	MC-30	MTurk	SEM-REL	SYN-REL
SVD	80	Mono	34.8	45.5	23.4	30.8	21.0	46.6	13.5	24.4
	48	Multi	58.1	65.3	52.7	**62.7**	**67.7**	62.1	23.4	33.2
RNN	80	Mono	23.6	35.6	17.5	26.2	47.7	32.9	4.7	18.2
	48	Multi	35.4	47.3	29.8	36.6	46.5	43.8	4.1	12.2
SG	80	Mono	**63.9**	69.9	**60.9**	54.6	62.8	**66.9**	**47.8**	**47.8**
	48	Multi	63.1	**70.4**	57.6	54.9	64.7	58.7	46.5	44.2

Table 2: Spearman's correlation (left) and accuracy (right) on different tasks. Bold indicates best result across all vector types. Mono: monolingual and Multi: multilingual.

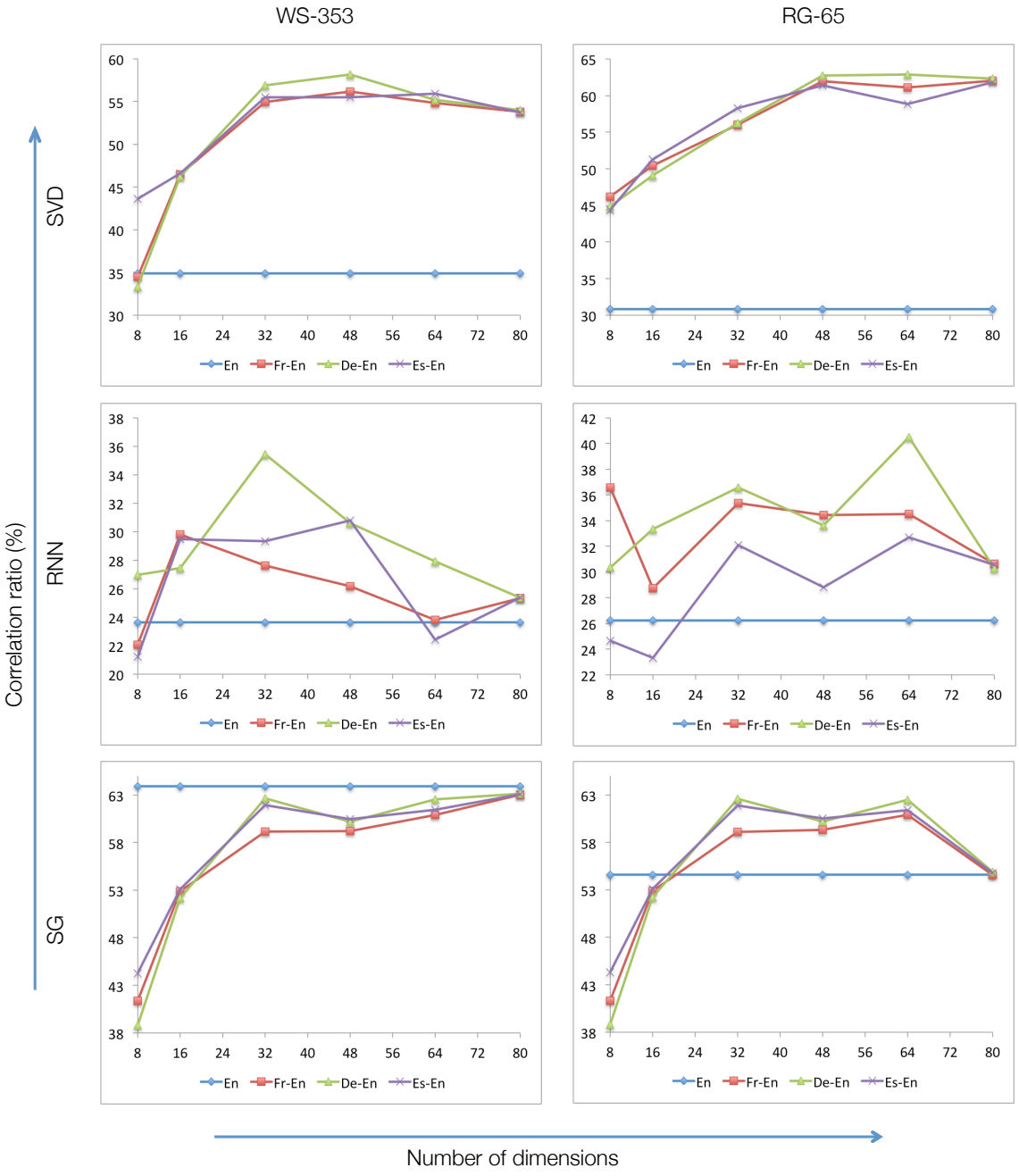

Figure 4: Performance as a function of vector length on word similarity tasks. The monolingual vectors always have a fixed length of 80, they are just shown in the plots for comparison.

saturate towards the end. For all the three language pairs the SVD vectors show uniform pattern of performance which gives us the liberty to use any language pair at hand. This is not true for the RNN vectors whose curves are significantly different for every language pair. SG vectors show a uniform pattern across different language pairs and the performance with multilingual context converges to the monolingual performance when the vector length becomes equal to the monolingual case ($k = 80$). The fact that both SG and SVD vectors have similar behavior across language pairs can be treated as evidence that semantics or information at a conceptual level (since both of them basically model word cooccurrence counts) transfers well across languages (Dyvik, 2004) although syntax has been projected across languages as well (Hwa et al., 2005; Yarowsky and Ngai, 2001). The pattern of results in the case of RNN vectors are indicative of the fact that these vectors encode syntactic information as explained in §6 which might not generalize well as compared to semantic information.

7 Related Work

Our method of learning multilingual word vectors is most closely associated to Zou et al. (2013) who learn bilingual word embeddings and show their utility in machine translation. They optimize the monolingual and the bilingual objective together whereas we do it in two separate steps and project to a common vector space to maximize correlation between the two. Vulić and Moens (2013) learn bilingual vector spaces from non parallel data induced using a seed lexicon. Our method can also be seen as an application of multi-view learning (Chang et al., 2013; Collobert and Weston, 2008), where one of the views can be used to capture cross-lingual information. Klementiev et al. (2012) use a multitask learning framework to encourage the word representations learned by neural language models to agree cross-lingually.

CCA can be used for dimension reduction and to draw correspondences between two sets of data.Haghighi et al. (2008) use CCA to draw translation lexicons between words of two different languages using only monolingual corpora. CCA has also been used for constructing monolingual word representations by correlating word vectors that capture aspects of word meaning and different types of distributional profile of the word (Dhillon et al., 2011). Although our primary experimental emphasis was on LSA based monolingual word representations, which we later generalized to two different neural network based word embeddings, these monolingual word vectors can also be obtained using other continuous models of language (Collobert and Weston, 2008; Mnih and Hinton, 2008; Morin and Bengio, 2005; Huang et al., 2012).

Bilingual representations have previously been explored with manually designed vector space models (Peirsman and Padó, 2010; Sumita, 2000) and with unsupervised algorithms like LDA and LSA (Boyd-Graber and Blei, 2012; Zhao and Xing, 2006). Bilingual evidence has also been exploited for word clustering which is yet another form of representation learning, using both spectral methods (Zhao et al., 2005) and structured prediction approaches (Täckström et al., 2012; Faruqui and Dyer, 2013).

8 Conclusion

We have presented a canonical correlation analysis based method for incorporating multilingual context into word representations generated using only monolingual information and shown its applicability across three different ways of generating monolingual vectors on a variety of evaluation benchmarks. These word representations obtained after using multilingual evidence perform significantly better on the evaluation tasks compared to the monolingual vectors. We have also shown that our method is more suitable for vectors that encode semantic information than those that encode syntactic information. Our work suggests that multilingual evidence is an important resource even for purely monolingual, semantically aware applications. The tool for projecting word vectors can be found at `http://cs.cmu.edu/~mfaruqui/soft.html`.

Acknowledgements

We thanks Kevin Gimpel, Noah Smith, and David Bamman for helpful comments on earlier drafts of this paper. This research was supported by the NSF through grant IIS-1352440.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009.

A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proc. of ACL*.

Jordan L. Boyd-Graber and David M. Blei. 2012. Multilingual topic models for unaligned text. *CoRR*, abs/1205.2657.

Kai-Wei Chang, Wen-tau Yih, and Christopher Meek. 2013. Multi-relational latent semantic analysis. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1602–1612, Seattle, Washington, USA, October. Association for Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March.

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.

S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*.

Paramveer S. Dhillon, Dean P. Foster, and Lyle H. Ungar. 2011. Multi-view learning of word embeddings via cca. In *NIPS*, pages 199–207.

Mona Talat Diab. 2003. *Word sense disambiguation within a multilingual framework*. Ph.D. thesis, University of Maryland at College Park, College Park, MD, USA. AAI3115805.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923.

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 Proceedings of ACL System Demonstrations*.

Helge Dyvik. 2004. Translations as semantic mirrors: from parallel corpus to wordnet. *Language and Computers*, 49(1):311–326.

Manaal Faruqui and Chris Dyer. 2013. An information theoretic approach to bilingual word clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 777–783, Sofia, Bulgaria, August.

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

J.R. Firth. 1957. A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, pages 1–32.

Gene H. Golub and Charles F. Van Loan. 1996. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, Baltimore, MD, USA.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proc. of ACL*.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11:11–311.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING*.

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

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, June. Association for Computational Linguistics.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Andriy Mnih and Geoffrey Hinton. 2008. A scalable hierarchical distributed language model. In *In NIPS*.

Frederic Morin and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In *AISTATS05*, pages 246–252.

Jerome L. Myers and Arnold D. Well. 1995. *Research Design & Statistical Analysis*. Routledge, 1 edition, June.

Yves Peirsman and Sebastian Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 921–929, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 337–346, New York, NY, USA. ACM.

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.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633, October.

Eiichiro Sumita. 2000. Lexical transfer using a vector-space model. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 425–431, Stroudsburg, PA, USA. Association for Computational Linguistics.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, page 11. Association for Computational Linguistics.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research*, pages 141–188.

Peter D. Turney. 2001. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, pages 491–502, London, UK, UK. Springer-Verlag.

Peter D. Turney. 2006. Similarity of semantic relations. *Comput. Linguist.*, 32(3):379–416, September.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, November.

Ivan Vulić and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1613–1624, Seattle, Washington, USA, October. 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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bing Zhao and Eric P. Xing. 2006. Bitam: Bilingual topic admixture models for word alignment. In *In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL06*.

Bing Zhao, Eric P. Xing, and Alex Waibel. 2005. Bilingual word spectral clustering for statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText '05, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, Seattle, Washington, USA, October. Association for Computational Linguistics.