Learning principled bilingual mappings of word embeddings while preserving monolingual invariance

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group, University of the Basque Country (UPV/EHU)
{mikel.artetxe, gorka.labaka, e.agirre}@ehu.eus

Abstract

Mapping word embeddings of different languages into a single space has multiple applications. In order to map from a source space into a target space, a common approach is to learn a linear mapping that minimizes the distances between equivalences listed in a bilingual dictionary. In this paper, we propose a framework that generalizes previous work, provides an efficient exact method to learn the optimal linear transformation and yields the best bilingual results in translation induction while preserving monolingual performance in an analogy task.

1 Introduction

Bilingual word embeddings have attracted a lot of attention in recent times (Zou et al., 2013; Kočiský et al., 2014; Chandar A P et al., 2014; Gouws et al., 2014; Gouws and Søgaard, 2015; Luong et al., 2015; Wick et al., 2016). A common approach to obtain them is to train the embeddings in both languages independently and then learn a mapping that minimizes the distances between equivalences listed in a bilingual dictionary. The learned transformation can also be applied to words missing in the dictionary, which can be used to induce new translations with a direct application in machine translation (Mikolov et al., 2013b; Zhao et al., 2015).

The first method to learn bilingual word embedding mappings was proposed by Mikolov et al. (2013b), who learn the linear transformation that minimizes the sum of squared Euclidean distances for the dictionary entries. Subsequent work has proposed alternative optimization objectives to learn better mappings. Xing et al. (2015) incorporate length normalization in the training of word embeddings and try to maximize the cosine similarity instead, introducing an orthogonality constraint to preserve the length normalization after the projection. Faruqui and Dyer (2014) use canonical correlation analysis to project the embeddings in both languages to a shared vector space.

Beyond linear mappings, Lu et al. (2015) apply deep canonical correlation analysis to learn a nonlinear transformation for each language. Finally, additional techniques have been used to address the hubness problem in Mikolov et al. (2013b), both through the neighbor retrieval method (Dinu et al., 2015) and the training itself (Lazaridou et al., 2015). We leave the study of non-linear transformation and other additions for further work.

In this paper, we propose a general framework to learn bilingual word embeddings. We start with a basic optimization objective (Mikolov et al., 2013b) and introduce several meaningful and intuitive constraints that are equivalent or closely related to previously proposed methods (Faruqui and Dyer, 2014; Xing et al., 2015). Our framework provides a more general view of bilingual word embedding mappings, showing the underlying connection between the existing methods, revealing some flaws in their theoretical justification and providing an alternative theoretical interpretation for them. Our experiments on an existing English-Italian word translation induction and an English word analogy task give strong empirical evidence in favor of our theoretical reasoning, while showing that one of our models clearly outperforms previous alternatives.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2289–2294,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

2 Learning bilingual mappings

Let X and Z denote the word embedding matrices in two languages for a given bilingual dictionary so that their ith row X_{i*} and Z_{i*} are the word embeddings of the ith entry in the dictionary. Our goal is to find a linear transformation matrix W so that XW best approximates Z, which we formalize minimizing the sum of squared Euclidean distances following Mikolov et al. (2013b):

$$\arg\min_W \sum_i \|X_{i*}W - Z_{i*}\|^2$$

Alternatively, this is equivalent to minimizing the (squared) Frobenius norm of the residual matrix:

$$\arg\min_W \|XW - Z\|_F^2$$

Consequently, W will be the so called least-squares solution of the linear matrix equation $XW = Z$. This is a well-known problem in linear algebra and can be solved by taking the Moore-Penrose pseudoinverse $X^+ = \left(X^TX\right)^{-1}X^T$ as $W = X^+Z$, which can be computed using SVD.

2.1 Orthogonality for monolingual invariance

Monolingual invariance is needed to preserve the dot products after mapping, avoiding performance degradation in monolingual tasks (e.g. analogy). This can be obtained requiring W to be an orthogonal matrix ($W^TW = I$). The exact solution under such orthogonality constraint is given by $W = VU^T$, where $Z^TX = U\Sigma V^T$ is the SVD factorization of Z^TX (cf. Appendix A). Thanks to this, the optimal transformation can be efficiently computed in linear time with respect to the vocabulary size. Note that orthogonality enforces an intuitive property, and as such it could be useful to avoid degenerated solutions and learn better bilingual mappings, as we empirically show in Section 3.

2.2 Length normalization for maximum cosine

Normalizing word embeddings in both languages to be unit vectors guarantees that all training instances contribute equally to the optimization goal. As long as W is orthogonal, this is equivalent to maximizing the sum of cosine similarities for the dictionary entries, which is commonly used for similarity computations:

$$\arg\min_W \sum_i \left\| \frac{X_{i*}}{\|X_{i*}\|}W - \frac{Z_{i*}}{\|Z_{i*}\|} \right\|^2$$
$$= \arg\max_W \sum_i \cos\left(X_{i*}W, Z_{i*}\right)$$

This last optimization objective coincides with Xing et al. (2015), but their work was motivated by an hypothetical inconsistency in Mikolov et al. (2013b), where the optimization objective to learn word embeddings uses dot product, the objective to learn mappings uses Euclidean distance and the similarity computations use cosine. However, the fact is that, as long as W is orthogonal, optimizing the squared Euclidean distance of length-normalized embeddings is equivalent to optimizing the cosine, and therefore, the mapping objective proposed by Xing et al. (2015) is equivalent to that used by Mikolov et al. (2013b) with orthogonality constraint and unit vectors. In fact, our experiments show that orthogonality is more relevant than length normalization, in contrast to Xing et al. (2015), who introduce orthogonality only to ensure that unit length is preserved after mapping.

2.3 Mean centering for maximum covariance

Dimension-wise mean centering captures the intuition that two randomly taken words would not be expected to be semantically similar, ensuring that the expected product of two random embeddings in any dimension and, consequently, their cosine similarity, is zero. As long as W is orthogonal, this is equivalent to maximizing the sum of dimension-wise covariance for the dictionary entries:

$$\arg\min_W \|C_mXW - C_mZ\|_F^2$$
$$= \arg\max_W \sum_i \text{cov}\left(XW_{*i}, Z_{*i}\right)$$

where C_m denotes the centering matrix

This equivalence reveals that the method proposed by Faruqui and Dyer (2014) is closely related to our framework. More concretely, Faruqui and Dyer (2014) use Canonical Correlation Analysis (CCA) to project the word embeddings in both languages to a shared vector space. CCA maximizes the dimension-wise covariance of both projections (which is equivalent to maximizing the covariance of a single projection if the transformations are constrained to be orthogonal, as in our case) but adds an implicit restriction to the two mappings, making different dimensions have the same variance and be uncorrelated among themselves[1]:

$$\begin{aligned} &\arg\max_{A,B} \sum_i \operatorname{cov}\left(XA_{*i}, ZB_{*i}\right) \\ &\text{s.t.} \quad A^T X^T C_m X A = B^T Z^T C_m Z B = I \end{aligned}$$

Therefore, the only fundamental difference between both methods is that, while our model enforces monolingual invariance, Faruqui and Dyer (2014) do change the monolingual embeddings to meet this restriction. In this regard, we think that the restriction they add could have a negative impact on the learning of the bilingual mapping, and it could also degrade the quality of the monolingual embeddings. Our experiments (cf. Section 3) show empirical evidence supporting this idea.

3 Experiments

In this section, we experimentally test the proposed framework and all its variants in comparison with related methods. For that purpose, we use the translation induction task introduced by Mikolov et al. (2013b), which learns a bilingual mapping on a small dictionary and measures its accuracy on predicting the translation of new words. Unfortunately, the dataset they use is not public. For that reason, we use the English-Italian dataset on the same task provided by Dinu et al. (2015)[2]. The dataset contains monolingual word embeddings trained with the word2vec toolkit using the CBOW method with negative sampling (Mikolov et al., 2013a)[3]. The English embeddings were trained on a 2.8 billion word corpus (ukWaC + Wikipedia + BNC), while the 1.6 billion word corpus itWaC was used to train the Italian embeddings. The dataset also contains a bilingual dictionary learned from Europarl, split into a training set of 5,000 word pairs and a test set of 1,500 word pairs, both of them uniformly distributed in frequency bins. Accuracy is the evaluation measure.

Apart from the performance of the projected embeddings in bilingual terms, we are also interested in the monolingual quality of the source language embeddings after the projection. For that purpose, we use the word analogy task proposed by Mikolov et al. (2013a), which measures the accuracy on answering questions like "what is the word that is similar to *small* in the same sense as *biggest* is similar to *big*?" using simple word vector arithmetic. The dataset they use consists of 8,869 semantic and 10,675 syntactic questions of this type, and is publicly available[4]. In order to speed up the experiments, we follow the authors and perform an approximate evaluation by reducing the vocabulary size according to a frequency threshold of 30,000 (Mikolov et al., 2013a). Since the original embeddings are the same in all the cases and it is only the transformation that is applied to them that changes, this affects all the methods in the exact same way, so the results are perfectly comparable among themselves. With these settings, we obtain a coverage of 64.98%.

We implemented the proposed method in Python using NumPy, and make it available as an open source project[5]. The code for Mikolov et al. (2013b) and Xing et al. (2015) is not publicly available, so we implemented and tested them as part of the proposed framework, which only differs from the original systems in the optimization method (exact solution instead of gradient descent) and the length normalization approach in the case of Xing et al. (2015) (postprocessing instead of constrained training). As for the method by Faruqui and Dyer (2014), we used their original implementation in Python and MATLAB[6], which we extended to cover cases where the dictionary contains more than one entry for the same word.

[1] While CCA is typically defined in terms of correlation (thus its name), correlation is invariant to the scaling of variables, so it is possible to constrain the canonical variables to have a fixed variance, as we do, in which case correlation and covariance become equivalent

[2] http://clic.cimec.unitn.it/~georgiana.dinu/down/

[3] The context window was set to 5 words, the dimension of the embeddings to 300, the sub-sampling to 1e-05 and the number of negative samples to 10

[4] https://code.google.com/archive/p/word2vec/

[5] https://github.com/artetxem/vecmap

[6] https://github.com/mfaruqui/crosslingual-cca

	EN-IT	EN AN.
Original embeddings	-	76.66%
Unconstrained mapping	34.93%	73.80%
+ length normalization	33.80%	73.61%
+ mean centering	38.47%	73.71%
Orthogonal mapping	36.73%	76.66%
+ length normalization	36.87%	76.66%
+ mean centering	39.27%	76.59%

Table 1: Our results in bilingual and monolingual tasks.

	EN-IT	EN AN.
Original embeddings	-	76.66%
Mikolov et al. (2013b)	34.93%	73.80%
Xing et al. (2015)	36.87%	76.66%
Faruqui and Dyer (2014)	37.80%	69.64%
Our method	39.27%	76.59%

71.81

40.53% 67.17 71.74

Table 2: Comparison of our method to other work.

3.1 Results of our framework

The rows in Table 1 show, respectively, the results for the original embeddings, the basic mapping proposed by Mikolov et al. (2013b) (cf. Section 2) and the addition of orthogonality constraint (cf. Section 2.1), with and without length normalization and, incrementally, mean centering. In all the cases, length normalization and mean centering were applied to all embeddings, even if missing from the dictionary.

The results show that the orthogonality constraint is key to preserve monolingual performance, and it also improves bilingual performance by enforcing a relevant property (monolingual invariance) that the transformation to learn should intuitively have. The contribution of length normalization alone is marginal, but when followed by mean centering we obtain further improvements in bilingual performance without hurting monolingual performance.

3.2 Comparison to other work

Table 2 shows the results for our best performing configuration in comparison to previous work. As discussed before, (Mikolov et al., 2013b) and (Xing et al., 2015) were implemented as part of our framework, so they correspond to our uncostrained mapping with no preprocessing and orthogonal mapping with length normalization, respectively.

As it can be seen, the method by Xing et al. (2015) performs better than that of Mikolov et al. (2013b) in the translation induction task, which is in line with what they report in their paper. Moreover, thanks to the orthogonality constraint their monolingual performance in the word analogy task does not degrade, whereas the accuracy of Mikolov et al. (2013b) drops by 2.86% in absolute terms with respect to the original embeddings.

Since Faruqui and Dyer (2014) take advantage of CCA to perform dimensionality reduction, we tested several values for it and report the best (180 dimensions). This beats the method by Xing et al. (2015) in the bilingual task, although it comes at the price of a considerable degradation in monolingual quality.

In any case, it is our proposed method with the orthogonality constraint and a global preprocessing with length normalization followed by dimension-wise mean centering that achieves the best accuracy in the word translation induction task. Moreover, it does not suffer from any considerable degradation in monolingual quality, with an anecdotal drop of only 0.07% in contrast with 2.86% for Mikolov et al. (2013b) and 7.02% for Faruqui and Dyer (2014).

When compared to Xing et al. (2015), our results in Table 1 reinforce our theoretical interpretation for their method (cf. Section 2.2), as it empirically shows that its improvement with respect to Mikolov et al. (2013b) comes solely from the orthogonality constraint, and not from solving any inconsistency.

It should be noted that the implementation by Faruqui and Dyer (2014) also length-normalizes the word embeddings in a preprocessing step. Following the discussion in Section 2.3, this means that our best performing configuration is conceptually very close to the method by Faruqui and Dyer (2014), as they both coincide on maximizing the average dimension-wise covariance and length-normalize the embeddings in both languages first, the only difference being that our model enforces monolingual invariance after the normalization while theirs does change the monolingual embeddings to make different dimensions have the same variance and be uncorrelated among themselves. However, our model performs considerably better than any configuration from Faruqui and Dyer (2014) in both the monolingual and the bilingual task, supporting our hypothesis that these two constraints that are implicit in their method are not only conceptually confusing, but also have a negative impact.

4 Conclusions

This paper develops a new framework to learn bilingual word embedding mappings, generalizing previous work and providing an efficient exact method to learn the optimal transformation. Our experiments show the effectiveness of the proposed model and give strong empirical evidence in favor of our reinterpretation of Xing et al. (2015) and Faruqui and Dyer (2014). It is the proposed method with the orthogonality constraint and a global preprocessing with length normalization and dimension-wise mean centering that achieves the best overall results both in monolingual and bilingual terms, surpassing those previous methods. In the future, we would like to study non-linear mappings (Lu et al., 2015) and the additional techniques in (Lazaridou et al., 2015).

Acknowledgments

This research was partially supported by the European Commision (QTLeap FP7-ICT-2013-10-610516), a Google Faculty Award, and the Spanish Ministry of Economy and Competitiveness (TADEEP TIN2015-70214-P). Mikel Artetxe enjoys a doctoral grant from the Spanish Ministry of Education, Culture and Sports.

A Proof of solution under orthogonality

Constraining W to be orthogonal ($W^TW = I$), the original minimization problem can be reformulated as follows (cf. Section 2.1):

$$\begin{aligned}
&\arg\min_W \sum_i \|X_{i*}W - Z_{i*}\|^2 \\
&= \arg\min_W \sum_i \left(\|X_{i*}W\|^2 + \|Z_{i*}\|^2 - 2X_{i*}WZ_{i*}^T\right) \\
&= \arg\max_W \sum_i X_{i*}WZ_{i*}^T \\
&= \arg\max_W \ \mathrm{Tr}\left(XWZ^T\right) \\
&= \arg\max_W \ \mathrm{Tr}\left(Z^TXW\right)
\end{aligned}$$

In the above expression, $\mathrm{Tr}(\cdot)$ denotes the trace operator (the sum of all the elements in the main diagonal), and the last equality is given by its cyclic property. At this point, we can take the SVD of Z^TX as $Z^TX = U\Sigma V^T$, so $\mathrm{Tr}\left(Z^TXW\right) = \mathrm{Tr}\left(U\Sigma V^TW\right) = \mathrm{Tr}\left(\Sigma V^TWU\right)$. Since V^T, W and U are orthogonal matrices, their product V^TWU will also be an orthogonal matrix. In addition to that, given that Σ is a diagonal matrix, its trace after an orthogonal transformation will be maximal when the values in its main diagonal are preserved after the mapping, that is, when the orthogonal transformation matrix is the identity matrix. This will happen when $V^TWU = I$ in our case, so the optimal solution will be $W = VU^T$.

References

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 *Advances in Neural Information Processing Systems 27*, pages 1853–1861.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR2015), workshop track.*

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.

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.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2014. Bilbowa: Fast bilingual distributed representations without word alignments. *arXiv preprint arXiv:1410.2455.*

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, pages 224–229.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. 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, pages 270–280.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256.

Min-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*, pages 151–159.

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, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Michael Wick, Pallika Kanani, and Adam Pocock. 2016. Minimally-constrained multilingual embeddings via artificial code-switching. In *Thirtieth AAAI conference on Artificial Intelligence (AAAI)*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.

Kai Zhao, Hany Hassan, and Michael Auli. 2015. Learning translation models from monolingual continuous representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1527–1536.

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.