Learning Bilingual Word Representations by Marginalizing Alignments

Tomáš Kočiský **Karl Moritz Hermann** **Phil Blunsom**

Department of Computer Science
University of Oxford
Oxford, OX1 3QD, UK
{tomas.kocisky,karl.moritz.hermann,phil.blunsom}@cs.ox.ac.uk

Abstract

We present a probabilistic model that simultaneously learns alignments and distributed representations for bilingual data. By marginalizing over word alignments the model captures a larger semantic context than prior work relying on hard alignments. The advantage of this approach is demonstrated in a cross-lingual classification task, where we outperform the prior published state of the art.

1 Introduction

Distributed representations have become an increasingly important tool in machine learning. Such representations—typically continuous vectors learned in an unsupervised setting—can frequently be used in place of hand-crafted, and thus expensive, features. By providing a richer representation than what can be encoded in discrete settings, distributed representations have been successfully used in many areas. This includes AI and reinforcement learning (Mnih et al., 2013), image retrieval (Kiros et al., 2013), language modelling (Bengio et al., 2003), sentiment analysis (Socher et al., 2011; Hermann and Blunsom, 2013), frame-semantic parsing (Hermann et al., 2014), and document classification (Klementiev et al., 2012).

In Natural Language Processing (NLP), the use of distributed representations is motivated by the idea that they could capture semantics and/or syntax, as well as encoding a continuous notion of similarity, thereby enabling information sharing between similar words and other units. The success of distributed approaches to a number of tasks, such as listed above, supports this notion and its implied benefits (see also Turian et al. (2010) and Collobert and Weston (2008)).

While most work employing distributed representations has focused on monolingual tasks, multilingual representations would also be useful for several NLP-related tasks. Such problems include document classification, machine translation, and cross-lingual information retrieval, where multilingual data is frequently the norm. Furthermore, learning multilingual representations can also be useful for cross-lingual information transfer, that is exploiting resource-fortunate languages to generate supervised data in resource-poor ones.

We propose a probabilistic model that simultaneously learns word alignments and bilingual distributed word representations. As opposed to previous work in this field, which has relied on hard alignments or bilingual lexica (Klementiev et al., 2012; Mikolov et al., 2013), we marginalize out the alignments, thus capturing more bilingual semantic context. Further, this results in our distributed word alignment (DWA) model being the first probabilistic account of bilingual word representations. This is desirable as it allows better reasoning about the derived representations and furthermore, makes the model suitable for inclusion in higher-level tasks such as machine translation.

The contributions of this paper are as follows. We present a new probabilistic similarity measure which is based on an alignment model and prior language modeling work which learns and relates word representations across languages. Subsequently, we apply these embeddings to a standard document classification task and show that they outperform the current published state of the art (Hermann and Blunsom, 2014b). As a by-product we develop a distributed version of FASTALIGN (Dyer et al., 2013), which performs on par with the original model, thereby demonstrating the efficacy of the learned bilingual representations.

2 Background

The IBM alignment models, introduced by Brown et al. (1993), form the basis of most statistical machine translation systems. In this paper we base our alignment model on FASTALIGN (FA), a vari-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 224–229,
Baltimore, Maryland, USA, June 23-25 2014.

ation of IBM model 2 introduced by Dyer et al. (2013). This model is both fast and produces alignments on par with the state of the art. Further, to induce the distributed representations we incorporate ideas from the log-bilinear language model presented by Mnih and Hinton (2007).

2.1 IBM Model 2

Given a parallel corpus with aligned sentences, an alignment model can be used to discover matching words and phrases across languages. Such models are an integral part of most machine translation pipelines. An alignment model learns $p(\mathbf{f}, \mathbf{a}|\mathbf{e})$ (or $p(\mathbf{e}, \mathbf{a}'|\mathbf{f})$) for the source and target sentences $\mathbf{e}$ and $\mathbf{f}$ (sequences of words). $\mathbf{a}$ represents the word alignment across these two sentences from source to target. IBM model 2 (Brown et al., 1993) learns alignment and translation probabilities in a generative style as follows:

$$p(\mathbf{f}, \mathbf{a}|\mathbf{e}) = p(J|I) \prod_{j=1}^{J} p(a_j|j, I, J)\, p\big(f_j|e_{a_j}\big)\,,$$

where $p(J|I)$ captures the two sentence lengths; $p(a_j|j, I, J)$ the alignment and $p\big(f_j|e_{a_j}\big)$ the translation probability. Sentence likelihood is given by marginalizing out the alignments, which results in the following equation:

$$p(\mathbf{f}|\mathbf{e}) = p(J|I) \prod_{j=1}^{J} \sum_{i=0}^{I} p(i|j, I, J)\, p(f_j|e_i)\,.$$

We use FASTALIGN (FA) (Dyer et al., 2013), a log-linear reparametrization of IBM model 2. This model uses an alignment distribution defined by a single parameter that measures how close the alignment is to the diagonal. This replaces the original multinomial alignment distribution which often suffered from sparse counts. This improved model was shown to run an order of magnitude faster than IBM model 4 and yet still outperformed it in terms of the BLEU score and, on Chinese-English data, in alignment error rate (AER).

2.2 Log-Bilinear Language Model

Language models assign a probability measure to sequences of words. We use the log-bilinear language model proposed by Mnih and Hinton (2007). It is an n-gram based model defined in terms of an energy function $\mathrm{E}(w_n; w_{1:n-1})$. The probability for predicting the next word w_n given its preceding context of $n-1$ words is expressed using the energy function

$$\mathrm{E}(w_n; w_{1:n-1}) = -\left(\sum_{i=1}^{n-1} r_{w_i}^T C_i\right) r_{w_n} - b_r^T r_{w_n} - b_{w_n}$$

as $p(w_n|w_{1:n-1}) = \frac{1}{Z_c} \exp\left(-\mathrm{E}(w_n; w_{1:n-1})\right)$ where $Z_c = \sum_{w_n} \exp\left(-\mathrm{E}(w_n; w_{1:n-1})\right)$ is the normalizer, $r_{w_i} \in \mathbb{R}^d$ are word representations, $C_i \in \mathbb{R}^{d \times d}$ are context transformation matrices, and $b_r \in \mathbb{R}^d$, $b_{w_n} \in \mathbb{R}$ are representation and word biases respectively. Here, the sum of the transformed context-word vectors endeavors to be close to the word we want to predict, since the likelihood in the model is maximized when the energy of the observed data is minimized.

This model can be considered a variant of a log-linear language model in which, instead of defining binary n-gram features, the model learns the features of the input and output words, and a transformation between them. This provides a vastly more compact parameterization of a language model as n-gram features are not stored.

2.3 Multilingual Representation Learning

There is some recent prior work on multilingual distributed representation learning. Similar to the model presented here, Klementiev et al. (2012) and Zou et al. (2013) learn bilingual embeddings using word alignments. These two models are non-probabilistic and conditioned on the output of a separate alignment model, unlike our model, which defines a probability distribution over translations and marginalizes over all alignments. These models are also highly related to prior work on bilingual lexicon induction (Haghighi et al., 2008). Other recent approaches include Sarath Chandar et al. (2013), Lauly et al. (2013) and Hermann and Blunsom (2014a, 2014b). These models avoid word alignment by transferring information across languages using a composed sentence-level representation.

While all of these approaches are related to the model proposed in this paper, it is important to note that our approach is novel by providing a probabilistic account of these word embeddings. Further, we learn word alignments and simultaneously use these alignments to guide the representation learning, which could be advantageous particularly for rare tokens, where a sentence based approach might fail to transfer information.

Related work also includes Mikolov et al. (2013), who learn a transformation matrix to

reconcile monolingual embedding spaces, in an l_2 norm sense, using dictionary entries instead of alignments, as well as Schwenk et al. (2007) and Schwenk (2012), who also use distributed representations for estimating translation probabilities. Faruqui and Dyer (2014) use a technique based on CCA and alignments to project monolingual word representations to a common vector space.

3 Model

Here we describe our distributed word alignment (DWA) model. The DWA model can be viewed as a distributed extension of the FA model in that it uses a similarity measure over distributed word representations instead of the standard multinomial translation probability employed by FA. We do this using a modified version of the log-bilinear language model in place of the translation probabilities $p(f_j|e_i)$ at the heart of the FA model. This allows us to learn word representations for both languages, a translation matrix relating these vector spaces, as well as alignments at the same time.

Our modifications to the log-bilinear model are as follows. Where the original log-bilinear language model uses context words to predict the next word—this is simply the distributed extension of an n-gram language model—we use a word from the source language in a parallel sentence to predict a target word. An additional aspect of our model, which demonstrates its flexibility, is that it is simple to include further context from the source sentence, such as words around the aligned word or syntactic and semantic annotations. In this paper we experiment with a transformed sum over k context words to each side of the aligned source word. We evaluate different context sizes and report the results in Section 5. We define the energy function for the translation probabilities to be

$$\mathrm{E}(f, e_i) = -\left(\sum_{s=-k}^{k} r_{e_{i+s}}^T T_s\right) r_f - b_r^T r_f - b_f \quad (1)$$

where $r_{e_i}, r_f \in \mathbb{R}^d$ are vector representations for source and target words $e_{i+s} \in V_E, f \in V_F$ in their respective vocabularies, $T_s \in \mathbb{R}^{d \times d}$ is the transformation matrix for each surrounding context position, $b_r \in \mathbb{R}^d$ are the representation biases, and $b_f \in \mathbb{R}$ is a bias for each word $f \in V_F$.

The translation probability is given by $p(f|e_i) = \frac{1}{Z_{e_i}} \exp\left(-\mathrm{E}(f, e_i)\right)$, where $Z_{e_i} = \sum_f \exp\left(-\mathrm{E}(f, e_i)\right)$ is the normalizer.

In addition to these translation probabilities, we have parameterized the translation probabilities for the null word using a softmax over an additional weight vector.

3.1 Class Factorization

We improve training performance using a class factorization strategy (Morin and Bengio, 2005) as follows. We augment the translation probability to be $p(f|e) = p(c_f|e)\, p(f|c_f, e)$ where c_f is a unique predetermined class of f; the class probability is modeled using a similar log-bilinear model as above, but instead of predicting a word representation r_f we predict the class representation r_{c_f} (which is learned with the model) and we add respective new context matrices and biases. Note that the probability of the word f depends on *both* the class and the given context words: it is normalized only over words in the class c_f.

In our training we create classes based on word frequencies in the corpus as follows. Considering words in the order of their decreasing frequency, we add word types into a class until the total frequency of the word types in the currently considered class is less than $\frac{\text{total tokens}}{\sqrt{|V_F|}}$ and the class size is less than $\sqrt{|V_F|}$. We have found that the maximal class size affects the speed the most.

4 Learning

The original FA model optimizes the likelihood using the expectation maximization (EM) algorithm where, in the M-step, the parameter update is analytically solvable, except for the λ parameter (the diagonal tension), which is optimized using gradient descent (Dyer et al., 2013). We modified the implementations provided with CDEC (Dyer et al., 2010), retaining its default parameters.

In our model, DWA, we optimize the likelihood using the EM as well. However, while training we fix the counts of the E-step to those computed by FA, trained for the default 5 iterations, to aid the convergence rate, and optimize the M-step only. Let θ be the parameters for our model. Then the gradient for each sentence is given by

$$\frac{\partial}{\partial\theta} \log p(\mathbf{f}|\mathbf{e}) = \sum_{k=1}^{J} \sum_{l=0}^{I} \left[\frac{p(l|k, I, J)\, p(f_k|e_l)}{\sum_{i=0}^{I} p(i|k, I, J)\, p(f_k|e_i)} \cdot \frac{\partial}{\partial\theta} \log(p(l|k, I, J)\, p(f_k|e_l)) \right]$$

where the first part are the counts from the FA model and second part comes from our model.

We compute the gradient for the alignment probabilities in the same way as in the FA model, and the gradient for the translation probabilities using back-propagation (Rumelhart et al., 1986). For parameter update, we use ADAGRAD as the gradient descent algorithm (Duchi et al., 2011).

5 Experiments

We first evaluate the alignment error rate of our approach, which establishes the model's ability to both learn alignments as well as word representations that explain these alignments. Next, we use a cross-lingual document classification task to verify that the representations are semantically useful. We also inspect the embedding space qualitatively to get some insight into the learned structure.

5.1 Alignment Evaluation

We compare the alignments learned here with those of the FASTALIGN model which produces very good alignments and translation BLEU scores. We use the same language pairs and datasets as in Dyer et al. (2013), that is the FBIS Chinese-English corpus, and the French-English section of the Europarl corpus (Koehn, 2005). We used the preprocessing tools from CDEC and further replaced all unique tokens with UNK. We trained our models with 100 dimensional representations for up to 40 iterations, and the FA model for 5 iterations as is the default.

Table 1 shows that our model learns alignments on part with those of the FA model. This is in line with expectation as our model was trained using the FA expectations. However, it confirms that the learned word representations are able to explain translation probabilities. Surprisingly, context seems to have little impact on the alignment error, suggesting that the model receives sufficient information from the aligned words themselves.

5.2 Document Classification

A standard task for evaluating cross-lingual word representations is document classification where training is performed in one and evaluation in another language. This tasks require semantically plausible embeddings (for classification) which are valid across two languages (for the semantic transfer). Hence this task requires more of the word embeddings than the previous task.

Languages	Model		
	FA	DWA $k=0$	DWA $k=3$
ZH|EN	49.4	48.4	48.7
EN|ZH	44.9	45.3	45.9
FR|EN	17.1	17.2	17.0
EN|FR	16.6	16.3	16.1

Table 1: Alignment error rate (AER) comparison, in both directions, between the FASTALIGN (FA) alignment model and our model (DWA) with k context words (see Equation 1). Lower numbers indicate better performance.

We mainly follow the setup of Klementiev et al. (2012) and use the German-English parallel corpus of the European Parliament proceedings to train the word representations. We perform the classification task on the Reuters RCV1/2 corpus. Unlike Klementiev et al. (2012), we do not use that corpus during the representation learning phase. We remove all words occurring less than five times in the data and learn 40 dimensional word embeddings in line with prior work.

To train a classifier on English data and test it on German documents we first project word representations from English into German: we select the most probable German word according to the learned translation probabilities, and then compute document representations by averaging the word representations in each document. We use these projected representations for training and subsequently test using the original German data and representations. We use an averaged perceptron classifier as in prior work, with the number of epochs (3) tuned on a subset of the training set.

Table 2 shows baselines from previous work and classification accuracies. Our model outperforms the model by Klementiev et al. (2012), and it also outperforms the most comparable models by Hermann and Blunsom (2014b) when training on German data and performs on par with it when training on English data.[1] It seems that our model learns more informative representations towards document classification, even without additional monolingual language models or context information. Again the impact of context is inconclusive.

[1]From Hermann and Blunsom (2014a, 2014b) we only compare with models equivalent with respect to embedding dimensionality and training data. They still achieve the state of the art when using additional training data.

Model	en → de	de → en
Majority class	46.8	46.8
Glossed	65.1	68.6
MT	68.1	67.4
Klementiev et al.	77.6	71.1
BiCVM ADD	**83.7**	71.4
BiCVM BI	83.4	69.2
DWA ($k = 0$)	82.8	**76.0**
DWA ($k = 3$)	83.1	75.4

Table 2: Document classification accuracy when trained on 1,000 training examples of the RCV1/2 corpus (train→test). Baselines are the majority class, glossed, and MT (Klementiev et al., 2012). Further, we are comparing to Klementiev et al. (2012), BiCVM ADD (Hermann and Blunsom, 2014a), and BiCVM BI (Hermann and Blunsom, 2014b). k is the context size, see Equation 1.

5.3 Representation Visualization

Following the document classification task we want to gain further insight into the types of features our embeddings learn. For this we visualize word representations using t-SNE projections (van der Maaten and Hinton, 2008). Figure 1 shows an extract from our projection of the 2,000 most frequent German words, together with an expected representation of a translated English word given translation probabilities. Here, it is interesting to see that the model is able to learn related representations for words *chair* and *ratspräsidentschaft* (presidency) even though these words were not aligned by our model. Figure 2 shows an extract from the visualization of the 10,000 most frequent English words trained on another corpus. Here again, it is evident that the embeddings are semantically plausible with similar words being closely aligned.

6 Conclusion

We presented a new probabilistic model for learning bilingual word representations. This distributed word alignment model (DWA) learns both representations and alignments at the same time. We have shown that the DWA model is able to learn alignments on par with the FASTALIGN alignment model which produces very good alignments, thereby determining the efficacy of the learned representations which are used to calculate

Figure 1: A visualization of the expected representation of the translated English word *chair* among the nearest German words: words never aligned (green), and those seen aligned (blue) with it.

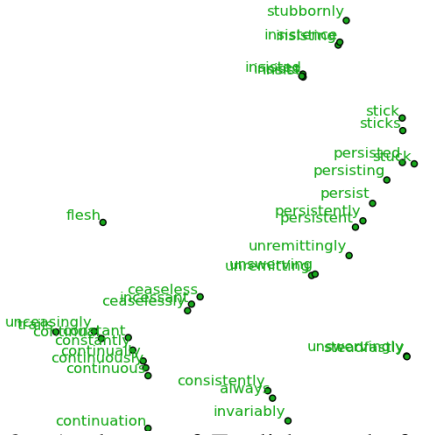

Figure 2: A cluster of English words from the 10,000 most frequent English words visualized using t-SNE. Word representations were optimized for $p(\text{zh}|\text{en})$ ($k = 0$).

word translation probabilities for the alignment task. Subsequently, we have demonstrated that our model can effectively be used to project documents from one language to another. The word representations our model learns as part of the alignment process are semantically plausible and useful. We highlighted this by applying these embeddings to a cross-lingual document classification task where we outperform prior work, achieve results on par with the current state of the art and provide new state-of-the-art results on one of the tasks. Having provided a probabilistic account of word representations across multiple languages, future work will focus on applying this model to machine translation and related tasks, for which previous approaches of learning such embeddings are less suited. Another avenue for further study is to combine this method with monolingual language models, particularly in the context of semantic transfer into resource-poor languages.

Acknowledgements

This work was supported by a Xerox Foundation Award and EPSRC grant number EP/K036580/1. We acknowledge the use of the Oxford ARC.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, February.

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

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of ICML*.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, July.

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

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of NAACL-HLT*.

Manaal Faruqui and Chris Dyer. 2014. Improving Vector Space Word Representations Using Multilingual Correlation. In *Proceedings of EACL*.

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

Karl Moritz Hermann and Phil Blunsom. 2013. The Role of Syntax in Vector Space Models of Compositional Semantics. In *Proceedings of ACL*.

Karl Moritz Hermann and Phil Blunsom. 2014a. Multilingual Distributed Representations without Word Alignment. In *Proceedings of ICLR*.

Karl Moritz Hermann and Phil Blunsom. 2014b. Multilingual Models for Compositional Distributional Semantics. In *Proceedings of ACL*.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic Frame Identification with Distributed Word Representations. In *Proceedings of ACL*.

Ryan Kiros, Richard S Zemel, and Ruslan Salakhutdinov. 2013. Multimodal neural language models. In *NIPS Deep Learning Workshop*.

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

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*.

Stanislas Lauly, Alex Boulanger, and Hugo Larochelle. 2013. Learning multilingual word representations using a bag-of-words autoencoder. In *NIPS Deep Learning Workshop*.

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

Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of ICML*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*.

Frederic Morin and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 246–252.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323:533–536, October.

A P Sarath Chandar, M Khapra Mitesh, B Ravindran, Vikas Raykar, and Amrita Saha. 2013. Multilingual deep learning. In *Deep Learning Workshop at NIPS*.

Holger Schwenk, Marta R. Costa-jussa, and Jose A. R. Fonollosa. 2007. Smooth bilingual n-gram translation. In *Proceedings of EMNLP-CoNLL*.

Holger Schwenk. 2012. Continuous space translation models for phrase-based statistical machine translation. In *Proceedings of COLING: Posters*.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of EMNLP*.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL*.

L.J.P. van der Maaten and G.E. Hinton. 2008. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *Proceedings of EMNLP*.