

---

# Multilingual Distributed Representations without Word Alignment

---

**Karl Moritz Hermann and Phil Blunsom**

Department of Computer Science

University of Oxford

Oxford, OX1 3QD, UK

{karl.moritz.hermann, phil.blunsom}@cs.ox.ac.uk

## Abstract

Distributed representations of meaning are a natural way to encode covariance relationships between words and phrases in NLP. By overcoming data sparsity problems, as well as providing information about semantic relatedness which is not available in discrete representations, distributed representations have proven useful in many NLP tasks. Recent work has shown how compositional semantic representations can successfully be applied to a number of monolingual applications such as sentiment analysis. At the same time, there has been some initial success in work on learning shared word-level representations across languages. We combine these two approaches by proposing a method for learning distributed representations in a multilingual setup. Our model learns to assign similar embeddings to aligned sentences and dissimilar ones to sentence which are not aligned while not requiring word alignments. We show that our representations are semantically informative and apply them to a cross-lingual document classification task where we outperform the previous state of the art. Further, by employing parallel corpora of multiple language pairs we find that our model learns representations that capture semantic relationships across languages for which no parallel data was used.

## 1 Introduction

Distributed representations of words are increasingly being used to achieve high levels of generalisation within language modelling tasks. Successful applications of this approach include word-sense disambiguation, word similarity and synonym detection (e.g. [10, 27]). Subsequent work has also attempted to learn distributed semantics of larger structures, allowing us to apply distributed representation to tasks such as sentiment analysis or paraphrase detection (i.a. [1, 3, 12, 14, 21, 25]). At the same time a second strand of work has focused on transferring linguistic knowledge across languages, and particularly from English into low-resource languages, by means of distributed representations at the word level [13, 16].

Currently, work on compositional semantic representations focuses on monolingual data while the cross-lingual work focuses on word level representations only. However, it appears logical that these two strands of work should be combined as there exists a plethora of parallel corpora with aligned data at the sentence level or beyond which could be exploited in such work. Further, sentence aligned data provides a plausible concept of semantic similarity, which can be harder to define at the word level. Consider the case of alignment between a German compound noun (e.g. “Schwerlastverkehr”) and its English equivalent (“heavy goods vehicle traffic”). Semantic alignment at the phrase level here appears far more plausible than aligning individual tokens for semantic transfer.

Using this rationale, and building on both work related to learning cross-lingual embeddings as well as to compositional semantic representations, we introduce a model that learns cross-lingual embeddings at the sentence level. In the following section we will briefly discuss prior work in these two fields before going on to describe the bilingual training signal that we developed for learning multilingual compositional embeddings. Subsequently, we will describe our model in greater detail as well as its training procedure and experimental setup. Finally, we perform a number of evaluations and demonstrate that our training signal allows a very simple compositional vector model to outperform the state of the art on a task designed to evaluate its ability to transfer semantic information across languages. Unlike other work in this area, our model does not require word aligned data. In fact, while we evaluate our model on sentence aligned data in this paper, there is no theoretical requirement for this and technically our algorithm could also be applied to document-level parallel data or even comparable data only.

## 2 Models of Compositional Distributed Semantics

In the case of representing individual words as vectors, the distributional account of semantics provides a plausible explanation of what is encoded in a word vector. This follows the idea that the meaning of a word can be determined by “the company it keeps” [11], that is by the context it appears in. Such context can easily be encoded in vectors using collocational methods, and is also underlying other methods of learning word embeddings [7, 20].

For a number of important problems, semantic representations of individual words do not suffice, but instead a semantic representation of a larger structure—e.g. a phrase or a sentence—is required. This was highlighted in [10], who proposed a mechanism for modifying a word’s representation based on its individual context. The distributional account of semantics can, due to sparsity, not be applied to such larger linguistic units. A notable exception perhaps is Baroni and Zamparelli [1], who learned distributional representations for adjective noun pairs using a collocational approach on a corpus of unprecedented size. The bigram representations learned from that corpus were subsequently used to learn lexicalised composition functions for the constituent words.

Most alternative attempts to extract such higher-level semantic representations have focused on learning composition functions that represent the semantics of a larger structure as a function of the representations of its parts. [21] provides an evaluation of a number of simple composition functions applied to bigrams. Applied recursively, such approaches can then easily be reconciled with the co-occurrence based word level representations. There are a number of proposals motivating such recursive or deep composition models. Notably, [3] propose a tensor-based model for semantic composition and, similarly, [4] develop a framework for semantic composition by combining distributional theory with pregroup grammars. The latter framework was empirically evaluated and supported by the results in [12]. More recently, various forms of recursive neural networks have successfully been used for semantic composition and related tasks such as sentiment analysis. Such models include recursive autoencoders [24], matrix-vector recursive neural networks [25], untied recursive neural networks [14] or convolutional networks [15].

### 2.1 Multilingual Embeddings

Much research has been devoted to the task of inducing distributed semantic representations for single languages. In particular English, with its large number of annotated resources, has enjoyed most attention. Recently, progress has been made at representation learning for languages with fewer available resources. Klementiev et al. [16] described a form of multitask learning on word-aligned parallel data to transfer embeddings from one language to another. Earlier work, Haghighi et al. [13], proposed a method for inducing cross-lingual lexica using monolingual feature representations and a small initial lexicon to bootstrap with. This approach has recently been extended by [18, 19], who developed a method for learning transformation matrices to convert semantic vectors of one language into those of another. It was demonstrated that this approach can be applied to improve tasks related to machine translation. Their CBOW model is also worth noting for its similarities to the composition function used in this paper. Using a slightly different approach, [29], also learned bilingual embeddings for machine translation. It is important to note that, unlike our proposed system, all of these methods require word aligned parallel data for training.

Two recent workshop papers deserve mention in this respect. Both Lauly et al. [17] and Sarath Chandar et al. [23] propose methods for learning word embeddings by exploiting bilingual data, not unlike the method proposed in this paper. Instead of the noise-contrastive method developed in this paper, both groups of authors make use of autoencoders to encode monolingual representations and to support the bilingual transfer.

So far almost all of this work has been focused on learning multilingual representations at the word level. As distributed representations of larger expressions have been shown to be highly useful for a number of tasks, it seems to be a natural next step to also attempt to induce these using cross-lingual data. This paper provides a first step in that direction.

### 3 Model Description

Language acquisition in humans is widely seen as grounded in sensory-motor experience [22, 2]. Based on this idea, there have been some attempts at using multi-modal data for learning better vector representations of words (e.g. [26]). Such methods, however, are not easily scalable across languages or to large amounts of data for which no secondary or tertiary representation might exist.

We abstract the underlying principle one step further and attempt to learn semantics from multilingual data. The idea is that, given enough parallel data, a shared representation would be forced to capture the common elements between sentences from different languages. What two parallel sentences have in common, of course, is the semantics of those two sentences. Using this data, we propose a novel method for learning vector representations at the word level and beyond.

#### 3.1 Bilingual Signal

Exploiting the semantic similarity of parallel sentences across languages, we can define a simple bilingual (and trivially multilingual) error function as follows: Given a compositional sentence model (CVM)  $\mathcal{M}_A$ , which maps a sentence to a vector, we can train a second CVM  $\mathcal{M}_B$  using a corpus  $\mathcal{C}_{A,B}$  of parallel data from the language pair  $A, B$ . For each pair of parallel sentences  $(a, b) \in \mathcal{C}_{A,B}$ , we attempt to minimize

$$E_{dist}(a, b) = \|a_{root} - b_{root}\|^2 \quad (1)$$

where  $a_{root}$  is the vector representing sentence  $a$  and  $b_{root}$  the vector representing sentence  $b$ .

#### 3.2 The BiCVM Model

A CVM learns semantic representations of larger syntactic units given the semantic representations of their constituents. We assume individual words to be represented by vectors ( $x \in \mathbb{R}^d$ ).

Previous methods employ binary parse trees on the data (e.g. [14, 25]) and use weighted or multiplicative composition functions. Under such a setup, where each node in the tree is terminal or has two children ( $p \rightarrow c_0, c_1$ ), a binary composition function could take the following form:

$$p = g(W^e[c_0; c_1] + b^e) \quad (2)$$

where  $[c_0; c_1]$  is the concatenation of the two child vectors,  $W^e \in \mathbb{R}^{d \times 2d}$  and  $b^e \in \mathbb{R}^d$  the encoding matrix and bias, respectively, and  $g$  an element-wise activation function such as the hyperbolic tangent. For the purposes of evaluation the bilingual signal proposed above, we simplify this composition function by setting all weight matrices to the identity and all biases to zero. Thereby the CVM reduces to a simple additive composition function:

$$a_{root} = \sum_{i=0}^{|a|} a_i \quad (3)$$

Of course, this is a very simplified CVM, as such a bag-of-words approach no longer accounts for word ordering and other effects which a more complex CVM might capture. However, for the purposes of this evaluation (and with the experimental evaluation in mind), such a simplistic composition function should be sufficient to evaluate the novel objective function proposed here.

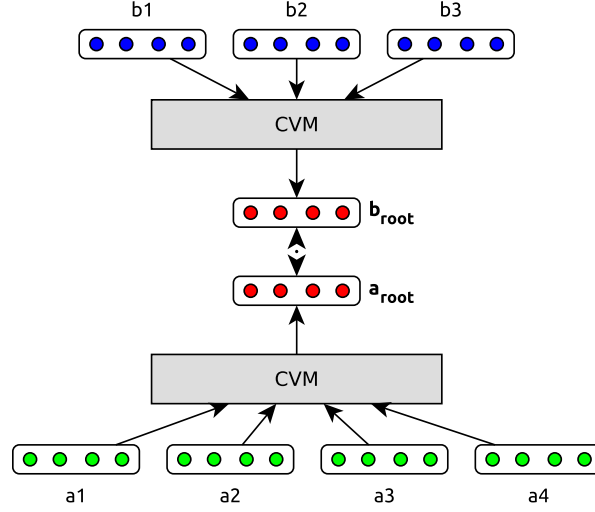


Figure 1: Description of a bilingual model with parallel input sentences  $a$  and  $b$ . The objective function of this model is to minimize the distance between the sentence level encoding of the bi-text. Principally any composition function can be used to generate the compositional sentence level representations. The composition function is represented by the *CVM* boxes in the diagram above.

Using this additive CVM we want to optimize the bilingual error signal defined above (Eq. 1). For the moment, assume that  $\mathcal{M}_A$  is a perfectly trained CVM such that  $a_{root}$  represents the semantics of the sentence  $a$ . Further, due to the use of parallel data, we know that  $a$  and  $b$  are semantically equivalent. Hence we transfer the semantic knowledge contained in  $\mathcal{M}_A$  onto  $\mathcal{M}_B$ , by learning  $\theta_{\mathcal{M}_B}$  to minimize:

$$E_{bi}(\mathcal{C}_{A,B}) = \sum_{(a,b) \in \mathcal{C}_{A,B}} E_{dist}(a, b) \quad (4)$$

Of course, this objective function assumes a fully trained model which we do not have at this stage. While this can be a useful objective for transferring linguistic knowledge into low-resource languages [16], this precondition is not helpful when there is no model to learn from in first place. We resolve this issue by jointly training both models  $\mathcal{M}_A$  and  $\mathcal{M}_B$ .

Applying  $E_{bi}$  to parallel data ensures that both models learn a shared representation at the sentence level. As the parallel input sentences share the same meaning, it is reasonable to assume that minimizing  $E_{bi}$  will force the model to learn their semantic representation. Let  $\theta_{bi} = \theta_{\mathcal{M}_A} \cup \theta_{\mathcal{M}_B}$ . The joint objective function  $J(\theta_{bi})$  thus becomes:

$$J(\theta_{bi}) = E_{bi}(\mathcal{C}_{A,B}) + \frac{\lambda}{2} \|\theta_{bi}\|^2 \quad (5)$$

where  $\lambda \|\theta_{bi}\|_1$  is the  $L_2$  regularization term.

It is apparent that this joint objective  $J(\theta_{bi})$  is degenerate. The models could learn to reduce all embeddings and composition weights to zero and thereby minimize the objective function. We address this issue by employing a form of contrastive estimation penalizing small distances between non-parallel sentence pairs. For every pair of parallel sentences  $(a, b)$  we sample a number of additional sentences  $n \in \mathcal{C}_B$ , which—with high probability—are not exact translations of  $a$ . This is comparable to the second term of the loss function of a large margin nearest neighbour classifier (see Eq. 12 in [28]):

$$E_{noise}(a, b, n) = [1 + E_{dist}(a, b) - E_{dist}(a, n)]_+ \quad (6)$$

where  $[x]_+ = \max(x, 0)$  denotes the standard hinge loss. Thus, the final objective function to minimize for the BICVM model is:

$$J(\theta_{bi}) = \sum_{(a,b) \in \mathcal{C}_{A,B}} \left( \sum_{i=1}^k E_{noise}(a, b, n_i) \right) + \frac{\lambda}{2} \|\theta_{bi}\|^2 \quad (7)$$

### 3.3 Model Learning

Given the objective function as defined above, model learning can employ the same techniques as any monolingual CVM. In particular, as the objective function is differentiable, we can use standard gradient descent techniques such as stochastic gradient descent, L-BFGS or the adaptive gradient algorithm AdaGrad [8]. Within each monolingual CVM, we use backpropagation through structure after applying the joint error to each sentence level node.

## 4 Experiments

### 4.1 Data and Parameters

All model weights were randomly initialised using a Gaussian distribution. There are a number of parameters that can influence model training. We selected the following values for simplicity and comparability with prior work. In future work we will investigate the effect of these parameters in greater detail. L2 regularization (1), step-size (0.1), number of noise elements (50), margin size (50), embedding dimensionality ( $d=40$ ). The noise elements samples were randomly drawn from the corpus at training time, individually for each training sample and epoch. We use the Europarl corpus (v7)<sup>1</sup> for training the bilingual model. The corpus was pre-processed using the set of tools provided by cdec<sup>2</sup> [9] for tokenizing and lowercasing the data. Further, all empty sentences as well as their translations were removed from the corpus.

We present results from two experiments. The BICVM model was trained on 500k sentence pairs of the English-German parallel section of the Europarl corpus. The BICVM+ model used this dataset in combination with another 500k parallel sentences from the English-French section of the corpus, resulting in 1 million English sentences, each paired up with either a German or a French sentence. Each language’s vocabulary used distinct encodings to avoid potential overlap.

The motivation behind BICVM+ is to investigate whether we can learn better embeddings by introducing additional data in a different language. This is similar to prior work in machine translation where English was used as a pivot for translation between low-resource languages [5].

We use the adaptive gradient method, AdaGrad [8], for updating the weights of our models, and terminate training after 50 iterations. Earlier experiments indicated that the BICVM model converges faster than the BICVM+ model, but we report results on the same number of iterations for better comparability<sup>3</sup>.

### 4.2 Cross-Lingual Document Classification

We evaluate our model using the cross-lingual document classification (CLDC) task of Klementiev et al. [16]. This task involves learning language independent embeddings which are then used for document classification across the English-German language pair. For this, CLDC employs a particular kind of supervision, namely using supervised training data in one language and evaluating without supervision in another. Thus, CLDC is a good task for establishing whether our learned representations are semantically useful across multiple languages.

We follow the experimental setup described in [16], with the exception that we learn our embeddings using solely the Europarl data and only use the Reuters RCV1/RCV2 corpora during the classifier training and testing stages. Each document in the classification task is represented by the average

<sup>1</sup><http://www.statmt.org/europarl/>

<sup>2</sup><https://github.com/redpony/cdec>

<sup>3</sup>These numbers were updated following comments in the ICLR open review process. Results for other dimensionalities and our source code for our model are available at <http://www.karlmoritz.com>.

Model	en $\rightarrow$ de	de $\rightarrow$ en
Majority Class	46.8	46.8
Glossed	65.1	68.6
MT	68.1	67.4
I-Matrix	77.6	71.1
BICVM	83.7	71.4
BICVM+	<b>86.2</b>	<b>76.9</b>

Table 1: Classification accuracy for training on English and German with 1000 labeled examples. Cross-lingual compositional representations (BICVM and BICVM+), cross-lingual representations using learned embeddings and an interaction matrix (I-Matrix) [16] translated (MT) and glossed (Glossed) words, and the majority class baseline. The MT and Glossed results are also taken from Klementiev et al. [16].

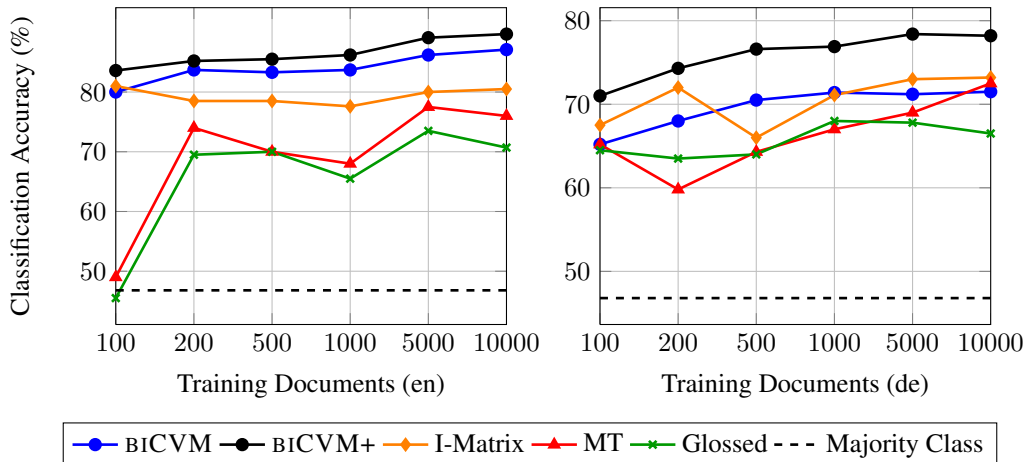


Figure 2: Classification accuracy for a number of models (see Table 1 for model descriptions). The left chart shows results for these models when trained on English data and evaluated on German data, the right chart vice versa.

of the  $d$ -dimensional representations of all its sentences. We train the multiclass classifier using the same settings and implementation of the averaged perceptron classifier [6] as used in [16].

We ran the CLDC experiments both by training on English and testing on German documents and vice versa. Using the data splits provided by [16], we used varying training data sizes from 100 to 10,000 documents for training the multiclass classifier. The results of this task across training sizes are shown in Figure 2. Table 1 shows the results for training on 1,000 documents.

Both models, BICVM and BICVM+ outperform all prior work on this task. Further, the BICVM+ model outperforms the BICVM model, indicating the usefulness of adding training data even from a separate language pair.

### 4.3 Visualization

While the CLDC experiment focused on establishing the semantic content of the sentence level representations, we also want to briefly investigate the induced word embeddings. In particular the BICVM+ model is interesting for that purpose, as it allows us to evaluate our approach of using English as a pivot language in a multilingual setup.

In Figure 3 we show the t-SNE projections for a number of English, French and German words. Of particular interest should be the right chart, which highlights bilingual embeddings between French and German words. Even though the model did not use any parallel French-German data during training, it still managed to learn semantic word-word similarity across these two languages.

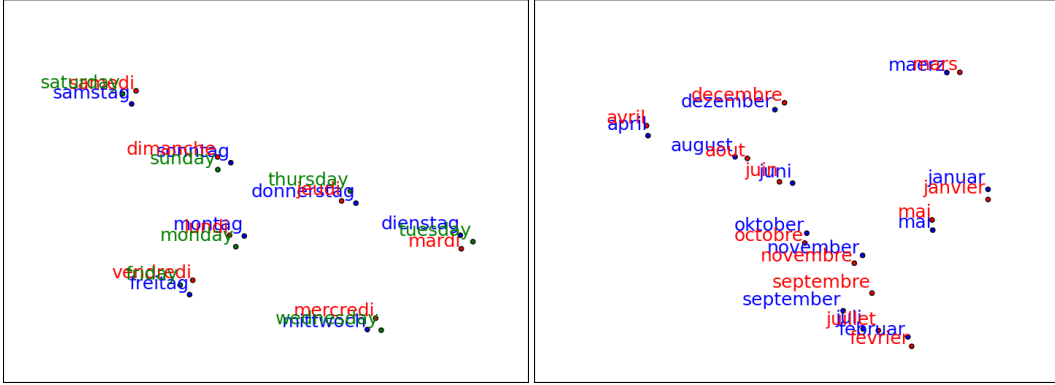


Figure 3: The left scatter plot shows t-SNE projections for a weekdays in all three languages using the representations learned in the BICVM+ model. Even though the model did not use any parallel French-German data during training, it still learns semantic similarity between these two languages using English as a pivot. To highlight this, the right plot shows another set of words (months of the year) using only the German and French words.

## 5 Conclusions

With this paper we have proposed a novel method for inducing cross-lingual distributed representations for compositional semantics. Using a very simple method for semantic composition, we nevertheless managed to obtain state of the art results on the CLDC task, specifically designed to evaluate semantic transfer across languages. After extending our approach to include multilingual training data in the BICVM+ model, we were able to demonstrate that adding additional languages further improves the model. Furthermore, using some qualitative experiments and visualizations, we showed that our approach also allows us to learn semantically related embeddings across languages without any direct training data.

Our approach provides great flexibility in training data and requires little to no annotation. Having demonstrated the successful training of semantic representations using sentence aligned data, a plausible next step is to attempt training using document-aligned data or even corpora of comparable documents. This may provide even greater possibilities for working with low-resource languages.

In the same vein, the success of our pivoting experiments suggest further work. Unlike other pivot approaches, it is easy to extend our model to have multiple pivot languages. Thus some pivots could preserve different aspects such as case, gender etc., and overcome other issues related to having a single pivot language.

As we have achieved the results in this paper with a relatively simple CVM, it would also be interesting to establish whether our objective function can be used in combination with more complex compositional vector models such as MV-RNN [25] or tensor-based approaches, to see whether these can further improve results on both mono- and multilingual tasks when used in conjunction with our cross-lingual objective function. Related to this, we will also apply our model to a wider variety of tasks including machine translation and multilingual information extraction.

## Acknowledgements

The authors would like to thank Alexandre Klementiev and his co-authors for making their datasets and averaged perceptron implementation available, as well as answering a number of questions related to their work on this task. This work was supported by EPSRC grant EP/K036580/1 and a Xerox Foundation Award.

## References

- [1] Marco Baroni and Roberto Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, 2010.
- [2] Paul Bloom. Precis of how children learn the meanings of words. *Behavioral and Brain Sciences*, 24:1095–1103, 2001.
- [3] Stephen Clark and Stephen Pulman. Combining symbolic and distributional models of meaning. In *Proceedings of AAAI Spring Symposium on Quantum Interaction*. AAAI Press, 2007.
- [4] Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *Lambek Festschrift. Linguistic Analysis*, 36:345–384, 2010.
- [5] Trevor Cohn and Mirella Lapata. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of ACL*, pages 728–735, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
- [6] Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of ACL-EMNLP*. Association for Computational Linguistics, 2002. doi: 10.3115/1118693.1118694.
- [7] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of ICML*, 2008.
- [8] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, July 2011. ISSN 1532-4435.
- [9] Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of ACL*, 2010.
- [10] K. Erk and S. Padó. A structured vector space model for word meaning in context. *Proceedings of EMNLP*, 2008.
- [11] J. R. Firth. A synopsis of linguistic theory 1930-55. 1952-59:1–32, 1957.
- [12] Edward Grefenstette and Mehrnoosh Sadrzadeh. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of EMNLP*, 2011.
- [13] Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-HLT*, 2008.
- [14] Karl Moritz Hermann and Phil Blunsom. The Role of Syntax in Vector Space Models of Compositional Semantics. In *Proceedings of ACL*, 2013.
- [15] Nal Kalchbrenner and Phil Blunsom. Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, 2013.
- [16] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing crosslingual distributed representations of words. In *Proceedings of COLING*, 2012.
- [17] Stanislas Lauly, Alex Boulanger, and Hugo Larochelle. Learning multilingual word representations using a bag-of-words autoencoder. In *Deep Learning Workshop at NIPS*, 2013.
- [18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, 2013.
- [19] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, 2013.
- [20] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Proceedings of INTERSPEECH*, 2010.
- [21] Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In *In Proceedings of ACL*, 2008.
- [22] D. Roy. Grounded spoken language acquisition: Experiments in word learning. *IEEE Transactions on Multimedia*, 5(2):197–209, June 2003. ISSN 1520-9210. doi: 10.1109/TMM.2003.811618.



- [23] A P Sarath Chandar, M Khapra Mitesh, B Ravindran, Vikas Raykar, and Amrita Saha. Multilingual deep learning. In *Deep Learning Workshop at NIPS*, 2013.
- [24] Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of EMNLP*, 2011.
- [25] Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP-CoNLL*, pages 1201–1211, 2012.
- [26] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Proceedings of NIPS*. 2012.
- [27] P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.
- [28] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, June 2009. ISSN 1532-4435.
- [29] Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of EMNLP*, 2013.