# Multilingual Models for Compositional Distributed Semantics

**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

We present a novel technique for learning semantic representations, which extends the distributional hypothesis to multilingual data and joint-space embeddings. Our models leverage parallel data and learn to strongly align the embeddings of semantically equivalent sentences, while maintaining sufficient distance between those of dissimilar sentences. The models do not rely on word alignments or any syntactic information and are successfully applied to a number of diverse languages. We extend our approach to learn semantic representations at the document level, too. We evaluate these models on two cross-lingual document classification tasks, outperforming the prior state of the art. Through qualitative analysis and the study of pivoting effects we demonstrate that our representations are semantically plausible and can capture semantic relationships across languages without parallel data.

## 1 Introduction

Distributed representations of words provide the basis for many state-of-the-art approaches to various problems in natural language processing today. Such word embeddings are naturally richer representations than those of symbolic or discrete models, and have been shown to be able to capture both syntactic and semantic information. Successful applications of such models include language modelling (Bengio et al., 2003), paraphrase detection (Erk and Padó, 2008), and dialogue analysis (Kalchbrenner and Blunsom, 2013).

Within a monolingual context, the distributional hypothesis (Firth, 1957) forms the basis of most approaches for learning word representations. In this work, we extend this hypothesis to multilingual data and joint-space embeddings. We present a novel unsupervised technique for learning semantic representations that leverages parallel corpora and employs semantic transfer through compositional representations. Unlike most methods for learning word representations, which are restricted to a single language, our approach learns to represent meaning across languages in a shared multilingual semantic space.

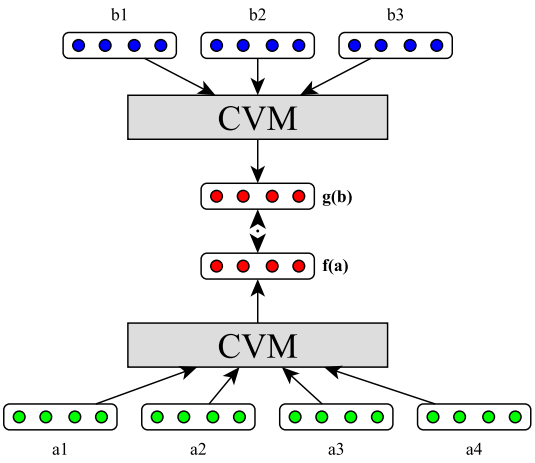


Figure 1: Model with parallel input sentences $a$ and $b$. The model minimises the distance between the sentence level encoding of the bitext. Any composition functions (CVM) can be used to generate the compositional sentence level representations.

We present experiments on two corpora. First, we show that for cross-lingual document classification on the Reuters RCV1/RCV2 corpora (Lewis et al., 2004), we outperform the prior state of the art (Klementiev et al., 2012). Second, we also present classification results on a massively multilingual corpus which we derive from the TED corpus (Cettolo et al., 2012). The results on this task, in comparison with a number of strong baselines, further demonstrate the relevance of our approach and the success of our method in learning multilingual semantic representations over a wide range of languages.



*Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 58–68,
Baltimore, Maryland, USA, June 23-25 2014. 

## 2 Overview

Distributed representation learning describes the task of learning continuous representations for discrete objects. Here, we focus on learning semantic representations and investigate how the use of multilingual data can improve learning such representations at the word and higher level. We present a model that learns to represent each word in a lexicon by a continuous vector in $\mathbb{R}^d$. Such distributed representations allow a model to share meaning between similar words, and have been used to capture semantic, syntactic and morphological content (Collobert and Weston, 2008; Turian et al., 2010, *inter alia*).

We describe a multilingual objective function that uses a noise-contrastive update between semantic representations of different languages to learn these word embeddings. As part of this, we use a compositional vector model (CVM, henceforth) to compute semantic representations of sentences and documents. A CVM learns semantic representations of larger syntactic units given the semantic representations of their constituents (Clark and Pulman, 2007; Mitchell and Lapata, 2008; Baroni and Zamparelli, 2010; Grefenstette and Sadrzadeh, 2011; Socher et al., 2012; Hermann and Blunsom, 2013, *inter alia*).

A key difference between our approach and those listed above is that we only require sentence-aligned parallel data in our otherwise unsupervised learning function. This removes a number of constraints that normally come with CVM models, such as the need for syntactic parse trees, word alignment or annotated data as a training signal. At the same time, by using multiple CVMs to transfer information between languages, we enable our models to capture a broader semantic context than would otherwise be possible.

The idea of extracting semantics from multilingual data stems from prior work in the field of semantic grounding. Language acquisition in humans is widely seen as grounded in sensory-motor experience (Bloom, 2001; Roy, 2003). Based on this idea, there have been some attempts at using multi-modal data for learning better vector representations of words (e.g. Srivastava and Salakhutdinov (2012)). Such methods, however, are not easily scalable across languages or to large amounts of data for which no secondary or tertiary representation might exist.

Parallel data in multiple languages provides an alternative to such secondary representations, as parallel texts share their semantics, and thus one language can be used to ground the other. Some work has exploited this idea for transferring linguistic knowledge into low-resource languages or to learn distributed representations at the word level (Klementiev et al., 2012; Zou et al., 2013; Lauly et al., 2013, *inter alia*). 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 attempt to induce these, too, cross-lingually.

## 3 Approach

Most prior work on learning compositional semantic representations employs parse trees on their training data to structure their composition functions (Socher et al., 2012; Hermann and Blunsom, 2013, *inter alia*). Further, these approaches typically depend on specific *semantic* signals such as sentiment- or topic-labels for their objective functions. While these methods have been shown to work in some cases, the need for parse trees and annotated data limits such approaches to resource-fortunate languages. Our novel method for learning compositional vectors removes these requirements, and as such can more easily be applied to low-resource languages.

Specifically, we attempt to learn semantics from multilingual data. The idea is that, given enough parallel data, a shared representation of two parallel sentences would be forced to capture the common elements between these two sentences. What parallel sentences share, of course, are their semantics. Naturally, different languages express meaning in different ways. We utilise this diversity to abstract further from mono-lingual surface realisations to deeper semantic representations. We exploit this semantic similarity across languages by defining a bilingual (and trivially multilingual) energy as follows.

Assume two functions $f : X \to \mathbb{R}^d$ and $g : Y \to \mathbb{R}^d$, which map sentences from languages $x$ and $y$ onto distributed semantic representations in $\mathbb{R}^d$. Given a parallel corpus $C$, we then define the energy of the model given two sentences $(a, b) \in C$ as:

$$E_{bi}(a, b) = \|f(a) - g(b)\|^2 \qquad (1)$$

We want to minimize $E_{bi}$ for all semantically equivalent sentences in the corpus. In order to prevent the model from degenerating, we further introduce a noise-constrastive large-margin update which ensures that the representations of non-aligned sentences observe a certain margin from each other. For every pair of parallel sentences $(a, b)$ we sample a number of additional sentence pairs $(\cdot, n) \in C$, where $n$—with high probability—is not semantically equivalent to $a$. We use these noise samples as follows:

$$E_{hl}(a, b, n) = [m + E_{bi}(a, b) - E_{bi}(a, n)]_+$$

where $[x]_+ = max(x, 0)$ denotes the standard hinge loss and $m$ is the margin. This results in the following objective function:

$$J(\theta) = \sum_{(a,b)\in\mathcal{C}} \left( \sum_{i=1}^{k} E_{hl}(a, b, n_i) + \frac{\lambda}{2}\|\theta\|^2 \right) \quad (2)$$

where $\theta$ is the set of all model variables.

### 3.1 Two Composition Models

The objective function in Equation 2 could be coupled with any two given vector composition functions $f, g$ from the literature. As we aim to apply our approach to a wide range of languages, we focus on composition functions that do not require any syntactic information. We evaluate the following two composition functions.

The first model, ADD, represents a sentence by the sum of its word vectors. This is a distributed bag-of-words approach as sentence ordering is not taken into account by the model.

Second, the BI model is designed to capture bigram information, using a non-linearity over bigram pairs in its composition function:

$$f(x) = \sum_{i=1}^{n} \tanh\left(x_{i-1} + x_i\right) \quad (3)$$

The use of a non-linearity enables the model to learn interesting interactions between words in a document, which the bag-of-words approach of ADD is not capable of learning. We use the hyperbolic tangent as activation function.

### 3.2 Document-level Semantics

For a number of tasks, such as topic modelling, representations of objects beyond the sentence level are required. While most approaches to compositional distributed semantics end at the word level, our model extends to document-level learning quite naturally, by recursively applying the composition and objective function (Equation 2) to compose sentences into documents. This is achieved by first computing semantic representations for each sentence in a document. Next, these representations are used as inputs in a higher-level CVM, computing a semantic representation of a document (Figure 2).

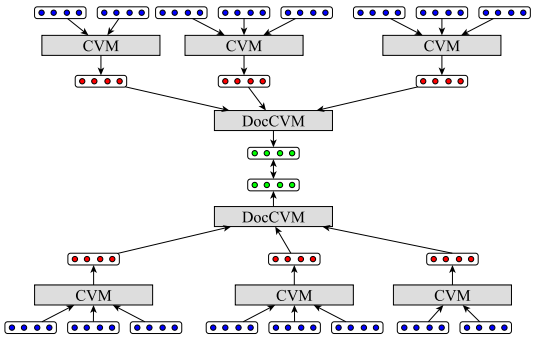


Figure 2: Description of a parallel document-level compositional vector model (DOC). The model recursively computes semantic representations for each sentence of a document and then for the document itself, treating the sentence vectors as inputs for a second CVM.

This recursive approach integrates document-level representations into the learning process. We can thus use corpora of parallel documents—regardless of whether they are sentence aligned or not—to propagate a semantic signal back to the individual words. If sentence alignment is available, of course, the document-signal can simply be combined with the sentence-signal, as we did with the experiments described in §5.3.

This concept of learning compositional representations for documents contrasts with prior work (Socher et al., 2011; Klementiev et al., 2012, *inter alia*) who rely on summing or averaging sentence-vectors if representations beyond the sentence-level are required for a particular task.

We evaluate the models presented in this paper both with and without the document-level signal. We refer to the individual models used as ADD and BI if used without, and as DOC/ADD and DOC/BI is used with the additional document composition function and error signal.

## 4 Corpora

We use two corpora for learning semantic representations and performing the experiments described in this paper.

The Europarl corpus v7[1] (Koehn, 2005) was used during initial development and testing of our approach, as well as to learn the representations used for the Cross-Lingual Document Classification task described in §5.2. We considered the English-German and English-French language pairs from this corpus. From each pair the final 100,000 sentences were reserved for development.

Second, we developed a massively multilingual corpus based on the TED corpus[2] for IWSLT 2013 (Cettolo et al., 2012). This corpus contains English transcriptions and multilingual, sentence-aligned translations of talks from the TED conference. While the corpus is aimed at machine translation tasks, we use the keywords associated with each talk to build a subsidiary corpus for multilingual document classification as follows.[3]

The development sections provided with the IWSLT 2013 corpus were again reserved for development. We removed approximately 10 percent of the training data in each language to create a test corpus (all talks with $id \geq 1{,}400$). The new training corpus consists of a total of 12,078 parallel documents distributed across 12 language pairs[4]. In total, this amounts to 1,678,219 non-English sentences (the number of unique English sentences is smaller as many documents are translated into multiple languages and thus appear repeatedly in the corpus). Each document (talk) contains one or several keywords. We used the 15 most frequent keywords for the topic classification experiments described in section §5.3.

Both corpora were pre-processed using the set of tools provided by cdec[5] for tokenizing and lowercasing the data. Further, all empty sentences and their translations were removed from the corpus.

## 5 Experiments

We report results on two experiments. First, we replicate the cross-lingual document classification task of Klementiev et al. (2012), learning distributed representations on the Europarl corpus and evaluating on documents from the Reuters RCV1/RCV2 corpora. Subsequently, we design a multi-label classification task using the TED corpus, both for training and evaluating. The use of a wider range of languages in the second experiments allows us to better evaluate our models' capabilities in learning a shared multilingual semantic representation. We also investigate the learned embeddings from a qualitative perspective in §5.4.

### 5.1 Learning

All model weights were randomly initialised using a Gaussian distribution ($\mu$=0, $\sigma^2$=0.1). We used the available development data to set our model parameters. For each positive sample we used a number of noise samples ($k \in \{1, 10, 50\}$), randomly drawn from the corpus at each training epoch. All our embeddings have dimensionality $d$=128, with the margin set to $m$=$d$.[6] Further, we use L2 regularization with $\lambda$=1 and step-size in $\{0.01, 0.05\}$. We use 100 iterations for the RCV task, 500 for the TED single and 5 for the joint corpora. We use the adaptive gradient method, AdaGrad (Duchi et al., 2011), for updating the weights of our models, in a mini-batch setting ($b \in \{10, 50\}$). All settings, our model implementation and scripts to replicate our experiments are available at `http://www.karlmoritz.com/`.

### 5.2 RCV1/RCV2 Document Classification

We evaluate our models on the cross-lingual document classification (CLDC, henceforth) task first described in Klementiev et al. (2012). 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 further supervision in another. Thus, CLDC can be used to establish whether our learned representations are semantically useful across multiple languages.

We follow the experimental setup described in Klementiev et al. (2012), with the exception that we learn our embeddings using solely the Europarl data and use the Reuters corpora only during for classifier training and testing. Each document in the classification task is represented by the average of the $d$-dimensional representations of all its sentences. We train the multiclass classifier using an averaged perceptron (Collins, 2002) with the same settings as in Klementiev et al. (2012).

[1] `http://www.statmt.org/europarl/`

[2] `https://wit3.fbk.eu/`

[3] `http://www.clg.ox.ac.uk/tedcldc/`

[4] English to Arabic, German, French, Spanish, Italian, Dutch, Polish, Brazilian Portuguese, Romanian, Russian and Turkish. Chinese, Farsi and Slowenian were removed due to the small size of those datasets.

[5] `http://cdec-decoder.org/`

[6] On the RCV task we also report results for $d$=40 which matches the dimensionality of Klementiev et al. (2012).

| Model | en → de | de → en |
|---|---|---|
| Majority Class | 46.8 | 46.8 |
| Glossed | 65.1 | 68.6 |
| MT | 68.1 | 67.4 |
| I-Matrix | 77.6 | 71.1 |
| $dim = 40$ | | |
| ADD | 83.7 | 71.4 |
| ADD+ | 86.2 | 76.9 |
| BI | 83.4 | 69.2 |
| BI+ | 86.9 | 74.3 |
| $dim = 128$ | | |
| ADD | 86.4 | 74.7 |
| ADD+ | 87.7 | 77.5 |
| BI | 86.1 | 79.0 |
| BI+ | **88.1** | **79.2** |

Table 1: Classification accuracy for training on English and German with 1000 labeled examples on the RCV corpus. Cross-lingual compositional representations (ADD, BI and their multilingual extensions), I-Matrix (Klementiev et al., 2012) translated (MT) and glossed (Glossed) word baselines, and the majority class baseline. The baseline results are from Klementiev et al. (2012).

We present results from four models. The ADD model is trained on 500k sentence pairs of the English-German parallel section of the Europarl corpus. The ADD+ model uses an additional 500k parallel sentences from the English-French corpus, resulting in one million English sentences, each paired up with either a German or a French sentence, with BI and BI+ trained accordingly. The motivation behind ADD+ and BI+ is to investigate whether we can learn better embeddings by introducing additional data from other languages. A similar idea exists in machine translation where English is frequently used to pivot between other languages (Cohn and Lapata, 2007).

The actual CLDC experiments are performed by training on English and testing on German documents and vice versa. Following prior work, we use varying sizes between 100 and 10,000 documents when training the multiclass classifier. The results of this task across training sizes are in Figure 3. Table 1 shows the results for training on 1,000 documents compared with the results published in Klementiev et al. (2012). Our models outperform the prior state of the art, with the BI models performing slightly better than the ADD models. As the relative results indicate, the addition of a second language improves model performance. It it interesting to note that results improve in both directions of the task, even though no additional German data was used for the '+' models.

### 5.3 TED Corpus Experiments

Here we describe our experiments on the TED corpus, which enables us to scale up to multilingual learning. Consisting of a large number of relatively short and parallel documents, this corpus allows us to evaluate the performance of the DOC model described in §3.2.

We use the training data of the corpus to learn distributed representations across 12 languages. Training is performed in two settings. In the *single* mode, vectors are learnt from a single language pair (en-X), while in the *joint* mode vector-learning is performed on all parallel sub-corpora simultaneously. This setting causes words from all languages to be embedded in a single semantic space.

First, we evaluate the effect of the document-level error signal (DOC, described in §3.2), as well as whether our multilingual learning method can extend to a larger variety of languages. We train DOC models, using both ADD and BI as CVM (DOC/ADD, DOC/BI), both in the *single* and *joint* mode. For comparison, we also train ADD and DOC models without the document-level error signal. The resulting document-level representations are used to train classifiers (system and settings as in §5.2) for each language, which are then evaluated in the paired language. In the English case we train twelve individual classifiers, each using the training data of a single language pair only. As described in §4, we use 15 keywords for the classification task. Due to space limitations, we report cumulative results in the form of F1-scores throughout this paper.

**MT System** We develop a machine translation baseline as follows. We train a machine translation tool on the parallel training data, using the development data of each language pair to optimize the translation system. We use the cdec decoder (Dyer et al., 2010) with default settings for this purpose. With this system we translate the test data, and then use a Naïve Bayes classifier[7] for the actual experiments. To exemplify, this means the $de{\rightarrow}ar$ result is produced by training a translation system from Arabic to German. The Arabic test set is translated into German. A classifier is then trained

[7]We use the implementation in Mallet (McCallum, 2002)

| Setting | Languages | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Arabic | German | Spanish | French | Italian | Dutch | Polish | Pt-Br | Roman. | Russian | Turkish |
| $en \rightarrow$ L2 | | | | | | | | | | | |
| MT System | **0.429** | **0.465** | **0.518** | **0.526** | **0.514** | **0.505** | **0.445** | **0.470** | **0.493** | 0.432 | 0.409 |
| ADD *single* | 0.328 | 0.343 | 0.401 | 0.275 | 0.282 | 0.317 | 0.141 | 0.227 | 0.282 | 0.338 | 0.241 |
| BI *single* | 0.375 | 0.360 | 0.379 | 0.431 | 0.465 | 0.421 | 0.435 | 0.329 | 0.426 | 0.423 | **0.481** |
| DOC/ADD *single* | 0.410 | 0.424 | 0.383 | 0.476 | 0.485 | 0.264 | 0.402 | 0.354 | 0.418 | 0.448 | 0.452 |
| DOC/BI *single* | 0.389 | 0.428 | 0.416 | 0.445 | 0.473 | 0.219 | 0.403 | 0.400 | 0.467 | 0.421 | 0.457 |
| DOC/ADD *joint* | 0.392 | 0.405 | 0.443 | 0.447 | 0.475 | 0.453 | 0.394 | 0.409 | 0.446 | **0.476** | 0.417 |
| DOC/BI *joint* | 0.372 | 0.369 | 0.451 | 0.429 | 0.404 | 0.433 | 0.417 | 0.399 | 0.453 | 0.439 | 0.418 |
| L2 $\rightarrow$ $en$ | | | | | | | | | | | |
| MT System | 0.448 | 0.469 | **0.486** | 0.358 | **0.481** | 0.463 | **0.460** | 0.374 | **0.486** | 0.404 | 0.441 |
| ADD *single* | 0.380 | 0.337 | 0.446 | 0.293 | 0.357 | 0.295 | 0.327 | 0.235 | 0.293 | 0.355 | 0.375 |
| BI *single* | 0.354 | 0.411 | 0.344 | 0.426 | 0.439 | 0.428 | 0.443 | 0.357 | 0.426 | 0.442 | 0.403 |
| DOC/ADD *single* | **0.452** | **0.476** | 0.422 | 0.464 | 0.461 | 0.251 | 0.400 | 0.338 | 0.407 | **0.471** | 0.435 |
| DOC/BI *single* | 0.406 | 0.442 | 0.365 | **0.479** | 0.460 | 0.235 | 0.393 | 0.380 | 0.426 | 0.467 | **0.477** |
| DOC/ADD *joint* | 0.396 | 0.388 | 0.399 | 0.415 | 0.461 | **0.478** | 0.352 | **0.399** | 0.412 | 0.343 | 0.343 |
| DOC/BI *joint* | 0.343 | 0.375 | 0.369 | 0.419 | 0.398 | 0.438 | 0.353 | 0.391 | 0.430 | 0.375 | 0.388 |

Table 2: F1-scores for the TED document classification task for individual languages. Results are reported for both directions (training on English, evaluating on L2 and vice versa). Bold indicates best result, underline best result amongst the vector-based systems.

| Training Language | Test Language | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Arabic | German | Spanish | French | Italian | Dutch | Polish | Pt-Br | Rom'n | Russian | Turkish |
| Arabic | | 0.378 | 0.436 | 0.432 | 0.444 | 0.438 | 0.389 | 0.425 | 0.420 | 0.446 | 0.397 |
| German | 0.368 | | 0.474 | 0.460 | 0.464 | 0.440 | 0.375 | 0.417 | 0.447 | 0.458 | 0.443 |
| Spanish | 0.353 | 0.355 | | 0.420 | 0.439 | 0.435 | 0.415 | 0.390 | 0.424 | 0.427 | 0.382 |
| French | 0.383 | 0.366 | 0.487 | | 0.474 | 0.429 | 0.403 | 0.418 | 0.458 | 0.415 | 0.398 |
| Italian | 0.398 | 0.405 | 0.461 | 0.466 | | 0.393 | 0.339 | 0.347 | 0.376 | 0.382 | 0.352 |
| Dutch | 0.377 | 0.354 | 0.463 | 0.464 | 0.460 | | 0.405 | 0.386 | 0.415 | 0.407 | 0.395 |
| Polish | 0.359 | 0.386 | 0.449 | 0.444 | 0.430 | 0.441 | | 0.401 | 0.434 | 0.398 | 0.408 |
| Portuguese | 0.391 | 0.392 | 0.476 | 0.447 | 0.486 | 0.458 | 0.403 | | 0.457 | 0.431 | 0.431 |
| Romanian | 0.416 | 0.320 | 0.473 | 0.476 | 0.460 | 0.434 | 0.416 | 0.433 | | 0.444 | 0.402 |
| Russian | 0.372 | 0.352 | 0.492 | 0.427 | 0.438 | 0.452 | 0.430 | 0.419 | 0.441 | | 0.447 |
| Turkish | 0.376 | 0.352 | 0.479 | 0.433 | 0.427 | 0.423 | 0.439 | 0.367 | 0.434 | 0.411 | |

Table 3: F1-scores for TED corpus document classification results when training and testing on two languages that do not share any parallel data. We train a DOC/ADD model on all $en$-L2 language pairs together, and then use the resulting embeddings to train document classifiers in each language. These classifiers are subsequently used to classify data from all other languages.

| Setting | Languages | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | English | Arabic | German | Spanish | French | Italian | Dutch | Polish | Pt-Br | Roman. | Russian | Turkish |
| Raw Data NB | 0.481 | 0.469 | 0.471 | 0.526 | 0.532 | 0.524 | 0.522 | 0.415 | 0.465 | 0.509 | 0.465 | 0.513 |
| Senna | 0.400 | | | | | | | | | | | |
| Polyglot | 0.382 | 0.416 | 0.270 | 0.418 | 0.361 | 0.332 | 0.228 | 0.323 | 0.194 | 0.300 | 0.402 | 0.295 |
| *single* Setting | | | | | | | | | | | | |
| DOC/ADD | 0.462 | 0.422 | 0.429 | 0.394 | 0.481 | 0.458 | 0.252 | 0.385 | 0.363 | 0.431 | 0.471 | 0.435 |
| DOC/BI | 0.474 | 0.432 | 0.362 | 0.336 | 0.444 | 0.469 | 0.197 | 0.414 | 0.395 | 0.445 | 0.436 | 0.428 |
| *joint* Setting | | | | | | | | | | | | |
| DOC/ADD | 0.475 | 0.371 | 0.386 | 0.472 | 0.451 | 0.398 | 0.439 | 0.304 | 0.394 | 0.453 | 0.402 | 0.441 |
| DOC/BI | 0.378 | 0.329 | 0.358 | 0.472 | 0.454 | 0.399 | 0.409 | 0.340 | 0.431 | 0.379 | 0.395 | 0.435 |

Table 4: F1-scores on the TED corpus document classification task when training and evaluating on the same language. Baseline embeddings are Senna (Collobert et al., 2011) and Polyglot (Al-Rfou' et al., 2013).

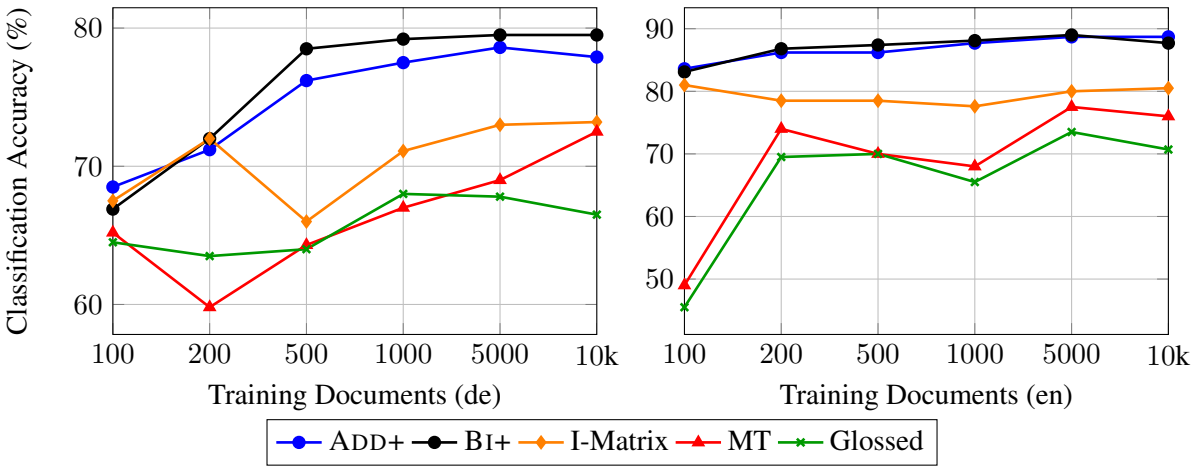


Figure 3: Classification accuracy for a number of models (see Table 1 for model descriptions). The left chart shows results for these models when trained on German data and evaluated on English data, the right chart vice versa.

on the German training data and evaluated on the translated Arabic. While we developed this system as a baseline, it must be noted that the classifier of this system has access to significantly more information (all words in the document) as opposed to our models (one embedding per document), and we do not expect to necessarily beat this system.

The results of this experiment are in Table 2. When comparing the results between the ADD model and the models trained using the document-level error signal, the benefit of this additional signal becomes clear. The *joint* training mode leads to a relative improvement when training on English data and evaluating in a second language. This suggests that the *joint* mode improves the quality of the English embeddings more than it affects the L2-embeddings. More surprising, perhaps, is the relative performance between the ADD and BI composition functions, especially when compared to the results in §5.2, where the BI models relatively consistently performed better. We suspect that the better performance of the additive composition function on this task is related to the smaller amount of training data available which could cause sparsity issues for the bigram model.

As expected, the MT system slightly outperforms our models on most language pairs. However, the overall performance of the models is comparable to that of the MT system. Considering the relative amount of information available during the classifier training phase, this indicates that our learned representations are semantically useful, capturing almost the same amount of information as available to the Naïve Bayes classifier.

We next investigate linguistic transfer across languages. We re-use the embeddings learned with the DOC/ADD *joint* model from the previous experiment for this purpose, and train classifiers on all non-English languages using those embeddings. Subsequently, we evaluate their performance in classifying documents in the remaining languages. Results for this task are in Table 3. While the results across language-pairs might not be very insightful, the overall good performance compared with the results in Table 2 implies that we learnt semantically meaningful vectors and in fact a joint embedding space across thirteen languages.

In a third evaluation (Table 4), we apply the embeddings learnt with out models to a monolingual classification task, enabling us to compare with prior work on distributed representation learning. In this experiment a classifier is trained in one language and then evaluated in the same. We again use a Naïve Bayes classifier on the raw data to establish a reasonable upper bound.

We compare our embeddings with the SENNA embeddings, which achieve state of the art performance on a number of tasks (Collobert et al., 2011). Additionally, we use the Polyglot embeddings of Al-Rfou' et al. (2013), who published word embeddings across 100 languages, including all languages considered in this paper. We represent each document by the mean of its word vectors and then apply the same classifier training and testing regime as with our models. Even though both of these sets of embeddings were trained on much larger datasets than ours, our models outperform these baselines on all languages—even outperforming the Naïve Bayes system on on several

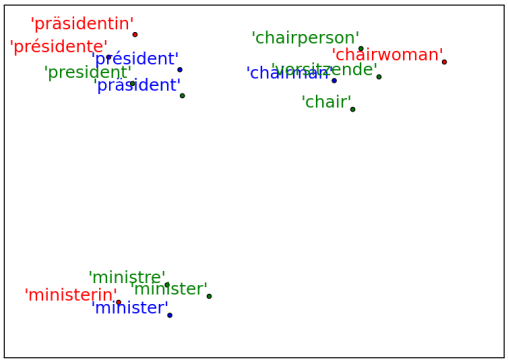


Figure 4: t-SNE projections for a number of English, French and German words as represented by the BI+ model. Even though the model did not use any parallel French-German data during training, it learns semantic similarity between these two languages using English as a pivot, and semantically clusters words across all languages.

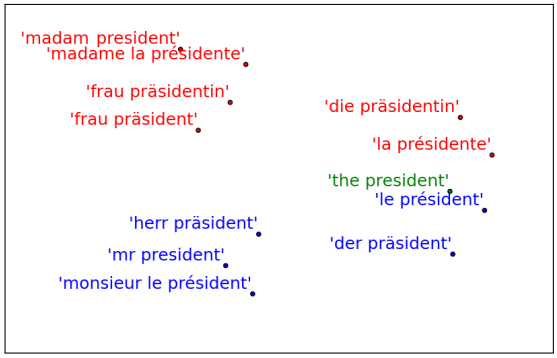


Figure 5: t-SNE projections for a number of short phrases in three languages as represented by the BI+ model. The projection demonstrates linguistic transfer through a pivot by. It separates phrases by gender (red for female, blue for male, and green for neutral) and aligns matching phrases across languages.

languages. While this may partly be attributed to the fact that our vectors were learned on in-domain data, this is still a very positive outcome.

### 5.4 Linguistic Analysis

While the classification experiments focused on establishing the semantic content of the sentence level representations, we also want to briefly investigate the induced word embeddings. We use the BI+ model trained on the Europarl corpus for this purpose. Figure 4 shows the t-SNE projections for a number of English, 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.

Going one step further, Figure 5 shows t-SNE projections for a number of short phrases in these three languages. We use the English *the president* and gender-specific expressions *Mr President* and *Madam President* as well as gender-specific equivalents in French and German. The projection demonstrates a number of interesting results: First, the model correctly clusters the words into three groups, corresponding to the three English forms and their associated translations. Second, a separation between genders can be observed, with male forms on the bottom half of the chart and female forms on the top, with the neutral *the president* in the vertical middle. Finally, if we assume a horizontal line going through *the president*, this line could be interpreted as a "gender divide", with male and female versions of one expression mirroring each other on that line. In the case of *the president* and its translations, this effect becomes even clearer, with the neutral English expression being projected close to the mid-point between each other language's gender-specific versions.

These results further support our hypothesis that the bilingual contrastive error function can learn semantically plausible embeddings and furthermore, that it can abstract away from mono-lingual surface realisations into a shared semantic space across languages.

## 6 Related Work

**Distributed Representations** Distributed representations can be learned through a number of approaches. In their simplest form, distributional information from large corpora can be used to learn embeddings, where the words appearing within a certain window of the target word are used to compute that word's embedding. This is related to topic-modelling techniques such as LSA (Dumais et al., 1988), LSI, and LDA (Blei et al., 2003), but these methods use a document-level context, and tend to capture the topics a word is used in rather than its more immediate syntactic context.

Neural language models are another popular approach for inducing distributed word representations (Bengio et al., 2003). They have received a lot of attention in recent years (Collobert and Weston, 2008; Mnih and Hinton, 2009; Mikolov et al., 2010, *inter alia*) and have achieved state of the art performance in language modelling. Collobert et al. (2011) further popularised using neural network architectures for learning word embeddings from large amounts of largely unlabelled data by showing the embeddings can then be used to improve standard supervised tasks.

Unsupervised word representations can easily be plugged into a variety of NLP related tasks. Tasks, where the use of distributed representations has resulted in improvements include topic modelling (Blei et al., 2003) or named entity recognition (Turian et al., 2010; Collobert et al., 2011).

**Compositional Vector Models** 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. Self-evidently, sparsity prevents the learning of such representations using the same collocational methods as applied to the word level. Most literature instead focuses on learning composition functions that represent the semantics of a larger structure as a function of the representations of its parts.

Very simple composition functions have been shown to suffice for tasks such as judging bigram semantic similarity (Mitchell and Lapata, 2008). More complex composition functions using matrix-vector composition, convolutional neural networks or tensor composition have proved useful in tasks such as sentiment analysis (Socher et al., 2011; Hermann and Blunsom, 2013), relational similarity (Turney, 2012) or dialogue analysis (Kalchbrenner and Blunsom, 2013).

**Multilingual Representation Learning** Most research on distributed representation induction has focused on single languages. English, with its large number of annotated resources, has enjoyed most attention. However, there exists a corpus of prior work on learning multilingual embeddings or on using parallel data to transfer linguistic information across languages. One has to differentiate between approaches such as Al-Rfou' et al. (2013), that learn embeddings across a large variety of languages and models such as ours, that learn joint embeddings, that is a projection into a shared semantic space across multiple languages.

Related to our work, Yih et al. (2011) proposed S2Nets to learn joint embeddings of tf-idf vectors for comparable documents. Their architecture optimises the cosine similarity of documents, using relative semantic similarity scores during learning. More recently, Lauly et al. (2013) proposed a bag-of-words autoencoder model, where the bag-of-words representation in one language is used to train the embeddings in another. By placing their vocabulary in a binary branching tree, the probabilistic setup of this model is similar to that of Mnih and Hinton (2009). Similarly, Sarath Chandar et al. (2013) train a cross-lingual encoder, where an autoencoder is used to recreate words in two languages in parallel. This is effectively the linguistic extension of Ngiam et al. (2011), who used a similar method for audio and video data. Hermann and Blunsom (2014) propose a large-margin learner for multilingual word representations, similar to the basic additive model proposed here, which, like the approaches above, relies on a bag-of-words model for sentence representations.

Klementiev et al. (2012), our baseline in §5.2, use a form of multi-agent learning on word-aligned parallel data to transfer embeddings from one language to another. Earlier work, Haghighi et al. (2008), proposed a method for inducing bilingual lexica using monolingual feature representations and a small initial lexicon to bootstrap with. This approach has recently been extended by Mikolov et al. (2013a), Mikolov et al. (2013b), who developed a method for learning transformation matrices to convert semantic vectors of one language into those of another. Is 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 ADD composition function used here. Using a slightly different approach, Zou et al. (2013), also learned bilingual embeddings for machine translation.

## 7 Conclusion

To summarize, we have presented a novel method for learning multilingual word embeddings using parallel data in conjunction with a multilingual objective function for compositional vector models. This approach extends the distributional hypothesis to multilingual joint-space representations. Coupled with very simple composition functions, vectors learned with this method outperform the state of the art on the task of cross-lingual document classification. Further experiments and analysis support our hypothesis that bilingual signals are a useful tool for learning distributed representations by enabling models to abstract away from mono-lingual surface realisations into a deeper semantic space.

## Acknowledgements

This work was supported by a Xerox Foundation Award and EPSRC grant number EP/K036580/1.

## References


R. Al-Rfou', B. Perozzi, and S. Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of CoNLL*.

M. Baroni and R. Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*.

Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, March.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

P. Bloom. 2001. Precis of how children learn the meanings of words. *Behavioral and Brain Sciences*, 24:1095–1103.

M. Cettolo, C. Girardi, and M. Federico. 2012. Wit$^3$: Web inventory of transcribed and translated talks. In *Proceedings of EAMT*.

S. Clark and S. Pulman. 2007. Combining symbolic and distributional models of meaning. In *Proceedings of AAAI Spring Symposium on Quantum Interaction*. AAAI Press.

T. Cohn and M. Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of ACL*.

M. Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of ACL-EMNLP*.

R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of ICML*.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

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

S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

C. Dyer, A. Lopez, J. Ganitkevitch, J. Weese, F. Ture, P. Blunsom, H. Setiawan, V. Eidelman, and P. Resnik. 2010. cdec: A Decoder, Alignment, and Learning framework for finite-state and context-free translation models. In *Proceedings of ACL*.

K. Erk and S. Padó. 2008. A structured vector space model for word meaning in context. *Proceedings of EMNLP*.

J. R. Firth. 1957. A synopsis of linguistic theory 1930-55. 1952-59:1–32.

E. Grefenstette and M. Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of EMNLP*.

A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-HLT*.

K. M. Hermann and P. Blunsom. 2013. The Role of Syntax in Vector Space Models of Compositional Semantics. In *Proceedings of ACL*.

K. M. Hermann and P. Blunsom. 2014. Multilingual Distributed Representations without Word Alignment. In *Proceedings of ICLR*.

N. Kalchbrenner and P. Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. *Proceedings of the ACL Workshop on Continuous Vector Space Models and their Compositionality*.

A. Klementiev, I. Titov, and B. Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING*.

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

S. Lauly, A. Boulanger, and H. Larochelle. 2013. Learning multilingual word representations using a bag-of-words autoencoder. In *Deep Learning Workshop at NIPS*.

D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, December.

A. K. McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of INTERSPEECH*.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *CoRR*.

T. Mikolov, Q. V. Le, and I. Sutskever. 2013b. Exploiting Similarities among Languages for Machine Translation. *CoRR*.

J. Mitchell and M. Lapata. 2008. Vector-based models of semantic composition. In *In Proceedings of ACL*.

A. Mnih and G. Hinton. 2009. A scalable hierarchical distributed language model. In *Proceedings of NIPS*.

J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. 2011. Multimodal deep learning. In *ICML*.

D. Roy. 2003. Grounded spoken language acquisition: Experiments in word learning. *IEEE Transactions on Multimedia*, 5(2):197–209, June.

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

R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of EMNLP*.

R. Socher, B. Huval, C. D. Manning, and A. Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP-CoNLL*, pages 1201–1211.

N. Srivastava and R. Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Proceedings of NIPS*.

J. Turian, L. Ratinov, and Y. Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of ACL*.

P. D. Turney. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, 44:533–585.

W.-T. Yih, K. Toutanova, J. C. Platt, and C. Meek. 2011. Learning discriminative projections for text similarity measures. In *Proceedings of CoNLL*.

W. Y. Zou, R. Socher, D. Cer, and C. D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of EMNLP*.