Jointly Learning to Embed and Predict with Multiple Languages

Daniel C. Ferreira[*] **André F. T. Martins**[†*♯] **Mariana S. C. Almeida**[*♯]
[*]Priberam Labs, Alameda D. Afonso Henriques, 41, 2º, 1000-123 Lisboa, Portugal
[†]Unbabel Lda, Rua Visconde de Santarém, 67-B, 1000-286 Lisboa, Portugal
[♯]Instituto de Telecomunicações, Instituto Superior Técnico, 1049-001 Lisboa, Portugal
{dcf,mla}@priberam.pt, {andre.martins}@unbabel.com

Abstract

We propose a joint formulation for learning task-specific cross-lingual word embeddings, along with classifiers for that task. Unlike prior work, which first learns the embeddings from parallel data and then plugs them in a supervised learning problem, our approach is one-shot: a single optimization problem combines a co-regularizer for the multilingual embeddings with a task-specific loss. We present theoretical results showing the limitation of Euclidean co-regularizers to increase the embedding dimension, a limitation which does not exist for other co-regularizers (such as the ℓ_1-distance). Despite its simplicity, our method achieves state-of-the-art accuracies on the RCV1/RCV2 dataset when transferring from English to German, with training times below 1 minute. On the TED Corpus, we obtain the highest reported scores on 10 out of 11 languages.

1 Introduction

Distributed representations of text (embeddings) have been the target of much research in natural language processing (Collobert and Weston, 2008; Mikolov et al., 2013; Pennington et al., 2014; Levy et al., 2015). Word embeddings partially capture semantic and syntactic properties of text in the form of dense real vectors, making them apt for a wide variety of tasks, such as language modeling (Bengio et al., 2003), sentence tagging (Turian et al., 2010; Collobert et al., 2011), sentiment analysis (Socher et al., 2011), parsing (Chen and Manning, 2014), and machine translation (Zou et al., 2013).

At the same time, there has been a consistent progress in devising "universal" multilingual models via cross-lingual transfer techniques of various kinds (Hwa et al., 2005; Zeman and Resnik, 2008; McDonald et al., 2011; Ganchev and Das, 2013; Martins, 2015). This line of research seeks ways of using data from resource-rich languages to solve tasks in resource-poor languages. Given the difficulty of handcrafting language-independent features, it is highly appealing to obtain rich, delexicalized, multilingual representations embedded in a shared space.

A string of work started with Klementiev et al. (2012) on learning bilingual embeddings for text classification. Hermann and Blunsom (2014) proposed a noise-contrastive objective to push the embeddings of parallel sentences to be close in space. A bilingual auto-encoder was proposed by Chandar et al. (2014), while Faruqui and Dyer (2014) applied canonical correlation analysis to parallel data to improve monolingual embeddings. Other works optimize a sum of monolingual and cross-lingual terms (Gouws et al., 2015; Soyer et al., 2015), or introduce bilingual variants of skip-gram (Luong et al., 2015; Coulmance et al., 2015). Recently, Pham et al. (2015) extended the non-compositional paragraph vectors of Le and Mikolov (2014) to a bilingual setting, achieving a new state of the art at the cost of more expensive (and non-deterministic) prediction.

In this paper, we propose an alternative joint formulation that learns embeddings suited to a particular task, together with the corresponding classifier for that task. We do this by minimizing a combination of a supervised loss function and a multilingual regularization term. Our approach leads to a convex optimization problem and makes a bridge between classical co-regularization approaches for semi-supervised learning (Sindhwani et al., 2005; Altun et al., 2005; Ganchev et al.,

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 2019–2028,
Berlin, Germany, August 7-12, 2016.

2008) and modern representation learning. In addition, we show that Euclidean co-regularizers have serious limitations to learn rich embeddings, when the number of task labels is small. We establish this by proving that the resulting embedding matrices have their rank upper bounded by the number of labels. This limitation does not exist for other regularizers (convex or not), such as the ℓ_1-distance and noise-contrastive distances.

Our experiments in the RCV1/RCV2 dataset yield state-of-the-art accuracy (92.7%) with this simple convex formulation, when transferring from English to German, without the need of negative sampling, extra monolingual data, or non-additive representations. For the reverse direction, our best number (79.3%), while far behind the recent `para_doc` approach (Pham et al., 2015), is on par with current compositional methods.

On the TED corpus, we obtained general purpose multilingual embeddings for 11 target languages, by considering the (auxiliary) task of reconstructing pre-trained English word vectors. The resulting embeddings led to cross-lingual multi-label classifiers that achieved the highest reported scores on 10 out of these 11 languages.[1]

2 Cross-Lingual Text Classification

We consider a **cross-lingual classification** framework, where a classifier is trained on a dataset from a source language (such as English) and applied to a target language (such as German). Later, we generalize this setting to multiple target languages and to other tasks besides classification.

The following data are assumed available:

1. A **labeled dataset** $\mathcal{D}_l := \{\langle \boldsymbol{x}^{(m)}, y^{(m)}\rangle\}_{m=1}^{M}$, consisting of text documents $\boldsymbol{x}$ in the source language categorized with a label $y \in \{1, \ldots, L\}$.
2. An **unlabeled parallel corpus** $\mathcal{D}_u := \{(\boldsymbol{s}^{(n)}, \boldsymbol{t}^{(n)})\}_{n=1}^{N}$, containing sentences $\boldsymbol{s}$ in the source language paired with their translations $\boldsymbol{t}$ in the target language (but no information about their categories).

Let V_{S} and V_{T} be the vocabulary size of the source and target languages, respectively. Throughout, we represent sentences $\boldsymbol{s} \in \mathbb{R}^{V_{\text{S}}}$ and $\boldsymbol{t} \in \mathbb{R}^{V_{\text{T}}}$ as vectors of word counts, and documents $\boldsymbol{x}$ as an average of sentence vectors. We assume that the unlabeled sentences largely outnumber the labeled documents, $N \gg M$, and that the number of labels L is relatively small. The goal is to use the data above to learn a classifier $h : \mathbb{R}^{V_{\text{T}}} \to \{1, \ldots, L\}$ for the target language.

This problem is usually tackled with a two-stage approach: in the first step, bilingual word embeddings $\boldsymbol{P} \in \mathbb{R}^{V_{\text{S}} \times K}$ and $\boldsymbol{Q} \in \mathbb{R}^{V_{\text{T}} \times K}$ are learned from $\mathcal{D}_u$, where each row of these matrices contains a Kth dimensional word representation in a shared vector space. In the second step, a standard classifier is trained on $\mathcal{D}_l$, using the source embeddings $\boldsymbol{P} \in \mathbb{R}^{V_{\text{S}} \times K}$. Since the embeddings are in a shared space, the trained model can be applied directly to classify documents in the target language. We describe next these two steps in more detail. We assume throughout an additive representation for sentences and documents (denoted ADD by Hermann and Blunsom (2014)). These representations can be expressed algebraically as $\boldsymbol{P}^\top \boldsymbol{x}, \boldsymbol{P}^\top \boldsymbol{s}, \boldsymbol{Q}^\top \boldsymbol{t} \in \mathbb{R}^K$, respectively.

Step 1: Learning the Embeddings. The cross-lingual embeddings $\boldsymbol{P}$ and $\boldsymbol{Q}$ are trained so that the representations of paired sentences $(\boldsymbol{s}, \boldsymbol{t}) \in \mathcal{D}_u$ have a small (squared) Euclidean distance

$$d_{\ell_2}(\boldsymbol{s}, \boldsymbol{t}) = \frac{1}{2}\|\boldsymbol{P}^\top \boldsymbol{s} - \boldsymbol{Q}^\top \boldsymbol{t}\|^2. \quad (1)$$

Since a direct minimization of Eq. 1 leads to a degenerate solution ($\boldsymbol{P} = \boldsymbol{0}, \boldsymbol{Q} = \boldsymbol{0}$), Hermann and Blunsom (2014) use instead a noise-contrastive large-margin distance obtained via negative sampling,

$$d_{\text{ns}}(\boldsymbol{s}, \boldsymbol{t}, \boldsymbol{n}) = [m + d_{\ell_2}(\boldsymbol{s}, \boldsymbol{t}) - d_{\ell_2}(\boldsymbol{s}, \boldsymbol{n})]_+, \quad (2)$$

where $\boldsymbol{n}$ is a random (unpaired) target sentence, m is a "margin" parameter, and $[x]_+ := \max\{0, x\}$. Letting J be the number of negative examples in each sample, they arrive at the following objective function to be minimized:

$$\mathcal{R}_{\text{ns}}(\boldsymbol{P}, \boldsymbol{Q}) := \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{J} d_{\text{ns}}(\boldsymbol{s}^{(n)}, \boldsymbol{t}^{(n)}, \boldsymbol{n}^{(n,j)}). \quad (3)$$

This minimization can be carried out efficiently with gradient-based methods, such as stochastic gradient descent or AdaGrad (Duchi et al., 2011). Note however that the objective function in Eq. 3 is not convex. Therefore, one may land at different local minima, depending on the initialization.

[1] We provide the trained embeddings at `http://www.cs.cmu.edu/~afm/projects/multilingual_embeddings.html`.

Step 2: Training the Classifier. Once we have the bilingual embeddings $\boldsymbol{P}$ and $\boldsymbol{Q}$, we can compute the representation $\boldsymbol{P}^\top \boldsymbol{x} \in \mathbb{R}^K$ of each document $\boldsymbol{x}$ in the labeled dataset $\mathcal{D}_l$. Let $\boldsymbol{V} \in \mathbb{R}^{K \times L}$ be a matrix of parameters (weights), with one column $\boldsymbol{v}_y$ per label. A linear model is used to make predictions, according to

$$\begin{aligned} \widehat{y} &= \text{argmax}_{y \in \{1,\ldots,L\}} \boldsymbol{v}_y^\top \boldsymbol{P}^\top \boldsymbol{x} \\ &= \text{argmax}_{y \in \{1,\ldots,L\}} \boldsymbol{w}_y^\top \boldsymbol{x}, \end{aligned} \quad (4)$$

where $\boldsymbol{w}_y$ is a column of the matrix $\boldsymbol{W} := \boldsymbol{PV} \in \mathbb{R}^{V_\text{S} \times L}$. In prior work, the perceptron algorithm was used to learn the weights $\boldsymbol{V}$ from the labeled examples in $\mathcal{D}_l$ (Klementiev et al., 2012; Hermann and Blunsom, 2014). Note that, at test time, it is not necessary to store the full embeddings: if $L \ll K$, we may simply precompute $\boldsymbol{W} := \boldsymbol{PV}$ (one weight per word and label) if the input is in the source language—or $\boldsymbol{QV}$, if the input is in the target language—and treat this as a regular bag-of-words linear model.

3 Jointly Learning to Embed and Classify

Instead of a two-stage approach, we propose to learn the bilingual embeddings and the classifier *jointly* on $\mathcal{D}_l \cup \mathcal{D}_u$, as described next.

Our formulation optimizes a combination of a **co-regularization function** $\mathcal{R}$, whose goal is to push the embeddings of paired sentences in $\mathcal{D}_u$ to stay close, and a **loss function** $\mathcal{L}$, which fits the model to the labeled data in $\mathcal{D}_l$.

The simplest choice for $\mathcal{R}$ is a simple Euclidean co-regularization function:

$$\begin{aligned} \mathcal{R}_{\ell_2}(\boldsymbol{P}, \boldsymbol{Q}) &= \frac{1}{N} \sum_{n=1}^{N} d_{\ell_2}(\boldsymbol{s}^{(n)}, \boldsymbol{t}^{(n)}) \\ &= \frac{1}{2N} \sum_{n=1}^{N} \|\boldsymbol{P}^\top \boldsymbol{s}^{(n)} - \boldsymbol{Q}^\top \boldsymbol{t}^{(n)}\|^2. \end{aligned} \quad (5)$$

An alternative is the ℓ_1-distance:

$$\mathcal{R}_{\ell_1}(\boldsymbol{P}, \boldsymbol{Q}) = \frac{1}{N} \sum_{n=1}^{N} \|\boldsymbol{P}^\top \boldsymbol{s}^{(n)} - \boldsymbol{Q}^\top \boldsymbol{t}^{(n)}\|_1. \quad (6)$$

One possible advantage of $\mathcal{R}_{\ell_1}(\boldsymbol{P}, \boldsymbol{Q})$ over $\mathcal{R}_{\ell_2}(\boldsymbol{P}, \boldsymbol{Q})$ is that the ℓ_1-distance is more robust to outliers, hence it is less sensitive to differences in the parallel sentences. Note that both functions in Eqs. 5–6 are jointly convex on $\boldsymbol{P}$ and $\boldsymbol{Q}$, unlike the one in Eq. 3. They are also simpler and do not require negative sampling. While these functions have a degenerate behavior in isolation (since they are both minimized by $\boldsymbol{P} = \boldsymbol{0}$ and $\boldsymbol{Q} = \boldsymbol{0}$), we will see that they become useful when plugged into a joint optimization framework.

The next step is to define the loss function $\mathcal{L}$ to leverage the labeled data in $\mathcal{D}_l$. We consider a log-linear model $P(y \mid \boldsymbol{x}; \boldsymbol{W}) \propto \exp(\boldsymbol{w}_y^\top \boldsymbol{x})$, which leads to the following logistic loss function:

$$\mathcal{L}_{LL}(\boldsymbol{W}) = -\frac{1}{M} \sum_{m=1}^{M} \log P(y^{(m)} \mid \boldsymbol{x}^{(m)}; \boldsymbol{W}). \quad (7)$$

We impose that $\boldsymbol{W}$ is of the form $\boldsymbol{W} = \boldsymbol{PV}$ for a fixed $\boldsymbol{V} \in \mathbb{R}^{K \times L}$, whose choice we discuss below.

Putting the pieces together and adding some extra regularization terms, we formulate our joint objective function as follows:

$$\begin{aligned} \mathcal{F}(\boldsymbol{P}, \boldsymbol{Q}) &= \mu \mathcal{R}(\boldsymbol{P}, \boldsymbol{Q}) + \mathcal{L}(\boldsymbol{PV}) \\ &\quad + \frac{\mu_\text{S}}{2} \|\boldsymbol{P}\|_\text{F}^2 + \frac{\mu_\text{T}}{2} \|\boldsymbol{Q}\|_\text{F}^2, \end{aligned} \quad (8)$$

where $\mu, \mu_\text{S}, \mu_\text{T} \geq 0$ are regularization constants. By minimizing a combination of $\mathcal{L}(\boldsymbol{PV})$ and $\mathcal{R}(\boldsymbol{P}, \boldsymbol{Q})$, we expect to obtain embeddings $\boldsymbol{Q}^*$ that lead to an accurate classifier h for the target language. Note that $\boldsymbol{P} = \boldsymbol{0}$ and $\boldsymbol{Q} = \boldsymbol{0}$ is no longer a solution, due to the presence of the loss term $\mathcal{L}(\boldsymbol{PV})$ in the objective.

Choice of $\boldsymbol{V}$. In Eq. 8, we chose to keep $\boldsymbol{V}$ fixed rather than optimize it. The rationale is that there are many more degrees of freedom in the embedding matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ than in $\boldsymbol{V}$ (concretely, $\mathcal{O}(K(V_\text{S} + V_\text{T}))$ versus $\mathcal{O}(KL)$, where we are assuming a small number of labels, $L \ll V_\text{S} + V_\text{T}$). Our assumption is that we have enough degrees of freedom to obtain an accurate model, regardless of the choice of $\boldsymbol{V}$. These claims will be backed in §4 by a more rigorous theoretical result. Keeping $\boldsymbol{V}$ fixed has another important advantage: it allows to minimize $\mathcal{F}$ with respect to $\boldsymbol{P}$ and $\boldsymbol{Q}$ only, which makes it a convex optimization problem if we choose $\mathcal{R}$ and $\mathcal{L}$ to be both convex—*e.g.*, setting $\mathcal{R} \in \{\mathcal{R}_{\ell_2}, \mathcal{R}_{\ell_1}\}$ and $\mathcal{L} := \mathcal{L}_{LL}$.

Relation to Multi-View Learning. An interesting particular case of this formulation arises if $K = L$ and $\boldsymbol{V} = \boldsymbol{I}_L$ (the identity matrix). In that case, we have $\boldsymbol{W} = \boldsymbol{P}$ and the embedding matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ are in fact weights for every pair of word and label, as in standard bag-of-word

models. In this case, we may interpret the co-regularizer $\mathcal{R}(\boldsymbol{P}, \boldsymbol{Q})$ in Eq. 8 as a term that pushes the *label scores* of paired sentences $\boldsymbol{P}^\top \boldsymbol{s}^{(n)}$ and $\boldsymbol{Q}^\top \boldsymbol{t}^{(n)}$ to be similar, while the source-based log-linear model is fit via $\mathcal{L}(\boldsymbol{W})$. The same idea underlies various semi-supervised co-regularization methods that seek agreement between multiple views (Sindhwani et al., 2005; Altun et al., 2005; Ganchev et al., 2008). In fact, we may regard the joint optimization in Eq. 8 as a generalization of those methods, making a bridge between those methods and representation learning.

Multilingual Embeddings. It is straightforward to extend the framework herein presented to the case where there are *multiple* target languages (say R of them), and we want to learn one embedding matrix for each, $\{\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_R\}$. The simplest way is to consider a sum of pairwise co-regularizers,

$$\mathcal{R}'(\boldsymbol{P}, \{\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_R\}) := \sum_{r=1}^{R} \mathcal{R}(\boldsymbol{P}, \boldsymbol{Q}_r). \quad (9)$$

If $\mathcal{R}$ is additive over the parallel sentences (which is the case for $\mathcal{R}_{\ell_2}$, $\mathcal{R}_{\ell_1}$ and $\mathcal{R}_{\text{ns}}$), then this procedure is equivalent to concatenating all the parallel sentences (regardless of the target language) and adding a language suffix to the words to distinguish them. This reduces directly to a problem in the same form as Eq. 8.

Pre-Trained Source Embeddings. In practice, it is often the case that pre-trained embeddings for the source language are already available (let $\bar{\boldsymbol{P}}$ be the available embedding matrix). It would be foolish not to exploit those resources. In this scenario, the goal is to use $\bar{\boldsymbol{P}}$ and the dataset $\mathcal{D}_u$ to obtain "good" embeddings for the target languages (possibly tweaking the source embeddings too, $\boldsymbol{P} \approx \bar{\boldsymbol{P}}$). Our joint formulation in Eq. 8 can also be used to address this problem. It suffices to set $K = L$ and $\boldsymbol{V} = \boldsymbol{I}_L$ (as in the multi-view learning case discussed above) and to define an auxiliary task that pushes $\boldsymbol{P}$ and $\bar{\boldsymbol{P}}$ to be similar. The simplest way is to use a reconstruction loss:

$$\mathcal{L}_{\ell_2}(\boldsymbol{P}, \bar{\boldsymbol{P}}) := \frac{1}{2}\|\boldsymbol{P} - \bar{\boldsymbol{P}}\|_{\text{F}}^2. \quad (10)$$

The resulting optimization problem has resemblances with the retrofitting approach of Faruqui et al. (2015), except that the goal here is to extend the embeddings to other languages, instead of pushing monolingual embeddings to agree with a semantic lexicon. We will present some experiments in §5.2 using this framework.

4 Limitations of the Euclidean Co-Regularizer

One may wonder how much the embedding dimension K influences the learned classifier. The next proposition shows the (surprising) result that, with the formulation in Eq. 8 with $\mathcal{R} = \mathcal{R}_{\ell_2}$, it makes absolutely no difference to increase K past the number of labels L. Below, $\boldsymbol{T} \in \mathbb{R}^{V_{\text{T}} \times N}$ denotes the matrix with columns $\boldsymbol{t}^{(1)}, \ldots, \boldsymbol{t}^{(N)}$.

Proposition 1. *Let $\mathcal{R} = \mathcal{R}_{\ell_2}$ and assume $\boldsymbol{T}$ has full row rank.*[2] *Then, for any choice of $\boldsymbol{V} \in \mathbb{R}^{K \times L}$, possibly with $K > L$, the following holds:*

1. *There is an alternative, low-dimensional, $\boldsymbol{V}' \in \mathbb{R}^{K' \times L}$ with $K' \leq L$ such that the classifier obtained (for both languages) by optimizing Eq. 8 using $\boldsymbol{V}'$ is the same as if using $\boldsymbol{V}$.*[3]
2. *This classifier depends on $\boldsymbol{V}$ only via the L-by-L matrix $\boldsymbol{V}^\top \boldsymbol{V}$.*
3. *If $\boldsymbol{P}^*, \boldsymbol{Q}^*$ are the optimal embeddings obtained with $\boldsymbol{V}$, then we always have $\text{rank}(\boldsymbol{P}^*) \leq L$ and $\text{rank}(\boldsymbol{Q}^*) \leq L$ regardless of K.*

Proof. See App. A.1 in the supplemental material. □

Let us reflect for a moment on the practical impact of Prop. 1. This result shows the limitation of the Euclidean co-regularizer $\mathcal{R}_{\ell_2}$ in a very concrete manner: when $\mathcal{R} = \mathcal{R}_{\ell_2}$, we only need to consider representations of dimension $K \leq L$.

Note also that a corollary of Prop. 1 arises when $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_L$, i.e., when $\boldsymbol{V}$ is chosen to have orthonormal columns (a sensible choice, since it corresponds to seeking embeddings that leave the label weights "uncorrelated"). Then, the second statement of Prop. 1 tells us that the resulting classifier will be the same as if we had simply set $\boldsymbol{V} = \boldsymbol{I}_L$ (the particular case discussed in §3). We will see in §5.1 that, despite this limitation, this classifier is actually a very strong baseline. Of course, if the number of labels L is large enough,

[2] This assumption is not too restrictive: it holds if $N \geq V_{\text{T}}$ and if no target sentence can be written as a linear combination of the others (this can be accomplished if we remove redundant parallel sentences).

[3] Let $\boldsymbol{P}^*, \boldsymbol{Q}^*$ and $\boldsymbol{P}'^*, \boldsymbol{Q}'^*$ be the optimal embeddings obtained with $\boldsymbol{V}$ and $\boldsymbol{V}'$, respectively. Since we are working with linear classifiers, the two classifiers are the same in the sense that $\boldsymbol{P}^*\boldsymbol{V} = \boldsymbol{P}'^*\boldsymbol{V}'$ and $\boldsymbol{Q}^*\boldsymbol{V} = \boldsymbol{Q}'^*\boldsymbol{V}'$.

this limitation might not be a reason for concern.[4] An instance will be presented in §5.2, where we will see that the Euclidean co-regularizer excels.

Finally, one might wonder whether Prop. 1 applies only to the (Euclidean) ℓ_2 norm or if it holds for arbitrary regularizers. In fact, we show in App. A.2 that this limitation applies more generally to Mahalanobis-Frobenius norms, which are essentially Euclidean norms after a linear transformation of the vector space. However, it turns out that for general norms such limitation does *not* exist, as shown below.

Proposition 2. *If $\mathcal{R} = \mathcal{R}_{\ell_1}$ in Eq. 8, then the analogous to Proposition 1 does not hold. It also does not hold for the ℓ_∞-norm and the ℓ_0-"norm."*

Proof. See App. A.3 in the supplemental material. □

This result suggests that, for other regularizers $\mathcal{R} \neq \mathcal{R}_{\ell_2}$, we may eventually obtain better classifiers by increasing K past L. As such, in the next section, we experiment with $\mathcal{R} \in \{\mathcal{R}_{\ell_2}, \mathcal{R}_{\ell_1}, \mathcal{R}_{\text{ns}}\}$, where $\mathcal{R}_{\text{ns}}$ is the (non-convex) noise-contrastive regularizer of Eq. 3.

5 Experiments

We report results on two experiments: one on cross-lingual classification on the Reuters RCV1/RCV2 dataset, and another on multi-label classification with multilingual embeddings on the TED Corpus.[5]

5.1 Reuters RCV1/RCV2

We evaluate our framework on the cross-lingual document classification task introduced by Klementiev et al. (2012). Following prior work, our dataset $\mathcal{D}_u$ consists of 500,000 parallel sentences from the Europarl v7 English-German corpus (Koehn, 2005); and our labeled dataset $\mathcal{D}_l$ consists of English and German documents from the RCV1/RCV2 corpora (Lewis et al., 2004), each categorized with one out of $L = 4$ labels. We used the same split as Klementiev et al. (2012): 1,000 documents for training, of which 200 are held out as validation data, and 5,000 for testing.

Note that, in this dataset, we are classifying documents based on their bag-of-word representations, and learning word embeddings by bringing the bag-of-word representations of parallel sentences to be close together. In this sense, we are bringing together these multiple levels of representations (document, sentence and word).

We experimented with the joint formulation in Eq. 8, with $\mathcal{L} := \mathcal{L}_{LL}$ and $\mathcal{R} \in \{\mathcal{R}_{\ell_2}, \mathcal{R}_{\ell_1}, \mathcal{R}_{\text{ns}}\}$. We optimized with AdaGrad (Duchi et al., 2011) with a stepsize of 1.0, using mini-batches of 100 Reuters RCV1/RCV2 documents and 50,000 Europarl v7 parallel sentences. We found no need to run more than 100 iterations, with most of our runs converging under 50. Our vocabulary has 69,714 and 175,650 words for English and German, respectively, when training on the English portion of the Reuters RCV1/RCV2 corpus, and 61,120 and 183,888 words for English and German, when training in the German portion of the corpus. This difference is due to the inclusion of words in the training data into the vocabulary. We do not remove any words from the vocabulary, for simplicity. We used the validation set to tune the hyperparameters $\{\mu, \mu_{\text{S}}, \mu_{\text{T}}\}$ and to choose the iteration number. When using $K = L$, we chose $\boldsymbol{V} = \boldsymbol{I}_L$; otherwise, we chose $\boldsymbol{V}$ randomly, sampling its entries from a Gaussian $\mathcal{N}(0, 0.1)$.

Table 1 shows the results. We include for comparison the most competitive systems published to date. The first thing to note is that our joint system with Euclidean co-regularization performs very well for this task, despite the theoretical limitations shown in §4. Although its embedding size is only $K = 4$ (one dimension per label), it outperformed all the two-stage systems trained on the same data, in both directions.

For the EN→DE direction, our joint system with ℓ_1 co-regularization achieved state-of-the-art results (92.7%), matching two-stage systems that use extra monolingual data, negative sampling, or non-additive document representations. It is conceivable that the better results of $\mathcal{R}_{\ell_1}$ over $\mathcal{R}_{\ell_2}$ come from its higher robustness to differences in the parallel sentences.

For the DE→EN direction, our best result (79%) was obtained with the noise-contrastive co-regularizer, which outperformed all systems except `para_doc` (Pham et al., 2015). While the accuracy of `para_doc` is quite impressive, note that it requires 500-dimensional embeddings

[4] For regression tasks (such as the one presented in the last paragraph of 3), instead of the "number of labels," L should be regarded as the number of output variables to regress.

[5] Our code is available at `https://github.com/dcferreira/multilingual-joint-embeddings`.

		K	EN→DE	DE→EN
I-Matrix	[KTB12]	40	77.6	71.1
ADD	[HB14]	40	83.7	71.4
ADD	[HB14]	128	86.4	74.7
BI	[HB14]	40	83.4	69.2
BI	[HB14]	128	86.1	79.0
BilBOWA	[GBC15]	40	86.5	75.0
Binclusion	[SSA15]	40	86.8	76.7
Bincl.+RCV	[SSA15] (‡)	40	**92.7**	84.4
CLC-WA	[SLLS15] (†)	40	91.3	77.2
`para_sum`	[PLM15] (†)	100	90.6	78.8
`para_doc`	[PLM15] (†)	500	**92.7**	**91.5**
Joint, $\mathcal{R}_{\ell_2}$		4	91.2	78.2
Joint, $\mathcal{R}_{\ell_1}$		4	**92.7**	76.0
Joint, $\mathcal{R}_{\ell_1}$		40	**92.7**	76.2
Joint, $\mathcal{R}_{\text{ns}}$		4	91.2	76.8
Joint, $\mathcal{R}_{\text{ns}}$		40	91.4	79.3

Table 1: Accuracies in the RCV1/RCV2 dataset. Shown for comparison are Klementiev et al. (2012) [KTB12], Hermann and Blunsom (2014) [HB14], Gouws et al. (2015) [GBC15], Soyer et al. (2015) [SSA15], Shi et al. (2015) [SLLS15], and Pham et al. (2015) [PLM15]. Systems marked with (†) used the full 1.8M parallel sentences in Europarl. The one with (‡) used additional target monolingual data from RCV1/RCV2. The bottom rows refer to our joint method, with Euclidean (ℓ_2), ℓ_1, and noise-contrastive co-regularization.

(hence many more parameters), was trained on more parallel sentences, and requires more expensive (and non-deterministic) computation at test time to compute a document's embedding. Our method has the advantage of being simple and very fast to train: it took less than 1 minute to train the joint-$\mathcal{R}_{\ell_1}$ system for EN→DE, using a single core on an Intel Xeon @2.5 GHz. This can be compared with Klementiev et al. (2012), who took 10 days on a single core, or Coulmance et al. (2015), who took 10 minutes with 6 cores.[6]

Although our theoretical results suggest that increasing K when using the ℓ_1 norm may increase the expressiveness of our embeddings, our results do not support this claim (the improvements in DE→EN from $K = 4$ to $K = 40$ were tiny). However, it led to a gain of 2.5 points when using negative sampling. For $K = 40$, this system is much more accurate than Hermann and Blunsom (2014), which confirms that learning the embeddings together with the task is highly beneficial.

[6]Coulmance et al. (2015) reports accuracies of 87.8% (EN→DE) and 78.7% (DE→EN), when using 10,000 training documents from the RCV1/RCV2 corpora.

5.2 TED Corpus

To assess the ability of our framework to handle multiple target languages, we ran a second set of experiments on the TED corpus (Cettolo et al., 2012), using the training and test partitions created by Hermann and Blunsom (2014), downloaded from `http://www.clg.ox.ac.uk/tedcorpus`. The corpus contains English transcriptions and multilingual, sentence-aligned translations of talks from the TED conference in 12 different languages, with 12,078 parallel documents in the training partition (totalling 1,641,985 parallel sentences). Following their prior work, we used this corpus both as parallel data ($\mathcal{D}_u$) and as the task dataset ($\mathcal{D}_l$). There are $L = 15$ labels and documents can have multiple labels.

We experimented with two different strategies:

- A one-stage system (*Joint*), which jointly trains the multilingual embeddings and the multi-label classifier (similarly as in §5.1). To cope with multiple target languages, we used a sum of pairwise co-regularizers as described in Eq. 9. For classification, we use multinomial logistic regression, where we select those labels with a posterior probability above 0.18 (tuned on vali-

dation data).

- A two-stage approach (*Joint w/ Aux*), where we first obtain multilingual embeddings by applying our framework with an auxiliary task with pre-trained English embeddings (as described in Eq. 10 and in the last paragraph of §3), and then use the resulting multilingual representations to train the multi-label classifier. We address this multi-label classification problem with independent binary logistic regressors (one per label), trained by running 100 iterations of L-BFGS (Liu and Nocedal, 1989). At test time, we select those labels whose posterior probability are above 0.5.

For the *Joint w/ Aux* strategy, we used the 300-dimensional GloVe-840B vectors (Pennington et al., 2014), downloaded from `http://nlp.stanford.edu/projects/glove/`.

Table 2 shows the results for cross-lingual classification, where we use English as source and each of the other 11 languages as target. We compare our two strategies above with the strong Machine Translation (MT) baseline used by Hermann and Blunsom (2014) (which translates the input documents to English with a state-of-the-art MT system) and with their two strongest systems, which build document-level representations from embeddings trained bilingually or multilingually (called DOC/ADD *single* and DOC/ADD *joint*, respectively).[7] Overall, our *Joint* system with ℓ_2 regularization outperforms both Hermann and Blunsom (2014)'s systems (but not the MT baseline) for 8 out of 11 languages, performing generally better than our ℓ_1-regularized system. However, the clear winner is our ℓ_2-regularized *Joint w/ Aux* system, which wins over all systems (including the MT baseline) by a substantial margin, for all languages. This shows that pre-trained source embeddings can be extremely helpful in bootstrapping multilingual ones.[8] On the other hand, the performance of the *Joint w/ Aux* system with ℓ_1 regularization is rather disappointing. Note that the limitations of $\mathcal{R}_{\ell_2}$ shown in §4 are not a concern here, since the auxiliary task has $L = 300$ dimensions (the dimension of the pre-trained embeddings). A small sample of the multilingual embeddings produced by the winner system is shown in Table 4.

Finally, we did a last experiment in which we use our multilingual embeddings obtained with *Joint w/ Aux* to train monolingual systems for each language. This time, we compare with a bag-of-words naïve Bayes system (reported by Hermann and Blunsom (2014)), a system trained on the Polyglot embeddings from Al-Rfou et al. (2013) (which are multilingual, but not in a shared representation space), and the two systems developed by Hermann and Blunsom (2014). The results are shown in Table 3. We observe that, with the exception of Turkish, our systems consistently outperform all the competitors. Comparing the bottom two rows of Tables 2 and 3 we also observe that, for the ℓ_2-regularized system, there is not much degradation caused by cross-lingual training versus training on the target language directly (in fact, for Spanish, Polish, and Brazilian Portuguese, the former scores are even higher). This suggests that the multilingual embeddings have high quality.

6 Conclusions

We proposed a new formulation which jointly minimizes a combination of a supervised loss function with a multilingual co-regularization term using unlabeled parallel data. This allows learning task-specific multilingual embeddings together with a classifier for the task. Our method achieved state-of-the-art accuracy on the Reuters RCV1/RCV2 cross-lingual classification task in the English to German direction, while being extremely simple and computationally efficient. Our results in the Reuters RCV1/RCV2 task, obtained using Europarl v7 as parallel data, show that our method has no trouble handling different levels of representations simutaneously (document, sentence and word). On the TED Corpus, we obtained the highest reported scores for 10 out of 11 languages, using an auxiliary task with pre-trained English embeddings.

Acknowledgments

We would like to thank the three anonymous reviewers. This work was partially supported by the European Union under H2020 project SUMMA, grant 688139, and by Fundação para a Ciência e Tecnologia (FCT),

[7] Note that, despite the name, the Hermann and Blunsom (2014)'s *joint* systems are not doing joint training as we are.

[8] Note however that, overall, our *Joint w/ Aux* systems have access to more data than our *Joint* systems and also than Hermann and Blunsom (2014)'s systems, since the pre-trained embeddings were trained on a large amount of English monolingual data. Yet, the amount of target language data is the same.

		Ara.	Ger.	Spa.	Fre.	Ita.	Dut.	Pol.	Br. Pt.	Rom.	Rus.	Tur.
MT Baseline	[HB14]	42.9	46.5	51.8	52.6	51.4	50.5	44.5	47.0	49.3	43.2	40.9
DOC/ADD *single*	[HB14]	41.0	42.4	38.3	47.6	48.5	26.4	40.2	35.4	41.8	44.8	45.2
DOC/ADD *joint*	[HB14]	39.2	40.5	44.3	44.7	47.5	45.3	39.4	40.9	44.6	47.6	41.7
Joint, $\mathcal{R}_{\ell_2}$, $K = 15$		41.8	46.6	46.6	46.0	48.7	52.5	39.5	40.8	47.6	44.9	47.2
Joint, $\mathcal{R}_{\ell_1}$, $K = 15$		44.0	44.7	49.4	40.1	46.1	49.4	35.7	43.5	40.5	42.2	43.4
Joint w/ Aux, $\mathcal{R}_{\ell_2}$, $K = 300$		**46.9**	**52.0**	**59.4**	**54.6**	**56.0**	**53.6**	**51.0**	**51.7**	**53.9**	**52.3**	**49.5**
Joint w/ Aux, $\mathcal{R}_{\ell_1}$, $K = 300$		44.0	40.4	40.4	39.5	38.6	38.1	43.2	36.6	35.1	44.3	44.4

Table 2: Cross-lingual experiments on the TED Corpus using English as a source language. Reported are the micro-averaged F_1 scores for a machine translation baseline and the two strongest systems of Hermann and Blunsom (2014), our one-stage joint system (*Joint*), and our two-stage system that trains the multilingual embeddings jointly with the auxiliary task of fitting pre-trained English embeddings (*Joint w/ Aux*), with both ℓ_1 and ℓ_2 regularization. Bold indicates the best result for each target language.

		Ara.	Ger.	Spa.	Fre.	Ita.	Dut.	Pol.	Br. Pt.	Rom.	Rus.	Tur.
BOW baseline	[HB14]	46.9	47.1	52.6	53.2	52.4	52.2	41.5	46.5	50.9	46.5	**51.3**
Polyglot	[HB14]	41.6	27.0	41.8	36.1	33.2	22.8	32.3	19.4	30.0	40.2	29.5
DOC/ADD Single	[HB14]	42.2	42.9	39.4	48.1	45.8	25.2	38.5	36.3	43.1	47.1	43.5
DOC/ADD Joint	[HB14]	37.1	38.6	47.2	45.1	39.8	43.9	30.4	39.4	45.3	40.2	44.1
Joint w/ Aux, $\mathcal{R}_{\ell_2}$, $K = 300$		48.6	**54.4**	57.5	**55.8**	**56.9**	**54.5**	46.1	**51.3**	**56.5**	**53.0**	49.5
Joint w/ Aux, $\mathcal{R}_{\ell_1}$, $K = 300$		**52.4**	47.8	**57.8**	50.0	53.3	52.3	**47.6**	49.0	49.2	51.4	50.9

Table 3: Monolingual experiments on the TED Corpus. Shown are the micro-averaged F_1 scores for a bag-of-words baseline, a system trained on Polyglot embeddings, the two strongest systems of Hermann and Blunsom (2014), and our *Joint w/ Aux* system with ℓ_1 and ℓ_2 regularization.

january_en	**science_en**	**oil_en**	**road_en**	**speak_en**
januari_nl	مولعلا_ar	óleo_pb	route_fr	spreken_nl
şubat_tr	ملعلا_ar	olie_nl	strada_it	fala_pb
gennaio_it	ciência_pb	petrolio_it	weg_nl	ملاكلا_ar
februarie_ro	science_fr	öl_de	drum_ro	gesproken_nl
ريارباف_ar	ştiinţa_ro	pétrole_fr	ريسلا_ar	habla_es
ianuarie_ro	wetenschap_nl	petrol_tr	estrada_pb	konuşma_tr
febrero_es	scienza_it	petróleo_es	drogi_pl	говорить_ru
janvier_fr	ciencia_es	طفنلا_ar	lopen_nl	horen_nl
ريانيد_ar	wissenschaft_de	petróleo_pb	strade_it	mowy_pl
janeiro_pb	científica_pb	petrol_ro	drodze_pl	vorbească_ro
enero_es	nauka_pl	aceite_es	wegen_nl	spreekt_nl
september_nl	bilim_tr	ropę_pl	yol_tr	ثيدحلا_ar
settembre_it	ştiinţa_ro	нефть_ru	camino_es	sprechen_de
septiembre_es	ştiinţă_ro	petrolul_ro	conduce_ro	ii_ro
september_de	nauki_pl	нефти_ru	andar_pb	discours_fr
ekim_tr	наука_ru	طفذ_ar	пути_ru	sentire_it
ربمتبس_ar	ملع_ar	ropy_pl	ريست_ar	contar_pb
febbraio_it	مولع_ar	تيز_ar	далеко_ru	себя_ru
septembrie_ro	scientifica_it	ulei_ro	yolculuk_tr	صخش_ar
setembro_pb	scienze_it	تيزلا	yola_tr	poser_fr

Table 4: Examples of nearest neighbor words for the multilingual embeddings trained with our *Joint w/ Aux* system with ℓ_2 regularization. Shown for each English word are the 20 closest target words in Euclidean distance, regardless of language.

through contracts UID/EEA/50008/2013, through the LearnBig project (PTDC/EEI-SII/7092/2014), and the GoLocal project (grant CMUPERI/TIC/0046/2014).

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. *arXiv preprint arXiv:1307.1662* .

Yasemin Altun, Mikhail Belkin, and David A. McAllester. 2005. Maximum Margin Semi-Supervised Learning for Structured Variables. In *Advances in Neural Information Processing Systems 18*. pages 33–40.

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

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proc. of the 16th Conference of the European Association for Machine Translation*. pages 261–268.

Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh M. Khapra, Balaraman Ravindran, Vikas Raykar, and Amrita Saha. 2014. An Autoencoder Approach to Learning Bilingual Word Representations.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proc. of Empirical Methods for Natural Language Processing*. pages 740–750.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proc. of the International Conference on Machine Learning*. ACM, pages 160–167.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12:2493–2537.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Transgram, Fast Cross-lingual Word-embeddings. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pages 1109–1113.

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.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proc. of Annual Meeting of the North-American Chapter of the Association for Computational Linguistics*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proc. of Annual Meeting of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kuzman Ganchev and Dipanjan Das. 2013. Cross-lingual discriminative learning of sequence models with posterior regularization. In *Proc. of Empirical Methods in Natural Language Processing*.

Kuzman Ganchev, Joao Graca, John Blitzer, and Ben Taskar. 2008. Multi-view learning over structured and non-identical outputs. In *Proc. of Conference on Uncertainty in Artificial Intelligence*.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. *Proceedings of the 32nd International Conference on Machine Learning (2015)* pages 748–756.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributed Semantics. *Proceedings of ACL* pages 58–68.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering* 11(3):311–325.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers (2012)* pages 1459–1474.

Philipp Koehn. 2005. Europarl: A parallel corpus

for statistical machine translation. *MT summit* 11.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. *International Conference on Machine Learning - ICML 2014* 32:1188–1196.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3:211–225.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research* 5:361–397.

D. C. Liu and J. Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming* 45:503–528.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual Word Representations with Monolingual Quality in Mind. *Workshop on Vector Modeling for NLP* pages 151–159.

André F. T. Martins. 2015. Transferring coreference resolvers with posterior regularization. In *ACL*.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proc. of Empirical Methods in Natural Language Processing*.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)* pages 1–12.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. pages 1532–1543.

Kaare Brandt Petersen and Michael Syskind Pedersen. 2012. The Matrix Cookbook.

Hieu Pham, Minh-Thang Luong, and Christopher D. Manning. 2015. Learning Distributed Representations for Multilingual Text Sequences. *Workshop on Vector Modeling for NLP* pages 88–94.

Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning Cross-lingual Word Embeddings via Matrix Co-factorization. *Annual Meeting of the Association for Computational Linguistics* pages 567–572.

Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. 2005. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML workshop on learning with multiple views*. Citeseer, pages 74–79.

Richard Socher, Jeffrey Pennington, and Eh Huang. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Conference on Empirical Methods in Natural Language Processing*. pages 151–161.

Hubert Soyer, Pontus Stenetorp, and Akiko Aizawa. 2015. Leveraging Monolingual Data for Crosslingual Compositional Word Representations. *Proceedings of the 2015 International Conference on Learning Representations (ICLR)* .

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP*. pages 35–42.

Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proc. of Empirical Methods for Natural Language Processing*. pages 1393–1398.