Multilingual Distributed Representations without Word Alignment

Karl Moritz Hermann and Phil Blunsom
Department of Computer Science, University of Oxford

TL;DR

Multilingual Distributed Representations

- Semantic-transfer across languages
- Compositional model means no need for word alignments
- Noise-contrastive max-margin approach for joint learning
- Approach applicable to any type of compositional vector model
- Concise model outperforming state of the art on several tasks

Extending the Distributional Hypothesis

The distributional hypothesis is used to learn word representations based on the context that these words appear in.

We extend this hypothesis to learn representations based on the *semantic* context of words, and use multilingual data to provide this context.

We use a compositional vector model to transfer semantics at the sentence level, thereby preventing bias from noisy alignments and capturing the broader semantic context.

The biCVM Objective Function

Strongly align representations of semantically equivalent sentences (a, b):

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

Maintain a margin between unaligned sentence pairs (a, n):

$$E_{noise}(a, b, n) = [1 + E_{dist}(a, b) - E_{dist}(a, n)]_+ \quad (2)$$

The resulting objective function for a parallel corpus $\mathcal{C}_{A,B}$:

$$J(\theta_{bi}) = \sum_{(a,b)\in\mathcal{C}_{A,B}} \left(\sum_{i=1}^{k} E_{noise}(a, b, n_i) \right) + \frac{\lambda}{2}\|\theta_{bi}\|^2 \quad (3)$$

The biCVM Model

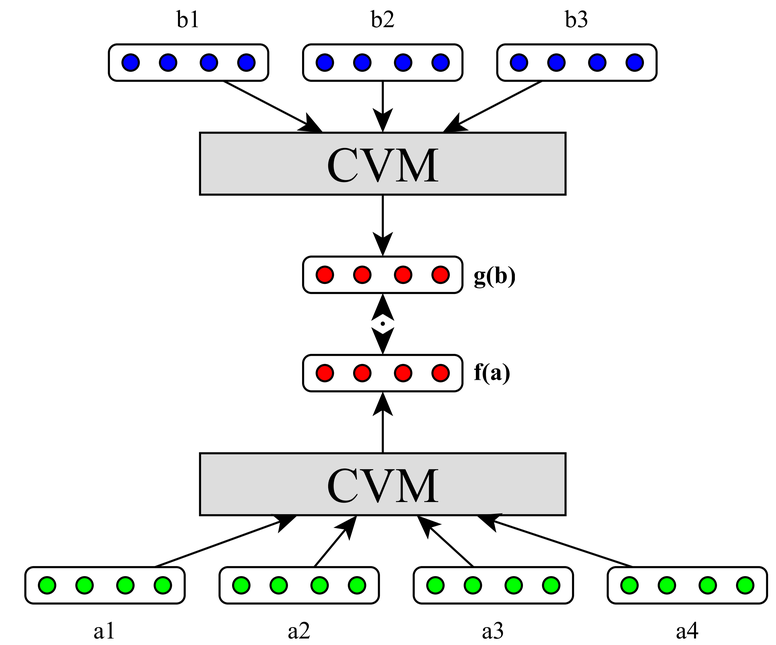

Figure: Schematic description of biCVM model applied to two sentences a and b, using composition vector models f and g.

Experimental Setup

1. **Representation Learning**
   We train our model on 500,000 parallel sentences from the Europarl corpus (v7, German-English section) using adaptive gradient descent and an L_2 regularizer.

   biCVM+ is additionally trained on English-French parallel data.

2. **Classifier training**
   Using the trained model, we learn representations for the documents in the RCV v2 corpus training data, and then train an averaged perceptron to classify these documents by label (4 labels). Settings as in Klementiev et al. (2010).

Cross-Lingual Document Classification

The BiCVM models outperform the prior state of the art on a cross-lingual document classification task, where a simple classifier is trained using data in one language and then evaluated on text from a second language.

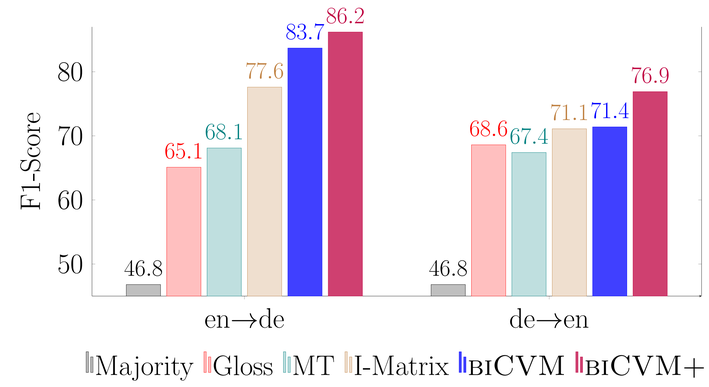

Figure: Results on the RCV cross-lingual document classification task.

Qualitative Analysis

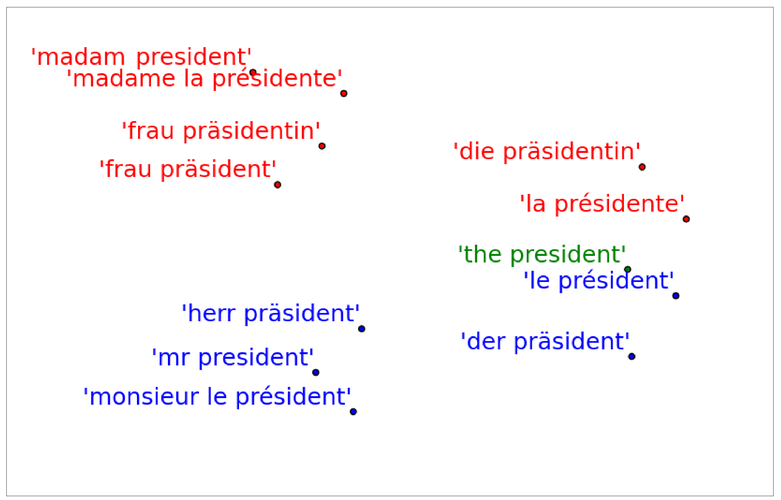

Figure: t-SNE projections from a model trained on two language pairs (English-German and English-French). The colours denote gender, with green denoting the gender neutral English form.