Multilingual Distributional Semantics

Benno Kruit Sara Veldhoen

January 13, 2015

Outline

Our first idea (and why it wouldn't work)

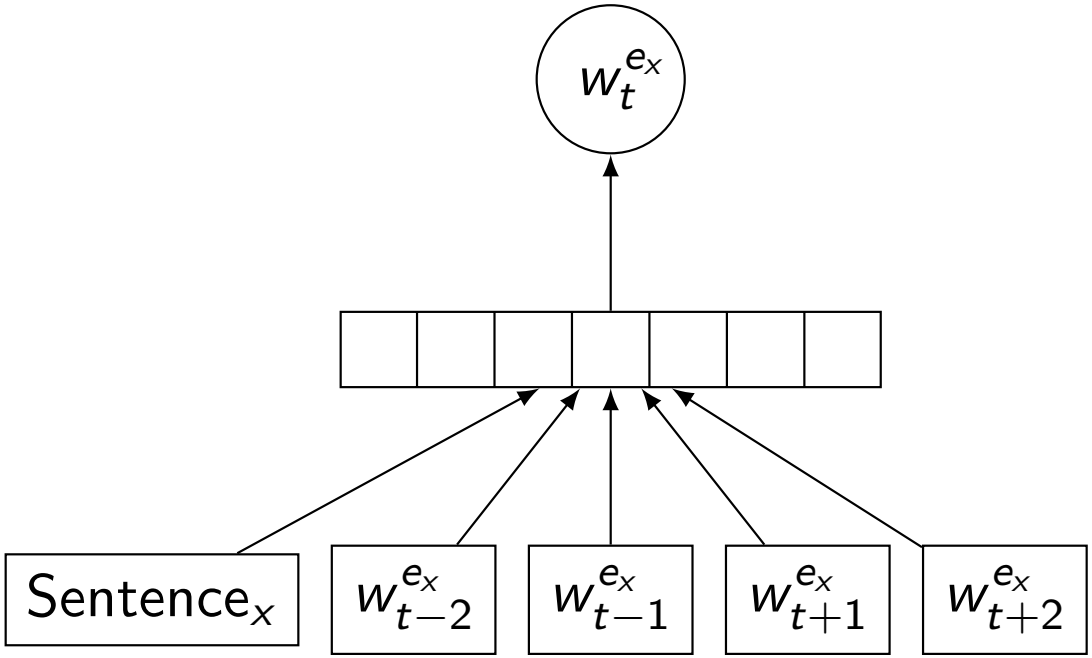


Figure : Bilingual distributed memory. The same architecture is trained with English context and word prediction replaced by the other language(s).

Our new idea

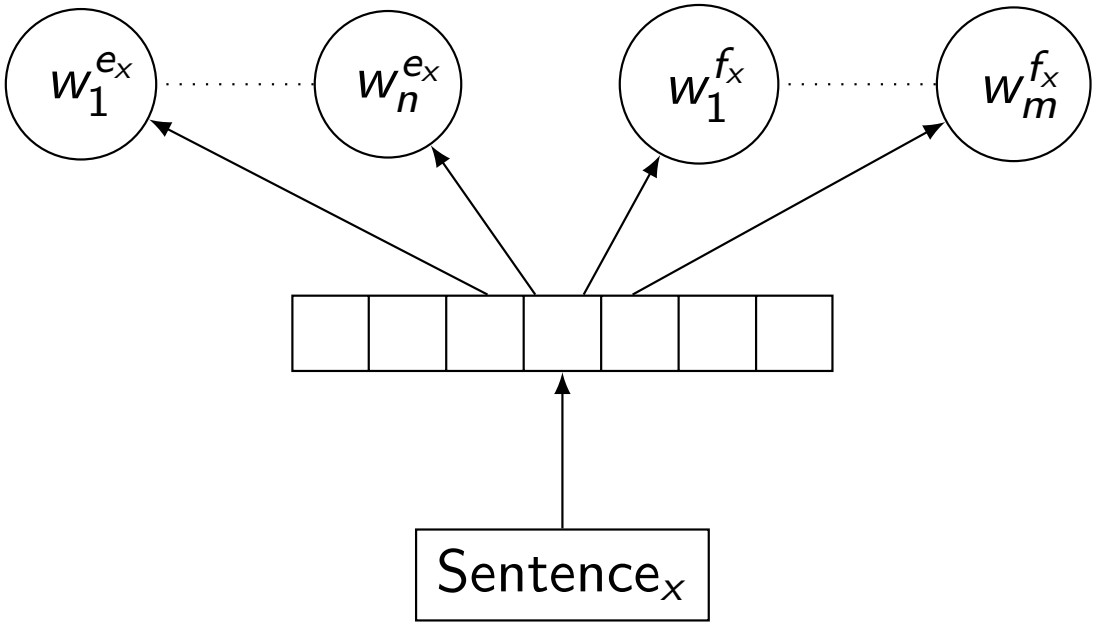


Figure : Bilingual dbow

Our new idea

- Training a single embedding for parallel sentences
- Word embeddings are not trained
- Can be extended to more than two languages

Our new idea

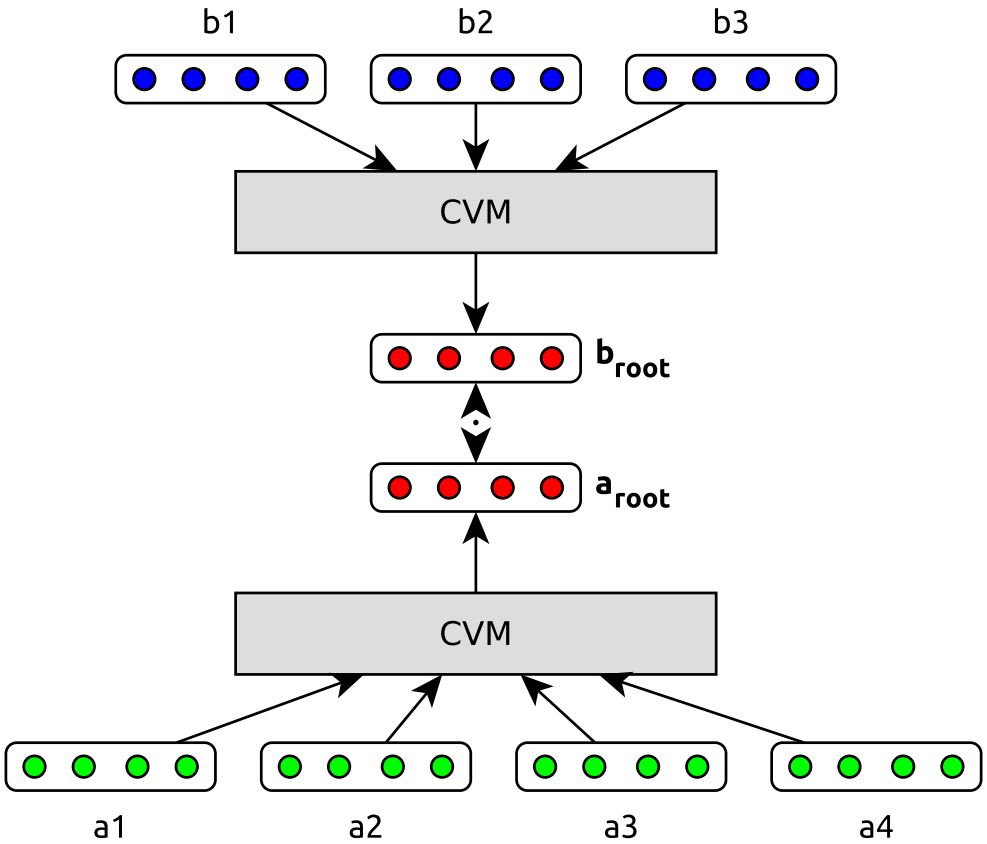
- Use the sentence embeddings to obtain word vector:

$$emb(w) = \frac{1}{freq(w, D)} \sum_{s \in D} freq(w, s) emb(s)$$

- Quite good performance (as we will see later)

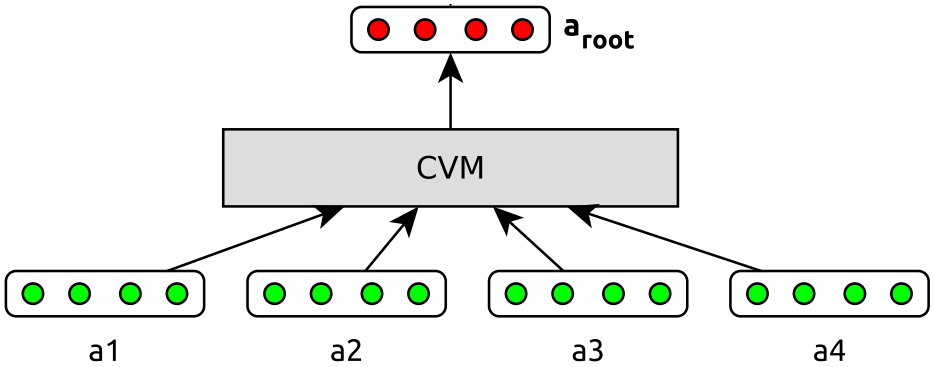
Our new idea

- Recall the model by Hermann and Blunsom:



Our new idea

- We could have a similar training procedure
- Only: we are not training the sentences, but assume fixed 'gold standard' sentence embeddings



- So, we could plug in any compositional model

Evaluation and results

- Training word embeddings: on Europarl data (50k or 500k sentences)
- Monolingual (English) evaluation: analogy task
- Crosslingual evaluation: document classification

Evaluation and results

Crosslingual Doccument classification:

- Given word embeddings, obtain document representation for train and test documents in all languages

$$emb(doc) = \sum_{w \in doc} idf(w) * emb(w)$$

- Train a classifier (averaged perceptron) on the training document representations for one language
- Test classifier performance on the test document representations for another language

Evaluation and results

RCV (Reuters) data:

- English-German
- Multiclass classification:
  each document is assigned a single class (topic)
- Performance measure: accuracy
- Baseline: majority class

Evaluation and results

TED data:

- Many languages
- Binary classification: each class (topic) has positive and negative examples
- Performance measure: F1 score
- Baseline: ??