Project Proposal
Word Embeddings in Multiple Languages

Sara Veldhoen Benno Kruit

December 16, 2014

1 Introduction

The aim of our project is to obtain a single vector space that captures the semantics of many different languages at the same time. Since such a vector space is less dependent on a specific language or culture, it is supposed to be a better representation of the real world concepts underlying language. Moreover, we suspect that rare words get a more reliable representation.

Multilingual word embeddings are useful for a variety of tasks. They are already used in Machine Translation, and could prove useful for cross-lingual information retrieval, parsing and semantic frame induction. Embedding many languages in the same means data from one language could improve word representations in other languages and mitigate sparsity. However, this is as of yet an open research question.

In section 2 we discuss existing approaches to this task. We will reproduce these experiments for the case of more than two languages, plus a new approach that we introduce in section 3.

The experiments we want to conduct are explained in section 4 plus the data we plan to use. For evaluation, we will use the methods from [5, 3]. We also plan to create a visualisation by mapping part of the semantic space to 2D and investigate the words in that area.

2 Related work

Some research has focused on this problem, using both different techniques to obtain word representations, and different approaches to the cross-lingual aspects. The evaluation methods applied also vary a lot.

2.1 Linear Mapping

According to Mikolov et al. [5], the vector space of word representations in different languages are geometrically similar, because words in languages are grounded in real world concepts. It is therefore possible to find a linear mapping between these vector spaces.

The approach is to first train word embeddings on large monolingual data for both languages separately, using the `word2vec` implementation. In the reported experiments, the so-called CBOW architecture is used, that predicts a word given its context in both directions. Notably, the authors also propose a way to include some phrases: multi-word expressions. This may prove useful for translation, as one multiple words can together express a concept that has a single word in another language.

Using a relatively small set of gold standard word translations, in this case obtained from Google Translate, a transformation matrix W is searched. The training objective is to minimize the distance between words that are translations of one another.

The evaluation is performed on a test set of gold-standard word translations, again from Google Translate. The word representation in the source language is transformed using W, and a ranked list of the nearest words in the target language is the output. The precision at ranks 1 and 5 is reported.

2.2 Multitask Learning

Distributed representations for a pair of languages are induced jointly by Klementiev and Titov [3]. Words in both languages are represented in a single vector space.

The induction is treated as a multitask learning problem where each task corresponds to a single word. The training influences other tasks depending on the task-relatedness. The latter is derived from co-occurrence statistics in bilingual parallel data: the number of alignment links between that word and its (supposed) translations.

The word representations are induced in a neural language model architecture. The n preceding words form the context, their representations are concatenated to form a context vector. The probability of the next word occuring is predicted from this vector. The training procedure aims to find the word representations that minimize the data (log) likelihood: $L(\theta) = \sum_{t=1}^{T} \log \hat{P}_\theta(w_t | w_{t-n+1:t-1})$.

The method is evaluated on a real-world task: crosslingual document classification. Topic annotations are available for documents in one of the languages, and the system predicts the topics of documents in the other language. The jointly induced word representation outperform two other approaches to the problem: glossing (where every word in the document is translated separately, based on word alignments) and Machine Translation.

2.3 Joint Learning from Sentence Embeddings

Unlike the previous approaches, Hermann and Blunsum [2] start from sentence alignments, which share the same semantics. The assumption is that some function can describe the composition of word embeddings into a sentence embedding. For the sake of argument, the authors use a simple bag-of-words additive interpretation of composition. The word embeddings are induced jointly for both languages from these sentence-embeddings, by minimizing the distance between both sums of word embeddings. In order to make sure the weights won't be reduced to zero, similarity between unaligned sentence embeddings is penalized.

The same evaluation as in [3] is applied, i.e. the document classification task. Furthermore, the authors present a graphical qualitative analysis. In [1], one of the authors expands this approach by evaluating on a larger number of language pairs.

3 A new approach

Le and Mikolov [4] have extended their monolingual `word2vec` model to create representations of sentences, paragraphs or documents. It uses the paragraph vector as a part of the context of each word in the paragraph (Figure 1). This way, the paragraph vector influences the learned representations of those words in the same way that their context words do.

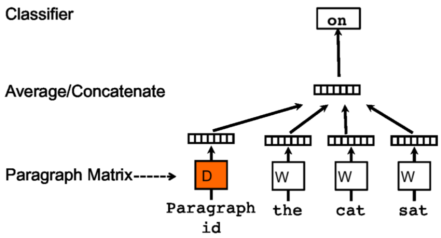

Figure 1: `word2vec` for paragraphs

This paragraph representation could also be used for encouraging similarity between two bitext sentences. In our novel approach, we will run the algorithm from [4], but using the same paragraph vector when training word vectors from parallel sentences. This is equivalent to concatenating the parallel sentences and training from the context windows that do not bridge the sentence boundary. Without using word alignments, the information from the words in the first language will create a representation for the sentence. Then, this sentence vector will influence the learning of word embeddings in the second language, again without word alignments.

The sentence representation therefore acts as a way to relate the word spaces in both languages. We hope this will create a word vector space that is trainable on both monolingual and parallel data, allowing for the mitigation of sparsity in all languages.

4 Experiments and evaluation

We are interested in comparing approaches to learn a word representation space for multiple languages. This vector space can be useful for a variety of tasks.

4.1 Training

The source code is available for both `paragraph2vec`[1] and `bicvm`[2]. The work in [2] can be replicated with no further adjustments to `bicvm`. Additionally, we should be able to train a vector space with more than two languages by tagging each word with a marker that specifies its language. The new approach described above will need some small changes to `paragraph2vec`, but these are minimal. We will implement and use the linear mapping algorithm described in section 2.1 as a baseline.

As some of the methods require sentence-aligned (or word-aligned) data, we use the Europarl corpus for training. We hope to be able to use many languages, but at least three or four. Possibly, Wikipedia or Reuters data can be used for extra monolingual training if needed.

4.2 Evaluation

The document classification task described in [3] is based on Reuters corpora, which are available in English, French, German, Italian and Spanish. These languages are also in the Europarl data. In this task, we can compare both approaches desribed above, along with the linear mapping baseline and glossing.

For visualizing the vector space, we will project a selection of words onto a plane and highlight semantic relationships. We will also visualize rare words and words with high variability across languages.

References

[1] Karl Moritz Hermann. *Distributed Representations for Compositional Semantics.* PhD thesis, 2014.

[2] Karl Moritz Hermann and Phil Blunsom. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*, 2013.

[3] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing crosslingual distributed representations of words. 2012.

[4] Quoc V. Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. *arXiv preprint arXiv:1405.4053*, 32, May 2014.

[5] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[1] Implemented as part of `https://github.com/piskvorky/gensim`

[2] `https://github.com/karlmoritz/bicvm`