Fancy Title

Benno Kruit
10576223
benno.kruit@student.uva.nl

Sara Veldhoen
10545298
sara.veldhoen@student.uva.nl

Abstract

This is an abstract

1 Introduction

2 Related Work

Some research has focused on the induction of multilingual word embeddings, 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. (Mikolov et al., 2013b), 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 (Klementiev et al., 2012). 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 (Hermann and Blunsom, 2013) 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 (Klementiev et al., 2012) is applied, i.e. the document classification task. Furthermore, the authors present a graphical qualitative analysis. In (Hermann, 2014), this approach is expanded by evaluating on a larger number of language pairs.

3 New approach

One model to induce word embeddings from (monolingual) text is called `word2vec` and was introduced in

Le and Mikolov (Le and Mikolov, 2014) have extended their monolingual `word2vec` model to create representations of word sequences (sentences, paragraphs or documents). It uses a *paragraph vector* as a part of the context of each word in the sequence (Figure 1). This way, the paragraph vector influences the learned representations of those words in the same way that their context words do.

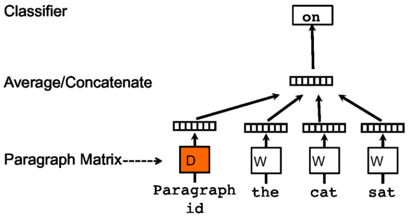

Figure 1: `word2vec` for paragraphs

This paragraph representation could also be used for encouraging similarity between two bitext sentences. In our novel approach, we will run the algorithm from (Le and Mikolov, 2014), but using the same paragraph vector when training word vectors from parallel sentences. The sentence representation therefore acts as a way to relate the word spaces in both languages, without using word alignments. 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.

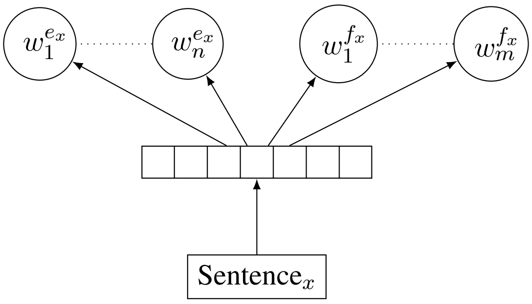

Figure 2: Bilingual dbow

We will explore at least two training methods:

- Sequentially training all sentence pairs. As a paragraph id, we use a single identifier for every sentence pair in the bitext. This is equivalent to concatenating the parallel sentences and training from the context windows that do not bridge the sentence boundary.

- A two-step process: First creating paragraph representations for each sentence pair from a fully trained monolingual model. The information from the words in the first language will create a representation for the sentence. Then, we fix the sentence representations and train the word spaces in each language using these vectors. These sentence vectors will influence the learning of word embeddings in the other languages. The error gradient for the sentence vector can either be distributed over the words or be discarded.

4 Experiments

Using the paragraph2vec dbow architecture, we obtain sentence embeddings for parallel sentences. The model is depicted in figure 2 From the embedding of a single parallel sentence representation, the network tries to predict all words that occur in the sentence either language. Note that no word embeddings are trained, only word indexes are predicted from the sentence embedding. The error is propagated back to train the sentence embeddings.

In order to test the quality of the multilingual sentence embeddings, we take the document representation to simply be the average of its sentence embeddings. These representations are then used to train and test the two document classification tasks.

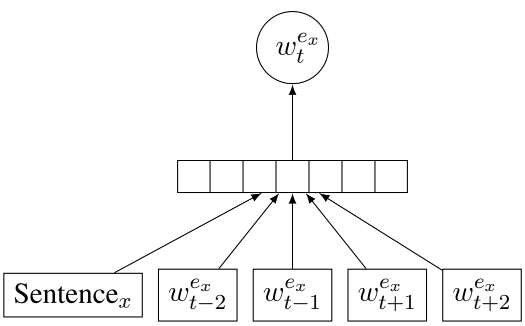

Figure 3: Bilingual distributed memory. The same architecture is trained with English context and word prediction replaced by the other language(s).

We explore how the resulting sentence embeddings can be used to induce word embeddings in two languages. The word embeddings are in the same space and anticipated to be aligned cross-lingually, because the sentence representations for bilingual sentences are equal.

In one setting, we define the word representation as the average of the embeddings of all sentences it occurs in. We evaluate the word embeddings that result from this.

Another training procedure relies on the previous experiments. It is based on the distributed memory training from paragraph2vec and is illustrated in figure 3. The sentence embeddings that resulted from the dbow training were used and kept fixed. The word word embeddings from the previous experiment to initialize the multilingual DM setting. The idea is to further refine the word embeddings using a smaller context. However, the training occurs independently for both languages and the commonality of the semantic space relies solely on the sentence embeddings.

5 Evaluation

It is not trivial to measure the quality of the multilingual word embeddings. The semantic space should be reliable for each language in isolation, and consistent across languages. Even the former is not easy to assess. In (Mikolov et al., 2013a), an analogy task is introduced to this aim.

The latter is evaluated on a real-world task of cross-lingual document classification. The former is evaluated

5.1 Document classification

In (Klementiev et al., 2012), a real-world task is introduced to this end: cross-lingual document classification. The task, that is also used in (Hermann and Blunsom, 2013), is based on Reuters corpora, which has topic-annotated documents. The evaluation data is available for English and German documents that belong to a single topic, and thus the gold standard can be represented by a one-hot vector.

Each document is represented by the average of the representations of its tokens (in (Klementiev et al., 2012), the average is weighted by idf score), or sentences (in (Hermann and Blunsom, 2013)). An averaged version of the perceptron algorithm is trained for document classification in one language, and tested on data in another resulting in a classification accuracy score. If the semantic space is coherent between languages, performance should not be much worse than monolingual document classification.

The WIT TED corpus (Cettolo et al., 2012) contains short documents with transcriptions and translations of TED talks, with topic annotations. The original distribution was aimed at machine translation, but (Hermann and Blunsom, 2014) propose it for a multilingual document classification task. The major advantage of this task is the availability of documents in many languages. It has documents in English paired in both directions with other languages, seven of which are also in the Europarl corpus: Spanish, French, German, Italian, Dutch, Portuguese, Polish, and Romanian.

The classification labels in this set are technology, culture, science, global issues, design, business, entertainment, arts, politics, education, art, health, creativity, economics, and biology. Note that contrary to the previous task, a document can have more than one topic annotation. A binary classifier is thus trained for each topic, using the same system as before. Performance is reported both as classification accuracy and F1 score. As the chance accuracy for majority class is quite high, since there are only few positive examples per class, F1 is more informative for comparing performance.

6 Results and discussion

7 Conclusion

References

[Cettolo et al.2012] Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association* *for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

[Hermann and Blunsom2013] Karl Moritz Hermann and Phil Blunsom. 2013. Multilingual distributed representations without word alignment.

[Hermann and Blunsom2014] Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributional Semantics. In *Proceedings of ACL*.

[Hermann2014] Karl Moritz Hermann. 2014. *Distributed Representations for Compositional Semantics*. Ph.D. thesis.

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

[Le and Mikolov2014] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents.

[Mikolov et al.2013a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Mikolov et al.2013b] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation.