EXPLOITING BILINGUAL WORD EMBEDDINGS TO ESTABLISH TRANSLATIONAL EQUIVALENCE

Anton Serjogin
Centre for Information and Speech Processing, LMU
anton.serjogin@gmail.com

15.05.2017

The main idea is to use Word Embedding for translation of specific bilingual domains without using a dictionary.

Information can be presented in different ways: it can be either an image, an audio or simply a text (word). The first two may represent the information quite richly, whereas the third one does not. Representing words as unique, discrete ids furthermore leads to data sparsity, and usually means that more data is needed in order to successfully train statistical models. By using word embedding, vector representations, such obstacles can be overcome. With Word Embedding each word can be represented as a vector. If the context of one word is similar to the other one, then these words have common semantic and syntactic similarities. Such vectors appear close to each other in a so-called high dimensional vector space. In order to reduce the number of dimensions in this vector space, you can use, for instance, clustering.

A number of implementations are presented, which are Word2Vec (presented by Google in 2013) and FastText (Facebook research). The difference between these two is the use of n-grams. Word2Vec learns vectors only for complete words found in the training corpus. FastText, on the other hand, learns vectors for the n-grams that are found within each word, as well as each complete word. Word2Vec features 2 models: CBOW (continuous bag of word) and Skip-Gram. Skip-gram: works well with small amounts of training data, represents well even rare words or phrases. CBOW: several times faster to train than the skip-gram, slightly better accuracy for the frequent words. A very important thing to remember is that models can produce wrong results, if the initial training set features bad examples. Aside from high dimensional vector models, there are also linear models, which can be used for prediction, such as Linear Regression.

The work has been done on two languages: English and German languages. In total, 4 different corpora were used: General (110M tokens), Medical Big (50M tokens), EMEA (4M tokens) and TED Talks (2M tokens). Next, for each corpus a little parallel corpus with an approximate of 5 000 was defined. For the first 1000 high frequency words that do not have a translation and do not appear in the parallel corpus, a mapping is done through the Regression Linear model.

The following steps remain to be done:

- Search for low-frequency word mapping
- Search for better, high-accuracy mappings
- Applying different types of models
- Evaluation of words that are not located in the corpora