Can word embeddings be used in application of morphosyntactic disambiguation task?

Adrianna Janik

Sages sp. z o.o. (Kodołamacz.pl)
ada.janik@gmail.com

Abstract

This article is an exploration of a model for choosing morphosyntactic disambiguation within given forms. The model is based on an idea of words embeddings combined with the graph theory. The evaluation was made using accuracy score of 85% for the test dataset. The embeddings were constructed by using morphosyntactic forms in place of words. This approach was pursued due to an observation that the idea of embeddings can be generalized for other sequence-like problems or, as in this case, different forms of words. This model is a demonstration of how the morphosyntatic embeddings can be used in a disambiguation task. Although it cannot be used as a stand-alone disambiguer yet, it can probably have a partial application in hybrid solutions. The problem of word embeddings is a relatively new and interesting trend in the deep learning studies. Word2Vec (Mikolov et al., 2013) and the idea of word embeddings originated in the domain of Natural Language Processing. As we can see the idea of words within the context of a sentence or a surrounding word window is universal. It can be applied to any problem dealing with sequences of related data. The starting points are the definitions. The first question is: what do we understand by *vocabulary*? The Mirriam-Webster dictionary defines it as: *'a list or collection of words(. . .)'*. And what about a *word*? According to Mirriam-Webster it is *'a speech sound or series of speech sounds that symbolizes and communicates a meaning usually without being divisible into smaller units capable of independent use'*. Now the whole question is, if we describe a word as a set of morphosyntactic tags representing the traditionally defined word and use this simpler model in the embeddings, what will be the final result? Giving the sequence of morphosyntactic forms as it occurs in a dataset we can extract some information about the syntax of the language. This model as vocabulary instead of raw words takes its morphosyntactic forms e.g.: 'subst:sg:acc:m3' or 'prep:nom' and for each adjacent pair of possible disambiguations in the sentence calculates similarity. Following that reasoning, for each sentence weighted a directed graph of disambiguations is created in such a way that the weight between two disambiguation-nodes is the similarity. As a result, with such a model we can analyze the graph in search for the shortest path between the beginning of the sentence and its end.

Keywords: word embeddings, graph theory, morphosyntactic disambiguation

1. References

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.