Abstract

The impact of inflection on word vectors

Dániel Lévai

supervised by

András Kornai, D.Sc.

Mathematicians and computational linguists have been researching for a long time how to embed words into high-dimensional vector while retaining the most of the meaning and the morphology of the words. In the recent years, there has been a huge improvement in natural language processing (NLP) tasks with the use of neural network based word embeddings, such as negative-sampling skip-gram models like word2vec (Mikolov et al., 2013a; Goldberg and Levy, 2014).

In this thesis, we will be analyzing the impact of the words by inflection. In section 2, we will present necessary definitions and theorems which will be used in the latter sections. In section 3, we will explain how the word vectors are created, and why Mikolov et al., (2013a) has a huge importance in NLP. In section 4, we will demonstrate a method to process the morphological analyses produced by existing tools to be used in the following sections. At the end of the section, the measurements treating vector length and log-frequency in Arora et al., (2015) are verified. In section 5, a general overview is presented of the clusters established by the vectors of words with the same morphological analyses. In section 6, a similarity measure is defined between clusters. In section 7, we will compare the clusters by the affixes, measuring the deviation caused by the affixes.

The main result of this thesis is the defined similarity measure, which is proven useful for comparing word clusters and verifying the coherence of existing clusters while offering a slightly different word clustering. Linguistically, according to the model used, the current grammatical categories of the words are well-founded, with only exception to one category.

2018