Abstract

Helping computers understand the meaning of sentences is one of the most challenging tasks in Natural Language Processing (NLP). According to the principle of compositionality, the meaning of a whole (e.g. a sentence) is a function of its parts and the way they are combined. Sentences are composed of words; thus the first step to create proper representations for sentences is to build appropriate representations for words.

We argue that the type of context and the size of corpus affect the quality of the resulting word embeddings. To verify this, we build three different semantic spaces using three different contexts (bag-of-words, syntactic, and character n-gram) and two corpora of different sizes (British National Corpus and Wikipedia). We evaluate the resulting embeddings in various tasks qualitatively and quantitatively. Then we define three composition models (summation/averaging, concatenation and multiplication) to build sentence representations using the previous word embeddings. These models are evaluated in two different tasks (phrase and sentence similarity) to see which model of word embeddings along with which composition model perform best.

The results suggest that among the three types of context, the syntactic one needs the largest corpus, while the character n-gram context the smallest corpus to produce quality word embeddings, while the bag-of-words is the fastest to build. Moreover, representing a sentence by averaging the vectors of its constituent words yields the best results in phrase and sentence level similarity tasks compared to the other methods of composition (i.e. concatenation and multiplication). Furthermore, the character n-gram creates the best embeddings to represent the sentence's constituent words in the task of paraphrase detection, especially when the size of the corpus is small.