Optimized Training and Evaluation of Arabic Word Embeddings

Jordan King, Lisa Singh, Eric Wang

November 3, 2015

Abstract

Word embeddings are an increasingly important tool for NLP tasks, especially those that require semantic understanding of words. Methodologies and properties of English word embeddings have been extensively researched, however little attention has been given to the production and application of Arabic word embeddings. Arabic is significantly more morphologically complex than English, and due to the many conjugations, suffixes, articles, and other grammar constructs, an Arabic word might not be a single word in the sense that it is in English. This has a significant effect on the training and application of Arabic word embeddings, as the embeddings are trained on unigram tokens. While there are a number of techniques to break down Arabic words through lemmatization and tokenization, the quality of resulting word embeddings must be investigated to understand the effects of these transformations. In this work, we investigate a number of preprocessing methods and training parameterizations to establish best practice strategies for training Arabic word embeddings. To enable this research, we required a semantic similarity task to evaluate the Arabic word embeddings. There is little work done to provide a large semantic similarity task in Arabic so we created a list of 1000 similarity scores for given Arabic word pairs using native Arabic speakers. With this semantic similarity task, we evaluated all training parameterizations by using the embeddings from a given method to obtain a similarity score to evaluate against the task. Additionally, we evaluated the word embeddings' ability to capture syntactic properties of words using a part of speech tagging task. Using these tasks, we were able to identify the training strategies that produce the best results for each task. We also offer a suite of Arabic NLP tools that we developed

alongside this work that attempts to fill the void of accessible open source Arabic NLP tools. Altogether, this work provides best practices for training Arabic word vectors, an open semantic similarity task developed by native Arabic speakers, and a python package of Arabic NLP tools.

1 Introduction

We are researching methods to create Arabic word embeddings, which are numerical representations of a given words meaning - both semantic meaning and syntactic meaning. These embeddings are obtained using machine learning algorithms that utilize the context a word appears in to infer its meaning. This works very well as words with similar meanings tend to be used in similar contexts. For example "I eat bread every night" and "I eat rice every night" is an example of how foods may appear in similar contexts. With enough text to process, we can train numerical vectors to learn that bread and rice appear in these "common-for-food" contexts. Similarly, we can learn syntactic relationships as different parts of speech appear in certain context patterns as well.
The end result of having good word embeddings is that we have a representation of the meaning of a word, without ever translating or looking at a dictionary of meanings. We can harvest the semantic and syntactic meaning straight from a corpus of natural written language. With accurate word embeddings, we can perform neat and useful operations to investigate the relationships between words. To list a few of these operations, we can measure the similarity of two words, identify which word from a set is least similar, and solve basic analogies. The classic party trick for word embeddings is to take (the numerically embeddings of) king, subtract man, and add woman. The resulting embedding is closest to the embedding for queen! Intuitively, this allows us to subtract the male gender meaning from kings embedding, add the female gender meaning, and end up with an embedding equivalent to queens embedding.
More about the algorithms and uses: https://code.google.com/p/word2vec/
We would like accurate Arabic word embeddings so we can interpret the general topics of discussion in Arabic media without using translation or ignoring some words belonging to a topic. Specifically, one task we are using the embeddings for is to learn what words are highly similar to words such as fear (in Arabic), and then compute the degree to which some media is using these words similar to fear in the context of some city. This information may help us predict when people will be forced to migrate from the city.