Optimized Training and Evaluation of Arabic Word Embeddings

Jordan King, Lisa Singh, Eric Wang

November 1, 2015

Abstract

There is a lot of work done on training English word vectors. However, Arabic is much more morphologically complex than English, and due to the many conjugations, suffixes, articles, and more, an Arabic word might not be a single word in the sense that it is in English. However, there isnt a single clear answer as to how these Arabic words can be broken apart. Some options are to tokenize the words by breaking off articles and affixes, to obtain the root word lemma (not the actual Arabic root), or just to leave the words be to preserve the full meaning. We are going to test all of these methods and more, but it doesnt do much good to test them without some evaluation. That is what this task is for. In order to evaluate the word embeddings that we get from some methodology, we would like to have a list of accurate and diverse similarity scores for word pairs. This way, we can use the embeddings from a given method to obtain a similarity score, and compare that against this human truth. The method that performs most similarly to the human truth will be considered the best at semantic similarity. Additionally, we hope to publish this list of word pairs so others can conduct similar research. There is little published that is even similar to the list we hope to create, and we hope this will help others further research Arabic natural language processing.

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.

2 Related Literature

3 Experiments

4 Conclusion