Comp. Ling. (WS 2018-19) - Final Project Idea

Idea: Extracting Synonyms and Antonyms from Embeddings

Objective

We know a word embedding contains lot of data related to that word. The objective here is to list possible synonyms and antonyms of input words using word embeddings.

Proposed Approach

Till date, there have been lots of different approaches to calculate word embeddings. The idea is to use two of the most popular word embedding systems **Word2Vec** (Mikolov et. al. 2013) and **GloVe** (Pennington et. al. 2014) and combine them both to get a concatenated word embedding to capture more information about a word. Next, in order to see how our embeddings look like, we can try to visualize the word embeddings in a multi-dimensional vector space (probably using some visualizing tools like **t-SNE**). Working with multi-dimensional data is really tough and not advisable, so we will try to use **dimensionality reduction** methods (maybe like **PCA**) to bring the data down to a workable dimension before starting to work on it.

Having done this, we can now either use **distance-based similarity functions** (like **Cosine similarity**) or maybe certain **clustering algorithms** (like **KNN**) to find the list of related words for any particular word in the reduced vector space. This list of words probably contains groups of related words (*synonyms, antonyms* and *other domain-specific related words*).

Now that we have a list of such "related" words, our idea is to use **WordNet** (Miller et. al. 1995, Fellbaum et. al. 1998) to figure out the word-sense of the word and possibly be able to find lists of synonyms and antonyms of each word. Then we can classify the previously gruoped words (from the embeddings) into *synonym-antonym-others* classes, hopefully.

Data

For this task, we are planning to use a dataset with a big vocabulary (like the *1-billion words*) or maybe a domain-specific dataset (like *Cornell movie reviews*).

NOTE: This is just an idea. It still needs to be verified if it is be feasible practically or if it is a good idea for the project. Please read it with a grain of salt!