Semantic textual similarity

Bruno Gavranović, Neven Miculinić and Stipan Mikulić

March, 2017

Topic description

Semantic textual similarity (STS) measures how similar two text snippets are. The task is, given two sentences, to determine their similarity score on a continuous scale from 0 to 5. The score of 5 denotes a perfect semantic equivalence. This task was offered in multiple tracks, covering various cross-lingual and monolingual pairs. Your task, however, is the track concerned with English monolingual pairs. Of course, you are encouraged to give a shot on other tracks as well.

Project plan

First and foremost we want to merge together the data from different sources and build vocabulary. We plan to develop simple baseline first and then build upon that. After reading referenced papers we decided that our pipeline will look something like the following:

1. Data preprocessing
    - Tokenization
    - POS tagging
    - Parsing (syntax)
    - ...
2. Transforming data to corresponding input for models
    - N-gram overlap
    - TF-IDF vector
    - Paragraph2Vec
3. Building different models
    - LSTM neural network
    - SVM

- Linear Regression

4. Evaluation

We will make baseline model as centroid vector of word embeddings from sentence.