CS6370: Natural Language Processing
Word similarity Assignment

February 24, 2018

1 Problem Statement

What renders a word its meaning? If it is the string of letters that gives a word its meaning, then why is 'wind' more similar to 'air' than to 'bind'? Word similarity estimation plays a crucial role in many NLP applications like Information Retrieval, Word Sense Disambiguation, Text categorization and Summarization.

In the class, we studied about different approaches to estimating word similarity. Now, it is time to get your hands-on.

- Knowledge-based methods
    1. WordNet based measures
    2. Explicit Semantic Analysis
- Introspective or Corpus based methods
    1. Latent Semantic Analysis
    2. Word2Vec

The Assignment mainly comprises of two parts:

1. WordSim-353 dataset contains a list of 353 word pairs with relevance scores given by human judges. Your task is to implement the above listed approaches and measure the correlation between estimated word similarities and WordSim-353.
2. Analyze and compare the different techniques. For example, why one of your techniques performs better than others? Explain each technique and identify its strengths and weaknesses by analysing the positive and negative results. Are there any word pairs that are difficult (easy) across all techniques?
3. Can you think of an hybrid technique? (optional, for additional credits).

2 Resources

You can use the following resources to implement the techniques:

1. Google n-grams for introspective approaches
   https://pypi.python.org/pypi/google-ngram-downloader
2. WordSim-353
   http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html
3. Word2Vec: Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013)
4. Further updates will be posted on Moodle

3 Evaluation

The evaluation of the assignment will be based on

1. Implementation (30%).
2. Performance (30%)
3. Analysis (40%).

4 Deliverables

1. Code (for all 4 approaches).
2. Well-written report that critically discusses the results and observations.

5 Deadlines

- February 24 - Assignment release date.
- March 12 - Progress meeting with TAs.
- March 20 - Final deadline (code and report).

Plagiarism in any form - report or code - is intolerable. Copying contents verbatim from any paper or even references is strictly not allowed. If you are found to engage in plagiarism, it will be treated very seriously and will result in a U grade (irrespective of other course evaluations). This may also be forwarded to the Dean, Academic Courses for further action. All your project reports and codes will be verified using plagiarism detector tools.