Use Extended Gloss Overlaps to Measure Words Relatedness

Shengxin Qian
ECE 590 Text Analysis

October 12, 2016

1 Structure of Code

1). Extract each pair of words from input file

2). Transform all characters of input words into lowercase, eliminate all punctuation and remove those whitespace at the front and the end. Use the same function clean their corresponding glosses

3). For each pair of input words, use their glosses, their hypernyms' glosses and their hyponyms' glosses to calculate the average semantic similarity.

2 Detail of semantic similarity calculation

1). Calculate the maximum common subsequence of input corresponding glosses of two words. Pronoun, preposition, article or conjunction will not be considered as effective words in subsequence.

2). Use the square of the length of maximum common subsequence as the score of two input glosses.

3). The definition of words, the hypernyms set and the hyponums set of each words were derived from nltk wordnet corpus.

3 Select Input and Output data

1). The input words were derived from the Rubenstein and Goodenough dataset(1965). This dataset contain 65 human labeled nouns.
2). There is a correlation between the rank of human labeled result and that of program's outputs.
3). "res.txt" is the program's output and "RG_word_original.txt" is the human labeled result.