Meaning and Computation: Exercise 1

It is preferable, although not mandatory, to solve this exercise by writing Python scripts to do the computations.

1. **Word sense disambiguation:** Read the paper on the Yarowsky algorithm -

   Yarowsky, D. ``Unsupervised Word Sense Disambiguation Rivaling Supervised Methods'' In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics.* Cambridge, MA, pp. 189-196, 1995.

   a. Pick a word and two of its senses (like the "plant" example we saw in class).
   b. Find two seed collocations for each of the senses. This can be based on common sense.
   c. Run a single iteration of the Yarowsky algorithm on the provided corpus, with context window of size 2 (without smoothing). Write down the list of frequent collocations and their rankings.
      Single iteration means: finding all instances of the word, classifying the known ones according to the seed, ranking them (if seeds for both senses are found, ignore this instance).
      Frequent collocations: you can decide how frequent a collocation has to be to be included.
      Infinity values: Note that it is likely that some of the scores will be +/- infinity.
   d. Select 5 sentences that contain the top collocation from the list for sense A, and 5 sentences for sense B (excluding the seed collocations). In what proportion of the cases was the predicted sense correct?
   e. The original Yarowsky algorithm addresses the case of two senses, by examining the log-likelihood ratio of the two senses for each collocation. Suggest an extension to the algorithm for three-way classification (a word with 3 senses).

2. **Thesaurus-based word similarity:**
   a. Extract from WordNet a sub-tree from the noun taxonomy (i.e., nodes should be nouns, and edges should be defined by hypernymy and hyponymy relations). The tree should contain at least 3 levels (that is, include at least one node that has grandchildren in the sub-tree) and 10 nodes. Write down the distance matrix between the nodes.

Note: it is advisable to use the WordNet search tool (see link below)

b. Compute the similarities between every two words according to Lin's similarity measure. Find five (5) examples of two pairs of words whose similarity measures under the tree-distance matrix makes more/less (choose one) sense to **you** than Lin's measure. Explain your answer.
    - For instance, if "tree" ended up more similar "box" than to "flower", and you find that unreasonable, the two pairs can be ("tree","box"), ("tree","flower"). Find 5 pairs of pairs like this.
    - Take corpus frequency counts from the provided list. Compute frequency counts only based on the sub-tree you've created. That is, do **NOT** compute the information content of a word based on the entire WordNet tree, sum only over descendants within the small sub-tree you created.

c. Ask another person (who is not taking part in the course), whether s/he agrees with your judgments from (b). In what percentage of the cases did you agree (this is called inter-annotator-agreement)?

3. The simplified LESK algorithm for word sense disambiguation operates as follows:
    - Given an ambiguous word *w* in a document, select a context window of size *n* (we'll take *n=2*). Predict *w*'s sense as the sense whose dictionary gloss has the largest overlap of content words with *w*'s context window.
    - Content words are here defined as verbs, nouns, adjectives and adverbs (but not determiners, pronouns, coordinating and subordinating conjunctions, cardinal numbers or prepositions)**.**
    - For instance, if the glosses of the two senses of "dog" are:
        i. A domesticated carnivorous mammal that typically has a long snout, an acute sense of smell, and a barking, howling, or whining voice.
        ii. A person regarded as unpleasant, contemptible, or wicked.

    then an instance of "dog" appearing next to words like "barking", instances of "dog" appearing next to "howling" and "snout" will be classified as indicating sense (i), while those appearing next to words like "wicked" or "unpleasant" will be classified as having sense (ii).

a. Consider the word you have selected in question 1, and the 10 sentences you have found in 1d. How many of them are correctly classified by this algorithm?

b. Discuss this algorithm's advantages and disadvantages relative to the Yarowksy algorithm.

Grades: Grades will be between 1 and 7. The exercise's grade will be 7% of the final grade.

Resources:

1. The Yarowsky paper: www.aclweb.org/anthology/P95-1026
2. WordNet search tool: http://wordnetweb.princeton.edu/perl/webwn
3. Word frequency list: see corpus_ex1.freq_list file in Moodle
4. Corpus for running the disambiguation algorithms: see corpus_ex1 file in Moodle

   Each line corresponds to a word, aside from these special symbols:
   <s>: beginning of sentence
   </s>: end of sentence
   <text id="">: beginning of document
   </text>: end of document