ASSIGNMENT 2

COMP 550, Fall 2020

Due: November 18^{th}, 2020 at 9:00pm.

You must do this assignment individually. You may consult with other students orally, but may not take notes or share code, and you must complete the final submission on your own.

Assignment

Question 1: Word Sense Disambiguation (100 points)

Implement and apply Lesk's algorithm to the publicly available data set of SemEval 2013 Shared Task #12 (Navigli and Jurgens, 2013), using NLTK's interface to WordNet v3.0 as your lexical resource. (Be sure you are using WordNet v3.0!) The relevant files are available on the course website. Starter code is also provided to help you load the data, though you should verify that the code is correct. More information on the data set can be found at `https://www.cs.york.ac.uk/semeval-2013/task12/`.

The provided code will load all of the cases that you are to resolve, along with their sentential context. Apply word tokenization and lemmatization (you have code to do this from A1) as necessary, and remove stop words.

As a first step, compare the following two methods for WSD:

1. The most frequent sense baseline: this is the sense indicated as #1 in the synset according to WordNet
2. NLTK's implementation of Lesk's algorithm (`nltk.wsd.lesk`)

Use accuracy as the evaluation measure. There is sometimes more than one correct sense annotated in the key. If that is the case, you may consider an automatic system correct if it resolves the word to any one of those senses. What do you observe about the results?

Next, develop **two additional methods** to solve this problem. One of them must use the idea of bootstrapping. This may require you to acquire additional texts in English. Since bootstrapping often requires you to specify knowledge about words using heuristics or by specifying a seed set, be sure that your method to start the bootstrapping process covers at least five different lexical items for which you are performing WSD.

The other may be any other method of your design. Feel free to use your creativity to find ways to improve performance, as long as you follow standard practices in experimentation! The two methods must be substantially different; they may not be simply the same method with a different parameter value. Make and justify decisions about any other parameters to the algorithms, such as what exactly to include in the sense and context representations, how to compute overlap, the use of the development set, which the starter code will load for you. You may use any heuristic, probabilistic model, or other statistical method that we have discussed in class.

Some issues and points to watch out for:

- The gold standard key presents solutions using lemma sense keys, which are distinct from the synset numbers that we have seen in class. You will need to convert between them to perform the

evaluation. This webpage `https://wordnet.princeton.edu/man/senseidx.5WN.html` explains what lemma sense keys are.

- The data set contains multi-word phrases, which should be resolved as one entity (e.g., latin_america). Make sure that you are converting between underscores and spaces correctly, and check that you are dealing with upper- vs lower-case appropriately.
- We are using instances with id beginning with d001 as the dev set, and the remaining cases as the test set, for simplicity. This is different from the setting in the original SemEval evaluation, so the results are not directly comparable.

Discuss the results of your experiments with the baselines and models. Also include a discussion of the successes and difficulties faced by the models. Include sample output, some analysis, and suggestions for improvements. The entire report, including the description of your model, must be no longer than **two pages**. Going beyond this length will result in a deduction of marks.

Your grade will depend on whether you adequately followed the guidelines above, whether you followed standard model design and experimental procedure during the development of your method, and on the quality of your report (both linguistic, and content).

What To Submit

Electronically: Submit a .pdf containing the report. For the programming part, you should submit one zip file called 'a2.zip' with your source code to myCourses under Assignment 2. If you use extra raw text, and the size of this extra text is greater than 1MB, submit a short sample instead so that your code will run.