Sun Jae (Jasmine) Lee

Thesis: Milestone #1

Problem Definition:

The task is to determine whether a given three-word noun phrase is left branching or right branching. I plan to adapt a model detailed in Vered Shwartz's paper "Paraphrase to Explicate: Revealing Implicit Noun-Compound Relations" that produces free text paraphrases to explicate relations between two nouns in a noun-compound. I aim to change this model so that it can take a three word noun phrase, explicate the relation between the first word and the rest of the noun phrase as well as between the last word and the rest of the noun phrase, then compare the confidence of the two in order to determine whether the noun phrase is left branching or right branching. The intuition behind this is that if the model comes up with a paraphrase for the relation between the first word and the last two words with higher probability, then it is likely that the last two words are more closely related, so we label the noun phrase as right branching. Similarly, if the model comes up with a paraphrase for the relation between the last word and the first two words with higher probability, then it is likely that the first two words are more closely related, so we label the noun phrase as left branching.

For evaluation, I will use precision, recall, and F-score. I elaborate more on "data.md" in my Milestone 2 folder.

For my data set, I will use a combined set of noun phrases from Penn TreeBank and from OntoNotes 5.0. Though the original Penn TreeBank used flat NPs, the work of Vadas and Curran (2007) annotated the full structure of NPs in the Penn TreeBank's WSJ corpus. With this, I was able to extract 10613 three-word noun phrases with gold labels. I also was able to extract 5419 three-word noun phrases with gold labels from OpenNotes. I will shuffle the Penn TreeBank examples with the OntoNotes examples and of these 16032 examples, I plan to use 80% for training, 10% for validation, and 10% for test.