LING571 – Hw8
WordNet

1. Overview

In this assignment we implemented a thesaurus-based approach to word sense disambiguation and, in particular, used Resnik similarity as the measure of similarity between word senses.

2. Methodology

An application was written in Python using NLTK to accomplish the goals of performing a set of analyses; taking care to not use the forbidden fruit (built-in Resnik similarity function).

In fact, while I did not use the built-in function in my final submission, I did use it for testing; and I was pleased to see that identical results (up the about 12 decimal places) were obtained whether using my internal similarity calculation or the built-in Resnik similarity function.

3. Challenges and Learning Outcomes

Technically, this project wasn't overly challenging. The code is relatively simple, just a bunch of loops, etc. The challenging part was trying to understand what the algorithm is trying to do; and how to adjust it to suit the objectives of the assignment.

There was a very subtle difference in implementation required to answer the two sections of the assignment:

- Processing human judgements
- Word sense disambiguation

Processing human judgements was relatively simple. It took a word w_0 and compared it to another word, say p and tried to identify the shared hypernym with the greatest information content. In other words, what is the most informative "common ground" between all senses of all senses of the word p and w_0. To achieve this, we maintain a record of the weights allocated to the various senses of w_0 considering each sense of p. We then return the maximum value of this collection to indicate the greatest similarity between senses of the two words.

Word sense disambiguation was a little more challenging. In this case, we were looking across a number of probe words $p_1, p_2, \dots, p_n$ to identify the most closely related sense of w_0. This ends up being an import distinction as we need to maintain a record of the weights allocated to the various senses of w_0 across *all* senses of p_i for every $i \in \{1, 2, \dots, n\}$. In this way, we can find the most informative sense of w_0 across all senses the probe words.

4. Results

The results compare relatively well to human judgments (around 72.4% correlation). I think this is a useful indicator but, personally, I don't consider this problem to be sufficiently solved. I think that a shortcoming of this approach is that an obscure sense of a word can be deemed very informative due to how frequency or infrequently the word is used in a particular corpus; however, the connection between the concepts semantically may by tangential. Considering the word "golf" for example, I think the most salient interpretation is of the sport. However, in a corpus that references cars a lot, a Volkswagen Golf (a brand of car) might be given more salience than a speaker of the language might warrant. Having said that, this shortcoming is not specific to this problem but rather to the whole class of problems that assume that a corpora is a representative samples of a language.

5. Insights

The algorithm seemed to fare much worse when used to predict the desired word sense from a series of other probe words. In this experiment, the algorithm was only able to correctly identify the correct sense of w_0 for around 50% of the samples.

I think the reason for this might be because, since we are allocating weight to the word senses of w_0 over all probe words, it is possible for some senses of w_0 to receive unexpectedly high weight due to some skewness in the corpus.

On the other hand, what is the common thread between "line" as in "news" and "line" as in "brainstorm"? The connection seems tentative to me as a native speaker of English. So, we should bear in mind that the algorithm will return the highest calculated sense of w_0; no matter how high (or low) that weight is; whereas people are much more likely to confess that we simply do not see a pattern in the probe words at all.

So, depending on the application of this algorithm it might be helpful to define an *informativeness threshold* below which items are reported as dissimilar.

6. Completeness

I completed the assignment; however, due to time constraints, I was unfortunately not able to attempt the extra-credit this time.