Sneha Shankar Narayan
CS646 — Fall 2014 **Preliminary Report for P3** snehas@cs.umass.edu

Summary of Project

I have decided to take the query expansion with thesaurus route for P3 and am looking at implementing a subset of the ideas in "Combining multiple evidence from different types of Thesaurus for query expansion" by Rila Mandala*, Takenobu Tokunaga, Hozumi Tanaka. [1]

The paper looks at automatic query expansion using thesauri and explains three ways to do it and looks at an approach that can combine the results of the three approaches. I'm trying to implement two of the approaches mentioned in the paper

Approach 1: The similarity between query terms and other words is determined using WordNet - a hand crafted thesaurus developed by Princeton. The paper uses the noun taxonomies to find the similarities between two words.

Approach 2: The paper also describes a head-modifier based thesaurus approach where in the collection is first parsed using the Apple pie parser to detect similar grammatical context and the similarity between two words is calculated depending on this semantic analysis as well.

The word-pair similarity values obtained from these two approaches are averaged and normalized and the similar words to a word given in the query is determined by using a threshold (which in the paper is experimented with 0.1) for the word-pair similarity value and the given query is expanded.

Evaluation: The community generated queries of the robust corpus is going to be used. The queries specified here will be expanded by the methodology mentioned above and sent to Galago. The results of the queries will be evaluated using the provided judgements, and compared against the non-query expanded results.

Current status:

- Finished reading the paper and noted down all the components that have to br set up and all the calculations that need to be made.
- Set up wordnet and understood how it works.
- Set up the apple pie parser and figuring out how it works in order to parse the robust corpus with it.
- Figured out how to use Galago's APIs to get the index dumps in order to calculate the similarity factors as specified in the paper.
- Next step is to write a python script to parse the queries that are input, append it with similar terms that are identified using wordnet and the head modifier thesaurus, and then send the query to galago. Using the same script as in the last step of P2, I can calculate the stats (like MAP) of this result and compare it with the existing result I have from P2.

References

[1] http://www.iro.umontreal.ca/~nie/IFT6255/mandala99combining.pdf