

Homework #7 (due on last day of class – 5/7/09)

Natural Language (100=15+15+35+35 points)

This assignment focuses on how various NLP technologies might be used to improve information retrieval.

Text Normalization (15 points)

The IR and NLP research communities have focused very little effort on pragmatic issues such as sentence boundary detection, non-sentence detection (fragments, titles, outline structure, etc.), abbreviation and acronym detection, restoration of missing diacritical marks (e.g., Schutze vs. Schütze), dialect standardization (e.g., color/colour), and punctuation detection and removal. (Spelling correction is a notable exception.) Qualitatively estimate the significance of not performing some document normalization on IR performance. Suggestion: consider how different results would (or would not) be depending on whether (a) no normalization is done vs. (b) perfect (human-quality) normalization is performed. For this question you should ignore stopword removal or stemming, which could be considered forms of normalization, but which we have covered in depth in the class.

Part-of-Speech (POS) Tagging (15 points)

Suppose you had a perfect part-of-speech tagger – one that could correctly determine the appropriate grammatical class of each word in a document or query. Argue whether this capability could be used to effectively enhance IR performance; explain your reasoning, and give examples if helpful.

Using WordNet (35 points)

Use the on-line version of WordNet (at <http://wordnet.princeton.edu/>) (hint: click on ‘use online’). Specifically look up the words {set, read, and blue}, {crucible, pizza, and hegemony}, {sprite, sprint, and dell}, and {photography, publish, and island}. Look up any other words you are interested in. From these observations, what conclusions can you draw about the utility of dictionary-based word-sense disambiguation for the purposes of IR? Clearly explain your observations and reasoning.

Retrieving with Good Sense (35 points)

Read Mark Sanderson’s paper ‘Retrieving with Good Sense’ (the paper is on the course website). In a few sentences describe Sanderson’s kalishnikov/banana experiment. Now explain what the goal of the experiment was and what was learned.

Sanderson gives an excellent survey of work in this field (WSD applied to IR). The most significant large-scale success he cites is work by Schutze and Pedersen. As far as I am aware nobody has reported success reproducing their positive results. Explain in some detail why this might be (e.g., give reasons why their result is not generally true, or why it is, but nobody else has duplicated their work). Feel free to cite strong counter-examples if you are aware of some work refuting my claim that their positive results have not be reproduced by others.