

Homework #7 (due on – 5/2/11)

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

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

Perfect Anaphora Resolution (15 points)

Deciding which named entities correspond to pronouns in a document is called co-reference resolution, or more specifically, anaphora resolution. Suppose that you had an NLP tool that could indicate for every pronoun, which proper noun(s) mentioned in the document are being referred to, without making any errors. Argue whether with such an ideal anaphora resolution tool, you could use it to improve retrieval performance.

Example: "A boozed-up Nicolas Cage was arrested early Saturday in New Orleans after he loudly argued with his wife — and then taunted cops, according to police and published reports." -> he: "Nicolas Cage", his: "Nicolas Cage".

Perfect 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 (i.e., whether a word is a noun, verb, adjective, etc...). 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 WordNet online’). Specifically look up the words {set, read, and blue}, {crucible, pizza, and hegemony}, {sprite, sprint, and dell}, and {photography, publish, and island}. Each set of words has some property that is indicative of whether WordNet might or might not be useful for improving information retrieval system performance by automatically disambiguating words. For each set of words, explain your observations from WordNet and based on 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 available 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 (word sense disambiguation applied to IR). The most significant large-scale success he cites is work by Schütze and Pedersen. As far as I am aware no other researchers have 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 replicated 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 been reproduced by others.