

Homework #7 (due on – 5/1/17)

Natural Language (100=15+40+25+20 points)

This assignment focuses on how various NLP technologies might be used to improve information retrieval. There is no programming component for this assignment.

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 "ideal" 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 a loud argument there with his wife — and then taunted cops, according to police and published reports." -> he: "Nicolas Cage", his: "Nicolas Cage"; there: New Orleans.

Exploring WordNet (40 points)

Use the on-line version of WordNet (at <http://wordnet.princeton.edu/>) (hint: click on 'use WordNet online'). Specifically look up the following sets of words: {set, read, and blue}, {crucible, pizza, and hegemony}, {sprite, sprint, and dell}, and {photography, publish, and island}. Each of these sets of words has a 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 of these four sets of words, characterize your observations of the set after looking up each term using WordNet, and explain what property they share that reflects on whether dictionary-based word sense disambiguation is likely to improve retrieval performance. Clearly explain your reasoning.

Retrieving with Good Sense (25 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. Argue in some detail about 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 counter-examples if you are aware of some work refuting my claim that their positive results have not been reproduced by others.

Web IR for Question Answering (20 points)

Come up with four or five questions that you believe will include a couple that are very easy to answer from the Internet and a couple that are not so easy to answer. (Examples: "How tall is the Eiffel Tower" vs. "What are the risks of starting a 529 plan before having any children?" or "Should I invest my life savings in Tesla stock?"). Submit each question to one or more of your favorite Web search engines (e.g., Bing, Google, Yahoo, or similar) and examine the top page of results. Describe which engine (or engines) you used, and how good the results are. Report any surprising results, impressive successes, or catastrophic failures.