SILS Biomedical Informatics Journal Club
INLS 279
Fall 2003

Date: September 2, 2003
Presenter: John MacMullen
Topic: Terminology problems in literature mining
Slides: http://ils.unc.edu/bioinfo/docs/20030902-MacMullen.ppt
Reading(s):

- Nenadic, G., Spasic, I., & Ananiadou, S. (2003). Terminology-driven mining of biomedical literature. Bioinformatics 19(8), 938-943. http://bioinformatics.oupjournals.org/cgi/reprint/19/8/938

Summary:

This paper presents a method for automatic term recognition and clustering in biomedical documents.

Discussion:

The main assumptions of the paper are that:

1. "knowledge encoded in textual documents is organized around sets of domain-specific terms, which are used as a basis for sophisticated knowledge acquisition." [938] *We discussed the ambiguity of terms such as: knowledge, encoded, sets, domain-specific, terms [vs concepts], sophisticated, and knowledge acquisition. This is a good illustration of the problems faced by NLP. We could disprove this assumption by demonstrating that knowledge is not encoded in domain-specific terms.*

2. "Terms represent the most important concepts in a domain and characterize documents semantically." [939] *We discussed whether this is true, because it is a huge assumption with no supporting evidence, and is controversial in the indexing and IR worlds. Is there always a 1-1 mapping between a term and a concept?*

3. "the basic problem is to recognize domain-specific concepts and to extract instances of specific relationships among them." [938] *Actually this is 3 problems: recognition, extraction, and association.*

We looked at an example to understand some of the problems in NLP associated with syntactics and semantics of automatic term recognition:

John gave Brad his JASIST article. He said it wouldn't be too hard to read.

Software has problems interpreting the meaning of this sentence for a number of reasons:

- *JASIST* is an acronym
- *his*, *he*, and *it* are ambiguous referents, illustrating the larger problem of anaphora. How to resolve objects and referents, particularly as the distance (in terms of the number of words) separating them grows is a huge challenge in NLP.
- *article* has homonyms that can create problems in non-domain-specific instances

This paper addresses one particular problem of NLP in general and biomedical literature mining specifically: how do we get software to recognize terms and terminology as distinct from "regular" words. (Terminology being a class of language where specialized terms and general terms with contextual meaning are combined for use in specific domains, e.g., biology.)

The workflow and system components described in the article:

- They begin with a specially-constructed corpus of 2,082 MEDLINE abstracts related to 'nuclear receptors', but they do not say how they created the corpus
- "Nuclear" is another good example of variation in sub-domain terminologies
- They are weak on how the termhood values are calculated from the 4 characteristics
- Their clustering approach uses contextual, syntactical, and lexical similarities between terms.
- Clustering perspectives: you need to have certain attributes or perspectives in mind in order to form clusters: age, height, weight, etc.

Other questions we raised in the discussion included:

- "a larger corpus does not have a proportionally higher number of acronyms" [942] True? How might we test this?
- "All term variants are considered jointly for the calculation of termhood" [942] What would happen if they weren't?
- In what ways is the hybrid similarity measure corpus dependent? [942]

Despite my criticisms, this paper is a good model of how to do a literature review:

1. state assumptions
2. state problem
3. state deficiencies of current approaches
4. illustrate current approach
5. illustrate current approach variants
6. describe proposed approach