SILS Biomedical Informatics Journal Club
INLS 279
Fall 2003

Date: October 7, 2003
Presenter: John MacMullen
Topic: Representation of biomedical sublanguages using symbolic notation
Slides: http://ils.unc.edu/bioinfo/docs/20031007-MacMullen.ppt
Reading(s):

- Harris, Zellig S. (2002). The structure of science information. Journal of Biomedical Informatics, 35, 215-221.

Summary:

This paper proposed a method for representing biomedical sublanguages using symbolic notation. The research goals the author was trying to address included how to structure scientific language to allow mathematical operations to be performed on it.

Discussion:

The paper presented some interesting variations on recurring themes in our discussions. We've been talking about ways to structure language – either *a priori* or after the fact – to allow better conceptual connections to be made between knowledge, and this paper can be seen as falling into that category as well.

This paper provided an interesting contrast to Shannon's quantitative model of information; Harris argues that his model provides "an information-theoretic approach to the *structure* of information, as against solely the *amount* of information" [217, emphasis added]. This is a very different approach from the Nenadic, Spasic, & Ananiadou paper, which was focused on advancing the state of the art of automatic term recognition.

One of the core premises of the paper is that we can probabilistically identify regular features of specialized language due to the natural co-occurrence of terms. Harris argues that meaning is not wholly intrinsic to terms, but is in part created by a term's higher than average co-occurrence with certain other terms.

This seems to essentially be an extension of Zipf's Law. Zipf's Law says that "the frequency of word occurrence is approximately inverse to the rank of that word by its number of occurances." In other words, a few terms (like 'the', 'a', 'it', etc.) occur very frequently, while most others occur rarely, so the frequency distribution of words in a text has a distinctive shape. Zipf, like Shannon, doesn't talk about content or meaning, but Harris does. ("Words with highest probability in respect to another word, or which otherwise can be shown structurally to have highest expectancy, add little or no information.") The significance of this is that Harris says we can omit the terms with little meaning in order to reduce the complexity of language.

Many questions were raised about Harris' symbolic notation, and the example where a text fragment was transformed from natural language into an intermediate marked-up version, and then into the compact syntactic and semantic notation.

- What are the sub-classes called? How were they arrived at? How extensible are they? How do we denote specific individual members of the set? (e.g., G-sub-001 = antigen)
- Would labeling these types result in a basic ontology framework? (eg G = independent molecular entity)
- Are individual letters appropriate to use in concept representation? Or should something more extensible be used? If we used numbers, for ex., could we eventually build a computable ontology? Would you essentially end up recreating MESH?
- This is a fairly straightforward approach to notation: use symbols to represent entities and relations. What's not clear is the transformation process when applying this to a large vocabulary and transforming real text. How do statements get encoded into 'Harrisyntax'? Is it done manually? He does not provide an algorithm for how the transformations were performed.
- Does this approach fall into the circularity problem Harris identified at the outset, where you can only define the language in terms of itself?
- Concept of "wellformedness" in natural language (vs math, computer languages, markup, etc.)
- Ability to move from a implies b to a_i-is-a-type-of-a and all a's have the set of properties p; the properties p include a certain action when b is present
- Wouldn't this approach result in a second language that's the same size as the first?

Other questions raised during the discussion included:

- Do Harris' arguments hold in an interdisciplinary environment (either a purposeful one, or a constructed one; i.e., when we do data/literature integration across subdisciplines which have different sublanguages)?
- He seems to be saying [219] that this approach is effective across languages; essentially a way to normalize content across languages – an "independent symbolic linguistic system". True?? What about, e.g., word order problems? [Italian / English example]
- Although this focuses on structuration, it is a very different approach from the ScholOnto project is doing. Are the two compatible?
- How might this concept be implemented in practice? (This is both an algorithm question and a policy question.)
- Who would apply it? (i.e., scientists/authors, indexers, librarians, etc.)
- What would some of the barriers to implementation be?