

15 February 2005
INLS 279 Bioinformatics Seminar
Patrick Herron
SILS, UNC Chapel Hill

Distilling Conceptual Connections from MeSH Co-Occurrences

Padmini Srinivasan, Dimitar Hristovski, originally presented by authors @ MEDINFO 2004

The goal of Srinivasan et al in distilling conceptual connections from MeSH co-occurrences is to analyze MeSH heading/subheading pair co-occurrences (MHSPCs), *e.g.*,

(diabetes / drug therapy) with (chemical / therapeutic use)
heading/concept subheading heading/concept subheading

The authors aim to perform such an extraction and analysis ultimately because they believe such pairs may aid in generating or revealing latent logical and conceptual connections across disciplines across all of MEDLINE—a way significantly more complex than the classic term-based ABC model of Swanson & Smalheiser. The standard of interestingness set forth in the current approach is the discovery of semantically meaningful MHSPCs, determined via chi-sq, that are relatively *domain independent*. Specifically, Srinivasan et al labored to automatically identify/extract key verbs associated with each MHSPC, and will use these verbs in a later study for the purpose of extracting highly related Ns and NPs. The grand vision of the authors is the automated construction of a knowledge base based on concept-based extraction schemes.

Most approaches to key verb extraction are highly domain dependent and expert-dependent as meaningful verbs would be manually picked in advance of finding co-occurrence of such meaningful verbs with these conceptual MHSPCs.

The foremost struggle with the paper noted in the class discussion was with the complexity of the approach. Given the complexity in combination with the brevity of the paper, it was only inevitable that some elements of Srinivasan's approach would remain unclear even after repeated careful readings of the paper. For example, it was not clear why MeSH terms used in the study were limited to strict heading/subheading pairs, or whether triples would be rendered into a set of three pairs, and so forth. The rationale for sticking to heading/subheading pairs was not altogether clear, though there was some sense that the modifying subheading would add some specificity (thereby arguably defeating some of the domain-independence intentions). The construction of verb profiles, particularly the formation of the verb profile vector, was not clear to the readers present for the presentation. Were profiles generated for each verb at the document level, or only at the corpus level? I surmised that the profiles were generated at the document level in all cases, but I found no agreement from the others present. Another important question arose, a likely victim of the paper's brevity: how did Srinivasan et al decide only 1 mil of the 60 mil possible MHSPCs were meaningful? In combination with the concern as to how the MeSH usage was limited to heading/subheading pairs, the latter question inevitably leads to some doubts as to the paper's use of randomness in its analysis. Finally, everyone was quite curious as to the role this verb-MHSPC collocation analysis task would have in the authors' greater research scheme. The particular aim is clear: the authors claim these verb profiles will provide useful constraints for extracting pair-associated nouns; which essentially means verbs will help pinpoint the nouns highly associated with the cross-domain MeSH-based conceptual connections.