Information-Theoretic Classification of SNOMED Improves the Organization of Context-Sensitive Excerpts from Cochrane Reviews

Lee Sam, MA[1], Tara Borlawsky, MA[1], Ying Tao[1], PhD, Jianrong Li[1], MS, Carol Friedman, PhD[2], Barry Smith, PhD[3], Yves A Lussier, MD[1]

[1]Ctr. for Biomed. Informatics Univ. of Chicago, IL;[2]Dept. of Biomed. Informatics, Columbia Univ.,NY, NY;[3]Univ. of Buffalo, Buffalo, NY;[4]OSU Med. Ctr. Info. Warehouse, Columbus, OH

Abstract. *The emphasis on evidence based medicine (**EBM**) has placed increased focus on finding timely answers to clinical questions in presence of patients. Using a combination of natural language processing for the generation of clinical excerpts and information theoretic distance based clustering, we evaluated multiple approaches for the efficient presentation of context-sensitive EBM excerpts.*

Motivation and Hypothesis. The emerging application of EBM to healthcare has brought with it a number of problems rooted in the need to access sources of information. However when physicians search the literature, they are often time constrained and are looking for concise, actionable information (*1, 2*). This study focuses on the problem of efficiently classifying context-sensitive excerpts (summaries of EBM knowledge) from complex clinical queries that include clinical context (e.g. Rx, Dx, past and familial history). **We hypothesized that ontological annotations over the excepts would provide the ability to reliably classify relevant excepts** with the knowledge that Hearst and others have established that the use of semantic organization can be useful when presenting multiple, orthogonal views of data (*3*). With this in mind, we systematically researched naïve and sophisticated approaches to semantically classifying the resulting list (e.g. simple table, native SNOMED classification, Information-theoretic approaches).

Background. In many cases, answers to clinical questions can be found, but are simply too time consuming or otherwise expensive to retrieve (*4*), making the specialized visualization of evidence for a clinical care setting a critical issue. Ely et al. developed a taxonomy of generic clinical questions (*5*) which we employed as a guide in the development of a generic presentation system for EBM. Using the Cochrane Systematic Reviews as a base corpus, we investigated a number of approaches for efficiently presenting results of clinically-oriented queries using 3,794 Cochrane reviews which generated 27,610 Knowledge Objects that we have previously annotated with the BioMedLEE Natural Language Processing system (*6*) and coded in 32,232 and 8,053 UMLS-coded diagnoses and drugs, respectively.

Methods. We implemented two naïve clustering methods (**NCMs**) based on the structure of SNOMED, a method based on clustering by the direct ancestor and a variant allowing for organization by arbitrary ancestor. Additionally, we researched and developed the variable granularity information theoretic approach (**VGIT**).VGIT uses Jiang's information theoretic distance measure (*7*) to determine semantic similarity between nodes in the SNOMED tree. We compared these methods according to two metrics (i) a **q**ualitative **e**xamination (**QE**) of the semantic parity between organizing classes, and (ii) a metric of **c**lustering **p**erformance (**CP**) that takes into account the breadth and depth of displayed information and is directly applicable to compare trees and lists.

Results. The VGIT method showed better performance with regard to the clustering metric due to a more optimal ratio of excerpts clustered in comparison to the number of resultant clusters. For example, in a benchmark query of "diabetes mellitus", chosen for its wide breadth of results, VGIT was significantly better in both QE and CP than the two naïve clustering methods in a wide range or results, however no differences were found in extreme ranges where clusters were defined by highly specific (non clustered = list with no parent) of highly general terms (one class with all terms under it as a list). *Limitations and Future work:* The study could be improved by systematically bootstrapping queries and verifying orders of magnitude more potential query results. We intend to conduct such a systematic study in the future.

Conclusion. Evaluating the classification using a cluster comparison approach, we concluded that the VGIT approach was optimal among the compared organizing methods in presenting results for rapid evaluation in comparison to using the hierarchical structure of the SNOMED terminology according to our clustering metric.

References

1.J. W. Ely, et al., *JAMIA* **12**, 217 (2005).
2.M. L. Thompson, *Bull Med Libr Assoc* **85**, 187 (1997).
3.M. Hearst *et al.*, *Comm of the ACM* **45**, 42 (2002).
4.R. Smith, *Bmj* **313**, 1062 (Oct 26, 1996).
5.J. W. Ely *et al.*, *Bmj* **321**, 429 (Aug 12, 2000).
6.T. Borlawsky, C. Friedman, Y. A. Lussier, *AMIA*, 56 (2006).
7.J. J. Jiang, Conrath, David W., (ROCLING), Tapei, Taiwan 1997.