Towards (Semi-)automatic Generation of Bio-medical ontologies

Vipul Kashyap, Ph.D.[1], Cartic Ramakrishnan, M.S.[2], Thomas C Rindflesch, Ph.D.[1]
[1]National Library of Medicine, Bethesda, MD
[2]LSDIS Lab, Department of Computer Science, UGA, Athens, GA
kashyap@nlm.nih.gov

Introduction

The design and construction of domain specific ontologies and taxonomies requires allocation of huge resources in terms of cost and time. These efforts are human intensive and we need to explore ways of minimizing human involvement and other resources. In the biomedical domain, we seek to leverage resources such as the UMLS®[1] Metathesaurus and NLP-based applications such as MetaMap[2] in conjunction with statistical clustering techniques, to (partially) automate the process. This is expected to be useful to the team involved in developing MeSH and other biomedical taxonomies to identify gaps in the existing taxonomies, and to be able to quickly bootstrap taxonomy generation for new research areas in biomedical informatics.

Approach

Our approach is illustrated in **Figure 1**, and has the following major components:

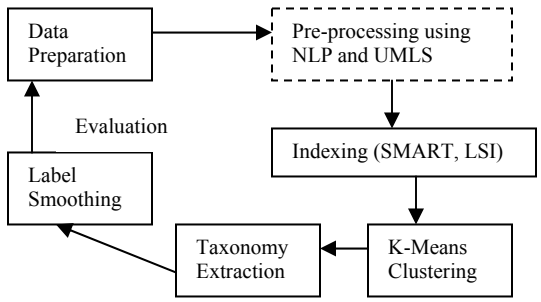
Data Set Preparation: We chose a subset of the MeSH[4] hierarchy and those abstracts from PubMed that are annotated with descriptors from the chosen subset of the MeSH hierarchy.

Preprocessing: POS tagging, followed by extraction of nouns, adjectives (words and phrases) will be applied to the abstracts.

Indexing: The collection is indexed using SMART[5] (or LSI), generating document vectors.

Clustering: Initially the whole document set is split into two clusters using K means clustering[6]. Iteratively, a cluster is chosen based on some criteria (e.g., size, intra-cluster variance) and split into two clusters, inducing a new level at each iteration. The intra-cluster variance of each cluster and its centroid is computed.

Figure 1: Approach for Taxonomy Generation



Taxonomy Extraction: Based on the intra-cluster variance, the taxonomical structure is extracted from the clusters. The centroid vectors of the clusters are mapped to the most significant terms in the document vector to get an initial labeling.

Label Smoothing: Simple NLP techniques like root form analysis and combining adjectives and nouns into phrases are adopted to come up with meaningful labels.

Evaluation

Since we created the data set based on a subset of the MeSH hierarchy, we shall compare the generated taxonomy with the MeSH hierarchy. Evaluation functions based on graph isomorphism techniques shall be designed. We also plan to have our taxonomies evaluated by domain experts.

Discussion

The goal of the Taxonomy Generation project is to design, experiment with and test algorithms for semi-automatic generation of bio-medical taxonomies. Statistical techniques are suitable for unsupervised content-based clustering of documents. NLP techniques such as POS tagging, adjectival modifications and other appropriate approaches might be needed to label the clusters properly. However, most NLP techniques (except tagging and underspecified parsing), though precise are not adaptable across multiple information domains. We propose to develop hybrid algorithms that will combine the best and appropriate features for both the types of technologies. Initially we plan to use domain independent approaches and evaluate the results before we move ahead with domain specific approaches such as using the UMLS Metathesaurus and the SPECIALIST lexicon and tools[3].

References

1. Lindberg D, Humphreys B, McCray A. The Unified Medical Language System. Methods Inf Med 1993:32(4):281-91.
2. Aronson A. R.., Rindflesch T C, Query Expansion using the UMLS Metathesaurus. Proceedings of the AMIA Annual Fall Symposium, 1997:485-9.
3. McCray A, Srinivasan S, Browne A. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care 1994:235-9.
4. MeSH. Medical Subject Headings. Bethesda (MD): National Library of Medicine, 2003.
5. Salton G, Editor. The SMART Retrieval System – Experiments in Automatic Document Retrieval. Prentice Hall Inc., Englewood Cliffs, NJ 1971.
6. Zhang Y, Karypis G. Criterion functions for Document Clustering, Technical Report, U. Minn. Computer Science, #TR-01-40.