UMLS Content Views Appropriate for NLP Processing of the Biomedical Literature vs. Clinical Text

Dina Demner-Fushman, MD, PhD, James G. Mork, MSc,
Sonya E. Shooshan, MLS, and Alan R. Aronson, PhD

U.S. National Library of Medicine, Bethesda, MD 20894

Abstract

Identification of medical terms in free text is a first step in such Natural Language Processing (NLP) tasks as automatic indexing of biomedical literature and extraction of patients' problem lists from the text of clinical notes. Many tools developed to perform these tasks use biomedical knowledge encoded in the Unified Medical Language System (UMLS) Metathesaurus. Continuing exploration of automatic approaches to creation of subsets (UMLS content views) which can support NLP processing of both the biomedical literature and clinical text, we found suppression of highly ambiguous terms in the conservative AutoFilter data view can replace manual filtering for literature applications and suppression of two character mappings in the same data view achieves acceptable performance in clinical applications.

Background

As the Unified Medical Language System® (UMLS®) Metathesaurus® has grown, the task of effectively using its knowledge has become more challenging. We continued exploring the UMLS content views customized for automatic indexing of biomedical publications (the NLM Medical Text Indexer, MTI) and entity extraction from clinical narrative for automatic clinical question answering (CQA) [1].

We developed three approaches – *conservative*, *moderate*, and *aggressive* – designed to systematically remove more and more Metathesaurus strings. The *conservative* approach deleted some short strings that might contribute to the overall ambiguity. The *moderate* approach removed specific source vocabularies that potentially introduce ambiguous and/or incomplete concept senses. The *aggressive* approach removed strings based on their degree of ambiguity.

We applied the string removal approaches to the three best views identified in the previous study – AutoFilter, AllFilter, and Minimal (formerly called *Aggressive*). The modified views were evaluated using two datasets: a randomly chosen subset of 10,000 MEDLINE citations indexed in 2007, and 356 sentences randomly selected from de-identified discharge summaries [2]. (See Figure 1)

Data views

- AutoFilter: MetaMap's strict model without the manual ambiguity filtering
- AllFilter: MetaMap's strict model with the manual ambiguity filtering
- Minimal: All Metathesaurus strings that are a proper substring of another string in the same concept are removed respecting word boundaries

Data view modifications

- C*onservative*: removal of UMLS concepts of 2 characters, 3 characters, and 3 character consonants
- *Moderate* removal of HL7, RXNORM, LNC, and all three combined
- *Aggressive*: removal of 2+ through 10+ ambiguity

Document collections

- 2008 LNCV document collection: 10,000 MEDLINE citations
- Clinical text collection: 356 random de-identified discharge sentences obtained from the Laboratory for Computational Physiology, Massachusetts Institute of Technology

Figure 1 - Data views, modifications applied to each view, and document collections

Results and Discussion

The best MTI results (31% precision at 54% recall) were achieved using the *aggressive modification* of the AutoFilter view (7+ ambiguity suppression). This approach can automate some of the ambiguity study that we now do manually. The conservative approach (suppression of two character mappings) in the same data view achieves 89.5% precision at 78.6% recall for clinical applications.

We were able to construct fully automatic content views that perform at least as well as manually constructed views. Our experiments suggest, however, that content views need to be constructed for each specific task and sub-language (text type).

References

[1] Aronson AR et al. Methodology for Creating UMLS Content Views Appropriate for Biomedical Natural Language Processing. *Proc AMIA Symp*. 2008;:21-5.

[2] Saeed M et al. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Comput Cardiol. 2002;29:641-4.