A New Tool to Identify Key Biomedical Concepts in Text Documents, with Special Application to Curriculum Content

Joshua C. Denny, Jeffrey D. Smithers, Anderson Spickard, III, M.D., M.S., Randolph A. Miller, M.D.
Department of Biomedical Informatics
Vanderbilt University Medical Center, Nashville, TN

The Vanderbilt KnowledgeMap (KM) project created a series of tools for conceptual mapping and analysis of medical text documents, using a novel approach based on both the NLM UMLS Metathesaurus and heuristic, approximate NLP techniques. The authors compared the performance of the KM concept matcher with the National Library of Medicine's MetaMap (MM) using a selected subset of curriculum documents (for which domain experts had identified key concepts manually). The two systems identified key medical concepts with a recall of 86% (KM) and 81% (MM). The precision was 92% for KM and 89% for MM.

Introduction

Natural Language Processing technique (NLP) and the Unified Medical Language System (UMLS) Metathesaurus have been applied to identify and extract "key" concepts from a broad range of biomedical text. "Understanding" medical curriculum content represents a difficult challenge for developers. Curricular documents come in many formats: full-text transcriptions; detailed, textual outlines; extracts from slide presentations; and, text that is broken into quasi-arbitrary heading and subheading markers. We describe the preliminary evaluation at the Vanderbilt School of Medicine of the KnowledgeMap (KM) system to identify medical concepts in curriculum content.

Methods

The KM system uses lexical tools developed from SPECIALIST for the normalization and processing of both documents and the UMLS Metathesaurus. The KM document parser identifies sentences and extracts outline information. Using the QuickTag (Cogilex, R&D, Inc.) part-of-speech tagger, the KM system identifies noun phrases. KM uses MRSTY semantic type information to mediate approximate NLP techniques. The system also incorporates an integrated acronym extractor and uses disambiguation techniques at both the document-level and phrase-level.

The authors evaluated the ability of KM and NLM's MetaMap to identify "important" concepts in a selected subset of documents from the first two years of medical school. The authors identified all medical concepts in five lecture documents to serve as the gold standard after establishing consistency (Kappa 0.75) in marking meaningful terms with content-experts for the lectures. A meaningful term was defined as either a meaningful word that describes a medical concept or a meaningful phrase that when reduced to the word level loses its meaning. Examples include the terms "heart," "lung," "disease" or the phrases "congestive heart failure," "heart lung machine," and "Stevens Johnson Syndrome."

The KM document parser processed the five gold standard documents into sentences. Each sentence was then submitted individually to both the KM concept indexer and to MetaMap. Two authors identified the meaningful terms in the gold standard, and, blinded to the identity of the concept indexers, determined from their output the number of meaningful terms "matched" (to calculate recall) and the number of incorrectly identified concepts not in the gold standard document set (to calculate precision).

Results

The five documents contained a total of 1954 meaningful terms. Results are shown in Figure 1.

Discussion

The authors are encouraged by the level of performance of the KnowledgeMap system, still in the early stages of its development. Use of the KM system to identify concepts within documents and between documents shows significant potential to help educators to locate, integrate, evaluate, and iteratively improve medical school curriculum content.

Indexer	Recall	False Positives	Precision
MetaMap	**1580 (81%)**	**195**	**89%**
KnowledgeMap	**1677 (86%)**	**146**	**92%**

Figure 1: Results from five test documents containing a total of 1954 meaningful terms