

# 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, MD, MS - Randolph A. Miller, MD

DEPARTMENT OF BIOMEDICAL INFORMATICS, VANDERBILT UNIVERSITY MEDICAL CENTER

## INTRODUCTION

Natural Language Processing techniques (NLP) and the Unified Medical Language System (UMLS) Metathesaurus have been applied to identify and extract medical 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 evaluation at the Vanderbilt School of Medicine of the KnowledgeMap (KM) system to identify medical concepts in curriculum content. The ultimate goal of the KM project is to provide a set of tools that improve the development and integration of medical curriculum.

## METHODS

### KM Concept Indexer

The KM system was designed to identify biomedical concepts in medical school curricular documents using the UMLS Metathesaurus. KM normalizes Metathesaurus concept names and document text using lexical tools developed from UMLS' SPECIALIST Lexicon. We developed a sentence parser to extract sentences from many document formats, including outlines. KM then identifies noun phrases using a part of speech tagger from Cogilex, R&D, Inc. Unique features of KM's algorithm include its ability to recognize concepts across prepositions and conjunctions using approximate NLP techniques (Figure 1) along with an acronym extractor that uses document-specific definitions for concept matching. KM also generates word variants for non-matching words using author-derived methods.

Since many concepts are ambiguous, KM attempts to resolve ambiguity using context- and document- level techniques. Examples of document-level disambiguation include favoring candidate concepts that are previously-seen in the document and those that co-occur with other document concepts in MEDLINE abstracts.

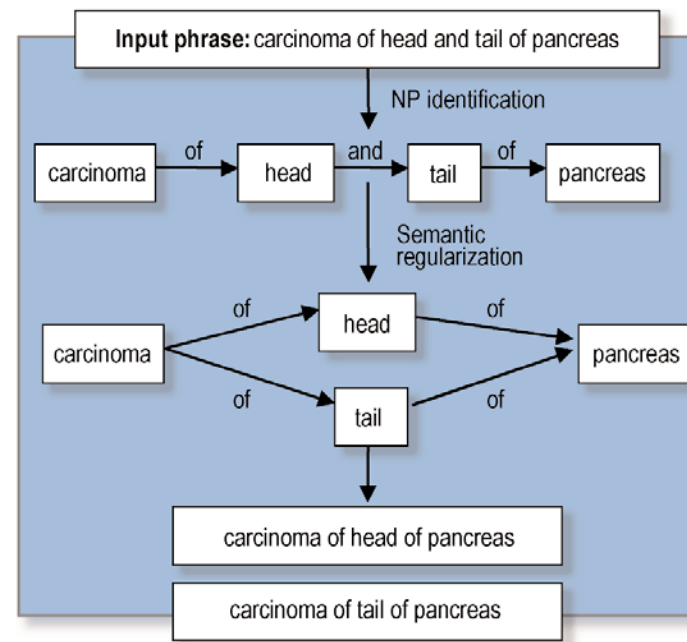


Figure 1. Example NLP processing

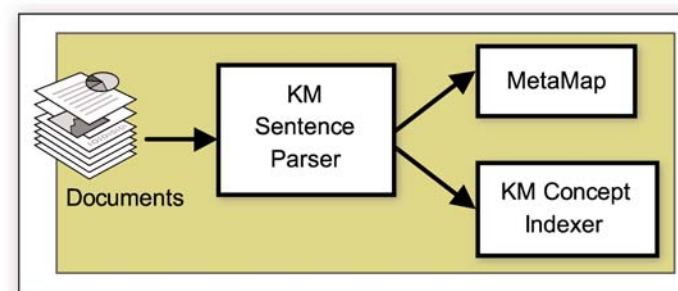


Figure 2.. Study Design

### Document Markup and Comparison

We 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 pilot study consisted of five documents. After some adjustments in our algorithm, we then studied ten documents equally chosen from both preclinical years (termed the "definitive" study). In each study, we:

- Defined a "meaningful term" (MT) as either a word or phrase that irreducibly describes a medical concept. Examples: heart, scoliosis, congestive heart failure, Stevens Johnson Syndrome
- Verified our consistency with content-experts in MT selection (Kappa 0.75).
- Identified MTs a priori for each study document
- Parsed each document (a requirement for optimal MetaMap performance) using the KM sentence parser, and then sent the parsed sentences to both MetaMap and KM (Figure 2).
- Standardized the outputs of both indexers using a script that masked the identity of each indexer (Figure 3).
- Scored each output simultaneously while blinded to the identity of the indexer.

## RESULTS

The results of concept indexing evaluation for KM and MetaMap are shown in Table 1. KM had a recall of 86% in the pilot study and 82% in the definitive study. MetaMap had a recall of 81% and 78% in the pilot and definitive studies respectively.

## DISCUSSION & FUTURE DIRECTIONS

We 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 potential to help educators to locate, integrate, evaluate, and iteratively improve medical school curriculum content. We are now developing a web-based system for display and concept-based navigation of the medical curriculum using KM.

| Concept Mapping for file: GNEG.html                                    |   |
|--|---|
| Input Sentence #1: 1. Drugs Used to Treat gram - Negative Infections   |   |
| A  | B   |
| C0013227:Drugs (Pharmaceutical Preparations) [Pharmacologic Substance] | C0449889:Drug used (Drug used) [Functional Concept]   |
| C0042153:Using (utilization) [Quantitative Concept]                    | C0085423:Gram-negative bacterial infection (Gram-Negative Bacterial Infections) [Disease or Syndrome] |
| C0439208:Gram (Gram) [Quantitative Concept]                            | C0332154:Treat (Received therapy or drug for) [Functional Concept]                                    |
| C0205160:Negative (Negative) [Finding]                                 |   |
| C0021311:Infections <1> (Infection) [Disease or Syndrome]              |   |

Figure 3. Standardized output for a document. A=MetaMap, B=KM

| Concept Identifier             | Pilot Study (n=5) | Definitive Study (n=10) |
|--------------------------------|-------------------|-------------------------|
| Gold Standard Meaningful Terms | 1955              | 4281                    |
| <b>MetaMap</b>                 |                   |                         |
| Recall                         | 1580 (81%)        | 3325 (78%)              |
| Precision                      |                   | 85%                     |
| <b>KM</b>                      |                   |                         |
| Recall                         | 1677 (86%)*       | 3510 (82%)*             |
| Precision                      |                   | 89%*                    |

Table 1. Recall and Precision for MetaMap and KM

\*P < 0.01

