

Knowledge Requirements for Automated Inference of Medical Textbook Markup

Daniel C. Berrios, MD, MPH,^{1,2} Andrew Kehler, PhD,³

Lawrence M. Fagan, MD, PhD²

¹Veterans Affairs Palo Alto Health Care System Palo Alto, CA

²Stanford Medical Informatics, Stanford University School of Medicine, Stanford, CA.

³SRI International, Menlo Park, CA.

ABSTRACT

Indexing medical text in journals or textbooks requires a tremendous amount of resources. We tested two algorithms for automatically indexing nouns, noun-modifiers, and noun phrases, and inferring selected binary relations between UMLS® concepts in a textbook of infectious disease. Sixty-six percent of nouns and noun-modifiers and 81% of noun phrases were correctly matched to UMLS® concepts. Semantic relations were identified with 100% specificity and 94% sensitivity. For some medical sub-domains, these algorithms could permit expeditious generation of more complex indexing.

INTRODUCTION

High-precision information retrieval (IR) requires an index that is more sophisticated than keywords, term frequencies, concepts (1), or context models (2). Automated or semi-automated generation of complex indexes is a primary research goal in IR. The frame-based MedIndex system (3,4) provides suggestions for human indexers, but still requires a large amount of human time and effort during the indexing process. The goal of our experiments was to examine the amount of knowledge, effort, and human guidance required by an inferencing algorithm for automatically indexing a medical textbook.

Our research group designed MYCIN II (5), a computer-based textbook information retrieval system. MYCIN II is based on a model of user-submitted queries (the query model), and text markup (indexing) of sections of text. Knowledge engineers generate a query model and text markup in interactive processes (6). The concepts and relations in the query model dictate the structure of the text markup. A major obstacle to use of the system is the time required for generation of the markup, which can take several hours per textbook chapter even when performed by experts.

We examined the possibilities for reducing this bottleneck. Specifically, we wished to automate the process of creating text markup. The current markup model consists of concepts (corresponding to 1998 UMLS® semantic types), values (UMLS concepts), and binary relations (from the UMLS Semantic Network), which we term a "relational index." This paper describes algorithms and resources for automatically identifying markup in a medical textbook.

METHODS

In order to create text markup, MYCIN II requires that all relevant concepts (classes of values, e.g., "pharmacologic substance"), values ("amoxicillin"), and relations ("treats") be marked in every sentence of the target text deemed appropriate for indexing by the domain expert. We used the 1998 UMLS knowledge base in conjunction with nominal phrase identification software (Lexical Corp.) and information extraction software (Marmot/Badger, U. Mass, Center for Intelligent Information Retrieval) to achieve these goals. The UMLS knowledge base provided semantic type and relational knowledge for concepts and values, and was accessed via PERL routines through its web API interface. To identify nominal compounds, we employed a tool that uses a very long barrier word list, with exception handling by handcrafted rules (7). The information extraction software was used principally for its tokenization, syntactic parsing, and semantic tagging capabilities. PERL routines were authored to perform the relational inference as described below.

We experimented on a single chapter from a textbook on infectious disease. The editor of the textbook marked up the entire chapter by hand. We conducted two separate experiments. In the first, the entire chapter was analyzed to determine how well concepts and concept values could be identified automatically using the UMLS. The goal was to determine the amount of additional domain (medicine) and sub-domain (infectious disease)

```

Look up word for semantic type in UMLS® Metathesaurus
If semantic type found, STOP. Return first semantic type found.
If no semantic type, see if word has a "base" form in UMLS® using the SPECIALIST lexicon (stems
plurals, etc).
    If base form, see if base form has a semantic type.
    If base form has semantic type, STOP. Return first semantic type found.
    If base form has no semantic type, END.

```

Figure 1. The "First Type" algorithm for semantic type determination.

specific knowledge required in order to employ a completely automated semantic typing algorithm. The second experiment was designed to determine the extent to which relations between concepts that are implicit in text could be inferred using relations like those in the UMLS semantic network and a simple algorithm.

We limited the second experiment to eight consecutive sentences in the Drug Therapy section of the same chapter because of the time required to manually review the relations that have been inferred. In one of these sentences, the author mentioned "Table 2," which lists a set of antibiotic susceptibilities for a particular organism. The textbook editor felt that all of the susceptibilities listed in Table 2 should have been inferred as relations at this point in the text. However, the IE tool has no mechanism for incorporating tabular or other non-textual data when such a reference appears in the text. We therefore excluded the markup associated with this table from our analyses.

Concept and Phrase Identification

The information extraction (IE) software requires a method for identifying all multi-word phrases. For these experiments, we were only concerned with correctly identifying noun phrases. We incorporated the nominal-phrase-recognition component of the Metaphrase program (4) (Lexical Technology, Inc.) into the parsing module of Marmot/Badger. Once the augmented parsing module of the IE software had tagged parts of speech as nouns or noun phrases, we employed an ad-hoc "First Type" algorithm to identify concepts (semantic types) and concept values (UMLS Metathesaurus concepts) (Fig. 1). In an identified concept had multiple semantic types in the UMLS Metathesaurus, the first returned semantic type was deemed the correct one.

Inferring Relations Automatically

For this limited experiment, we developed a very small markup model, consisting of only three semantic types and three binary relations between them (Fig 2,A). Certain terms that occurred in the text should have had one of these three semantic types, but were not found as such in the 1998 UMLS (Figure 2,B; see previous section). Because this experiment was solely concerned with automatically inferring relations between all valid medical concepts, these terms and their types were added to a local, domain-specific semantic dictionary that was used only for this second experiment.

Two of the three relations in the markup model were in the semantic network of the 1998 UMLS (Figure 2,C). There was no relation in the UMLS® semantic network linking the types "Substance" and "Organism Attribute," so this relation, which denotes antibiotic susceptibility, was created and stored in a local relations dictionary.

Concepts and their semantic types were grouped according to the clauses in the text in which they occurred (Fig 3, A). All possible instances of semantic relations between concepts were then generated (Fig 3, B). A relation was considered possible only if 1) the relation occurred in the UMLS semantic network or the local relations dictionary, and 2) concepts compatible with both semantic types in the binary relation occurred in the same clause. Compatible concepts were defined as those whose semantic types exactly matched or were descendants of the types in the binary relation.

For example, the sentence "*The organism is susceptible to quinolones, as well as tetracyclines, trimethoprim-sulfamethoxazole, chloramphenicol and rifampin.*" contains the concepts "susceptible", "quinolones," "tetracyclines," "trimethoprim-sulfamethoxazole," "chloramphenicol" and "rifampin" in one clause. "Susceptible" has the semantic type "Organism Attribute" and "quinolones" has the type "Organic Chemical",

A

| Concept (Semantic Type) | Example Values |
|-------------------------|---|
| Organism Attribute | Antibiotic resistance, Methicillin resistance |
| Laboratory Procedure | Bacterial sensitivity tests, Laboratory culture |
| Substance | Nafcillin, lincomycin, chloramphenicol |

B

| Concept (Semantic Type) | Example Values |
|-------------------------|--|
| Organism Attribute | susceptible, sensitive, resistant, moderately resistant, |
| Laboratory Procedure | susceptibility, susceptibilities, sensitivities, sensitivity |
| Substance | b-lactam |

C

| Concept | Relation* | Concept |
|--------------------|--------------------------|----------------------|
| Substance | has sensitivity result** | Organism Attribute |
| Substance | assessed for effect by | Laboratory Procedure |
| Organism Attribute | measured by | Laboratory Procedure |

*Inverse relations were also included, but are not shown.

**Not in the 1998 UMLS® Semantic Network

Figure 2. A "Relational Markup (Indexing) Model. A: Concepts and some example concept values. B: Concept for which values were present in the text, but not in the 1998 UMLS®. C: The three binary relations included in the model.

which is a child of the type "Substance." Consequently, an instance of the first relation in Figure 2C could be instantiated: "has_sensitivity_result(susceptible, quinolones)".

Possible relations were filtered according to the concepts and relations occurring in the markup model (Fig 3, C). This final group of concepts and relations was then hand-checked for accuracy by comparison to markup manually generated by an editor of the textbook.

RESULTS

The entire chapter contained 2,432 words, including 29 correctly identified noun/noun-modifying phrases, 842 instances of nouns, and 242 instances of noun modifiers. Of these, 486 were unique nouns and noun modifiers. Eleven terms were misidentified as nouns or noun modifiers (e.g., "harbor", "sole", "displays") due to bad parses. Thus, the positive predictive value was 98%. There were only 63 different semantic types identified in the entire chapter; the top ten types are listed in Table 1.

All the noun/noun-modifying phrases identified had only one correct semantic type, returned by the phrase-identification software. The

software failed to identify seven phrases (19%), almost all of which were indicators of genus and species of bacteria. The First Type algorithm returned a semantic type for 291 (66%) of the identified nouns and noun-modifiers. Of these, only ten (3%) had significant polysemy (e.g., "cultures", "patent", "strain").

Of the 189 instances of First Type failure, 58 (31%) were highly specialized words or phrases, including 16 (9%) hyphenated or contracted forms ("eikenella-like", "a-hemolytic", "sulfa-trimethoprim") and 9 (5%) abbreviations, proper names, and alternative drug names. For 33 (57%) of these domain-specific terms (e.g., "coinfecting")

Table 1. The ten most frequent semantic types.

| Semantic Type | Count |
|--------------------------|-------|
| Qualitative Concept | 32 |
| Bacterium | 25 |
| Quantitative Concept | 24 |
| Functional Concept | 19 |
| Organic Chemical | 18 |
| Temporal Concept | 16 |
| Intellectual Product | 14 |
| Body Part, Organ, or ... | 13 |
| Pharmacologic Substance | 13 |
| Disease or Syndrome | 11 |

“empirical”), no appropriate terms in the Metathesaurus were found by approximate match or manual search. The remaining 140 instances of First Type algorithm failure consisted of 103 domain-nonspecific (i.e., common English) terms

(e.g., “reason,” “excellent”). The parsing module relies on a local English dictionary for part-of-speech tagging, and these terms were simply not present.

Twenty-nine instances of relations were

Target Text:

Sentence 3: It is usually susceptible to b-lactam antibiotics such as penicillin, amoxicillin and ampicillin, but is resistant to penicillinase-resistant penicillin (dicloxacillin, methicillin, nafcillin), and displays variable susceptibilities to cephalosporins.

A

| Clause | Semantic type CONCEPT VALUE |
|--------|---|
| 1 | Antibiotic B-LACTAM Pharmacologic Substance AMPICILLIN Organism Attribute SUSCEPTIBLE Organic Chemical PENICILLIN Pharmacologic Substance AMOXICILLIN Antibiotic ANTIBIOTICS |
| 2 | Pharmacologic Substance NAFCILLIN Pharmacologic Substance METHICILLIN Organic Chemical PENICILLIN Organism Attribute RESISTANT Pharmacologic Substance DICLOXACILLIN |
| 3 | Laboratory Procedure SUSCEPTIBILITIES Qualitative Concept VARIABLE Organic Chemical CEPHALOSPORINS |

B

| | | |
|------------------------------------|--|-------------------------------------|
| Antibiotic=B-LACTAM | isa | Pharmacologic Substance=AMPICILLIN |
| Antibiotic=B-LACTAM | interacts with | Pharmacologic Substance=AMPICILLIN |
| Antibiotic=B-LACTAM | has sensitivity result | Organism Attribute=SUSCEPTIBLE |
| Antibiotic=B-LACTAM | interacts with | Organic Chemical=PENICILLIN |
| Antibiotic=B-LACTAM | isa | Pharmacologic Substance=AMOXICILLIN |
| Antibiotic=B-LACTAM | interacts with | Pharmacologic Substance=AMOXICILLIN |
| Antibiotic=B-LACTAM | isa | Antibiotic=ANTIBIOTICS |
| Antibiotic=B-LACTAM | interacts with | Antibiotic=ANTIBIOTICS |
| Pharmacologic Substance=AMPICILLIN | has_sensitivity_result_for specific organism | Organism Attribute=SUSCEPTIBLE |
| ... | | |

C

| | | |
|------------------|--------------------------------|----------------|
| B-LACTAM | has sensitivity result | SUSCEPTIBLE |
| AMPICILLIN | has sensitivity result | SUSCEPTIBLE |
| SUSCEPTIBLE | is sensitivity result for drug | PENICILLIN |
| SUSCEPTIBLE | is sensitivity result for drug | AMOXICILLIN |
| SUSCEPTIBLE | is sensitivity result for drug | ANTIBIOTICS |
| NAFCILLIN | has sensitivity result | RESISTANT |
| METHICILLIN | has sensitivity result | RESISTANT |
| RESISTANT | is sensitivity result for drug | DICLOXACILLIN |
| SUSCEPTIBILITIES | assesses effect of | CEPHALOSPORINS |
| SUSCEPTIBILITIES | assesses effect of | CEPHALOSPORINS |

Figure 3. Developing a relational index. A. Concept values and their semantic types are grouped by sentence and clause number in which they occur. B. Possible relations between concepts within clauses are generated (only a portion shown). C. Those relations specified by the markup model are extracted.

inferred in the target text. All were deemed to be correctly inferred (100% specificity/precision). Two instances of relations were not automatically inferred, but were authored by the textbook editor (94% sensitivity/recall). Both of these relations were conveyed in the text using the polysemantic term "activity" (e.g., "The activity of macrolides is poor.").

CONCLUSIONS

We conducted experiments to determine the amount of this knowledge-dense indexing that could be generated automatically. We studied a simple "First Type" algorithm for concept and phrase semantic typing and a rudimentary inference method for semantic relations. The semantic typing algorithm looks for lexical matches of words or their base forms in the UMLS. The inferencing method instantiates any possible relation between concepts as allowed in the UMLS Semantic Network and compatible with a relational markup model.

The First Type algorithm identified two-thirds of the noun and noun-modifying concepts in our test chapter, and assigned a correct semantic type to the overwhelming majority of them. Many of the remaining concepts were not domain-specific terms, and could have been identified with a larger English dictionary. The semantic types of the relatively few domain-specific terms not found in the UMLS Metathesaurus could be resolved by human knowledge engineering in an interactive process, or by developing a small domain-specific dictionary.

Our simple method to infer relations produced good results. However, these results must be considered in the context of our experimental design. For example, the relations we included in the experiment were deemed important for markup by the authors of the textbook. However, we experimented on a limited section of the chapter. Therefore, the inference method requires human direction in the text areas to be indexed and the specific relations it is allowed to infer. Furthermore, we had to override some of the semantic type definitions in the UMLS with domain-specific definitions (e.g., assigning "sensitive" the type "Laboratory Test Result"). To the extent that such manual knowledge engineering is required, our methods might require considerably more labor in other domains. Finally, we chose only to index binary relations in the textbook, largely because these are the types of relations modeled in the UMLS semantic network.

Acknowledgements

Supported by a National Library of Medicine Training Grant LM07033, by the Veterans Affairs Office of Academic Affairs and Health Services Research and Development Service Research funds and the Office of the Chief Information Officer, and National Cancer Institute Contract N44-CO5-61025 to Lexical Technologies, Inc. We wish to thank David D. Sherertz of Lexical Technologies and Victor L. Yu, MD of the Univ. of Pittsburgh, Dept. of Infectious Disease. Also, Ms. Kathleen Jones provided invaluable assistance.

References

1. Hersh WR, Hickam DH. A comparison of retrieval effectiveness for three methods of indexing medical literature. *The American Journal of the Medical Sciences* 1992;303(5):292-300.
2. Purcell GP, Shortliffe EH. Contextual models of clinical publications for enhancing retrieval from full-text databases. In: Gardner RM, editor. *Nineteenth Annual Symposium on Computer Applications in Medical Care*; 1995; New Orleans, LA: Hanley & Belfus, Inc.; 1995. p. 851-57.
3. Humphrey SM. Research on interactive knowledge-based indexing: the MedIndEx prototype. *Proc Annu Symp Comput Appl Med Care*1989:527-533.
4. Humphrey SM. Indexing biomedical documents: from thesaural to knowledge-based retrieval systems. *Artificial Intelligence in Medicine* 1992;4:343-371.
5. Kim DK, Fagan LM, Jones KT, Berrios DC, Yu VL. MYCIN II: design and implementation of a therapy reference with complex content-based indexing. *Proc AMIA Symp* 1998:175-179.
6. Berrios DC, Kehler A, Kim DK, Fagan LM, Yu VL. Automated Text Markup for Information Retrieval from an Electronic Textbook of Infectious Disease. *Proc AMIA Symp* 1998:975.
7. Tuttle MS, Olson NE, Keck KD, Cole WG, Erlbaum MS, Sherertz DD, et al. Metaphrase: an aid to the clinical conceptualization and formalization of patient problems in healthcare enterprises. *Methods Inf Med* 1998;37:373-383.