Using POSTDOC to recognize biomedical concepts in medical school curricular documents

By Steven L. Kanter, M.D.
Director, Office of Medical Education

Randolph A. Miller, M.D.
Professor of Medicine

Marie Tan, M.D.*

Jonathan Schwartz, M.D.
Fellow, Section of Medical Informatics

University of Pittsburgh School of Medicine†
Pittsburgh, Pennsylvania 15261

Recognition of the biomedical concepts in a document is prerequisite to further processing of the document: medical educators examine curricular documents to discover the coverage of certain topics, detect unwanted redundancies, integrate new content, and delete old content; and clinicians are concerned with terms in patient medical records for purposes ranging from creation of an electronic medical record to identification of medical literature relevant to a particular case. POSTDOC (POSTprocessor of DOCUMENTs) is a computer application that (1) accepts as input a free-text, ASCII-formatted document and uses the Unified Medical Language System (UMLS) Metathesaurus to recognize relevant main concept terms; (2) provides term co-occurrence data and thus is able to identify potentially increasing correlations among concepts within the document; and (3) retrieves references from MEDLINE files based on user identification of relevant subjects. This paper describes a formative evaluation of POSTDOC’s ability to recognize UMLS Metathesaurus biomedical concepts in medical school lecture outlines. The “precision” and “recall” varied over a wide range and were deemed not yet acceptable for automated creation of a database of concepts from curricular documents. However, results were good enough to warrant further study and continued system development.

INTRODUCTION

The University of Pittsburgh School of Medicine recently implemented the Physicians in Two Thousand (P.I.T.T.) curriculum, a goal-oriented, integrated, interdisciplinary program for all students. The new curriculum uses several different modalities of learning, including lectures, problem-based learning (PBL), tutorials, debates, dissection, laboratory experiences, and field trips. Each topic-based course in the first-year basic science block crosses traditional disciplinary lines and is taught by faculty members from various departments throughout the medical center.

In this setting, it is difficult for a given faculty member to know what was taught about his or her topic elsewhere in the curriculum. This difficulty is exacerbated by the interdisciplinary nature of the courses. For example, suppose a lecturer plans to develop a concept that requires a prior understanding of the anatomy of the conduction system of the heart. One can no longer infer that such a topic has been covered simply by knowing that the students have

* Dr. Tan completed her work on this project as a Fellow in the Section of Medical Informatics. Her current address is Department of Pediatrics, Allegheny General Hospital of Pittsburgh, Pittsburgh, Pennsylvania 15212.
† Address all correspondence to Steven L. Kanter, M.D., Director, Office of Medical Education, M-211 Scaife Hall, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15261.
completed the basic anatomy course. In fact, in the P.I.T.T. curriculum, this topic is not covered in the basic anatomy course but is integrated into a ten-week body-fluid homeostasis course offered early in the second year.

Most faculty members are unable to spend hours reviewing lengthy syllabi, lecture outlines, textbooks, and other materials to find this information; they need the ability to retrieve a list of topics included in lecture outlines and other curricular documents, preferably limiting retrieval to a specified topic. Members of the Curriculum Committee need to locate topics in the curriculum as a first step in determining adequate coverage of subjects. Other individuals need to browse a list of topics from curricular documents to find unwanted redundancies.

A review of the literature revealed that several schools of medicine and dentistry responded to these issues by initiating development of curricular content databases [1–9]. Most specified a unit of instruction (e.g., a single lecture or laboratory session); created a system for representing the knowledge in that unit by selection of keywords or preparation of a summary; and entered keywords or other material into a database management system, text file management system, or both. The MeSH vocabulary often was used for keyword selection, and at least one school was anticipating use of the Unified Medical Language System (UMLS) [10].

As previously discussed by Kanter, there are several problems with this approach [11]. Text-file searches have problems with precision and recall. A few keywords per lecture are not adequate to represent the knowledge in an instructional unit, and the use of faculty to select keywords significantly adds to the cost. The rapid discovery of new knowledge in biomedicine necessitates frequent updates to the content of a medical curriculum, making maintenance of a keyword database difficult, expensive, and time-consuming. Adding to the problem is the lack of a controlled vocabulary that can capture the concepts expressed in medical curricular materials. MeSH was designed for indexing the biomedical literature and is not well suited for educational purposes.

The goal of creating a database of curricular content that could be maintained and updated in a cost-effective manner with minimal reliance on faculty and student time raised the question, What range of precision and recall could be obtained by automatic recognition of biomedical concepts in curricular documents using current technologies? This question was pursued using the National Library of Medicine’s UMLS and POSTDOC (POSTprocessor of DOCUMENTs), a lexical matching technique developed at the University of Pittsburgh School of Medicine. The UMLS is an experimental series of relational tables listing biomedical concepts in its Metathesaurus and their interrelationships in its Semantic Network. POSTDOC is a descendant of CHARTLINE, a project developed to provide clinically relevant references from MEDLINE, based on terms from a patient’s medical record [12]. This article examines the capability of POSTDOC to recognize biomedical concepts represented in the UMLS Metathesaurus in lecture outlines from six courses in the basic science block.

METHODS

The POSTDOC algorithm

The algorithms used by POSTDOC and its predecessor, CHARTLINE, have been described previously [13]. The POSTDOC term-recognition algorithms represent a refinement of those used in CHARTLINE. Briefly, the program processes an ASCII-text document in three passes. During the first pass, the program retains only those words or their singular, plural, or possessive variants that appear in at least one term somewhere within the UMLS Metathesaurus [14]. If a word is not found within any term in the Metathesaurus, POSTDOC ignores it. For example, consider the phrase, “list three functions of the pentose phosphate pathway.” The words retained for analysis would be list, three, function, pentose, phosphate, and pathway.

In the second phase, POSTDOC processes the Metathesaurus words identified in the order in which they appeared in the source document. If two or more successive words recognized within the source document appear together in one or more Metathesaurus terms (e.g., pentose and phosphate occur together in a term, but list and three do not), the words are combined for further consideration. Single words also are retained if their adjacent recognized words from the source document are not found in any Metathesaurus terms along with that single word. Retained terms, referred to as candidate terms, are processed further using a number of heuristic rules. First, if the words in the candidate term represent more than 50% of the words in the actual Metathesaurus term, then the candidate term and its potential UMLS matching term are retained. In the example given, the text words “pentose phosphate” represent two thirds of the Metathesaurus term, “pentose phosphate shunt.” Second, if the candidate term has ten or fewer UMLS potential matching terms, it is deemed selective enough to be considered further (as opposed to being nonspecific and matching too many UMLS terms). Finally, one or more best terms are selected from all potential UMLS Metathesaurus terms, favoring terms that contain all the words in the candidate term but contain the fewest words overall. The resulting list of best terms are called “recognized terms.” In our example, “pentose pathway” appears as a term in the Metathesaurus; however, the POSTDOC program...
maps this phrase to "pentose phosphate shunt," its preferred lexical name.

The final pass of POSTDOC processing identifies, for a given recognized term, the other recognized terms from the document that co-occur with it in the literature. Using the UMLS Metathesaurus co-occurrence of terms table, it is possible to determine the number of articles in the MeSH-indexed biomedical literature in the past decade (more than 3,000 journals are indexed annually) that discuss two given MeSH concepts as "main concepts."

**Processing medical curriculum documents**

Textual material was selected systematically from faculty-prepared syllabi used in courses of the P.I.T.T. curriculum. The third lecture outline was picked from the course syllabus from each of six courses in the twenty-five-week basic science block. Each curricular document was prepared and analyzed as follows: an ASCII file of the text of the first two pages of the document was created. Headers, footers, instructions to the class, reading lists, and faculty names were removed, essentially leaving the educational objectives and content outline or descriptions for further analysis.

Stop-words were removed. The list of stop-words included student, outline, objective, content, atlas, reading, and textbook. The ASCII file was printed and presented to a faculty member who selected terms in the document that represented biomedical concepts pertinent to the topic and important to the students. Any sequence of contiguous words could be chosen. Selected terms could contain prepositions (e.g., carcinoma of the thyroid) but not conjunctions or disjunctions and were not to cross or include periods, commas, semicolons, colons, slashes, or parentheses. One of the authors independently selected all terms in each document meeting the criteria above, with the lone difference that selection was not limited to terms pertinent to the topic and important to students.

The text (ASCII file) then was processed by POSTDOC. The result was a list of terms from the 210,000 lexical terms (representing 130,000 main concepts) in the UMLS Metathesaurus that the program identified as most likely best terms. Term counts were determined as follows: (1) the number of curricular document terms judged by the faculty member to represent a biomedical concept (= true positives + false negatives); (2) the number of best terms found in the UMLS by POSTDOC (= true positives + false positives); (3) the number of "complete matches" (a curricular document term selected by a faculty member was recorded as a "complete match" if it was found [word for word, with lexical variants and synonyms acceptable] in the UMLS by POSTDOC as a best term, based on input of the curricular document term itself); (4) the number of "partial matches" (a curricular document term selected by a faculty member that was not a "complete match" was recorded as a "partial match" if a subset of the term was found [with lexical variants and synonyms acceptable] in the UMLS by POSTDOC as a best term, based on input of the curricular document term itself; number of "complete matches" + number of "partial matches" = true positives).

The following values were derived for each curricular document: (1) the number of curricular document terms selected by the faculty member that were not found in the UMLS by POSTDOC (false negatives); (2) the number of best terms found in the UMLS by POSTDOC but not judged to be present in the curricular document (false positives); (3) the ratio of the number of terms found by both POSTDOC and the faculty member to the number of terms selected by the faculty member (referred to in this paper as "recall"); and (4) the ratio of the number of terms found by both POSTDOC and the faculty member to the number of terms found by POSTDOC (referred to as "precision").

**RESULTS**

Table 1 shows the counts obtained for curricular documents. A comparison of the first two columns shows that the number of terms selected by one of the authors is always larger than the number of terms selected by the course faculty, for the reason noted earlier.

Table 2 shows values for the curricular documents analyzed using the terms selected by one of the authors and the total number of matches; i.e., complete matches plus partial matches. The recall values vary from 19% in the cell signaling lecture to 78% in the genetics lecture. Many terms in the cell signaling lecture were not found by POSTDOC because they are basic electrophysical and electrophysiological...
terms not present in the UMLS Metathesaurus. Many terms in the genetics lecture were present in the Metathesaurus, probably because they have been hot topics during the last decade. The precision values range from 42% to 80%.

Table 3 shows the recall and precision values for the curricular documents analyzed using the terms selected by the faculty members and the total number of matches. The recall values ranged from 17% to 73% and are similar to the values in Table 2 except for the cell structure document. The precision values were lower than in Table 2, as would be expected, because the selection of terms was more discriminating.

DISCUSSION AND CONCLUSION

Recognition and representation of the biomedical concepts in a document is prerequisite to further examination, processing, or treatment of the document for medical purposes. Medical educators examine the representation of terms in curricular documents to discover the coverage of certain topics and to detect omissions and unwanted redundancies [15–20]. Clinicians are concerned with the representation of terms in patient medical records [21] for purposes ranging from creation of an electronic medical record to identification of medical literature relevant to a particular case [22]. Although there is a place for manual extraction of key concepts (e.g., selection of a few keywords from a controlled vocabulary to represent a lecture outline or selection of diagnosis-related groups and current procedural terminology codes to represent a patient’s hospital discharge summary), medical informatics researchers are seeking to tap the potential of automated recognition of biomedical concepts in medical documents.

If most of the terms in curricular documents could be identified as Metathesaurus concepts, it would be possible to provide the faculty with a set of tools to retrieve lists of topics, identify redundancies, and perform other tasks that address previously mentioned concerns. Although the current version of POSTDOC did not produce uniformly acceptable results, its performance on some curricular documents was good enough to warrant further study. Improvements in recall for some content areas likely would be obtained if coverage of those areas in the UMLS Metathesaurus were improved. For example, anatomic concepts were not represented in sufficient detail in the thesaurus for curricular documents. Because the Metathesaurus is an experimental product that grows and evolves over time, future versions are likely to contain a greater number and variety of terms.

Improvements in precision would require improvements in POSTDOC’s methodology. For example, in the specialized tissue lecture outline, POSTDOC found the statement, “... share a common chorion and amnion are rare, suggesting ...” and produced a best term of “amnion-chorion procoagulant.” A reduction in mismatches such as this one would result in fewer false positives, thus improving precision. If POSTDOC had not used the terms chorion and amnion in a multiple-word best match, the terms could have been recognized correctly, thus increasing the number of true positives and improving recall. Additionally, the POSTDOC algorithm contained several heuristic simplifications that resulted in less than optimal behavior in some circumstances.

POSTDOC’s ability to identify potentially interesting correlations within a document, as well as between the document and the biomedical literature, can offer intelligent support to curriculum planners charged with integration of material throughout the curriculum. It also can support individual lecturers.
or small-group facilitators who frequently request information regarding topics students have covered in prior sessions. Future applications of this feature of POSTDOC might include an “idea generator” for faculty members preparing a lecture outline or PBL case. Co-occurrence data also could be useful in identifying pertinent references (as POSTDOC can do with a patient medical record), as well as in supporting comprehensive review of pertinent material and relationships. Medical students also could benefit from the ability to browse cross-references among lectures as well as PBL case objectives and text.

The UMLS Semantic Network allows classification of concepts by semantic type, and the hierarchical nature of the network provides the capability of generating broader concepts. This could be used to better address broad-based queries, such as retrieving a list of all concepts related to anatomy in a set of curricular documents, a real concern of faculty in the P.I.T.T. curriculum, who are charged with overseeing adequate coverage and integration of broad subject areas. Other hierarchies, such as those found in the Diagnostic and Statistical Manual, a classification of mental disorders, or SNOMED, may also be useful in this regard.

Automated location of relevant patient cases for use in PBL sessions would be a useful tool. This project is feasible at the University of Pittsburgh because more than 200,000 electronic patient records are available online (subject to appropriate security considerations) through the Medical Archival System [23]. POSTDOC technology, as well as the method of clustering similar terms based on the semantic types of their co-occurring terms, as described in Kanter, would be valuable in developing this tool [24].

ACKNOWLEDGMENTS

The authors wish to thank Carole Coffee, Ph.D.; Cynthia Lance-Jones, Ph.D.; Dione Cahillane; Lisa Kozen; Leslie Levine; and Nancy Stabryla for their valuable assistance in this project.

REFERENCES

10. Mattern, Computer databases.
13. Ibid.
15. Curry, op. cit.
17. Mattern, Computer databases.
18. Rozinski, op. cit.
22. Miller, op. cit.
24. Kanter, op. cit.