Research Paper ■

UMLS Concept Indexing for Production Databases:

A Feasibility Study

PRAKASH NADKARNI, MD, ROLAND CHEN, MD, CYNTHIA BRANDT, MD, MPH

Abstract **Objectives:** To explore the feasibility of using the National Library of Medicine's Unified Medical Language System (UMLS) Metathesaurus as the basis for a computational strategy to identify concepts in medical narrative text preparatory to indexing. To quantitatively evaluate this strategy in terms of true positives, false positives (spuriously identified concepts) and false negatives (concepts missed by the identification process).

Methods: Using the 1999 UMLS Metathesaurus, the authors processed a training set of 100 documents (50 discharge summaries, 50 surgical notes) with a concept-identification program, whose output was manually analyzed. They flagged concepts that were erroneously identified and added new concepts that were not identified by the program, recording the reason for failure in such cases. After several refinements to both their algorithm and the UMLS subset on which it operated, they deployed the program on a test set of 24 documents (12 of each kind).

Results: Of 8,745 matches in the training set, 7,227 (82.6 percent) were true positives, whereas of 1,701 matches in the test set, 1,298 (76.3 percent) were true positives. Matches other than true positive indicated potential problems in production-mode concept indexing. Examples of causes of problems were redundant concepts in the UMLS, homonyms, acronyms, abbreviations and elisions, concepts that were missing from the UMLS, proper names, and spelling errors.

Conclusions: The error rate was too high for concept indexing to be the only production-mode means of preprocessing medical narrative. Considerable curation needs to be performed to define a UMLS subset that is suitable for concept matching.

■ **J Am Med Inform Assoc.** 2001;8:80–91.

Affiliation of authors: Yale University School of Medicine, New Haven, Connecticut.

This work was supported in part by National Institutes of Health grants R01 LM06843-01 from the National Library of Medicine and U01 CA78266-02 from the National Cancer Institute.

The authors will provide the relational UMLS schema used in this study (plus UMLS data from sources that have not imposed restrictions on distribution) and the Microsoft Access front end (which includes Concept Locator) to anyone who makes a written request.

Correspondence and reprints: Prakash M. Nadkarni, MD, Center for Medical Informatics, Yale University School of Medicine, P.O. Box 208009, New Haven, CT 06520-8009; e-mail: <Prakash.Nadkarni@yale.edu>.

Received for publication: 2/29/00; accepted for publication: 7/31/00.

Free text, like that found in discharge summaries and progress notes, is an important part of the electronic patient record, because it captures nuances of information that coded information cannot. *Information retrieval* is the field of informatics concerned with the processing of free text, typically by domain-independent methods.[1,2] With the ubiquity of the World Wide Web (where most information is textual), information retrieval technology is now mainstream. Several vendors of relational database management systems have integrated information retrieval with their technologies. We emphasize that information retrieval is ancillary to, and does not replace, conventional means of querying patient data through relational tables.

Information retrieval relies on preprocessing a collection of documents to speed up subsequent retrieval of documents that are relevant to a user's query, based on keywords of interest contained in them.* General-purpose Web retrieval engines, such as Yahoo or Excite, index the *words* in documents. For documents belonging to a single domain such as medicine, however, word indexing does not leverage domain knowledge; for example, synonymous phrases are not automatically recognized. Searches of medical free text that is indexed only by word would require a user to manually specify synonymous forms or risk missing relevant documents.

Using concepts in a domain-specific thesaurus can enhance retrieval; that is, we can index the *concepts* identified in a document. Concept-identification approaches are discussed in the next section. For medical records, detection of a concept in a document does not in itself make that document relevant for that concept. The concept may refer to a finding that was looked for but absent or ruled out, or that occurred in the remote past. The recording of "significant negatives" is important in medicine, and robust handling of negation in narrative is still an open problem in information retrieval.

The Unified Medical Language System (UMLS) Metathesaurus (the world's largest domain-specific thesaurus) of the National Library of Medicine (NLM)[3] has been the focus of much research. The present work explores the use of the UMLS for a high-specificity algorithm suitable for automated concept matching (and thereby, indexing) of medical free text.† We quantify the algorithm's success rate with sample data and quantify instances of failure. We use failure analysis to refine both the algorithm and the UMLS subset that it relies on, so that future efforts to improve the success rate may be directed appropriately and optimally.

Background

Approaches to Querying Medical Text Using Controlled Thesauri

One way to use a thesaurus in querying medical text is through close integration of the thesaurus with the query program. When a user specifies a query in terms of one or more keywords of interest, synonyms of those keywords are located and added to the original query, thereby expanding (or broadening) it.

Studies using medical vocabularies for query broadening by intensive generation of lexical variants (including synonyms, abbreviations and acronyms, and morphological variants) have been carried out by the natural language processing group at the NLM and are described in Aronson et al.,[9] Aronson and Rindflesch,[10] Divita et al.,[11] and Rindflesch and Aronson.[12,13] An interesting approach described by Aronson et al.[9] uses a program called MetaMap. Metamap transforms the text in a document by limited syntactic analysis that recognizes simple noun phrases. After variant generation, the resulting phrases are matched to UMLS concepts, where possible, and then replaced with the preferred form of the matched concept (thereby reducing variability in the text). The transformed text, termed "surrogate text" by the authors, is then indexed with a retrieval system to allow query. The work described by Aronson et al.[9] used the well-known SMART retrieval system,[14] whereas the work described by Aronson and Rindflesch[10] used a more recent system, INQUERY.[15]

Srinivasan[16,17] has described an alternative approach that integrates thesaurus-derived document markup (though not the thesaurus itself). This method, termed *retrieval feedback* by the author, has been evaluated with a MEDLINE collection. It relies on the fact that two kinds of vocabularies are used to index MEDLINE documents. The first is the Medical Subject Headings (MeSH), a controlled vocabulary that is part of the UMLS; trained human indexers who carefully read the document's abstract have, historically, performed MeSH indexing. In addition, the documents are indexed by non-stop-words in the document title and abstract, which constitute a relatively uncontrolled vocabulary. In the retrieval feedback approach, documents that are returned with a high relevance rank in response to a user's query are selected, and the MeSH and non-MeSH keywords associated with them are used to expand the original query.

Concept Identification in Medical Narrative

We classify methods used to identify concepts in medical narrative into two categories, phrase-based and sentence-based. We discuss each in turn.

Phrase-based Methods

Phrase-based concept-identification methods use natural language processing to scan narrative and identify word and phrases of interest. These are then used to search the thesaurus. Most research has utilized noun phrases, as in the work of Elkin et al.[18] Aronson

* This process, termed *indexing*, is described in Appendix 1, which appears as supplemental material to this article in *JAMIA Online,* at www.jamia.org.

† An overview of the UMLS schema is provided in Appendix 2, which also appears as supplemental material to this article in *JAMIA Online.*

and Rindflesch's MetaMap program, summarized earlier, is an augmentation of the Elkin approach. A popular freeware natural language processing package for phrase recognition is the Xerox part-of-speech tagger[19]; this technology (used by Metamap, among others) has recently been commercialized as the LinguistX package.[20] Another commercial tagger is CLARIT, whose use has been described by Spackman and Hersh[21] and Evans et al.[22]

A criticism of the use of noun phrases alone is that, in medical narrative, many concepts can be identified correctly only through other parts of speech that are close to the noun phrase; for example, "blood pressure was *greatly elevated*" implies hypertension as opposed to blood pressure alone. Verb phrases such as "surgically resected" are intrinsically meaningful; the UMLS includes a large number of non-noun concepts. Furthermore, the same concept may be divided across two noun phrases, as in "hypertension is secondary to renal disease," which indicates renal hypertension.

Syntactic structure within a document determines the level of sophistication needed by a parser to successfully match concepts. Thus, in the above example, if the phrase "greatly elevated blood pressure" were encountered instead, it might be successfully matched because all four words constitute a single noun phrase with a terminal head ("pressure"). Aronson et al.[9] show that even a relatively simple parsing approach, "under-specified syntactic analysis" (identification of simple noun phrases with the head rightmost) is adequate in many cases.

In an attempt to overcome the limitations of single noun phrases, alternative approaches have attempted to use larger units of text; we will shortly illustrate, however, that these run against the limits of computational intractability. We believe that the phrase-based approach is fundamentally sound, and have used this approach for the work described in this paper.

Sentence-based Approaches

To address the problem of single concepts being split across multiple phrases, sentence-based approaches process a larger unit of text at a time. The approaches described to date rely on simple elimination of stop words and do not use part-of-speech tagging.

The SAPHIRE family of algorithms devised by Hersh and Greenes[23–25] exemplifies these approaches. The earliest SAPHIRE algorithm matched substrings of stemmed input text to stemmed concepts in a thesaurus,[23] making multiple passes across a block of text to identify all concepts. Its sensitivity was vulnerable to the order of words in a phrase in the text, which needed to be the same as in the thesaurus to match. A newer, word-order-insensitive algorithm, first mentioned by Hersh et al.[24] and later described by Hersh and Hickman,[25] permuted the order of individual words in input text. It processed input documents a line at a time, up to each carriage return. To prevent concepts from being split by carriage returns, some carriage returns were first removed through a filtering program, so that the indexing process in effect processed the data a sentence at a time.

Some limitations of sentence-based approaches are as follows:

- In contrast to phrase-based approaches, sentence-based approaches err toward reduced specificity (i.e., more false positives). If a sentence contains multiple concepts, permuting words may spuriously generate valid concepts that were not implied in the original text. For example, the text segment "spleen rupture and normal stomach" (in an emergency surgery note) will match the concept of stomach rupture.
- While processing data a sentence at a time greatly improves recognition of concepts that are split across phrases, it cannot guarantee complete success. In the (admittedly artificial) example, "*Blood pressure* was recorded in the supine position. It was found to be *greatly elevated*," the concept of hypertension is split across two sentences.
- Finally, sentence-based approaches have the potential to be extremely machine-intensive.‡

Partial Matches in Concept Indexing

In 1995, Hersh and Leone[30] described a completely new SAPHIRE algorithm for interactive query of the UMLS. This algorithm allows partial matches and returns concepts in descending order of relevance; an elegant Web implementation can currently be accessed via http://www.ohsu.edu/cliniweb/saphint/. While it is appropriate for interactive UMLS query, however, we find this algorithm unsuitable for automated concept indexing of medical text, because there is no obvious computational strategy for eliminating false positive partial matches that pass the SAPHIRE threshold. If false positives are more numerous than true positives, then most entries in the concept index will be misleading.

‡ Sentence-based approaches are discussed in Appendix 3, which appears as supplemental material to this article in *JAMIA Online*, at www.jamia.org.

In a preliminary experiment, we implemented this algorithm and tested it with a surgery note containing the term "ligamentum flavum" (a ligament that connects adjacent vertebrae). Apart from the exact match "ligamentum flavum," we also got more than 30 partial matches for each pair of adjacent vertebrae—"C1/C2 ligamentum flavum," "C2/C3 ligamentum flavum," and so on, up to the coccyx. (The operation site, the lumbar spine, was noted two sentences previously in the narrative note.) This experience caused us to lean toward high specificity in our concept-matching approach, even at the cost of some sensitivity.

Thesaurus Issues: Composite vs. General Concepts

In the UMLS, which depends on its source vocabularies for comprehensiveness, *general* (primitive, atomic) concepts as well as relatively specific, *composite* concepts are formed by combining two or more general concepts. Thus, "carcinoma of pancreas" is a composite of "carcinoma" and "pancreas." The inclusion of composite terms depends on the source-vocabulary curators. Thus, "digitalis-induced atrial fibrillation" does not exist; if encountered in text, it can only match two separate concepts, "digitalis" and "atrial fibrillation." The most specific concept cannot always be matched. In the example "hypertension is secondary to renal disease," phrase-based approaches will miss "renal hypertension," instead matching "hypertension" and "renal disease" separately.

Sometimes, composite concepts exist, but the concepts from which they are derived are missing. For example, "seizure activity," an electroencephalographic finding, is missing, although "monitor for seizure activity" and "seizure activity not present" exist. When a particular general concept is missing from the thesaurus, false positive matches may result as an artifact of the concept-matching algorithm. A more specific concept (which is a child of the general concept) may be matched erroneously to a phrase in the text simply because it provides the closest (or a unique) match from all the concepts in the thesaurus.

Ambiguous Terms: Homonyms, Acronyms and Abbreviations

Homonyms are strings that map to multiple concepts. For example, "anesthesia" refers to loss of sensation as a clinical finding or to a procedure ancillary to surgery. Without contextual (i.e., domain plus syntactic) knowledge, it is difficult to match the phrase to the correct concept. To disambiguate the word "immunology"—which can refer to study of a biological function, a family of laboratory procedures, or a biomedical occupation—Rindflesch and Aronson[13] used a set of rules based on patterns in the enclosing sentence. Scaling up this approach is a daunting task, however; the 1999 UMLS lists 13,688 ambiguous term entries. Other methods, however, that are less labor intensive than manually devising rules (e.g., machine learning) have yet to be explored.

Much research in word-sense disambiguation tends to yield solutions that are highly domain-specific and nongeneralizable. However, Aronson et al.[9] describe a potentially powerful and generalizable approach, by which the contents of the UMLS Semantic Network (every concept has one or more semantic types) might be used to disambiguate homonymous concepts on the basis of the semantic types of adjacent, nonambiguous concepts.

Some acronyms and abbreviations in the UMLS are also words in their own right; e.g., "PEG" for polyethylene glycol and "cAMP" for cyclic adenosine monophosphate. While acronyms in published biomedical literature might be recognized by case, we found that case was too inconsistent to be relied on in medical notes. The UMLS's coverage of common abbreviations is not complete; missing, for example, is "VTach" for "ventricular tachycardia." There is currently no way to query the UMLS data for all instances of abbreviations or acronyms, because such terms are not explicitly flagged.

Methods

The experiment was divided into two phases. The first, *training* phase involved refinement of the concept-matching algorithm (and curation of the UMLS data on which it relied) by the first two authors, using a set of 50 discharge summaries and 50 surgery notes. These were obtained from the Veterans Administration Medical Center in West Haven, Connecticut, and were uploaded into a database table as text ("memo") fields. The notes spanned several specialties; for example, surgery notes spanned ophthalmology, neurosurgery, cardiac surgery, orthopedics, and general surgery.

The training phase was important in enabling us to identify the *range* of conditions under which concept matching could fail or be otherwise problematic. We used two different types of document to test our algorithm over a greater range of medical subdomains. The two documents types also differed significantly in structure. Surgery notes were typically very telegraphic, with sentences conveying the facts

rather than possessing a fully formed grammatical structure. Discharge notes varied widely in structure. Some were terse, whereas others were highly verbose and contained enough explanatory text to be understood by a non-medical reader.[§]

In the second, *test* phase, we tested the algorithm with 24 new documents (12 discharge summaries, 12 surgery notes). An independent expert (the third author) first manually identified concepts in these notes, also recording negation of a concept where present. Subsequently, the third author inspected the output of the concept-identification program for these documents, and performed a failure analysis. The numbers obtained here provided a more realistic estimate of how the matching algorithm would perform in practice.

We first provide an overview of the steps performed in the training phase:

1. A list of stop-words was obtained from an electronic source[8] for use in other steps of our experiment. As will be described, we had to alter this list several times during the course of our experiment.
2. A Microsoft SQL Server database had been previously created to store a relational version of the UMLS 99 Metathesaurus. Only the English subset of UMLS was used. Extra tables were added to this database to store data and results for the present study, and a subset of the UMLS data was created for use in concept matching. We also created a Microsoft Access front end to access the database from our desktop machines over a local network.
3. The notes were preprocessed to remove standard headings (e.g., "diagnosis," "follow-up"). Each note was then written to a text file, which was then processed with a commercial phrase-identification program (described shortly). The program's output, which consists of delimited text fields, was imported programmatically into the database. The imported data were used for two different purposes.

   For the first, we created rich text format (RTF) equivalents of each note. (RTF is a machine-independent format originally defined by Microsoft.) We then programmatically color-coded different parts of the note text on the basis of phrases present in the parsed output. The purpose of the color-coding was to allow easy visual identification of possible problems with the phrase recognizer as well as with our own concept-matching code. The color-coding scheme is described later in this section.

   For the second, the phrases were processed with a concept-finding algorithm to identify matches from the UMLS subset, and matches were written to tables in the database.
4. Finally, the first two authors inspected each note visually along with the matches. Where our algorithm failed to recognize relevant concepts, these were manually added to the matches list and flagged as false negatives. (This study was not concerned with quantifying agreement between authors about which concepts were relevant. Therefore, the rule for resolving differences was that a concept was relevant if either author deemed it to be.) Each automatically matched concept was inspected manually for correctness in the context of its occurrence in the document, and problems were flagged with codes indicating the nature of failure.

We underwent several rounds of iteration. Thus, the output of an earlier stage of the experiment typically revealed shortcomings of the existing strategy and suggested obvious algorithmic or curation refinements.

For the test phase, the third author marked concepts of interest in each document through a macro that operated on the plain-ASCII document after it was imported into Microsoft Word. This macro (associated with a function key) allowed electronic highlighting of selected text by the addition of a yellow background. All the test documents were manually marked up this way and saved to disk. Then, a Microsoft Access program written by the third author opened each document in turn, identified phrases highlighted in yellow, and wrote each phrase, along with the document ID and the byte offset relative to the start of the document, to a database table. The test documents were then processed by the concept-recognition program, which also records byte offsets for matched concepts. The list of manually flagged concepts and automatically detected concepts were then visually compared side-by-side and sorted by document ID and byte offset, so that failure and problems in matching could be detected easily. The original documents were also inspected if the context of a particular phrase in a document had to be determined.[¶]

[§] Part of a discharge summary is illustrated in Appendix 4, Figure 1, which appears as supplemental material to this article in *JAMIA Online*, at www.jamia.org.

[¶] Details of each step, including the concept-matching algorithm, are also provided in Appendix 4, which appears online.

Table 1 ■

The Results of Concept Matching

Match Type	Training Set (100 documents)			Test Set (24 documents)	
	No. of Matches	Percentage	Distinct Concepts	No. of Matches	Percentage
True positive	7,227	82.6%	2,268	1,298	76.3%
Redundant UMLS concept	490	5.6%	209	119	7.0%
Homonym	481	5.5%	127	45	2.6%
UMLS general concept missing	158	1.8%	86	38	2.2%
Concept not in UMLS	127	1.5%	31	42	2.5%
FP, acronym/abbrev	83	0.9%	51	15	0.9%
FN, variant not in UMLS	41	0.5%	16	44	2.6%
FN, inferable by ctx/expert	38	0.4%	12	31	1.8%
FN, acronym/abbreviation/elision	29	0.3%	6	37	2.2%
Concept not useful for indexing	25	0.3%	7	6	0.4%
Too many non-stop-words	25	0.3%	25	7	0.4%
FN, spelling/grammar error	8	0.1%	8	0	0.0%
FN/FP, proper name	10	0.1%	10	19	1.1%
FP, spelling/grammar error	3	0.0%	3	0	0.0%
TOTALS:	**8,745**		**2,859**	**1,701**	

NOTES: The three columns indicate category of match, the number of matches for each category, and the number of distinct concepts matched. FN indicates false negative; FP, false positive. The number of negated concepts in the test set was 110.

Results

The 100 notes used in the *training set* of documents contained 1.12 MB of text, with an average of 11,200 bytes and 1,800 words per note. Discharge notes were distinctly longer than surgery notes. It took an average of 55 seconds to process each note completely and recognize concepts. Much of this time involved database accesses over a local area network; the phrase recognition program, running locally, took less than a second per note. Time requirements might have been reduced somewhat if both the concept-indexing program and the database server had resided on the same machine. The characteristics of the test set are contrasted with those of the training set at the end of this section.

The results of the matching process for both training and test sets are summarized in Table 1. The columns in this table are as follows:

- *The category of match* ("Match Type"). The abbreviations FN and FP indicate false negative and false positive, respectively.
- *The number of matches for each category for both training and test sets, with percentages, after tallying matches for each note.* The total number of matches was 8,745 in the training set and 1,701 in the test set.
- *The number of distinct concepts matched across all notes for each category (training set only).* This is useful for analyzing failures without counting the same instance twice. Thus, the phrase "retrograde cold blood cardioplegia," a concept that probably should be in the UMLS but is currently missing, was seen in several open-heart surgery notes. The total number of distinct concepts matched was 2,859.

All categories other than "true positive" indicate problems either with vocabulary contents and curation or with the algorithm. Our definition of "true positive" was simply that the individual phrase matched to the correct concept, regardless of negation or tense. Therefore, "true positive" may or may not be the same as "relevant." For example, if a user were looking for patients presenting with alcoholism, a match to "alcoholism" would mean nothing per se. However, if the user were looking for all patients screened or interviewed for a history of alcoholism (whether it was actually present), every instance of "alcoholism" might well be relevant.

Match Failures

Failures to match unambiguously can be grouped into three categories:

- *Non-recognition due to the tagging / the noun-phrase method of targeting candidates*. Examples of these failures are spelling and grammatical errors in the text, and proper names. (As previously mentioned, grammatical errors cause subtle errors in the FET's phrase tagging process.) Although not discussed here, an artifact of the noun phrase method is that when a single concept is spread across two or more phrases, the matching process will match to two or more separate general concepts rather than one composite concept. Thus, as discussed previously, in the segment "hypertension is secondary to renal disease," the noun phrase approach would match the concepts "hypertension" and "renal disease" rather than the concept "renal hypertension." This is not necessarily bad, but it means that if a production system were to be consulted by a user searching for the concept of renal hypertension, it would need to consult the MRREL table of the UMLS, find the immediate "parent" concepts for renal hypertension, and expand the user's original query.
- *Problems due to UMLS content*. This category contains redundant UMLS concepts, term variants missing from the UMLS, missing general concepts (where a specific variant is present but the more general form is not), missing concepts, and concepts that are present in the UMLS but not useful for concept indexing.
- *Limitations of the matching algorithm*. Examples are homonyms, acronyms and abbreviations, phrases that are too long for the algorithm, and elided forms ("incomplete" phrases with one or more missing words). Some elided forms do not occur in the UMLS at all, whereas others need domain expertise to disambiguate in the context of their occurrence, as discussed shortly. Acronyms and abbreviations are partly a thesaurus problem; some are present in the UMLS, but others are not.

We discuss these categories in more details below; the numbers and percentages in parentheses refer to the training set. The numbers for the test set are not recapitulated in the text, but salient features are discussed toward the end of this section.

Redundant UMLS Concepts

Redundant UMLS concepts (5.6 percent, 490 matches) were cases in which a phrase or subphrase matched more than one concept even after disambiguation was attempted. Filtering on patterns like NOS (not otherwise specified) and NEC (not elsewhere classified) eliminated many but not all redundant concepts. For example, "spinal tap" and "spinal puncture" are two separate UMLS concepts even though they should be a single concept.

Another problem is noun-adjective variants. For example, "fibrosis" and "fibrotic" are two separate concepts, as are "necrotic" and "necrosis." From the concept-recognition viewpoint, the adjective is merely a variant of the noun form. Other instances include identical concepts with variations in spelling of the preferred term, e.g., "jaundice" and "Jaundice" (uppercase "J"). (The duplicate entries for "jaundice" appear to be due to a curation error; details of the two concepts are almost identical, except that "jaundice" has an associated definition, whereas "Jaundice" has none.) Only the last category is recorded in UMLS's ambiguous-terms list.

From the curation viewpoint, a genuine danger of recording two concepts instead of one is that the related concepts, recorded in the MRREL table of the UMLS, can be inconsistent with each other. Thus, the concept "necrosis" has the siblings "edema" and "gangrene," whereas "necrotic," the adjective form, has the child "gangrene," and "edema" is not associated with it. There is a continuing effort at the NLM to merge and eliminate duplicated concepts. The files merged.cui and deleted.cui, which record the changes made in this respect, are part of every annual UMLS release. This problem should, therefore, progressively abate with future releases.

Homonyms and Term Variants Not in the UMLS

Homonyms (5.5 percent, 481 matches) were described earlier in the Background section; we discuss term variants not found in the UMLS (0.47 percent, 41 matches) here.

In several cases, multiple matches for a phrase could not be disambiguated because the "default" concept that should match a word or phrase, if it stands by itself, was not recorded in the UMLS. For example, "xiphoid" in narrative typically refers to the "xiphoid bone" (a part of the sternum), whereas "flu vaccine" refers to "influenza vaccine." Some of these were abbreviations, such as IVP for "Intravenous Pyelography procedure" and CT for "X-Ray Tomography, Computed," respectively. (Currently UMLS records only "C.A.T." as a term for the procedure, even though "CT" is a more widely used abbreviation today.) Some verb forms of procedures are missing ("cardioverted," "cauterized"); the former verb is also missing from the SPECIALIST lexicon, which is part of the UMLS distribution.

Concepts Not in UMLS and Missing General Concepts

Concepts that are not in UMLS (1.45 percent, 127 matches) and missing general concepts (1.8 percent, 158 matches) were determined by manual inspection, either when a phrase did not match any concept or when it matched a wrong concept. The UMLS depends on its source vocabularies for comprehensiveness, and because vocabulary development has been driven by specific needs, such as publication indexing, diagnosis, and billing, some domains in medicine are under-represented. Thus, the concept of "relocation" as the opposite of "dislocation" (pertaining to a joint) is absent, although the UMLS records relocation of patients and of cardiac valves. Some missing concepts are compound words (e.g., "zygomaticofrontal"), verb forms of medication administration (e.g., heparinize/heparinization, coumadinize, digitalize), and adjective forms of procedures (e.g., Dopplerable). Most specialized surgical instruments are not recorded, nor are many descriptive psychiatry terms (e.g., "hyper-arousal"). Missing general concepts were discussed earlier.

Acronyms, Abbreviations, and Elided Forms

Acronyms resulted in false negatives (0.33 percent, 29 matches) when they were present in nonstandard forms that were not recorded in the UMLS. Abbreviations also caused false positives (0.95 percent, 83 matches) when they were identical to non-abbreviated words (e.g., "RAT," which refers to the animal or to recurrent acute tonsillitis). Elided forms led to false negatives (e.g., "white count" for "total white blood cell count," and "differential" for "differential white blood cell count"). Similarly, while the phrase "cocaine and alcohol dependence" implied the two concepts "cocaine dependence" and "alcohol dependence," the less specific concept, "cocaine," was identified instead. Disambiguation of some elided forms (e.g., "superior thyroid," which could refer either to the artery or to the vein) requires domain expertise. (In the context of carotid plaque-removal surgery, the phrase "superior thyroid" is more likely to refer to the artery.)

Concepts Not Useful for Indexing

Although we programmatically eliminated all suppressable synonyms as well as forms preceded by an acronym, several concepts of marginal utility (0.29 percent, 25 matches) were not removed. Examples were "In Blood" and "Stroke work right." These can be eliminated only through laborious manual curation.

Phrases Too Long for Algorithm (Too Many Non-stop-words)

Twenty-five phrases across all 100 notes were flagged as more than five non-stop-words long. In all cases but one ("chronic post traumatic stress disorder symptoms"), the phrases had conjunctions missing ("left hip open reduction [with] internal fixation") or were poorly phrased ("large left hemisphere MCA distribution stroke"). More important, the concepts embodied in the unprocessed phrase were present elsewhere in the note, where they were successfully matched. We regard this failure rate as acceptable, given efficiency considerations. In other words, the cut-off of five appears to provide reasonable computational efficiency with good coverage for the vast majority of phrases.

In production mode, rather than abandoning such phrases entirely, it might be desirable simply to attempt to match each individual word in the phrase to a concept. Although general matches are less useful than specific ones, they are better than no matches at all. For this study, however, we needed to determine the suitability of our arbitrarily chosen cut-off. (In an earlier stage of the experiment, we used a cut-off of four. This was found to be too low; among other concepts, "non-insulin dependent diabetes mellitus" was unprocessed.)

Spelling Errors and Proper Names

Spelling errors (0.12 percent, 11 matches) caused concepts to be missed as well as spuriously matched if the misspelling was a valid word in the thesaurus (e.g., "ilium" for "ileum"). Proper names (0.11 percent, 10 matches) posed a problem that we have not yet solved. For efficiency, it is desirable to filter out from a document most proper names—like that of the patient, which occurs repeatedly—prior to concept matching. In this way, the problem of trying to concept-match last names that are also words in the thesaurus (e.g., Black, Ward) is also bypassed.

The IBM FET program does a very good job of recognizing proper names. The problem, however, is that certain medically important names (e.g., Alzheimer, Romberg, Charcot) are also eliminated, and concepts containing them would never be matched. The only solution we can think of—manual creation of a list of such names, to be consulted in the preprocessing step—is significantly labor intensive but is probably necessary for production operation. Such an approach must also disambiguate such concept occurrences from instances in which a patient coinci-

dentally has a medically important last name. (For example, when processing each note, a program can access the patient ID associated with it and use the ID to access patient demographic data, including name information.)

Salient Characteristics of the Test Document Set

Table 1 shows that the test set of documents differed somewhat from the training set in that the documents were shorter on average and each document contained fewer concepts. This was because they represented subdomains of medicine with a different frequency compared with the training set. Thus, cardiothoracic, vascular, and neurosurgical conditions were relatively over-represented in both discharge summaries and surgery notes.

The frequency distribution of the types of matches (discussed shortly) also differed significantly between the training and test sets ($P < 0.0001$ by the chi-square test). For example, missing variants of terms, elided forms, unrecognized acronyms, and unrecognized proper names were over-represented.

This is understandable; for example, documents in surgical specialties tend to contain proportionally more proper names that refer to instruments or surgical techniques. The difference in frequency distributions does not affect the validity of the results; the objective of the exercise is to see how the concept-matching algorithm performs when subjected to documents with different characteristics. The relative frequencies of different types of failures are less important than the fact that these failures must be systematically identified and categorized if we are to devise strategies that can address them.

Of the 1,298 "true positive" matches in the test set, 110 (8.5 percent) actually represented negation of a concept (with a condition being absent, denied by the patient, or ruled out). Although our test set of 24 documents is small, this relatively modest percentage seems to indicate that failure to handle negation robustly may not, by itself, make production concept indexing non-viable. This hypothesis needs to be tested with larger amounts of data; it may not apply universally and may well be false in the case of particular concepts that are routinely sought (but infrequently present) in a case.

In our test data, the words "lymphadenopathy" and "complications" were negated three times each; the former may be important for searching. If greater weight is given to documents in which a particular concept occurs more than once, as is likely if the concept is a significant theme of the document, then documents in which the concept is negated will get less weighting, because negation of a concept hardly ever needs to be stated more than once in a note.

Discussion

Although, in our results, the overall incidence of true positive matches (82.6 percent in the training set, 76.3 percent in the test set) appears superficially impressive, it also means that roughly one index entry in five had some problem that would manifest in production mode. In our opinion, accuracy needs to be much higher for concept indexing to be used in production mode. Certainly, a one-in-five error rate would be unacceptable for OCR (optical character recogition) software or for a human typist. Furthermore, concepts are rarely looked for in isolation; usually, a user is doing a Boolean search (e.g., show documents that contain concepts X and Y and Z) or a "vector-model" search (e.g., rank documents by relevance based on these three concepts). If a single concept has only a 0.8 chance of being a true positive, then the Boolean combination of three concepts has a chance of only 0.8^3, or 0.41, of being true positive. In addition, given the caveat that "true positive" is not necessarily the same as "relevant" (because of negation), the proportion of genuinely relevant documents for a given query may be somewhat lower still.

Concept indexing solves some problems but raises others. Previous experiments with it have not always yielded encouraging results. On the basis of experiments conducted with the SMART system, Salton et al.[33] have asserted that indexing with words is superior to indexing on phrases in a controlled vocabulary. Several experiments conducted by Hersh et al.,[24,25,34,35] have indicated that concept indexing with the earlier versions of SAPHIRE was somewhat less effective, with respect to retrieval, than indexing with traditional word-based methods. Although traditional word indexing, as performed with off-the-shelf software, also has its limitations (chiefly with synonyms), word indexing does not make any claims to intelligence. On the other hand, concept indexing, which aims to partly address the problem of the *meaning* of text, implicitly does make such claims. Therefore, users may well react to perceived lapses in concept indexing much more negatively. Our experience also indicates that concept indexing alone is not sufficiently viable to support robust querying of medical record data.

However, approaches that combine concepts in the UMLS (or the MeSH, one of its major components) with word-indexing approaches have been more

encouraging. Thus, the work of Aronson et al.[9] reported a modest improvement with the combined approach than with word indexing or concept indexing alone. Srinivasan's work on retrieval feedback using MeSH terms[16] reported a significant improvement with a combined strategy. In a subsequent paper, Aronson and Rindflesch,[10] while reviewing and validating Srinivasan's work with further experiments, concluded that an optimal strategy would be to combine their own approach (MetaMap) with a retrieval feedback approach.

We envisage a somewhat different approach to integration of concept indexing with word indexing. We propose a user interface that allows query explicitly by words, concepts, or both. If by both, a query would be a two-step process. The UMLS concepts would be matched to the keywords in the user's query and displayed to the user, who could then select the concepts of interest. The retrieval engine would then do both a word search and a concept search, giving greater weight to documents that matched both words and concepts. We have not yet created such an interface or retrieval engine; the task should prove to be an interesting software challenge, and many issues (such as the weighting scheme) need to be resolved.

MetaMap's intensive variant-generation phase yields more sensitivity but less specificity than our own algorithm. In their 1993 paper, Rindflesch and Aronson[12] describe false positives due to variant generation artifacts; thus, "base" (a variant of "foundation") maps to "inorganic chemical" because of the homonym phenomenon. As previously stated, our algorithm is biased toward high specificity for the concept match because of our hypothesis (to be tested in future work) that the accompanying word index might provide the requisite sensitivity.

Conclusions and Future Directions

The present work is concerned with the important prerequisite of highly specific concept matching, without which a concept index is of little use. Our algorithm is, of course, tailored to the UMLS, and some of the problems encountered might not apply to the much smaller controlled vocabularies of other domains.

Many of the categories of false negative, false positive, and ambiguous matches with concept indexing need to be addressed through curation of the UMLS subset used for matching. In 1992, Hersh and Hickam[25] expressed the hope that future editions of the UMLS would be enhanced to be useful for concept indexing. However, because the UMLS has to support varied audiences, enhancements for one audience might be deleterious for another. Two UMLS enhancements—creation of the ambiguous terms/strings tables and the flagging of suppressable synonyms—are aimed specifically at the concept-matching audience, but clearly much more needs to be done. Some issues are highlighted below.

First, for now, ambiguous terms are the Achilles heel of concept indexing. The ambiguous-entries tables in the UMLS do not currently list acronyms and abbreviations. As stated earlier, we have generated a table of terms with non-unique stemmed forms. Although many entries in this list are of the "NOS/NEC" variety, the list also includes abbreviations that are identical to non-abbreviated words. However, this list is so large (49,800 stemmed forms from 132,300 terms, incorporating 111,300 unique concepts) that we have currently been able to access it only programmatically—to look for similar forms, for example, with and without "NEC." We will eventually need to process the list manually or devise clever ways of processing it algorithmically.

One admittedly ad hoc strategy to deal with ambiguous terms is to treat them as "pseudo-concepts," assigning them IDs beyond the range of UMLS concepts proper, and index them with these IDs. (The method of assigning IDs beyond the UMLS range is widely used to maintain local vocabularies, as described in Rocha et al.,[36] for example) The table of ambiguous terms must be available to the query process, so that if the user specifies such terms in the query expression, the program can warn the user that the matches might lack specificity. The alternative approach of Aronson et al.,[9] which seeks to use the UMLS Semantic Network for disambiguation, was cited earlier in the Background section.

Second, to make the existing term list more useful for concept matching, it might be necessary to store the "default concept" against a term if the term is encountered as an isolated noun phrase. For example, CT and IVP by themselves would imply X-Ray Computed Tomography and Intravenous Pyelography, respectively.

Third, in addition to incorporating more concepts for medical subdomains that are currently under-represented in the UMLS (e.g., orthopedics), the vocabulary may need to be expanded by the incorporation of more high-level, general concepts that are currently missing from the UMLS, even though they form parts of composite concepts. The creation of algorithms to

identify potential higher-level concepts from the existing contents of the UMLS is an open problem.

Concurrently, the suitability for concept indexing of many highly composite concepts present in the UMLS needs to be carefully assessed. This applies especially to concepts derived from sources such as ICD-9 and ICD-10. For example, several concepts are very similar, being distinguished from each other by the presence of one or more negations. Thus, we have "acute gastrojejunal ulcer with hemorrhage, without mention of obstruction" vs. "acute gastrojejunal ulcer with hemorrhage and obstruction," and so on. These are formed from the general concepts "acute," "gastrojejunal ulcer," "hemorrhage,"and "obstruction."

The question is whether the general concepts are useful enough by themselves. In medical narrative, the clinical condition codified by the composite concepts would be described over multiple phrases and possibly over multiple sentences, and would hardly ever be matched by a phrase-based approach. One way to curate ICD codes is to go to the original source and use the decimal nomenclature (there are fewer decimals for higher-level, general concepts) as the basis for creation of subsets.

Finally, it would be very useful if future editions of the UMLS explicitly recorded, against each string for a term, whether the term was an abbreviated form (e.g., an acronym) or not. This would greatly reduce the incidence of false positive and false negative matches. Currently, it requires a human curator to perform this interpretation. Unfortunately, the SPECIALIST lexicon, a component of the UMLS distribution that records abbreviations explicitly, is currently too limited in this matter. Many of the problematic abbreviations (e.g., PEG for polyethylene glycol) are not recorded in SPECIALIST.

The authors thank Dr. Michael Hehenberger of IBM Corporation for providing the Intelligent Miner for Text software at no cost. Dr. Betsy Humphreys of the National Library of Medicine provided valuable information about the UMLS, as well as pointers to previous work. The test data was made available through Dr. H. David Stein, Forrest W. Levin and Dr. Joseph Erdos of the VAMC, West Haven, Connecticut. Finally, they thank the National Library of Medicine for making a valuable resource like the UMLS freely available, and making this research possible.

References ■

1. Salton G. Automatic text processing: the transformation, analysis, and retrieval of information by computer. Reading, Mass: Addison-Wesley, 1989.
2. Hersh WR. Information Retrieval: A Health Care Perspective. New York: Springer-Verlag, 1996.
3. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993;32:281–91.
4. Salton G, Wu H, Yu CT. Measurement of Term Importance in Automatic Indexing. J Am Soc Inf Sci. 1981;32(3):175–86.
5. Wilbur WJ, Yang Y. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Comput Biol Med. 1996;26(3):209–22.
6. National Library of Medicine. UMLS Knowledge Sources. 9th ed. Bethesda, Md: NLM, 1999.
7. Porter MF. An algorithm for suffix stripping. Program. 1980;14(3):130–7.
8. Baeza-Yates R, Frakes WB. Information Retrieval: Data Structures and Algorithms. Englewood Cliffs, NJ: Prentice Hall, 1993.
9. Aronson A, Rindflesch T, Browne A. Exploiting a large thesaurus for information retrieval. Proc RIAO '94. 1994:197–216.
10. Aronson AR, Rindflesch TC. Query expansion using the UMLS Metathesaurus. Proc AMIA Annu Fall Symp. 1997:485–9.
11. Divita G, Browne AC, Rindflesch TC. Evaluating lexical variant generation to improve information retrieval. In: Proc AMIA Annu Symp. 1998:775–9.
12. Rindflesch TC, Aronson AR. Semantic processing in information retrieval. Proc 17th Annu Symp Comput Appl Med Care. 1993:611–5.
13. Rindflesch TC, Aronson AR. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. Proc Annu Symp Comput Appl Med Care. 1994:240–4.
14. Salton G. The SMART Retrieval System: Experiments in Automatic Document Processing. Englewood Cliffs, NJ: Prentice Hall, 1971.
15. Callan J, Croft W, Harding S. The INQUERY retrieval system. Proc Int Conf Database Expert Syst Appl. 1992:347–56.
16. Srinivasan P. Retrieval feedback in MEDLINE. J Am Med Inform Assoc. 1996;3(2):157–67.
17. Srinivasan P. Query Expansion and MEDLINE. Inf Proc Manage. 1996;32(4): 431–43.
18. Elkin PL, Cimino JJ, Lowe HJ, et al. Mapping to MeSH: the art of trapping MeSH equivalence from within narrative text. Proc 12th Symp Comput Appl Med Care. 1988:185–190.
19. Cutting D, Pedersen J. The Xerox Part-of-Speech Tagger. Palo Alto, Calif: Xerox Corporation, 1994. Available freely online for downloading as LISP code from: ftp://parcftp.xerox.com:/pub/tagger/tagger-1-0.tar.Z:.
20. Inxight Corporation. LinguistX Platform [product summary]. Palo Alto, Calif: Inxight Corp., 1999. Available at: http://www.inxight.com/Products/Developer/Platform.
21. Spackman K, Hersh W. Recognizing noun phrases in medical discharge summaries: an evaluation of two natural language parsers. ProcAMIA Annu Fall Symp. 1996:155–8.
22. Evans D, Brownlow N, Hersh W, Campbell E. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. Proc AMIA Annu Fall Symp. 1996:388–92.
23. Hersh W, Greenes R. SAPHIRE—an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. Comput Biomed Res. 1990; 23(5): 410–25.
24. Hersh W, Hickam D, Leone T. Words, concepts, or both: optimal indexing units for automated information retrieval. Proc 16th Annu Symp Comput Appl Med Care. 1992:644–8.
25. Hersh W, Hickam D. A comparison of retrieval effectiveness for three methods of indexing medical literature. Am J Med Sci. 1992;303(5):292–300.

26. Gordon H. Discrete Probability. (Undergraduate Texts in Mathematics.) New York: Springer-Verlag, 1997.
27. Aho A, Ullman J. Foundations of Computer Science. "C" ed. New York: WH Freeman, 1995.
28. Hersh W, Leen T, Rehfuss P, Malveau S. Automatic prediction of trauma registry procedure codes from emergency room dictations. MedInfo '98. 1998:665–9.
29. Miller R, Myers JD. Quick medical reference (QMR) for diagnostic assistance. MD Comput. 1986;3(5):34–48.
30. Hersh WR, Leone TJ. The SAPHIRE server: a new algorithm and implementation. Proc 19th Annu Symp Comput Appl Med Care. 1995:858–62.
31. Huff S, Rocha R, McDonald C, et al. Development of the Logical Observations Identifiers, Names, and Codes (LOINC) vocabulary. J Am Med Inform Assoc. 1998;5(3):276–92.
32. Nadkarni PM. Concept Locator: a client-server application for retrieval of UMLS Metathesaurus concepts through complex Boolean query. Comput Biomed Res. 1997;30: 323–36.
33. Salton GB, Buckley C, Smith M. On the application of syntactic methodologies in automatic text analysis. Inf Proc Manage. 1990;26(1):73–92.
34. Hersh W, Hickam D. A comparison of two methods for indexing and retrieval from a full-text medical database. Med Decis Making. 1993;13(3):220–6.
35. Hersh W, Hickam D. Information retrieval in medicine: the SAPHIRE experience. MedInfo '95. 1995:1433–7.
36. Rocha RA, Huff SM, Haug P, Warner HR. Designing a controlled medical vocabulary server: the VOSER project. Comput Biomed Res. 1994;27(6):472–507.