# Recognizing New Medical Knowledge Computationally

Stuart J. Nelson, M.D., Department of Internal Medicine
Medical College of Georgia, Augusta, Georgia

William G. Cole, Ph.D., Department of Medical Education
University of Washington, Seattle, Washington

Mark S. Tuttle, Nels E. Olson, and David D. Sherertz,
Lexical Technology, Inc., Alameda, California

*Can new medical knowledge be recognized computationally? We know knowledge is changing, and our knowledge-based systems will need to accommodate that change in knowledge on a regular basis if they are to stay successful. Computational recognition of these changes seems desirable. It is unlikely that low level objects in the computational universe, bits and characters, will change much over time, higher level objects of language, where meaning begins to emerge, may show change. An analysis of ten arbitrarily selected paragraphs from the Medical Knowledge Self-Assessment Program of the American College of Physicians was used as a test bed for nominal phrase recognition. While there were words not known to Meta-1.2, only 8 of the 32 concepts new to the primary author were pointed to by new words. Use of a barrier word method was successful in identifying 23 of the 32 new concepts. Use of co-occurrence (in sentences) of putative nominal phrases may reduce the amount of human effort involved in recognizing the emergence of new relationships.*

## INTRODUCTION

Medical knowledge is constantly changing and growing. Such a statement invokes little interest or dispute. We can note that much concern in medical education is expressed toward teaching the skills necessary for ongoing continuing education. We recognize that one of the major problems facing any information retrieval system or knowledge based system is that of adapting to the constant changes seen in medical knowledge. No system can succeed without addressing this problem. Yet we rely almost entirely on noncomputational methods and tools for the early steps in addressing this problem.

As a knowledge based system designed to provide tools for improving information retrieval, the UMLS must of necessity be concerned with this update problem. Particularly when considering the development and maintenance of a thesaurus of biomedical naming, we need to ask how much the existing naming systems evolve, and how rapidly do the naming systems adapt to new knowledge? The design of a distributed maintenance environment [1] would be guided by knowing the rate and behavior of this knowledge change. A second and equally compelling question is whether new knowledge reflected in names of new things, or in new relationships between things already named?

Our concern is with whether what is new can be recognized computationally from current techniques of processing of text. Will we be able to recognize important new concepts computationally? If our knowledge evolves largely as naming new concepts, do these concepts consist largely of words that have previously been involved in medical names, or are they innovations, either neologisms or words not previously used in biomedical contexts? Can we derive an estimate of the magnitude of the problem? How much human judgment will be required to maintain an up-to-date thesaurus of biomedical names?

Consider as one example "platelet Fc receptor." While the name itself seems to be new, and to name a recently recognized entity, the words themselves are not. As a counterexample, consider "taxol". It is a new name for something newly discovered. How many of our changes occur in each of those ways?

Methods for recognizing computationally that a new relationship between two well-known concepts is being described need to be developed as well. Changing or new patterns of co-occurrence of concepts might be useful in recognizing new knowledge about relationships. That is, suppose two concepts occur in the same sentence. Such proximity implies that there is some significant relationship between these concepts. If such proximity had not been previously noted, this might signify emerging knowledge about the relationship between the two concepts.

| COMPONENT | EXAMPLE | TIME VARIABILITY | CONTRIBUTION TO MEANING |
|---|---|---|---|
| Bits | '0', '1' | None | None |
| Characters | a, B, 2 | Little | None |
| Morphemes | oculo | Modest | Beginning |
| Words | flare | Moderate | Basic semantic components |
| Nominal Phrases | serum sodium | ? | Core concepts |
| Sentences | . . . | Widely | Complex concepts,<br>Relationships between simple concepts |
| Paragraphs | . . . | | Extended thought |

Table 1. A simple computational view of the structure of language.

Computational, i.e., syntactical, components of language are listed in increasing order of organization. Meaning emerges at higher levels of organization. Bits are not likely to change in their definition in our lifetimes. By themselves they contribute little to meaning. Looking at a series of zeros and ones is unlikely to be edifying to us. Characters may change somewhat over time, e.g., if 16 bit characters are introduced in order to accommodate diacritical marks in other languages, etc. By themselves, their presence does not give something meaning. It is only when characters are clumped together in familiar patterns that we recognize as morphemes that we can begin to impute a meaning. Words provide our first basic building blocks of language, but a given word can participate in many names of the core concepts of medicine (nominal phrases). Acronyms and abbreviations can be thought of as words formed without the traditional morphemic derivation. In our present culture, they may play a special role as harbingers of new knowledge. Nominal phrases, short phrases which name the concepts central to our concern, are the concern of naming systems., e.g., MeSH, SNOMED, ICD-9, and CPT. The ability to recognize these nominal phrases reliably in an automatic fashion is an important goal in language processing. Their variability over time, with the development of new knowledge, is a major concern. Sentences express relationships between concepts and define new complex concepts. Paragraphs, chapters, and books participate in the expression of higher levels of extended thought, including relationships among relationships, the clarification of ideas not easily expressed in single sentences, etc.

## A Simple Structure For Language

In order to look at this subject it is helpful to define a structure of language as represented computationally. Language can be represented as a hierarchy of objects, analogous to the hierarchy of objects in the natural world [2]. We can represent this hierarchy as shown in Table 1. What does this hierarchy of language tells us about the basic processibility of language by machines? A prediction would be that each higher level of organization of language would be less processible. That is, it is easier to recognize a word than a nominal phrase. It is easier to recognize a nominal phrase than to attribute a meaning to it. Still more difficult is the recognition of relationships between concepts. It is analogous to interpreting the meaning of a simple declarative sentence of the form "A relates to B."

## The Nominal Phrase Problem

Words have been the principal objects addressed in computational processing of language over the past few years. However, they participate in multiple meanings, and a reader is highly dependent on context to decode the meaning of a given word. Projects aimed at inferring context computationally (e.g., the Lex Project at UCSF [3]) have not been widely adopted. A simple noun phrase naming a single concept, which may be a single word (e.g., "electron") or several words in combination, appears to be the smallest piece of language which has a given distinct meaning. To distinguish this type of noun phrase from all of the possible noun phrases we will call these phrases "nominal phrases."

Most of the naming systems incorporated in the Metathesaurus have used these short nominal phrases to name concepts central to biomedicine. While some concepts may not be expressible by a single short phrase, Zipf's law claims that, if in fact they are central to the universe of discourse, at some point they will be defined in terms of a short phrase.

How can we define and recognize these nominal phrases? Some rules for recognizing a nominal phrase include *a*, that the phrase is "short," *b*, that the phrase names something in biomedicine, *c*, that the phrase does not attempt to distinguish the particular item named from other items with the same name (this rule generally eliminates a lot of difficult noun phrases, including many of those containing verbal clauses, e.g., "the man who built the Eiffel tower"), *d*, that modifiers are in general omitted, and *e*, that the phrase is at the basic level of discourse. With these rules, it is assumed that we as humans can look at a short phrase and determine if it names a concept. The problem has another subtlety, it is only by convention that we can agree that a given phrase names a concept of sufficient import to be of interest.

The same string of words (or characters) may represent more than one name. The problem of polysemy, where, for example, "Arizona" can be both the name of a state and the species name of a bacterium, is a case where the very distinct meanings are represented by the same string. The reader must resort to higher levels of organization (that is, contextual information) to determine exactly which of the possible meanings is being named.

More subtle is the problem of gradation of meaning of a given nominal phrase. Metonymy is so common that we often don't recognize that we are using it. For example, "serum sodium" can mean the ions circulating in the blood, the test of the concentration of those ions circulating in the blood, as well as the result of that test.

In these situations, where the exact meaning of a nominal phrase is not self-evident, we humans use contextual clues to help us decode the meaning expressed in a given phrase. Computationally, this particular task is a stumbling block. It is only in limited domains that there has been much success in defining context.

Nevertheless, the computational recognition of nominal phrases will be an important step forward in our ability to process large volumes of text. The concept-matching algorithm of Sapphire [4] can be used if one has a thesaurus of known nominal phrases. And it is possible that other matching algorithms can be used. The limitation of such an approach would be that it is incapable of processing text and suggesting new candidates for consideration. Vries [5], using large volumes of textual material from patient charts, and a "semantic net expansion", has been able to identify potential candidates. The limitation of his approach has been its dependence on high volume occurrences of the nominal phrase.

A different method of computational recognition of potential nominal phrases has been examined here. This method, described by Tersmette, et al, [6] uses an extended list of barrier words to "chunk" the text. That is, from a corpus of textual material a list of the most frequently occurring words is made. Those words judged to be major potential contributors to medical meaning are removed from the list. A chunk is generated from text by finding all words occurring in sequence, uninterrupted by barrier words or punctuation marks. The hypothesis is that these chunks will include many, if not all, of the nominal phrases of interest.

**Changing Knowledge**

How might changes of knowledge be understood in this model of the computational structure of language? One hypothesis might be that new knowledge is predominantly discovery of new relationships between existing concepts. If so, the language of medicine as reflected in the nominal phrases would change little. The few concepts that are new will be more abstract, nominalizations of verbs, rather than naming new physical or chemical objects.

If, on the other hand, the discovery of new knowledge is predominantly that of finding new objects, and the relationships remain similar to previously described relationships, new knowledge will be largely represented as new names for things. The degree to which these new names can be recognized is problematic. If new words are employed in these names of new concepts, the occurrence of words (new to our lexicon) would be a signal that new concepts were being introduced and discussed. On the other hand, if new names use words previously used in our lexicon, recognition of these new nominal phrases will require phrase recognition and review. We postulate that the nominal phrases in known naming systems are unlikely to reflect recent changes in naming because of new knowledge, and thus represent a static view of medical knowledge.

## METHODS AND RESULTS

The electronic version of the Medical Knowledge Self-Assessment Program (MKSAP) of the American College of Physicians [7] was used as a test bed for analysis of the language of emerging medical knowledge. The UMLS Metathesaurus (Meta-1.3), a thesaurus of biomedical concepts used in one or more naming systems, was one source used for naming information. An electronic copy of the text of Scientific American Medicine (SAM) [8] was used as a source as well. Meta-1.3 and the entire MKSAP had 10,801 non numeric words in common, using the UMLS standard code for word detection and removing the "words" which were strictly numeric. There were a total of 87,965 words in Meta-1.2, and a total of 24,344 in MKSAP. A large portion of the unrecognized words in the MKSAP were names of authors and of journals, including abbreviations.

Ten paragraphs were chosen arbitrarily from the text of the MKSAP. One paragraph was chosen

| PHRASE SELECTION METHOD | MATCHES TO META-1.3 |
|---|---|
| Manual (229 unique phrases) | 116 (51%) |
| Chunk List 1 (471 chunks) | 128 (27%) |
| Chunk List 2 (436 chunks) | 130 (30%) |
| Chunk List 3 (402 chunks) | 132 (33%) |
| Common to Manual and List 1 (146 phrases) | 85 (58%) |
| Common to Manual and List 2 (148 phrases) | 84 (57%) |
| Common to Manual and List 3 (144 phrases) | 85 (59%) |

Table 2.
Results of matching putative nominal phrases from MKSAP to Meta-1.3

from each of ten areas of discussion of the MKSAP. Introductory paragraphs were not chosen, and the paragraphs were expository. Words from these paragraphs, as defined by the UMLS standard code for word detection, were matched against the words occurring in the Metathesaurus. Of the 760 words in the 10 paragraphs, 608 words matched to Meta-1.2.

**Finding New Nominal Phrases**

Following the rules for nominal phrase recognition listed above, each of the ten paragraphs was analyzed by the first author, who selected those phrases which appeared to have relevance. The nominal phrases selected from those paragraphs were matched to Meta-1.3 concept names. Results of these matches are shown in Table 2.

The barrier word method as outlined by Tersmette, et. al, [6] was used to generate a sample of "chunks" of text which were candidates for consideration as possible nominal phrases. Three lists of barrier words were created using the entire MKSAP and SAM as the text sources. These lists were generated by finding the words of highest frequency usage. Those words which appeared likely to have some content information (e.g., "therapy") were removed from the high frequency list to form the barrier word list. List 1 was created from the 250 words of highest frequency occurrence, and was a total of 181 words. List 2, 337 words, was created from the 500 most frequently occurring words, and List 3, 619 words, from the 1000 most frequent.

Using each list, chunks (Chunk Lists 1, 2, and 3 from Barrier Word Lists 1, 2, and 3, respectively) were generated from the ten paragraphs of MKSAP. Each of the chunks was treated as a phrase, and matched against Meta-1.3. Results of that matching are shown in Table 2.

Of the 229 concepts identified in the MKSAP sample, 32 (14%) of them were new to SJN. Two of these concepts had names which were present in Meta-1.3 The other 30 did not match to Meta-1.3. Six of the concepts were single words, five of those 6 were acronyms. Only 8 of the 32 were pointed to by the occurrence, in the nominal phrase, of a word not previously seen in Meta-1.3.. A total of 23 of the new concepts occurred in Chunk list 1; 22 of the concepts could also be found in Chunk list 2; and 21 could be found in Chunk List 3.

Extrapolating, on the basis of the percentage of chunks in the 10 sample paragraphs, to the number of concepts in the MKSAP suggests that there are around 16,400 concepts discussed in the MKSAP. Of these concepts, approximately 2,200 will be new.

**Finding New Relationships**

Preliminary work indicated that out of 103,107 chunk-occurrences in MKSAP, and 552,412 chunk-occurrences in SAM, 414,493 of those occurrences were of chunks which occurred in both SAM and MKSAP. 31% of the chunk-occurrences in MKSAP were of chunks that could be matched to Meta-1.3. 97% of the chunks appearing in MKSAP that matched to Meta-1.3 also appeared in SAM.

The regularities of the text (after we had removed the references) in MKSAP and in SAM allowed us to separate, with a high degree of confidence, the text into sentences. Defining putative nominal phrases as those chunks, occurring in MKSAP, in SAM, and matching to Meta-1.3, we created a list of pairs of putative nominal phrases which co-occurred in one or more sentences in SAM (thus, a co-occurrence). The appearance of a co-occurrence was used as a marker for the presence of a relationship between the concepts named by those phrases. There were a total of 114,986 unique co-occurrences in SAM, and 31,205 unique co-occurrences in MKSAP. 17,641 (57%) of the co-occurrences in MKSAP did not occur in SAM.

There were a total of 187 co-occurrences present in the sample 10 paragraphs of MKSAP Those not in

SAM were examined to ascertain if the co-occurrences indicated an important, potentially new, relationship. Each of the 85 co-occurrences was classed into one of four categories: trivial and obvious (7), new and important (27), important but well-known (36), or coincidental (15).

## CONCLUSIONS AND DISCUSSION

The finding that 14% of the concepts in the sample paragraphs were new to the primary author suggests that MKSAP was a good source for looking for new biomedical concepts. If there were many more than that 14%, the MKSAP might become extremely difficult to understand. A framework of understanding using more well-known concepts must be built for a reader to understand these new concepts. The fact that few of those concepts were in existing naming systems tends to confirm our hypothesis that the naming systems will lag behind the development of new knowledge.

Most of the new concepts are formed from words previously used in nominal phrases. Looking for new concepts by looking for new words may not be a successful strategy, except for acronyms and other abbreviations. Recognizing them may be a helpful technique in identifying new concepts.

The barrier word technique of identifying "chunks" as potential nominal phrases seems to be useful for reducing the effort involved in recognizing concepts in a corpus of textual material, and warrants further investigation. As many as 50% or more of the chunks identified were recognizable phrases which matched the Meta-1.3 or the list of manually identified concepts. When chunks could be found in more than one source, the likelihood that this chunk represented a concept increased. While even with refinements the barrier word technique may not ever be successful in identifying all possible concepts, it may be helpful as a "quick and dirty", empirical technique to use when trying to identify the concepts in a large corpus of text.

Computing co-occurrences of putative nominal phrases identified by the barrier word technique may have benefit in attempting to recognize new relationships.

Although the value of concept-based (or nominal phrase-based) indexing of textual material remains to be demonstrated (for example, see [9] and [10]), many of us believe it will assist in achieving higher precision without sacrificing recall. Computational means for recognizing nominal phrases remains important, not only for retrieval of knowledge from text, but also in identifying important patient attributes from textual records. It appears possible that computational tools can be used to identify emerging new and important concepts in medicine. None of the methods is foolproof, but they may assist in lowering the amount of human effort involved.

## ACKNOWLEDGMENTS

We wish to thank Frank Davidoff, MD, FACP, Robert Spena, DSW, Thomson Kuhn, and the American College of Physicians for allowing us to have an electronic copy of the Medical Knowledge Assessment Program.

References

[1.] M. Tuttle, W. Hole, S. Nelson, R. Irons, D. Sherertz, N. Olson, O. Munist, W. Sperzel, M. Erlbaum, L. Fuller. Distributed Metathesaurus Enhancement: How Might It Work? AMIA 1992 Spring Congress, p. 52.
[2]. M.S. Blois. Information and Medicine. University of California Press, Berkeley, 1984
[3]. W.G. Cole, P.A. Michael, J.G. Stewart, M.S. Blois. Automatic Classification of Medical Text: The Influence of Publication Form. In: R.A. Greenes (ed.) Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care. IEEE, New York, 1988:196200
[4]. W.R. Hersh, D.H. Hickam. A Comparison of Retrieval Effectiveness for Three Methods of Indexing Medical Literature. Am J Med Sci.. 303:292300, 1992
[5]. J. Vries. Discussion presented at UMLS Contractors Meeting, National Library of Medicine, Bethesda, March 23, 1993.
[6]. K.W.F. Tersmette, A.F. Scott, G.W. Moore, N.W. Matheson, R.E. Miller. Barrier Word Method for Detecting Molecular Biology Multiple Word Terms. In: R.A. Greenes (ed.) Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care. IEEE, New York, 1988:20711
[7]. Medical Knowledge Self-Assessment Program IX. American College of Physicians, Philadelphia, 1991
[8} E. Rubinstein, D. D. Federman [eds] Scientific American Medicine. Scientific American, Inc. New York, 1993
[9] D.B. McCarn, C.M. Lewis. A Mathematical Model of Retrieval System Performance. J Am Soc Inf Sci. 41:495500, 1990
[10]. S.M. Humphrey. Indexing Biomedical Documents: From Thesaural to Knowledge-based Retrieval Systems. Artificial Intelligence in Medicine. 4:34371, 1992