The Semantic Structure of the UMLS Metathesaurus

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

Lloyd F. Fuller, Mark S. Erlbaum, Mark S. Tuttle, David D. Sherertz, and Nels E. Olson
LTI UMLS Group
Lexical Technology, Inc., Alameda California

ABSTRACT

Meta-1.1, the UMLS metathesaurus, represents medical knowledge in the forms of names of concepts and links between those concepts. The representations of the semantic neighborhood of a concept can be thought of as dimensions of the property of semantic locality and include term information (broader, narrower, or otherwise related), the contextual information (parent-child, siblings in a hierarchy), the semantic types, and the co-occurrence data (links discovered empirically from concepts used to index the medical literature.) The degree of redundancy of each of these dimensions was investigated by reviewing the extent of multiple presentations of concepts which appear as related to a given concept. The degree of overlap was surprisingly small. While the co-occurrence data finds some of the links represented by other dimensions, those links are but minute fractions of the vast amount of co-occurrence derived links. Because parent-child relationships are often subsumptive (or categorical) in nature, it might be expected that siblings usually share the same semantic types. While true in the aggregate, the wide variance in percent of types shared may reflect the intended usages of the source vocabularies. Noun phrases were extracted from the definitions of 40 concepts in Meta-1 in order to assess systematically the coverage of important concepts by Meta-1, and to assess whether the links between these definitional concepts, which may have special value, and the concept being defined were indeed present. Out of 161 of these definitional concepts, 29 were not represented in Meta-1, and 37 of those represented in Meta-1 had no direct link to the concept they were defining. 95 of the definitional concepts were found to be directly linked to the concept they were defining. A program (a name-server) which could find the entry or entries of interest to a user, given a term and an indication of the intention of the user, could conceivably be developed to exploit the semantic structure presently in Meta-1. Strategies for a name-server might be different if the term sought had similar or different semantic types as the entered term. The lack of complete coverage of definitional concepts, and the lack of important links indicate that efforts should continue to expand the coverage of the Metathesaurus and to search for creative ways of recognizing the significant links between concepts.

INTRODUCTION, MOTIVATION, AND DEFINITION OF THE PROBLEM

The UMLS Metathesaurus contains a wide variety of semantic information about biomedical concepts. While the need to understand the syntactic structure of the metathesaurus is unquestioned, we assert that understanding of the semantic structure of Meta-1.1 will be vital to its successful exploitation in a complete system. Since the aim of the UMLS is to facilitate the access to electronically retrievable biomedical knowledge sources of many types [1], and thus to facilitate collaboration, understanding of this semantic structure will be essential to collaborative work using UMLS tools. Furthermore, the degree of completeness and consistency found in the semantic information in the metathesaurus will be fundamental to meeting user's expectations in any such system.

What does the semantic structure of the metathesaurus consist of? The semantic content of the metathesaurus is made up of concepts, their names and the relations between concepts. The semantic structure is the model of the real world that these names and relations present. Our definition would include (1) the semantic network and the relationships permitted by the semantic network, (2) the hierarchical information represented by each of the source vocabularies, (3) the representation of the semantic neighborhood of a concept, and (4) any labelled links between concepts (whether hierarchical or not). Not included in this definition are such syntactical considerations as the structure of the database, syntactic categories to which terms have been assigned, character sets, and how a given term for a concept comes to be known as the preferred term. Previous discussions of aspects of the semantic structure and content of Meta-1.1 have included a description of the scope and structure of the UMLS semantic network [2], as well as a description of the dimensions of semantic locality expressed in Meta-1 [3].

What will understanding of the semantic structure enable us to do? If we have a reasonable notion of how to find the **name** of an object (in the real world) in the metathesaurus, our use of the metathesaurus would be facilitated. Consider a program which, when given a term by a user, and given a source vocabulary (or, perhaps equivalently, an expected use for the term, or, as another equivalent indicator of intention, the users outlook),

0195-4210/92/$5.00

returns a set of candidate terms from the vocabulary of choice (determined by this intentionality), together with some indication of the relation of the candidate terms to the term entered. We could call such a program a name-server. Which of the dimensions of semantic locality, or aspects of the semantic structure, would it be most worthwhile for such a name-server to pursue? Should the name-server pursue related terms of related terms (even if we know that "relatedness" is not a transitive function), and thereby risking sacrificing relevance for the sake of getting something, or should some other dimension of locality be investigated? How does one develop a "distance measure" of locality that can translate these dimensions into something comparable?

The Unified Medical Language System (UMLS) Metathesaurus (Meta-1.1) is organized by the property of semantic locality. The dimensions of semantic locality in Meta-1.1 include semantic types, term information, contextual information, and co-occurrence data.[2] The semantic types and the permissible relationships between types, constitute the semantic network. While some of the links between concepts, notably several thousand of the links between child and parent in MeSH, have been labelled, in most of the cases where links exist, the relationship is not labelled.

In this way the metathesaurus is quite different from many attempts at knowledge representation [4,5]. Meta-1.1 is not a "fully-instantiated semantic network." There is much semantic information available in Meta-1.1, but in a somewhat different form, one that seems currently not to be well understood or exploited. While the metathesaurus is an ontology of biomedicine, or is evolving into one (perhaps someday to be authoritative), it does not, as many ontologies do, depend on a semantic network as the principal means of representing objects and relationships within it. Further, it does not require a knowledge representation language to express all of this structure. It represents (1) how things are named in medicine (thus what objects and events are significant in biomedicine), and (2) links between one thing and another. If the relationship between two things has been categorized, that categorization of the relationship is recorded as well.

How effectively does this semantic structure represent the real world, and how much contribution will this structure make to the success of the UMLS knowledge sources? Such a question may be impossible to answer in a general way. As a first step, we might recognize the problem as one of completeness and consistency. Are the important concepts of biomedicine in the metathesaurus? Are the important relationships represented? These questions might be answered when talking about completeness. Are the expectations, generated by finding that some relationship is represented (say between A and B), that another relationship should be represented (the one between B and C), being fulfilled? That is, is the metathesaurus consistent in the degree of granularity of representation of relationships?

While these low-level rules, that if a concept is important you can find it, and if a relationship is important it will be there, can provide a beginning to assessment of the effectiveness of the metathesaurus, the global question of the efficacy of the semantic structure cannot be fully assessed without considering the intended use in a given project. The additional contribution made by any one aspect of the structure, and the value of that structure can only be assessed in software developed for a given purpose. Yet exploration of the knowledge represented may be worthwhile, if only in motivating other potential user/developers to consider how they might creatively exploit these dimensions of semantic locality.

We are attempting to explore the semantic structure of Meta-1.1 by performing a number of explorations of degree of overlap of the dimensions, of the redundancy of represented links, and exploring whether concepts known to be intimately linked (from the definitions) to a concept are to be found in the semantic neighborhood. In addition to facilitating the use of the Metathesaurus, the understanding gained from these explorations may aid in planning the continued improvement of the metathesaurus. For example, should more effort be placed in adding new concepts (into this structure), or does the structure need enhancement to achieve a sufficiently useful representation of the semantic neighborhood? Most of the semantic information in Meta-1.1 is that which has been represented in a source vocabulary. (The co-occurrence data is an exception, having been calculated empirically from the MEDLINE tapes.) Would future versions of the Metathesaurus be enhanced by adding relationships not represented in vocabularies, or by adding more concepts?

A high level question, particularly when considering the problem of distance measures, is to what degree these dimensions of semantic locality are orthogonal, or to what degree they overlap. Is carrying all of this information about each concept redundant, or are each of these dimensions different in the types of relationships between two concepts that they might represent? It is not immediately clear whether orthogonality may hinder or help with the problem of providing a distance measure. Some degree of redundancy or overlap may provide a desirable robustness in representation, insuring that closely related terms are not neglected or otherwise lost in the noise of many terms being represented as related.

A second high level question is whether the represented links represent the semantic neighborhood sufficiently. If, for example, concepts which might naturally be thought to be closely related cannot be located in the same semantic neighborhood, then the representations may not be achieving their desired goal. Further additions to the metathesaurus, and setting priorities for metathesaurus expansion, might be predicated on how well the semantic neighborhoods of certain concepts of larger interest are represented. For example, are clinical findings, their links to diseases and to body structures, sufficiently well represented in the

metathesaurus? If not, then the addition of clinical terminologies which do represent these links well might be a high priority addition to the metathesaurus. The strength of the metathesaurus will be found not only in how well it lumps things which are similar to each other together in a neighborhood, but in how well it shows the other significant relationships, and how consistently it does so. If Mycobacterium tuberculosis is linked to tuberculosis, should not Clostridium difficile be linked to pseudomembranous enterocolitis?

EACH OF THE DIMENSIONS OF LOCALITY IMPLIES A RELATIONSHIP

Each of the dimensions of semantic locality can be thought of as either directly stating or implying relationships of a given concept with other concepts in Meta-1.1. While term information represents information about relationships between strings of characters, such as lexical variant or synonymous relationships, it also lists concepts (entries) considered related in the Reviewed Related Term (RRT) field. In some instances these relationships are labelled, indicating that the two concepts are similar, with one being broader in meaning (but otherwise similar) than the other. Contextual information shows where the concept occurs in a source hierarchy, together with its parents, and siblings. The relationship with a parent is often that of an "is-a" relationship, but some other vertical relationships have been labelled as well. While not entirely true, it is frequently useful to think of non-labelled parent-child relationships as being of some type of relationship which involves subsumption. That is, in some sense the parent concept is broader in meaning than the child. Co-occurring data do not have any label on the link between concepts; the fact that two concepts have both been used to index the same article implies an empirically discovered relationship. Two entries with the same semantic type have an implied relationship, that of "similar to".

A NOTE ABOUT DEFINITIONS

It might be helpful to review what we mean by several terms. We frequently use the word "concept" interchangeably with the word "entry", as the fundamental precept of Meta-1.1 is one concept-one entry. A link is an indication of a relationship between two concepts, whether or not that relationship between two concepts has been characterized. "Term" means one name (note that there may be several) of a concept, whether or not that term is itself in the Metathesaurus.

METHODS

A first step towards investigating the degree of redundancy in representations of links between concepts was taken by calculating the overlap between term data, co-occurrence data, and contextual data. The term data represents not only synonyms, but also related concepts. A link between two concepts was counted as appearing in the term data if the second concept appeared as a reviewed related term (RRT) of the first concept. (Because synonymy and lexical variance represent relationships between terms in the same synonym class, they do not appear as links to other concepts.) Pairs of identifiers (MC#) were written to a line, each line thus represented a link. Duplicate lines were removed.

A similar process was used with the co-occurrence data; lines consisting of MC# pairs, representing links between concepts, were generated, and duplicates removed. For contextual data, (information about the occurrence of the concept in a source hierarchy), child-parent links between terms were extracted, and translated into links between MC# pairs, with elimination of duplicate lines. A similar process was used for sibling data. After joining together the results of these steps, bidirectional relationships were created, and duplicate lines removed. After generation of these files, a simple *comm* command found those lines which were duplicate, and thus represented redundant links.

In order to evaluate the degree of overlap between semantic type assignments and contextual data, a somewhat different methodology was used. For each position in a hierarchy, the siblings, and the types assigned to each concept, were tabulated. The proportion of members of each sibling group having a given semantic type could then be calculated and averaged.

A third investigation attempted to discover if noun phrases in definitions of entries represented concepts in the Metathesaurus, and if those concepts were linked to the entry from whose definition the concept had been derived. Forty concepts were chosen, twenty of which were diseases (where one would expect the links to be most dense) and another twenty were not. Entries were chosen by finding the first entry having a definition and beginning with two letters chosen in advance on the basis of an acrostic. For each two letter key, two entries were found, one a disease and the other with some other semantic type. Noun phrases which appeared to represent significant concepts were extracted from the definition. Subsequently the MetaCard browser of Meta-1.1 was used to search for the concept represented by the noun phrase, and to attempt to discover if a link between that concept and the entry was present. This methodology appeared to be successful in identifying direct links, but may have missed indirect (related of related) links, largely because of the combinatorial explosion implied by reviewing all of the co-occurring terms of both the entry and the concept represented by the noun phrase.

RESULTS

The first investigation into the degree of redundancy in Meta-1.1 revealed the following: There are 427,079 links between entries represented in contextual data (CXT). There are 33,347 links between entries represented in the term (RRT) data. There are 4,881,189 links between entries represented in co-occurrence data (COT). There are 1374 links between 2 concepts

TABLE 1
Location Of Occurrence of Definitional Concepts within Semantic Neighborhood of Concept

GROUP	HT	STY	SIB	RT	COT	NL	NIM	TOTAL
Disease	17	0	2	6	23	24	13	85
Non-disease	14	5	2	8	18	13	16	76
Totals	31	5	4	14	41	37	29	161

Legend:
HT = Hierarchical term (Concept present, as ancestor, in one or more hierarchical trees of the entry)
STY= Semantic type (Concept in definition was the semantic type of entry)
SIB = Sibling (Concept present as sibling of entry)
RT = Related Term (Concept present as reviewed related term of entry)
COT = Co-occurring Term (Concept present in MEDLINE co-occurrences of entry)
NL = No direct link between concept identified by noun phrase and entry
NIM = Concept identified by noun phrase not found in Meta-1.1

represented in both CXT and RRT data. This overlap is 4% of the total number of RRT links, and 0.3% of the CXT links. There are 105,974 links between two concepts represented in both COT and CXT data. This overlap is 25% of the CXT links, and 2% of the COT links. There are 4247 relations between 2 concepts which are represented in both RRT and COT. This overlap is 12.7% of the RRT links, and 0.09% of the COT links. Only 856 links between 2 concepts were represented in all three sets of data.

The second investigation, into the degree to which the semantic types of siblings in a hierarchy were the same, found that to a large degree, the semantic types of siblings in a hierarchy were the same. The average proportion of siblings with a common semantic type was 0.705. The variance of that average, over the 7316 sibling groups, was 0.128. The proportion of types in common varied widely depending on the particular tree, ranging from 51% to 96%. Not surprisingly, in organism classification trees the type assignments tend to be very predictable, and with a high proportion (95%) in common. In trees where the concepts are less central to biomedicine, the proportion of type assignments in common appeared to be less. From the perspective of types, it seemed that some types (e.g., "virus") were almost always assigned to every member of the sibling groups that any one concept given that type participated in; other types tended to occur only sporadically in sibling groups (e.g., "experimental model of disease" was in common in a sibling group only 38% of the time.)

85 noun phrases were identified in the definitions of the twenty diseases. Of these, 48 (or 56%) of the concepts represented were found as direct links to the disease. In the non-disease category, 76 noun phrases were identified in the definitions of the 20 concepts, and 47 (61%) of them were found as direct links to the entry. In the diseases 13 (15%) of the concepts represented by the noun phrases could not be found in Meta-1.1, for non-diseases the same number was 16 (21%). In both the disease and non-disease category, the co-occurring data provided the greatest number of direct links. For the most part, the concepts with direct links found in the co-occurring data were concepts where the co-occurrence with the entry had occurred frequently. The results are presented in tabular form in Table 1.

DISCUSSION

While the COT links represent non-trivial fractions of the RRT and CXT links, in fact the preponderance of those relations are not represented as co-occurrence links. While, at first inspection, it does not appear to be true that those co-occurrences which occur in high frequency are represented by the RRT or CXT relations, this hypothesis needs further evaluation and testing. The high frequency COT links certainly represent important links between entries in Meta-1.1, links which have been derived from empirical data. Inspection of the links confirms their importance; the fact that two concepts have not been identified as related by the sources used in making the Metathesaurus or have not been named as belonging to the same hierarchy does not make them any less significant.

Every knowledge representation, including the semantic network of Meta-1, has its own particular view of the world. The idea that siblings in a hierarchy should share the same semantic type seems obvious, particularly when one considers the broadness of the categories of the Meta-1.1 semantic network. Yet the data show how different even two closely related schemes can be. The degree of commonality seen here might be useful in two ways. One, it can provide a method for review of the semantic type assignments; the idea being that if one sibling has an assignment much different (perhaps even from another portion of the network), that assignment might have been mistaken. Two, the semantic type assignments will no doubt clarify some of the issues those individuals maintaining the source vocabularies must confront.

What does this data imply about how a name-server might operate? The first and most important lesson is that any attempt to find the name of a desired concept must be able to explore multiple dimensions of the semantic neighborhood. The amount of redundancy in representing links is relatively small; each of the dimensions of semantic locality offers additional value in representing the semantic neighborhood of a concept.

Presumably, one would find the appropriate name for the concept by entering a term somehow related to the concept in question, and letting a name-server suggest candidate terms. The server would thus be quite dependent on intention. Knowing in which source vocabulary the concept needs to be expressed would certainly help limit the search for the appropriate concept name. Further, if the server knew the semantic type of the concept in question, it could be limited to responding with candidate terms of the appropriate type or types. The large number of links in the COT could then be narrowed to a much more manageable number. Searches for terms of the same type as the entered term would probably be most productive if the search looked at hierarchical and sibling terms first. Where the type was known to be different, other strategies might be more useful. The high proportion of links to definitional concepts (those occurring in the definitions of terms) found in co-occurrence data suggests that this co-occurrence data might be a rich source of concepts.

The finding that 15 to 20% of concepts found in definitions are not in Meta-1.1 indicates that, not surprisingly, the metathesaurus is still incomplete. The lack of links between these definitional concepts present in Meta-1.1 and the concepts they helped define suggests that close attention to developing new links will also be important. These definitional links would appear to represent something more profound than an empirically discovered relationship with another concept. They are at the heart of meaning of the concept. However, because of the sheer number of concepts involved, it seems unlikely that anything other than an attempt to recognize them in an automated or semi-automated fashion would be justifiable.

The effectiveness of the present semantic structure of the metathesaurus in faithfully representing the semantic neighborhoods of concepts, in a manner which can be exploited by a number of different programs, will only be judged by time. It may very well be that a name server utility may be of assistance to a variety of applications. These present investigations provide an appraisal of the current status of the development of the structure.

ACKNOWLEDGMENTS

Each of the authors has a continuing intellectual debt to the late Marsden Scott Blois in shaping their approach to the problems presented by the UMLS project. We wish to thank Alexa McCray for comments on an earlier version of this paper. Part of this work was supported by NLM Contract NO1-LM-0-3515.-

REFERENCES

1. Lindberg DAB, Humphreys BL. The UMLS Knowledge Sources. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications to Medical Care. New York: IEEE Computer Society, 1990:121-25.

2. McCray AT Hole WT. The scope and structure of the UMLS semantic network. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications to Medical Care. New York: IEEE Computer Society, 1990:126-30.

3. Nelson SJ, Tuttle MS, Cole WG, Sherertz DD, Sperzel WD, Erlbaum MS, Fuller LF, Olson NE. From Meaning to Term: Semantic Locality in the UMLS Metathesaurus. In: Clayton PD, ed. Proceedings of the Fifteenth Annual Symposium on Computer Applications to Medical Care. New York: McGraw Hill, 1991:209-13.

4. Sowa JF (ed.) Principles of Semantic Networks. San Mateo: Morgan Kaufman Publishers, Inc., 1991.

5. Brachman RJ, Levesque HJ (eds.) Readings in Knowledge Representation. San Mateo: Morgan Kaufman Publishers, Inc., 1985.