Battling Scylla and Charybdis: the Search for Redundancy and Ambiguity in the 2001 UMLS Metathesaurus

James J. Cimino, M.D.
Department of Medical Informatics, Columbia University, New York, New York, USA

I previously developed methods for identifying cases of multiple synonymous concepts (redundancy) and concepts with multiple meanings (ambiguity) and applied them to the 1995 UMLS Metathesaurus. These methods use semantic approaches (including knowledge about word synonymy and the semantic types assigned to concepts) to complement the standard lexical approaches. In this paper, I describe the results of their application to the 2001 Metathesaurus and examine their implications for the evolution of the UMLS.

INTRODUCTION

The Metathesaurus of the Unified Medical Language System (UMLS) has been constructed by the National Library of Medicine (NLM) to bring together from multiple terminologies and organize them into a set of concepts.[1] Each concept in the Metathesaurus is intended to have a single, unique meaning and is assigned one or more Semantic Types from the accompanying Semantic Network. The intent of the Metathesaurus is to support the retrieval and integration of information from disparate sources. The NLM has contracted with Apelon (Mountain View, CA) to provide the mappings between terms from terminologies and concepts in the Metathesaurus that attempt to minimize the number of concepts with the same meaning (redundancy) and the number of concepts with multiple meanings (ambiguity).

Apelon employs a variety of lexical methods to match terms to concepts (to reduce redundancy) and provides human review to identify inappropriate matches (to reduce ambiguity).[2] The 2001 Metathesaurus comprises over 1.9 million strings (1.7 million unique strings) from 98 terminologies; over the twelve years since its first version, it has grown to include 797,359 concepts. Given the sheer magnitude of the Metathesaurus, complete manual review of each concept, let alone manual review of all disjoint pairs of concepts to detect redundancy missed by lexical methods, is not feasible. Therefore, automated semantic methods (that is, systematic approaches based on concept meaning) are desirable to supplement the automated lexical and manual methods. For example, Hole and Srinivasan have described a variety of methods for detecting redundancy, including lexical matching with normalized words.[3] McCray and coworkers have examined methods for analyzing the Semantic Network to further aid in the auditing process.[4]

The NLM provides the UMLS to interested parties as an experimental product, with agreement by those parties to evaluate it and provide feedback. Under that agreement, I developed methods for detecting redundancy and ambiguity, which I applied to the 1995 version of the UMLS. I found 3274 redundant concepts, 1817 ambiguous concepts, and 544 relationships between concepts that were inconsistent with their semantics.[5] As part of my continued evaluation of the UMLS, I updated these methods and reapplied them to the 2001 Metathesaurus.

METHODS

Metathesaurus Data Model

The UMLS data model considers terminologies to be composed of terms that are themselves a collection of one or more strings, codes and other attributes. Strings are mapped to lexical groups that have similar surface forms (for example, the same words with different order or capitalization). Lexical groups are then mapped to concepts based on the meanings of the strings they contain.* Each concept is assigned one or more semantic types from the Semantic Net, based on its intended meaning. These assignments are generally derived from the semantic types of the terms in the source terminologies. Thus, if a concept comprises multiple terms from multiple terminologies, it is possible for it to have multiple semantic types. Inter-term relationships (including parent-child relationships) from the source vocabularies provide inter-concept relationships. Figure 1 (based on UMLS documentation) shows examples of how concepts are composed of terms, assigned semantic types, and related to each other.

Mutual Exclusion to Detect Ambiguity

As shown in Figure 1, each concept in the Metathesaurus is assigned at least one semantic type; indeed, many concepts are assigned two or more types. Frequently, for example, a chemical concept

* Over 17,000 Lexical Groups are assigned to multiple concepts; Figure 1 shows one example.

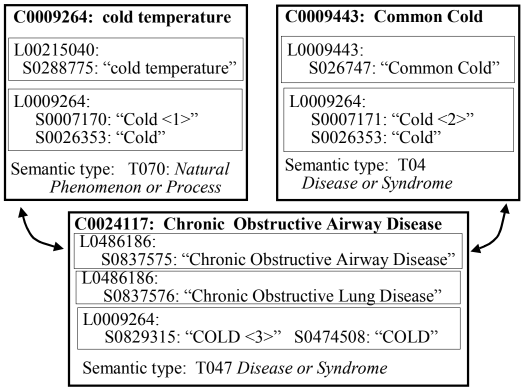
C0009264: cold temperature
L00215040:
S0288775: “cold temperature”
L0009264:
S0007170: “Cold <1>”
S0026353: “Cold”
Semantic type: T070: Natural Phenomenon or Process
C0009443: Common Cold
L0009443:
S026747: “Common Cold”
L0009264:
S0007171: “Cold <2>”
S0026353: “Cold”
Semantic type: T04 Disease or Syndrome
C0024117: Chronic Obstructive Airway Disease
L0486186:
S0837575: “Chronic Obstructive Airway Disease”
L0486186:
S0837576: “Chronic Obstructive Lung Disease”
L0009264:
S0829315: “COLD <3>” S0474508: “COLD”
Semantic type: T047 Disease or Syndrome

Figure 1: Metathesaurus conceptual schema. Concepts ("C" codes) contain one or more lexical groupings ("L" Codes) which, in turn, contain one or more strings ("S" Codes).

will be assigned one type based on its structure (e.g., *Carbohydrate*[*]) and one based on its activity (e.g., *Antibiotic*). However, many of the types are mutually exclusive; for example a concept cannot be both a plant and an animal or a substance and an event. Using the UMLS-supplied definitions (many of which list explicit situations of mutual exclusivity) for the 2001 Semantic Net, together with general knowledge of the world (such as the fact that plants can't be animals), I have derived what I believe to be a reasonable set of rules by which to determine if two types *can* be assigned to the same concept (Table 1).[**] These rules, provide a basis for semantic auditing of the Metathesaurus, since a concept assigned two mutually exclusive types is, by definition, ambiguous.[6]

The assignment of multiple, mutually exclusive semantic types to a concept suggests multiple meanings (i.e., ambiguity). Such assignment might occur through improper type assignment, but it may also be due to the mapping of lexically similar but nonsynonymous terms to the same concept. I therefore used the rules in Table 1 to identify potentially ambiguous concepts in the Metathesaurus.

Table 1: Rules used to determine which semantic types are not mutually exclusive (ME); unless otherwise addressed by the rules, any two semantic types are presumed to be ME. Codes in parentheses are Semantic Net Tree Addresses.

1: *Anatomical Abnormalities* (A1.2.2) and *Anatomical Structure* (A1.2) are not ME
2. *Manufactured Objects* (A1.3) are not ME with each other
3. Substances (A1.4) are not ME, except:
   a) *Element, Ion or Isotope* (A1.4.1.2.3) is ME with all other *Chemicals Viewed Structurally* (A1.4.1.2)
   b) *Inorganic Chemical* (A1.4.1.2.2) is ME with all *Organic Chemicals* (A1.4.1.2.1)
4. *Food* (A1.4.3) is not ME with any *Physical Object* (A1)
5. *Conceptual Entities* (A2) are not ME with each other, except:
   a) *Molecular Sequence* (A2.1.5.3) and *Geographic Area* (A2.1.5.4) are ME with each other and with *Body System* (A2.1.4.1), *Body Space or Junction* (A2.1.5.1) and *Body Location or Region* (A2.1.5.2)
6. *Body System* (A2.1.4.1), *Body Space or Junction* (A2.1.5.1) and *Body Location or Region* (A2.1.5.2) are not ME with *Anatomical Structure* (A1.2)
7. Events (B) are not ME with each other except:
   a) *Diagnostic Procedure* (A1.3.1.1) is ME with *Laboratory Procedure* (A1.3.1.2)
8. Ancestors are never ME with any of their descendants (this rule takes precedence)

Mutual Subsumption to Detect Redundancy

One way to detect redundancy is to look for two concepts that have the same name. An extension of this method takes advantage of the fact that concepts in the Metathesaurus may have many strings associated with them. I define *string subsumption* as the case where all the words in one string of a concept can be found among all the words of all the strings of a second concept. If two concepts each have a string that is subsumed by the other, I refer to these concepts as being *mutually subsumed*.[5]

In trying to determine if the words from one concept's string are contained in any of another concept's strings, I make use of two normalization methods to improve the chances of finding a match. The first method makes use of a set of keyword synonyms (e.g., "Renal = Kidney"), which was previously constructed[5] to identify concepts that "mutually subsume" each other. The second method uses the Metathesaurus normalized word index (not available in 1995). By using this composite method, I could detect, for example, that a concept called "Renal Disease" and a concept called "Kidney Diseases" subsume each other and are potentially redundant.

[*] In this paper, I will refer to concepts in **bold**, semantic types in *italics*, and strings in "quotes."

[**] It is simpler to list cases in which types *can* co-occur, than all the cases of mutual exclusivity.

Of course, it would not make sense to merge concepts with mutually exclusive semantic types, since that would cause ambiguity. I therefore use the rules in Table 1 to filter out concepts that should not be merged, based on their semantic types.[3,5] I quickly realized that I needed to add manual review to this process. For example, the concepts **cold temperature** and **Common Cold** are mutually subsumptive (they both contain the string "cold"). Since they are of not of mutually exclusive semantic types (by Rule 8: *Natural Phenomenon or Process* is an ancestor of *Disease or Syndrome*), my method would suggest these two concepts are potentially synonymous; manual review shows this to be false.

Analysis of Parent-Child Relationships

As shown in Figure 1, the Metathesaurus contains inter-concept relationships; these include Parent-Child relationships obtained from UMLS source terminologies. In the 1995 Metathesaurus, these relationships were not characterized further, and I treated them as "is-a" relationships for the purpose of comparing these relationships with the relationships between the semantic types of the parent and the child. According to the "is-a" assumption, the parent and child should have the same semantic type, or the type of the former should be an ancestor (in the Semantic Net hierarchy) of the latter. I found 544 pairs of terms for which this was not the case; examination of these pairs showed many instances of incorrect assignment of semantic types to concepts (as opposed to incorrect parent-child relationships).

In the 2001 Metathesaurus, some of the parent-child relationships are now labeled as having specific semantic relationships, including "is-a." I performed the same examination as on the 1995 Metathesaurus, this time restricting my examination to relationships labeled specifically as "is-a" relationships.

RESULTS

Detection of Ambiguity

The Metathesaurus contains 187,943 terms with multiple semantic types, with 217,985 pair-wise comparisons of 768 different semantic type pairs. Using the rules in Table 1, 391 pairs were considered not to be mutually exclusive (that is to say, they are allowed to occur), accounting for the vast majority (96%) of the concepts. Most of these (194 pairs accounting for 93% of the concepts) were classifications of multiple types of *Chemical.*

Table 2 shows some examples of the 8,082 concepts (4% of the concepts with multiple types; 1% of all concepts) with mutually exclusive semantic type assignments. Many appear to be due to ambiguous concepts. For example **C0035510: Toxicodendron** should probably be separated into two concepts; one that is a *Plant* (i.e., "Poison Ivy") and another that is a *Disease or Syndrome* (i.e., "Poison Ivy Dermatitis").

Table 2: Examples of multiple semantic type assignments that suggest ambiguity.

Body Part, Organ, or Organ Component and *Disease or Syndrome*
C0221219: Ectopic pancreas
C0223552: Fifth lumbar vertebra

Plant and *Disease or Syndrome*
C0035510: Toxicodendron

Alga and *Invertebrate*
C0015155: Euglena gracilis
C0032071: Plankton

Organism Attribute and *Diagnostic Procedure*
C0242789: Crown-Rump Length

Cell Function and *Biomedical Occupation or Discipline*
C0007608: Cell Movement

Invertebrate and *Disease or Syndrome*
C0030756: Lice Infestations

Injury or Poisoning and *Substance*
C0016542: Foreign Bodies

Mental or Behavioral Dysfunction and *Patient or Disabled Group*
C0013146: Drug Abuse

Genetic Function and *Biomedical Occupation or Discipline*
C0031325: Pharmacogenetics

Disease or Syndrome and *Patient or Disabled Group*
C0008715: Chronically Ill

Other reasons for violations of the rules may include incorrect type assignments, inappropriate rules, or incorrect definitions of semantic types. For example, 983 concepts violate Rule 3a; apparently because chemicals that contain an isotopic element have themselves been classified as *Element, Ion or Isotope.* Similarly, 929 concepts violate Rule 3b; apparently because organic chemicals that contain an "inorganic" atom have been classified as *Inorganic Chemical.*

Detection of Redundancy

When the 1.7 million strings in the Metathesaurus were compared to the normalized word index for 797,359 concepts, 22 million matches were found.

Of these, only 91,496 (45,748 pairs) were symmetrical – that is, all of the words in one string of each concept were found in any of the strings of the other concept. After excluding all pairs in which at least one of the semantic types of one concept was mutually exclusive with one of the semantic types of the other concept (according to the rules in Table 1), I found 38,140 pairs with mutually subsumed strings and compatible semantic types – a marked increase from the 5031 pairs found in 1995.

Determination of true redundancy is difficult. Many of the cases (14,236 pairs) are chemical concepts, in which the names are acronyms; true redundancy in these cases is difficult to ascertain (for example **C0673603 NPS-R-467** and **C0673604: NPS R-467**, both of which are *Organic Chemicals*). In other cases, the strings appear to be identical except for word order (such as the *Amino Acid, Peptide or Proteins* **C0673769: des-Arg(10)-(Leu(9))kallidin** and **C0673771: kallidin, des-Arg(10)-(Leu(9))-**), suggesting true redundancy. Confirmation will require careful review, since meanings of chemical names are usually sensitive to word order.

Analysis of the 23,904 nonchemical pairs also showed examples of apparent redundancy, such as the *Congenital Abnormalities* **C0266133: Congenital diverticulum of esophagus** and **C0555218: Congenital esophageal pouch**. However, the mutual subsumption method appears to have a low relevance rate when compared to my previous use of this method. A combination of factors is responsible for this change in performance. First of all, the keyword mapping has been enhanced using the normalized word index provided with the Metathesaurus. This has resulted in an increased match rate but at the cost of reduced specificity.

Second, the inclusion of additional strings in the 2001 Metathesaurus has reduced the usefulness of the mutual subsumption approach. When I examined the source strings for matching concepts, I found a marked increase in the number of incomplete term names included as synonyms of the concepts. For example, my method found that **C0011848: Diabetes Insipidus** and **C0687720: Central Diabetes Insipidus** have mutually subsumed strings. Obviously, the preferred name of the latter subsumes the preferred name of the former. Examination of the string source file (MRSO) explains the reverse subsumption: "Diabetes insipidus" is a synonym for "Central diabetes insipidus" in the 1999 Read Codes.

Another source of mutually subsumed strings appears to be from foreign language sources. For example, examination of the mutually subsumed pair **C0013005: Dolphins** and **C0325138: Whale, False Killer,** the former has the strings "FALSA BALEIA ASSASSINA" (from the Portuguese translation of MeSH) and "ORCA" (from the Spanish translation of MeSH), while the latter has the string "FALSA ORCA" (from the Portuguese translation of MeSH). In this case, the MeSH translation of "Dolphins" to "FALSA BALEIA ASSASSINA" is incorrect.

Analysis of Parent-Child Relations

Of the 9.6 million relationships in the Metathesaurus, 607,043 are parent-child relationships, of which 48,204 are "is-a" relationships. Examination of the semantic types of the concepts involved showed 2,868 cases in which the Semantic Net hierarchy could not account for the is-a relationships.

The most frequent pairs of semantic types involved in these relationships were *Body Location or Regions* (which is in the *Conceptual Entity* hierarchy) as parents of *Body Parts, Organs, or Organ Components* (which is in the *Physical Object* hierarchy). For example, **C00013769: Elbow** is the parent of **C0230353: Right elbow**. This pair suggests either that one of the concepts should be identified as the same semantic type as the other, or that the two semantic types should have a parent-child relationship in the Semantic Network. Over 75% of the 2,868 relationships involve similar disparities between physical and conceptual anatomic concepts.

Finally, in 142 cases, the semantic type assignments have not been done to the most specific appropriate level. For example, **C0004134: Ataxia** has the semantic type *Sign or Symptom* and is a parent of **C0751837: Gait Ataxia**, which has the less-specific semantic type *Finding*. This could be corrected by giving both concepts the type *Sign or Symptom*.

DISCUSSION

This study used a variety of semantic-based approaches to augment the lexical approaches used to audit the contents of the UMLS Metathesaurus. Despite the changes in the Metathesaurus over the past six years, these methods continue to identify potential problem areas on which to focus the attention of human reviewers.

My method for ambiguity detection continues to produce useful results. The number of concepts identified is relatively small (1% of the

Metathesaurus), allowing the NLM and Apelon to focus their resources for human review on specific problem areas. The results are also encouraging for the NLM: despite the increase in size of the Metathesaurus from 1995 to 2001, the number of apparently-ambiguous concepts actually decreased; many of the previous ambiguities[5] have disappeared.

I have updated my approach to detection of mutual-subsumption by taking advantage of the new normalized word index. Unfortunately, the specificity of the methods has decreased, resulting in an increase in mutually subsumed pairs by over 18-fold, while there does not appear to be a concomitant increase in detection of redundancy. Since these pairs require manual review, the change in the method has degraded performance; a repeat analysis, without the normalized string index, is indicated.

Even with the removal of the normalized string index, the string-subsumption method may have outlived its usefulness. The growth in the number of strings might be expected to help find redundancy but the inclusion of many "incomplete" names (as in the "diabetes insipidus" example) actually renders the method useless, since it relies on the asymmetric mapping (that is, one concept subsumes a second, but not vice versa) between more-general and more-specific terms to reduce the number of matches.

The string-subsumption method is further frustrated by the inclusion of erroneous mappings of strings in foreign languages into the Metathesaurus. It is my hope that the results of this study may help the NLM identify such errors and correct them.

My reliance on the structure of the Semantic Net means that problems in its hierarchy will be manifested as problems in my results. For example, the disjunction between concepts classified as *Body Location or Regions* and those classified as *Body Parts, Organs, or Organ Components* seems to be somewhat artificial, required only because the Semantic Net is a strict hierarchy. Recent work from the NLM draws the same conclusion and aggregates the various anatomical physical and conceptual entities into a single semantic group.[4]

A satisfying result of the work I have presented would be specific, quantitative statements about the occurrence of redundancy, ambiguity, incorrect semantic type assignments, and incorrect parent-child mappings. However, such results are not possible in this study, given the subjective nature of much of the information in the UMLS. Furthermore, I am not qualified to provide the necessary subjective review. A great deal of domain expertise will be needed for tasks such as deciding when two chemical name acronyms are synonymous. For other tasks, such as the appropriateness of semantic type assignments, only the NLM can make the proper judgments.[3] For example, are the sets of organic and inorganic chemicals mutually exclusive, or can a chemical containing, say, a mercury atom be classified as both?

The methods I present here lack the accuracy necessary for tasks such as the automated construction of the UMLS. Their results indicate that, by many metrics, the current approach of automated lexical and manual human processes may be "good enough." However, I believe my methods provide appropriate complementary ones that can be used to help focus the human review process. Despite the many instances of "false positives" and the need for final arbitration in many other cases to be passed on to the NLM, I believe I have found a number of specific problems that the NLM can readily address. To that end, I have passed my results on to the NLM as a contribution to the continuous process of improving this valuable national resource.

CONCLUSIONS

The use of relatively simple semantic methods continues to be a viable approach to augmenting the lexical methods used to manage the Metathesaurus.

Acknowledgments
I would like to thank Bill Hole, Alexa McCray, and Betsy Humphreys, of the NLM, for continuing to share the UMLS and for providing invaluable help in understanding its construction and contents. I would also like to thank Rachel Cimino and Rebecca Cimino for helping with the literary allusion in the title of this paper, and Andria Cimino for editorial assistance.

References

1. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. *JAMIA*. 1998;5(1):1-11.
2. Tuttle MS, Cole WG, Sherertz DD, Nelson SJ. Navigating to knowledge. *Meth Inform Med*. 1995;34:214-31.
3. Hole WT, Srinivasin S. Discovering missed synonymy in a large concept-oriented Metathesaurus. *JAMIA*, 2000;7 (suppl):354-358.
4. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS Semantic Types for reducing

coceptual complexity. Submitted to 2001 Medinfo.

5. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. *JAMIA*; 1998;5:41-51.

6. Cimino JJ. Desiderata for controlled medical vocabularies in the Twenty-First Century. *Meth Inform Med*; 1998;37(4-5):394-403.