Evaluation of a UMLS Auditing Process of Semantic Type Assignments

Huanying Gu, PhD,[1] George Hripcsak, MD,[2] Yan Chen, MS,[3,4] C. Paul Morrey, MS,[3] Gai Elhanan, MD,[5] James J. Cimino, MD,[2] James Geller, PhD,[3] Yehoshua Perl, PhD[3]
[1]SHRP, University of Medicine and Dentristry of New Jersey, Newark, NJ; [2]Columbia University, New York, NY; [3]New Jersey Institute of Technology, Newark, NJ; [4]BMCC, City University of NewYork, New York, NY; [5]3M Health Information Systems, CT

Abstract

The UMLS is a terminological system that integrates many source terminologies. Each concept in the UMLS is assigned one or more semantic types from the Semantic Network, an upper level ontology for biomedicine. Due to the complexity of the UMLS, errors exist in the semantic type assignments. Finding assignment errors may unearth modeling errors. Even with sophisticated tools, discovering assignment errors requires manual review. In this paper we describe the evaluation of an auditing project of UMLS semantic type assignments. We studied the performance of the auditors who reviewed potential errors. We found that four auditors, interacting according to a multi-step protocol, identified a high rate of errors (one or more errors in 81% of concepts studied) and that results were sufficiently reliable (0.67 to 0.70) for the two most common types of errors. However, reliability was low for each individual auditor, suggesting that review of potential errors is resource-intensive.

Introduction

The Unified Medical Language System (UMLS)[1] is a very large terminological system integrating more than 100 source terminologies. Each of the more than 1.3 million concepts is assigned one or more Semantic Types (STs), which are broad biomedical categories from the UMLS Semantic Network (SN).[2] Due to its size and complexity, it is unavoidable that errors have slipped into the UMLS. The assignment of STs to concepts is also error-prone, because the SN itself has well-known shortcomings.[3,4] Auditing the UMLS for errors is a very important task since many information systems in biomedicine utilize the UMLS. In a recent survey,[5] UMLS users suggested that the NLM spend an average 36% of a putative UMLS budget for auditing, which indicates its importance for users.

The *extent* of an ST is the set of the concepts that have been assigned this ST. Graphically, one may think of the extent of an ST as a box that contains all concepts that have been assigned this ST. Since each concept can be assigned one or more STs, some concepts occur in only one box, while others occur in several boxes. Conversely, the extent of each ST may contain concepts that are also assigned other STs and thus have different kinds of semantics (Figure 1(a)).

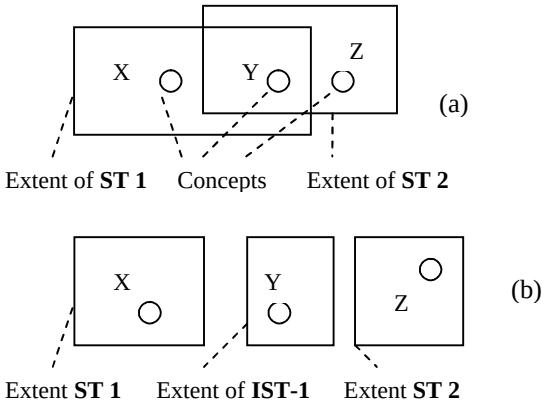

Figure 1: Extents of **(a)** overlapping STs; **(b)** RSTs

To provide a semantically uniform view of UMLS concepts, we have proposed a Refined Semantic Network[6,7] (RSN), with refined semantic types (RSTs). The RSTs include all STs of the UMLS SN, and additional intersection semantic types (ISTs). An IST is a combination of two or more STs for which there are concepts assigned exactly this combination. As a result, concepts assigned several STs in SN are assigned exactly one IST in RSN. Concepts assigned only one ST will be left in this ST's extent. Graphically, each RST gets its own box, and boxes do not overlap anymore (Figure 1(b)). For example, the concept e*xpressed breast milk* is assigned two STs **Body Substance** and **Food** of SN. In the RSN, it is assigned only one IST. Therefore, the extent of every RST in the RSN has a uniform semantics.

We create names for ISTs based on the STs of their extents, joined by the mathematical symbol for *intersection* (∩), e.g., **Body Substance ∩ Food.**[*] We denote the size of an IST's extent as *IST-n*, where *n* is the number of concepts assigned the relevant STs. Hence, IST-2 denotes an IST assigned 2 concepts. In previous research,[8] we found that concepts assigned

[*] Semantic types are written in bold; concepts, in italics.

ISTs with small extents (small ISTs) tend to have a large percentage of errors, with a probability of assignment errors of about 40% for IST-1 to IST-6. We are using the results from this previous study to define a priority order for auditing concepts. Previous auditing methods in the literature used ad hoc selection methods for inspection of concepts.[3,4] We are using a more principled approach. We note that the UMLS is too large for auditing all concepts, thus selecting concepts with a higher likelihood of errors has a great impact on the errors found. Nevertheless, if the resources for a large auditing project are available, our approach is also useful for a large scale QA effort.

Identifying errors in the ST assignments of a concept may also unearth other errors regarding this concept. For example, two ST assignments to a concept may indicate a case of ambiguity, where two UMLS source terminologies use the same term to refer to two different concepts, but the terms have been mapped to the same concept. Such ambiguity cases should be resolved in order to properly represent the actual semantics in the various source terminologies.

As a result of corrections in the ST assignments, an IST of small extent may disappear from the RSN, since there will be no concept left with such a combination. E.g., *intolerance function* is assigned two STs **Organism Function** and **Intellectual Product**. It is the only concept assigned the IST **Organism Function ∩ Intellectual Product.** However, *intolerance function* is not an intellectual product. Therefore, the IST **Organism Function ∩ Intellectual Product** will disappear from RSN. We expect that as a result of our study, many ISTs with small extents will similarly disappear, leaving in RSN only ISTs with a proper semantics. For example, the largest IST is **Organic Chemical ∩ Pharmacologic Substance**. This IST is assigned to all organic pharmacologic substances. Such an IST has a sound semantics. The RSN will help to streamline the process of assigning semantic types to new concepts in a way that will prevent many errors in assignments.

In a recent study,[5] UMLS users assigned their highest level of concern to the presence of erroneous ST assignments in the UMLS, out of six possible modeling errors. Missing ST assignments were rated as the second most problematic, following missing hierarchical relationships. Hence, errors in ST assignments are important to users. We therefore set out to review a large number of concepts assigned to small ISTs. Although our methods can draw attention to potential errors, only human experts can determine the veracity of semantic type assignments. In the present study, we explore this manual review process to measure its reliability and the effort required.

Methods

The first task of our study was to prepare the sample data set. This process consisted of two steps. First we identified ISTs of sizes between 1 and 6. Then we randomly selected concepts from the extents of these ISTs. Due to this random component the results are not reproducible, but there is no alternative to that other than auditing all 1.3 million concepts of the UMLS. Using the 2006AB edition of the UMLS Knowledge Sources (UMLSKS), we selected 70 concepts from 50 ISTs with extents of sizes from one to six. In some cases, we selected all concepts from the IST, while in others we selected a random subset, in order to keep the effort manageable.

We engaged four auditors, all of whom have training in medical informatics, with particular experience in medical terminology research. Two (JC and GE) are also experienced physicians and are referred to in this study as "experts," while the other two auditors (YC and HG) are referred to as "knowledge engineers."

Following our earlier auditing experience, we presented the auditors with the following data (if available in the UMLS) about each selected concept: the concept name, a list of its source terminologies, ST assignments, definitions, synonyms, parents and children (listed with STs in parentheses).

In the sample data set, the IST can be a combination of two or more STs. To accommodate all possible auditing results, we created an answer sheet, with eight possible choices, e.g., Semantic Type 1 error, ambiguity, no error, etc. Note that there may be more than one error for a concept and the answer sheet allows marking several boxes. For example, a concept may have listed two STs that are denoted in short as ST1 and ST2. An auditor may mark the corresponding boxes, if s/he thinks that both ST assignments are wrong (ST1 error and ST2 error, for short) and should be assigned instead another ST (to be inserted into the comments field of the answer sheet).

In the first round of the study, the four auditors independently reviewed all concepts in the sample data set. For each concept, they marked an answer sheet to indicate if they found one or more errors in the STs of that concept. In the second round, the four auditors' answers for each concept were aggregated and anonymized. The two expert auditors independently reviewed all the answers, marking

them as correct or incorrect. In a final round, the two experts consulted with each other and created a consensus reference standard.

To assess performance of the auditors, the knowledge engineers' answers from the first round were compared to the consensus reference standard. The experts' first round answers were compared only to the second round review by the opposite expert to avoid experts judging their own work. Performance was quantified by accuracy (proportion of all answers that matched the reference standard), recall (proportion of errors indicated in the reference standard that the auditor also reported), and precision (proportion of errors reported by the auditor that were also indicated in the reference standard). Ninety five percent confidence intervals were calculated for all estimates using the bootstrap method.[9]

The data were also stratified by error type. The prevalence of each error type was calculated. The reliability of the four auditors with respect to each error type was quantified. The specific agreement[10] and Cronbach's alpha reliability coefficient[11] were calculated for each error type.

Results

Table 1 shows the distribution of selected concepts in the sample data set. Note that we chose two samples of size IST-2.

IST Size	# of ISTs in UMLS 2006AB	# of ISTs Selecteded	# of Concepts in ISTs	# of Concepts Selected
IST-1	124	20	20	20
IST-2	68	5	10	10
IST-2	68	20	40	20
IST-3	37	2	6	5
IST-4	32	1	4	4
IST-5	26	1	5	5
IST-6	18	1	6	6
Total	287	50	91	70

Table 1: Distribution of selected concepts

We first present examples of the auditing process:
Concept 1: *Metaltite*; ST1: **Nucleic Acid, Nucleoside, or Nucleotide**, ST2: **Biomedical or Dental Material;**
Analysis: Metaltite® is a primer that is used to improve adhesion between resins and precious metals including many dental alloys. Metaltite® contains a thiouracil monomer. Two auditors considered the concept without errors while the two others considered ST1 erroneous. However, in the second round process both experts agreed that although the concept contains a thiouracil monomer, it is not a **Nucleic Acid, Nucleoside, or Nucleotide.** No further resolution was needed.

Concept 2: *AmericanIndianAlaskaNativeLanguages*; ST1: **Intellectual Product**, ST2: **Language;**
Analysis: One auditor voted for no ST errors, two considered it as an ST1 error (one of which considered it as well as a potential ambiguity) while the last auditor considered it as an ST2 error. After the second round, the two experts were still in disagreement. In their subsequent discussion, the experts agreed that although all languages are classified as intellectual products, this concept is an aggregating class and does not define a specific language. Therefore the experts resolved that this is an ST2 error. This decision allowed for the removal of the ambiguity classification. Note that the parent of *AmericanIndianAlaska-NativeLanguages* is *CodeSystem*, which has assigned the ST **Intellectual Product**. However, it does not have a parent-child relationship to the UMLS concept *Languages*, with ST **Language**. Thus, auditing the IST **Intellectual Product ∩ Language** exposed a missing parent-child relationship in addition to the erroneous ST.

Concept 3: *Endocardium*; ST1: **Body Part, Organ, or Organ Component**, ST2: **Tissue;**
Analysis: Three auditors reported an ST1 error and one an ST2 error. The experts were in disagreement on the concept and it was returned to them for discussion, during which they agreed that both STs were appropriate (i.e., no errors).

	1#	2#	3#	4#	5#	6#	7#	8#	Total
eng1	18	20	19	0	1	9	2	0	69
eng2	20	21	28	0	0	2	0	1	72
exp1	31	16	19	1	1	5	1	0	74
exp2	19	17	28	3	0	6	8	2	83

Table 2: Auditing results. "eng1" and "eng2" refer to the results of the two knowledge engineers; "exp1" and "exp2" refer to the two experts. The columns #1 to #8 refer to possible answers on the answer sheet.

Following the auditing protocol, the four auditors reviewed the sample. Each auditor marked at least one check box (1 to 8) per concept according to the errors found. Their choices are summarized in Table 2. Table 2 indicates some trends for specific auditors in comparison to others. Knowledge engineer 1 analyzed many cases (9) as ambiguity. Expert 2 added a new ST to many concepts (8) and had more cases of multiple errors. Expert 1 had more cases (31) as no errors. Knowledge engineer 2 and expert 2 had more cases (49 & 48) of wrong ST assignments.

Comparing the results from the experts, both of them agreed on 41 out of 70 concepts, of which 31 had errors. There are 29 disagreement cases, for which a consensus achieved showed 26 concepts with errors. There were errors in 57 (81%) of the 70 concepts. A report on these errors was submitted to the NLM. Interestingly, during the process, NLM changed information for five out of the 70 concepts. Three were removed and for two the ST assignments were changed.

The four auditors were moderately accurate (Table 3) and did not differ significantly from each other. The relatively high accuracy (about 0.88) mainly reflects the fact that most concepts had only one type of error, and auditors correctly marked most error types as absent. Recall and precision more clearly reflect true performance. On average, auditors detected 54% of the errors in the first round, and 56% of their assertions about errors were correct. Table 4 shows that ST1 and ST2 were the most common error types, and 81% of concepts had at least one error. Specific agreement, which indicates the likelihood of an arbitrarily chosen auditor agreeing with another one on a given error type, occurred about half the time for the frequent errors, and less often for the rarer ones.

Auditors	Accuracy (CI)	Recall (CI)	Precision (CI)
Eng1 (wrt consensus)	0.88 (0.85, 0.91)	0.52 (0.40, 0.64)	0.53 (0.41, 0.64)
Eng2 (wrt consensus)	0.89 (0.86, 0.92)	0.56 (0.45, 0.68)	0.59 (0.48, 0.71)
Expert1 (wrt expert2)	0.88 (0.85, 0.91)	0.59 (0.49, 0.70)	0.57 (0.47, 0.67)
Expert2 (wrt expert1)	0.86 (0.83, 0.90)	0.49 (0.37, 0.60)	0.54 (0.44, 0.64)

Table 3: Performance of the auditors

Error type	Prevalence of error in consensus standard (CI)	Specific agreement among 4 auditors in round one (CI)	Reliability per auditor and over all auditors
No error	0.19 (0.10, 0.27)	0.49 (0.39, 0.59)	0.28 0.60
ST1 error	0.29 (0.18, 0.39)	0.53 (0.41, 0.65)	0.37 0.70
ST2 error	0.40 (0.29, 0.51)	0.55 (0.45, 0.64)	0.34 0.67
ST3 error	0.04 (0.00, 0.09)	0.17 (0.00, 0.52)	–0.01 –0.03
ST4 error	0.01 (0.00, 0.04)	0.33 (0.00, 0.96)	0.33 0.67
Ambiguity	0.09 (0.02, 0.15)	0.12 (0.03, 0.20)	0.05 0.18
Add ST	0.03 (0.00, 0.07)	0.00 (0.00, 0.23)	–0.02 –0.11
Other error	0.00 (0.00, 0.00)	0.28 (0.00, 0.75)	0.22 0.53

Table 4: Properties by error type

Reliability is reported per rater (first reliability column in Table 4) or over all the raters (second reliability column). The per-rater reliability indicates how reliable a single auditor's answers are likely to be, whereas the reliability over all raters indicates how reliable an answer constructed from all the auditors' answers (e.g., by averaging or taking a vote) is likely to be.

The three most common error types (ST2, ST1, and none) achieved a reliability near 0.7 using the auditors' combined answers, but not when an individual auditor's answers are used. The less common error types have varying reliability, but the true value is difficult to tell given the rarity of the errors.

Discussion

As with our previous studies, examination of small ISTs has proven to be a productive way to identify errors in the UMLS, especially the assignment of erroneous semantic types. Our study uncovered fewer missing semantic type assignments, suggesting that this is less of a problem in the UMLS. However, concepts that have too few STs will tend to be in the extent of single STs (as opposed to ISTs), which were not included in the present study. The errors in this study are limited to ST assignments. However, the discovery of these errors can lead to the discovery of other errors. For example, for Concept 2 in the Results Section we identified one misapplication of the ST **Language**. This discovery, in turn, led to the discovery of a missing parent-child relationship in the Metathesaurus, which is due to a missing relationship in the source terminology. We note that the resolution for Concept 2: *AmericanIndianAlaskaNative-Languages* in the Results Section was done according to the context of its only source terminology HL7. According to the other UMLS sources, a group of languages is only assigned the ST **Language**.

The second round of expert review was included in our methodology to avoid having the experts judge their own work. If an expert were to make a mistake and could convince the other expert to make the same mistake in the consensus standard, then their estimated performance would be higher than it really is. Thus, the second round avoided biasing the results. Looking at the large number of errors found, one needs to remember that the sample being audited is taken from a very small (<0.1%) and highly selective sample of all UMLS concepts. Thus, these numbers are not representative for the whole UMLS.

The knowledge engineers performed similarly to the experts, perhaps because the group has been working together for a while, and all four may have acquired similar skills. However, even for our simple answer sheet with eight choices, overall reliability was low. A value of 0.7 is generally considered good.[12] We found that any given knowledge engineer is unlikely to produce reliable answers (up to 0.37 in our study). Only about half of the true errors will be detected, and only a little more than half of the reported errors will be correct. Note that reliability is the proportion of variance that is not due to error, and it can be defined even for a single auditor. It can only be estimated from a group of auditors, but using the Spearman-Brown prophecy formula[11], one can estimate what the reliability would be for different numbers of auditors, including one. This is a useful concept, because it tells you how good your process will be depending on how many auditors you choose to enlist. The reliability analysis indicates that if the answers of four knowledge engineers are combined then one can achieve adequate performance (reliability coefficient near 0.7, at least for the common errors). Although it is not ideal, two to three knowledge engineers together should be able to achieve moderate performance (reliability 0.5 to 0.6).

Thus, the manual review and correction of UMLS ST assignments is shown to be a labor-intensive process that requires multiple experts to produce reliable results. Nevertheless, UMLS users have confirmed that such corrections are important for their work. Our approach of examining small ISTs has detected a high percentage of errors, allowing us to focus limited resources on a small number of concepts. This focused scrutiny can lead to the correction of errors where they may be the densest, and also point to other errors, such as erroneous parent-child relationships.

Conclusions

The concepts that comprise the extents of small ISTs appear to be particularly prone to erroneous ST assignments. Evaluation of those ISTs with small extents allows auditors to focus their attention where it is most needed. This attention is labor-intensive, requiring multiple experts to achieve consensus. Our approach supports efficient completion of this task. With limited human resources for auditing, there is a tendency to divide auditing samples among auditors, to cover more concepts, rather than to assign the same sample to two or three auditors. Our analysis shows the (potential) shortcomings of this approach. More research evaluating the process of auditing concepts for various kinds of errors is needed to examine if indeed auditing should be done by groups of auditors, rather than a single one, to guarantee reliable results.

Acknowledgments

This work was partially supported by the United States National Library of Medicine under grant R 01 LM008445-01A2.

References

1. Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The Unified Medical Language System: An Informatics Research Collaboration. JAMIA 1998;5(1):1–11.
2. McCray AT. An Upper Level Oontology for the Biomedical Domain, Comp Funct Genom 2003;4:80–84.
3. Cimino JJ. Use of the Unified Medical Language System inpatient care at the Columbia-Presbyterian Medical Center. Methods Inf Med, 1995:34(1-2):158-164.
4. Mary V, Le Duff F, Mougin F, Le Beux P, Method for automatic management of the semantic network ambiguity in the UMLS: possible application for information retrieval on the Web. Stud Health Technol Inform. 2003:95:475-479.
5. Chen Y, Perl Y, Geller J, Cimino JJ. Analysis of a study of the users, uses, and future agenda of the UMLS. JAMIA 2007:14(2):221–231.
6. Gu H, Perl Y, Geller J, Halper M, Liu L, Cimino JJ. Representing the UMLS as an object-oriented database: modeling issues and advantages. JAMIA 2000: 7(1):66–80.
7. Geller J, Gu H, Perl Y, Halper M. Semantic Refinement and Error Correction in Large Terminological Knowledge Bases. Data & Knowledge Eng 2003;45;(1):1–32.
8. Gu H, Perl Y, Elhanan G, Min H, Zhang L, Peng Y. Auditing concept categorizations in the UMLS. Artif Intell Med. 2004 May; 31(1):29-44.
9. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
10. Hripcsak G, Heitjan D. Measuring agreement in medical informatics reliability studies. J Biomed Inform 2002;35:99–110.
11. Dunn G. Design and Analysis of Reliability Studies. New York: Oxford University Press, 1989.
12. Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF. A reliability study for evaluating information extraction from radiology reports. JAMIA 1999;6:143–50.