Published in final edited form as:

Methods Inf Med. 2016 ; 55(2): 158–165. doi:10.3414/ME14-01-0104.

Quality Assurance of UMLS Semantic Type Assignments to SNOMED CT Concepts with the Use of SNOMED CT Hierarchies

H. Gu¹, Y. Chen², Z. He³, M. Halper⁴, and L. Chen⁵

¹Computer Science Department, New York Institute of Technology, New York, NY

²Computer Information Systems Department, BMCC, CUNY, New York, NY

³Biomedical Informatics Department, Columbia University, New York, NY

⁴Information Technology Department, New Jersey Institute of Technology, Newark, NJ

⁵Science Department, BMCC, CUNY, New York, NY

Abstract

Background: The Unified Medical Language System (UMLS) is one of the largest biomedical terminological systems, with over 2.5 million concepts in its Metathesaurus repository. The UMLS's Semantic Network (SN) with its collection of 133 high-level semantic types serves as an abstraction layer on top of the Metathesaurus. In particular, the SN elaborates an aspect of the Metathesaurus's concepts via the assignment of one or more types to each concept. Due to the scope and complexity of the Metathesaurus, errors are all but inevitable in this semantic-type assignment process.

Objectives: To develop a semi-automated methodology to help assure the quality of semantic-type assignments within the UMLS.

Methods: The methodology uses a cross-validation strategy involving SNOMED CT's hierarchies in combination with UMLS semantic types. Semantically uniform, disjoint concept groups are generated programmatically by partitioning the collection of all concepts in the same SNOMED CT hierarchy according to their respective semantic-type assignments in the UMLS. Domain experts are then called upon to review the concepts in any group having a small number of concepts. It is our hypothesis that a semantic-type assignment combination applicable only to a very small number of concepts in a SNOMED CT hierarchy is an indicator of potential problems.

Results: The methodology was applied to the UMLS 2013AA release along with the SNOMED CT from January 2013. An overall error rate of 33% was found for concepts proposed by the quality-assurance methodology. Supporting our hypothesis, that number was four times higher than the error rate found in control samples.

Conclusion: The results show that the quality-assurance methodology can aid in effective and efficient identification of UMLS semantic-type assignment errors.

Keywords

Medical terminology; Quality assurance; UMLS; SNOMED CT; Semantic-type assignment; Auditing of terminologies; UMLS auditing

1. Introduction

The Unified Medical Language System (UMLS) [1] is a comprehensive biomedical terminological system that integrates more than 160 well established biomedical terminologies and ontologies. It provides a mapping structure among its sources and promotes the development of effective and interoperable biomedical information systems, such as electronic health records. The UMLS has become an invaluable resource for the biomedical community.

The UMLS has two major layers. The first is the Metathesaurus which is a repository of more than 2.5 million biomedical concepts. The second is the Semantic Network (SN), a compact abstraction layer above the Metathesaurus consisting of 133 broad categories called semantic types (STs). An important role of the SN is to elaborate a portion of the semantics of the concepts residing in the Metathesaurus. This is done by assigning each concept one or more STs. The assignments of STs to concepts also play a part in the integration of new terminologies into the UMLS. For example, the concept *ivy* with ST *Plant* can be differentiated from another concept with the same name when the second concept has been deemed to have the ST-assignment *Immunologic Factor*.

ST assignments, as described in [2], are carried out algorithmically where possible using information derived from the source. Subject-matter experts are called upon for review of such assignments as well as for additional manual ST assignments. Further reviews and revisions of the ST assignments derive from a relational model of the Metathesaurus.

Given that the UMLS is currently updated twice a year and that each new release involves a great deal of manual ST assignment work and review—not to mention the Metathesaurus's enormous scope and complexity—it is inevitable that ST errors and inconsistencies will be introduced into the UMLS. It is therefore imperative to formulate techniques to aid in its ongoing quality assurance (QA), which is an important part of any terminology's development cycle [3].

In this paper, we present a semi-automated methodology for QA of UMLS ST assignments. The methodology employs a cross-validation strategy involving hierarchies from SNOMED CT [4], which has previously been integrated into the UMLS. All concepts in the same SNOMED CT hierarchy are partitioned into disjoint groups according to their respective ST assignments in the UMLS. Each such group has its own unique combination of ST assignments, and all of them share an overarching broad meaning elaborated by the SNOMED CT hierarchy's category. Our basic premise is that reviewing concepts with such uniform semantics is more likely to be effective in identifying ST assignment errors than reviewing random concepts with disparate semantics. Furthermore, we hypothesize that if an ST-assignment combination only applies to a very small number of concepts (e.g., one, two,

or three) in a SNOMED CT hierarchy, it will be a useful indicator of potential problems. We demonstrate our cross-validation QA methodology by applying it to the 2013AA release of the UMLS along with the January 2013 version of SNOMED CT.

2. Background

2.1. SNOMED CT

SNOMED CT [4], a comprehensive healthcare terminology, is one of the sources of the UMLS. It uses description logic to model its approximately 300,000 concepts, which are arranged in 19 top-level hierarchies of IS-A relationships. Each concept resides in only one of the 19 hierarchies. However, a concept may be related to other concepts in any hierarchy in terms of non-IS-A (lateral) relationships. Each concept has a unique SNOMED Concept ID and a set of naming terms. A concept's fully specified name (FSN), one of its human-readable naming terms, ends with parenthesized text, called the *semantic tag*, denoting the semantic category to which the concept belongs. Each hierarchy in SNOMED CT generally has one semantic tag to represent its concepts' category. For example, all concepts in the SPECIMEN hierarchy have the same semantic tag *specimen*. Some hierarchies may have a few semantic tags that are used to categorize the concepts, but these are related and maintain semantic uniformity within the hierarchy. For example, the CLINICAL FINDING hierarchy has two semantic tags, *finding* and *disorder*, with the latter being a more specific category than the former.

When these concepts are integrated into the UMLS, they are assigned one or more STs. For instance, **infant care (regime/therapy)** was assigned the ST *Daily or Recreational Activity*. Overall, the ST-assignment process constitutes four steps, the first of which involves algorithmic suggestions derived from information in the source terminology. This is followed by domain experts' review and assignment, and then a review by National Library of Medicine (NLM) staff and contractors. Finally, a small NLM team carries out a review and makes required revisions [2]. The methodology developed in this paper helps identify ST assignment errors in the UMLS guided by the SNOMED CT hierarchies.

2.2. Related Work

The UMLS's extensive size and complexity make the introduction of errors nearly unavoidable when new terminologies are integrated or existing ones are revised. This is particularly true of ST assignments, which have previously been investigated for cases of erroneous omission and omission [5, 6, 7]. In [8], it was shown that UMLS concepts lacking the assignment of a particular ST could be effectively and efficiently identified by a search expanding outwards through parents and children in the extent of an ST. The performance of domain-expert auditors in their manual reviews of ST assignments was studied in [9]. As a preemptive measure to avoid the introduction of ST-assignment errors, a rule-based system called *adviseEditor* was developed to support the process of multiple ST assignments to the same concept [10]. In a longitudinal study on ST assignments [11], it was found that some illegitimate combinations of ST assignments were re-introduced after they had been eliminated in the previous releases of the UMLS, which indicates that ST assignments have not reached a stable state.

The methodology presented and carried out in this paper differs from previous work in that it exploits the structure of the actual source terminology itself (in this instance, SNOMED CT) to aid in the QA efforts of the UMLS. Concept groupings derived from the source serve as guides in the process of locating errors and inconsistencies.

In general, QA is a critical part of the development and maintenance of all biomedical terminologies, and indeed we find research dealing with systems besides the UMLS and SNOMED CT. For example, “evolutionary terminology auditing” has been applied in the quality assessment of the Gene Ontology [12]. The approach is based on determining how successive versions of a terminology do a better job in mimicking the structure of reality. An automated computational method was presented in [13] to audit symmetric concepts in FMA by leveraging self-bisimilarity and linguistic structure in concept names. To audit description-logic (DL) based terminologies, such as the National Cancer Institute thesaurus (NCIt) [14, 15, 16], it was proposed in [17] to partition a terminology’s classification into networks containing terms with identical DL roles, and organize the concepts into single-rooted sub-networks. A DL-based QA approach for LOINC was presented in [18], where LOINC was first transformed into an OWL DL representation and then mapped to SNOMED CT through the UMLS. A crowd-sourcing methodology to address the challenges of scalable ontology verification was described in [19]. Semantic Web technologies have been applied in the auditing of biomedical terminologies. For example, in [20], NCIt concepts and their properties were placed in an RDF triple store, through which the internal consistency of NCIt hierarchical and associative relations and their correspondence with relations in the UMLS Semantic Network were assessed.

3. Methods

Overall, our QA methodology seeks to identify groups of concepts within which the likelihood of finding ST assignment errors is relatively high. In this way, the domain expert’s review efforts can be focused appropriately. The concept groups identified will exhibit semantic uniformity. Such a characteristic is naturally found in the concepts of a specific SNOMED CT hierarchy given its overarching semantics. However, the number of concepts in a hierarchy can be quite large. For example, there are 97,399 concepts in CLINICAL FINDING. The goal is to narrow the focus quite a bit further.

The first step of the methodology is to determine for each SNOMED CT hierarchy H the entire set of STs assigned to its concepts in the context of the UMLS. We denote such a set $S_{\text{type}}(H)$. The notation $|S_{\text{type}}(H)|$ stands for the size (cardinality) of $S_{\text{type}}(H)$. As an illustration, consider Figure 1, where we see that SNOMED CT concepts C_2 and C_5 have been assigned ST_1 , C_1 , C_3 , and C_4 have been assigned ST_2 , and C_6 and C_7 have been assigned both ST_1 and ST_2 . Therefore, $S_{\text{type}}(H) = \{ST_1, ST_2\}$ and $|S_{\text{type}}(H)| = 2$. As a concrete example, RECORD ARTIFACT has 225 concepts that are assigned five different STs. Its set $S_{\text{type}}(\text{RECORD ARTIFACT}) = \{\text{Intellectual Product}, \text{Finding}, \text{Functional Concept}, \text{Health Care Activity}, \text{Manufactured Object}\}$, which happens to be the smallest such set among all SNOMED CT hierarchies. In another example, the 97,399 concepts of the hierarchy CLINICAL FINDING are assigned 85 different STs. Hence, $|S_{\text{type}}(\text{CLINICAL FINDING})| = 85$.

The second step of the methodology is to partition the SNOMED CT hierarchies into concept-group subsets such that each subset contains all concepts from a given hierarchy that have the exact same combination of ST assignments in the UMLS. We denote the complete set of such concept groups from a given hierarchy H as $S_{\text{group}}(H)$, with $|S_{\text{group}}(H)|$ being its size. From the illustration in Figure 1, we see that $S_{\text{group}}(H) = \{\{C_1, C_3, C_4\}, \{C_2, C_5\}, \{C_6, C_7\}\}$. The first group in $S_{\text{group}}(H)$ contains the three concepts that have been assigned ST_2 . The second group has the two concepts assigned ST_1 . And the third group has the two concepts assigned both ST_1 and ST_2 . Note that the three concept groups in $S_{\text{group}}(H)$ are all mutually disjoint and, in fact, collectively partition H .

As an example from SNOMED CT, the 97,399 concepts of hierarchy CLINICAL FINDING are partitioned into 88 disjoint groups ($|S_{\text{group}}(\text{CLINICAL FINDING})| = 88$). Among them, 27,482 concepts are in the group with the ST assignment of *Disease or Syndrome*, and 27,432 are in the group with the ST assignment *Finding*. The group with the combination ST assignment of *Congenital Abnormality* and *Disease or Syndrome* contains 1,103 concepts. There are only three concepts in the group with ST assignment *Embryonic Structure*.

As part of the second step, a list of subsets of $S_{\text{type}}(H)$ is tabulated representing the actual ST assignments associated with the members of $S_{\text{group}}(H)$. For example, for the hierarchy CLINICAL FINDING, the list would include $\{\textit{Disease or Syndrome}\}$, $\{\textit{Finding}\}$, $\{\textit{Congenital Abnormality, Disease or Syndrome}\}$, $\{\textit{Embryonic Structure}\}$, etc. It will be noted that the concepts in a group in $S_{\text{group}}(H)$ now have an identified, more refined uniform semantics, represented by the one subset of $S_{\text{type}}(H)$, as compared to its general hierarchical grouping in H .

The third step of the methodology exploits the newly elaborated refined semantics of the groups with a review carried out by a domain expert(s). It is advantageous to review concepts from such a semantically uniform group together as a unit to help in the verification of the correctness of their ST assignments. As we have seen above with the examples, the size of such a group is often much smaller than its hierarchy's. However, it can still be far too large for expert review purposes. For example, the group with ST assignment *Therapeutic or Preventive Procedure* from PROCEDURE hierarchy has 30,475 concepts. Reviewing this group manually is not practical. On the other hand, this group represents the ST assignment to 58% of PROCEDURE's total of 52,471 concepts. The widespread assignment of *Therapeutic or Preventive Procedure* probably had its rationale and is likely to be correct in most cases. Therefore, it is sensible to have the domain-expert reviewer ignore this concept group anyway.

So, in the third step, the review of concept groups for incorrect ST assignments in each hierarchy is limited to those with *small* numbers of concepts, namely, one to three concepts. (The choice of a threshold value of three for defining a "small" group is elaborated on in the Discussion section below.) Our hypothesis is that if a subset of $S_{\text{type}}(H)$ represents a unique semantic classification of a group with so few concepts in the hierarchy, those concepts have a greater likelihood of being classified incorrectly. For the PROCEDURE hierarchy, for example, only 17 groups in $S_{\text{group}}(\text{PROCEDURE})$ would be selected for review: 12 groups have only one concept, three groups have two concepts, and another two groups have three concepts. Since

each hierarchy can be treated as one large semantically uniform group, the SNOMED CT hierarchy's category can serve as an extra guideline for the domain experts when they carry out their review.

The three major steps of the QA methodology are illustrated in Figure 2. This process ultimately yields erroneous ST assignments with respect to SNOMED CT concepts residing in the UMLS.

4. Results

Our QA methodology was applied to the 2013AA release of the UMLS and the January 2013 release of SNOMED CT. Table 1 shows for each SNOMED CT hierarchy H (**Hierarchy (H)**), its number of UMLS concepts ($|\text{Concept}(H)|$), the number of STs assigned to its concepts ($|\text{S}_{\text{type}}(H)|$), the number of disjoint groups according to their ST assignments ($|\text{S}_{\text{group}}(H)|$), and the maximum size of concept groups in $\text{S}_{\text{group}}(H)$ ($\text{Max}(\text{S}_{\text{group}}(H))$). Let us note that when SNOMED CT's concepts were integrated into the UMLS, two or more concepts may have been merged into one concept. Thus, the total number of the UMLS concepts from the SNOMED CT, 292,666 (see Column 2 $|\text{Concept}(H)|$ of the table), is smaller than the number of concepts in SNOMED CT, which is about 297,000. Also note that for each pair of hierarchies H_i and H_j , $\text{S}_{\text{type}}(H_i)$ and $\text{S}_{\text{type}}(H_j)$ are not necessarily disjoint since an ST can be assigned to concepts in more than one hierarchy. For example, ST *Finding* is assigned to concepts in 16 out of the 19 SNOMED CT hierarchies. Only concepts in the hierarchies PHARMACEUTICAL, PHYSICAL FORCE, and PHYSICAL OBJECT do not have *Finding* assignments. Thus, the total number of STs assigned to concepts of all hierarchies (Column 3 $|\text{S}_{\text{type}}(H)|$) is 663, which is much larger than the total number of STs (133) in the UMLS Semantic Network.

Comparing the values of $|\text{S}_{\text{type}}(H)|$ and $|\text{S}_{\text{group}}(H)|$ (Columns 3 and 4) in Table 1, we notice that $|\text{S}_{\text{group}}(H)|$ is either greater than or equal to the corresponding $|\text{S}_{\text{type}}(H)|$. Thirteen out of the 19 hierarchies have $|\text{S}_{\text{group}}(H)| > |\text{S}_{\text{type}}(H)|$, since they not only contain the concept groups assigned one ST, but groups associated with two or more STs as well. For example, $|\text{S}_{\text{group}}(\text{STAGING AND SCALES})| = 12$ and $|\text{S}_{\text{type}}(\text{STAGING AND SCALES})| = 10$, which means the hierarchy STAGING AND SCALES is partitioned into 12 disjoint concept groups according to the ST assignments. The 12 subsets of $|\text{S}_{\text{type}}(\text{STAGING AND SCALES})|$ involve ten semantic types. Among these 12 groups, nine of them have just one ST assignment each. The other three groups were assigned two STs each. These dual ST assignments include $\{\text{Diagnostic Procedure, Intellectual Product}\}$, $\{\text{Quantitative Concept, Intellectual Product}\}$, and $\{\text{Manufactured Object, Intellectual Product}\}$. In some hierarchies, most of the concept groups were assigned two or more STs. In such a case, $|\text{S}_{\text{group}}(H)|$ is significantly larger than $|\text{S}_{\text{type}}(H)|$. For example, $|\text{S}_{\text{group}}(\text{PHARMACEUTICAL})|$ and $|\text{S}_{\text{group}}(\text{SUBSTANCE})|$ are more than four times greater than $|\text{S}_{\text{type}}(\text{PHARMACEUTICAL})|$ and $|\text{S}_{\text{type}}(\text{SUBSTANCE})|$, respectively. Note that the agreement of $|\text{S}_{\text{type}}(H)|$ and $|\text{S}_{\text{group}}(H)|$ does not strictly imply single ST assignments. For instance, $|\text{S}_{\text{type}}(\text{RECORD ARTIFACT})|$ and $|\text{S}_{\text{group}}(\text{RECORD ARTIFACT})|$ have the same value five. However, four concept groups have single ST assignments, while one group was assigned $\{\text{Manufactured Object, Intellectual Product}\}$.

A total of 292,666 SNOMED CT concepts were partitioned into 1,001 disjoint groups, whose sizes range from one to over 30,000, as shown in Table 1. Of these, 488 groups containing one to three concepts were chosen for review. These contained a total of 720 concepts whose ST assignments were evaluated for correctness. The concept reviews were carried out by two of the authors (YC and LC), both of whom have extensive experience in terminology QA; they have a medical and a biochemistry background, respectively. For any concept whose ST assignment was deemed incorrect, a possible remedy is proffered. The main results are given in Table 2, in ascending order of overall concept-error percentages. A total of 33% ($= 241/720$) of the reviewed concepts were found with ST-assignment errors, as seen at the bottom of the right-most column.

Our QA methodology is based on the hypothesis that concepts residing in small-size groups exhibiting uniform semantics have a higher likelihood of harboring ST assignment errors. In order to verify the hypothesis, we also reviewed a randomly selected control sample of concepts from each hierarchy. The number of concepts in the control sample for a given hierarchy is based on the total number of reviewed concepts in the groups of size one to three from that hierarchy. The latter quantity is denoted $N_{\text{rev}}(H)$ for hierarchy H . The size of the control sample for hierarchy H is given by the formula:

$$[(N_{\text{rev}}(H) / 30) + 1] \cdot 20$$

This formula was chosen to yield values somewhat close to the size of the collective review sets derived automatically by our methodology. For example, the PHARMACEUTICAL hierarchy has a total of 116 concepts in its groups of size one to three. (I.e., $N_{\text{rev}}(\text{PHARMACEUTICAL}) = 116$.) Therefore, the control sample has size 80 ($= [(116 / 30) + 1] \cdot 20$). See Table 3.

Table 3 shows the results of the reviews of the concepts from the groups of size one to three in our QA methodology and those from the control samples. The average error rate of 33% for groups from our QA methodology versus an 8% error rate for the control samples confirms our hypothesis (p value < 0.0001 , two-tailed Fisher's test).

5. Discussion

5.1 Interpretation and Evaluation

In the process of applying our methodology, we identified three kinds of ST-assignment errors that occurred for the SNOMED CT concepts: (1) entirely incorrect ST assignment; (2) partially incorrect ST assignment; and (3) incomplete ST assignment.

An entirely incorrect ST assignment is when all STs assigned to a concept are wrong. Most of the erroneous ST assignments discovered were of this variety. For example, SNOMED CT concept **group T** (not to be confused with concept groups in our methodology) in hierarchy QUALIFIER VALUE was assigned *Educational Activity*. The assignment is different from all the siblings of the concept, such as **group A**, **group B**, and so on, that were either assigned *Classification* or *Intellectual Product*. The error might have been introduced when **group T** was made a synonym of **sensitive training group** and **T-Group** on integration into the UMLS. In another example, **Non-human body structure** in hierarchy BODY STRUCTURE

was assigned *Entity*, which is one of the roots of the ST hierarchy in the Semantic Network. Few concepts should be assigned that broad ST according to the usage notes in [21]. Furthermore, the concept has a broader ST than its parent **Anatomical Structural** does, which indicates an error. It should be assigned the same ST as its parent: *Anatomical Structure*. An example from the PHARMACEUTICAL hierarchy is the concept **bee venom** with assigned STs *Biologically Active substance* and *Hazardous or Poisonous Substance*. Both contradict each other and are not correct classifications of **bee venom**, which should have ST *Pharmacological Substance* and *Immunologic Factor*.

A partially incorrect ST assignment can occur when concepts are assigned two or more STs, some of which are correct and some of which are incorrect. This can be seen for **glycoprotein** and **3.8S Alpha-2 glycoprotein**, the only two concepts in a group in SUBSTANCE hierarchy that are assigned the two STs *Amino Acid, Peptide, or Protein* and *Carbohydrate*. Since both concepts are glycoproteins, only *Amino Acid, Peptide, or Protein* is a correct classification. The *Carbohydrate* assignment should be removed. These two concepts would then belong to the group with the ST assignment *Amino Acid, Peptide, or Protein*. The group in SUBSTANCE hierarchy with the ST assignments of *Amino Acid, Peptide, or Protein* and *Carbohydrate* would cease to exist. In another example, concept **caprolactam** of SUBSTANCE hierarchy has the ST assignment *Indicator, Reagent, or Diagnostic Aid* and *Antibiotic*. The latter is incorrect. Instead, the concept should be assigned *Indicator, Reagent, or Diagnostic Aid* and *Organic Chemical*. One more example is **Anabolic steroid preparation** in the PHARMACEUTICAL hierarchy. It is classified as *Mental or Behavioral Dysfunction, Steroid*, and *Pharmacologic Substance*, but it is not *Mental or Behavioral Dysfunction*. That ST assignment should be removed.

Incomplete ST assignment occurs when an ST assignment does not completely elaborate the high-level semantics of a concept. For example, concept **Unilateral upper motor neurone lesion** of the SITUATION WITH EXPLICIT CONTENT hierarchy has only one ST *Pathologic Function*. The assignment omits the meaning that the concept inherits from its parent **Unilateral clinical finding**, which has ST *Finding*. In another example, both concepts **Lactase** and **Papain** of the PHARMACEUTICAL hierarchy are assigned *Pharmacologic Substance* and *Enzyme*. The assignment of ST *Amino Acid, Peptide, or Protein* is missing and should be added since not all enzymes belong to *Amino Acid, Peptide, or Protein*.

The overall results of our methodology show that most of the hierarchies have between 10% and 40% incorrect ST assignments for concepts in groups of size one to three. Only 2% and 3% of concepts are found incorrectly classified with ST assignments in hierarchies QUALIFIER VALUE and SPECIAL CONCEPT, respectively. Over 40% of concepts reviewed in the hierarchies PHARMACEUTICAL, SUBSTANCE, RECORD ARTIFACT, and SITUATION WITH EXPLICIT CONTENT were found to have ST assignment errors. The average 33% error-rate proves that our QA methodology is indeed effective in identifying ST assignment errors.

We reviewed only the concepts residing in small-size groups. The threshold value of three for small groups was chosen due to a preliminary study that was performed using groups of size up to seven. The results of that preliminary work indicated that the larger groups had a drop-off in terms of error yield. Moreover, the threshold of three simplifies the review work

of the domain expert, particularly when the concepts are analyzed collectively. Table 3 shows that the average error rate of 33% for groups from our QA methodology is significantly higher than the 8% error rate for the control samples. Among the control samples from the 19 hierarchies, ST-assignment errors were identified in only 11. Two hierarchies, SUBSTANCE and PHARMACEUTICAL, have the highest error rates of 14% and 20%, respectively. Note that the majority of concept groups from the hierarchies SUBSTANCE and PHARMACEUTICAL have ST assignments involving two to four STs. In our previous research, we had demonstrated that concepts assigned multiple STs are more complex and have a higher error likelihood [22]. The current results agree with those previous results.

In [23], an ST-assignment auditing approach was proposed based on analyses of concepts assigned multiple so-called meta-semantic types (a kind of higher-level abstraction grouping). There, an error yield of 29% of the reviewed concepts was obtained—only slightly below the average error rate of 33% achieved herein. However, considering that the concepts we reviewed in the present work were strictly from the SNOMED CT, which is widely regarded as one of the most well designed terminologies in the UMLS—and thus expected to have fewer ST-assignment errors to begin with—we can consider the results of the QA methodology in this paper to be more than just a marginal improvement over the results reported in [23].

5.2 Limitations

In our methodology, we inspected the groups with one to three concepts for all hierarchies without regard to the size differences of the hierarchies. However, the number of concepts in the respective SNOMED CT hierarchies differs greatly and can vary from 175 in PHYSICAL FORCE to 97,399 in CLINIC FINDING hierarchy; the former is only about 0.2% the size of the latter. The three-concept groups account for 1.7% of the concepts in PHYSICAL FORCE, but only 0.003% of CLINIC FINDING. Thus, one ST assignment for a group of concepts will be more reasonable in a small hierarchy than another ST assignment for an equal size group in a large hierarchy. Our methodology can be modified to take the hierarchy size into consideration. Instead of selecting fixed-size groups of concepts for review, we could define proportional sizes based on the total number of concepts in the hierarchy.

Due to the size and complexity of the UMLS, assuring the quality of the whole, or even a substantial portion, of it is an overwhelming task. In addition, the QA resources are limited. Utilizing the characteristics of the SNOMED CT whose concepts are divided into 19 disjoint hierarchies, with each hierarchy having its own semantic category, our methodology let the domain-expert reviewer focus on relatively small parts of the UMLS with higher chances of discovering errors. As currently formulated, our methodology is only applicable to concepts derived from the SNOMED CT. However, it can be extended easily to review concepts originating from other source terminologies if the source terminologies have similar features to the SNOMED CT.

Although our methodology focused the domain-expert reviewers' efforts on concepts with higher error rates, the control samples still exhibited errors at a rate of 8% on average. If we project that 8% error rate to the whole SNOMED CT, we would expect about 24,000 ST-assignment errors overall. The majority of these ST assignment errors cannot be found using

our methodology alone. Multiple QA techniques would need to be employed for this purpose. Our methodology can be seen as another tool in the QA tool-chest of the UMLS. We would emphasize, though, that the domain-expert review is greatly facilitated by the presentation of the targeted concepts within small groups in our approach.

5.3 Future Work

In our QA methodology, we partition the concepts of a hierarchy into the semantically uniform concept groups for the sake of guiding the domain expert's review. All concepts in a group share the same ST assignments, which are subsets of $S_{\text{type}}(H)$. However, concepts in such a group do not necessarily have the same semantic tag since not all hierarchies in the SNOMED CT have just one semantic tag associated with their concepts. Some hierarchies have two or more semantic tags. Therefore, in a future study, we plan to carry out a "double partitioning" based on shared ST assignments and semantic tags. Besides utilizing the SNOMED CT hierarchies, we relied on a domain expert's knowledge and judgment. We will further investigate algorithms that can programmatically detect the presence of possible ST assignment errors by checking the semantic consistence between a group's ST assignments and semantic tags using the Semantic Network hierarchy of the UMLS.

6. Conclusions

We presented a QA methodology to semi-automatically identify UMLS semantic-type assignment errors. The methodology was based on semantically uniform concept groups. SNOMED CT's hierarchies and UMLS semantic types were used to programmatically generate semantically uniform, disjoint concept groups. Concepts in such groups of small size were reviewed rigorously and shown to exhibit higher rates of semantic-type assignment errors in comparison to control samples. The results confirmed that the methodology is efficient and effective in locating semantic-type assignment problems. Overall, our introduced methodology augments the suite of tools available for the QA of semantic-type assignments within the UMLS.

References

- [1]. Humphreys BL, Lindberg DAB, Schoolman HM, and Barnett GO. The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Informatics Association*, 1998; 5(1):1–11. [PubMed: 9452981]
- [2]. Alexa TM and William TH. The Scope and Structure of the First Version of the UMLS Semantic Network. *Proceedings of Fourteenth Annual Symposium on Computer Applications in Medical Care*; 1990; 126–130
- [3]. Min H, Perl Y, Chen Y, Halper M, Geller J, and Wang Y. Auditing as part of the terminology design life cycle. *Journal of the American Medical Informatics Association*, 2006; 13(6):676–690. [PubMed: 16929044]
- [4]. IHTSDO: SNOMED CT, <<http://www.ihtsdo.org/snomed-ct>>
- [5]. Gu H, Perl Y, Elhanan G, Min H, Zhang L, and Peng Y. Auditing concept categorizations in the UMLS. *Artificial Intelligence in Medicine*; 2004; 31(1):29–44. [PubMed: 15182845]
- [6]. Gu H, Hripcsak G, Chen Y, Morrey CP, Elhanan G, Cimino JJ, et al. Evaluation of a UMLS auditing process of semantic type assignments. In Teich JM, Suermondt J Hripcsak G, editors. *Proc 2007 AMIA annual symposium*; 2007; 294–298.

- [7]. Chen Y, Gu H, Perl Y, Geller J, and Halper M. Structural group auditing of a UMLS semantic type's extent. *Journal of Biomedical Informatics*; 2009; 42(1):41–52. [PubMed: 18619563]
- [8]. Chen Y, Gu H, Perl Y, Halper M, and Xu J. Expanding the extent of a UMLS semantic type via group neighborhood auditing. *Journal of the American Medical Informatics Association*; 2009; 16(5):746–757. [PubMed: 19567802]
- [9]. Gu H, Elhanan G, Perl Y, Hripcsak G, Cimino JJ, Xu J, Chen Y, Geller J, and Morrey C.P. A study of terminology auditors' performance for UMLS semantic type assignments. *Journal of Biomedical Informatics*; 2012; 45(6):1042–1048. [PubMed: 22687822]
- [10]. Geller J, He Z, Perl Y, Morrey CP, and Xu J. Rule-based support system for multiple UMLS semantic type assignments. *Journal of Biomedical Informatics*; 2013; 46(1):97–110. [PubMed: 23041716]
- [11]. He Z, Morrey CP, Perl Y, Elhanan G, Chen L, Chen Y, and Geller J. Sculpting the UMLS refined semantic network. *Online Journal of Public Health Informatics*; 2014; 6(2):e181. [PubMed: 25422719]
- [12]. Ceusters W Applying evolutionary terminology auditing to the Gene Ontology. *Journal of Biomedical Informatics*; 2009; 42(3): 518–529. [PubMed: 19162233]
- [13]. Luo L, Mejino JL Jr, and Zhang GQ. An analysis of FMA using structural self-bisimilarity. *Journal of Biomedical Informatics*; 2013; 46(3): 497–505. [PubMed: 23557711]
- [14]. NCI Thesaurus: <http://ncit.nci.nih.gov/>
- [15]. Sioutos N, de Coronado S, Haber M, Hartel F, Shaiu W, and Wright L. NCI Thesaurus: A Semantic Model Integrating Cancer-Related Clinical and Molecular Information. *Journal of Biomedical Informatics*; 2007; 40(1): 30–43. [PubMed: 16697710]
- [16]. Frago G, de Coronado S, Haber M, Hartel F, and Wright L. Overview and Utilization of the NCI Thesaurus. *Comparative and Functional Genomics*; 2004; 5(8): 648–654. [PubMed: 18629178]
- [17]. Min H, Perl Y, Chen Y, Halper M, Geller J, and Wang Y. Auditing as part of the terminology design life cycle. *Journal of the American Medical Informatics Association*; 2006; 13(6):676–90. [PubMed: 16929044]
- [18]. Adamusiak T and Bodenreider O. Quality assurance in LOINC using Description Logic. *Proc 2012 AMIA Annual Symposium*; 2012: 1099–108.
- [19]. Mortensen JM, Minty EP, Januszyk M, Sweeney TE, Rector AL, Noy NF, and Musen MA. Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT. *Journal of the American Medical Informatics Association*; 2014. In press.
- [20]. Mougin F and Bodenreider O. Auditing the NCI Thesaurus with Semantic Web Technologies. *Proc 2008 AMIA Annual Symposium*; 2008: 500–4.
- [21]. Definition of UMLS Semantic Types. [cited August 27, 2014]; Available from: <http://semanticnetwork.nlm.nih.gov/Download/RelationalFiles/SRDEF>
- [22]. Geller J, Gu H, Perl Y, and Halper M. Semantic refinement and error correction in large terminological knowledge bases. *Data Knowledge Engineering*; 2003; 45(1):1–32.
- [23]. Gu H, Perl Y, Elhanan G, Min H Zhang L, and Peng Y. Auditing concept categorizations in the UMLS. *Artificial Intelligence in Medicine*; 2004; 31(1):29–44. [PubMed: 15182845]

Fig. 1:
Semantic-type assignments and resulting disjoint concept groups of a SNOMED hierarchy

Figure 2:
The process of the QA methodology

Table 1:

Statistics for SNOMED CT hierarchies

Hierarchy (H)	Concept(H)	S _{type} (H)	S _{group} (H)	Max(S _{group} (H))
Body structure	30,996	52	54	17,594
Clinic finding	97,399	85	88	27,482
Environment or geographical location	1,707	15	17	815
Event	3,632	21	22	3,019
Linkage concept	1,102	51	51	213
Observable entity	8,221	58	60	3,630
Organism	32,319	20	20	11,607
Pharmaceutical	15,570	34	148	8,410
Physical force	175	10	10	126
Physical object	4,501	19	21	3,554
Procedure	52,471	34	35	30,475
Qualifier value	8,994	82	83	1,469
Record artifact	225	5	5	209
Situation with explicit content	3,233	26	27	2,460
Social context	4,768	20	20	3,770
Special concept	802	42	43	110
Specimen	1,327	17	17	581
Staging and scales	1,286	10	12	1,214
Substance	23,938	62	268	3,652
Total:	292,666	663	1,001	N/A

Table 2:

Results of reviewing concepts in groups of size one to three (# Size-1/2/3 Groups: number of groups in the hierarchy with 1/2/3 concepts, respectively)

Hierarchy	# Size-1 Groups	# Errors	Error %	# Size-2 Groups	# Errors	Error %	# Size-3 Groups	# Errors	Error %	Overall Error %
Qualifier value	21	1	5%	6	0	0%	8	0	0%	2%
Special concept	11	1	9%	6	0	0%	4	0	0%	3%
Observable entity	53	4	8%	2	0	0%	3	0	0%	6%
Linkage concept	31	3	10%	7	0	0%	1	0	0%	6%
Social context	4	1	25%	1	0	0%	2	0	0%	8%
Physical force	3	1	33%	1	0	0%	1	0	0%	13%
Specimen	5	0	0%	4	2	25%	0	0	n/a	15%
Event	4	1	25%	2	0	0%	3	3	33%	24%
Body structure	11	4	36%	4	2	25%	1	0	0%	27%
Clinic Finding	16	4	25%	9	8	44%	3	0	0%	28%
Procedure	12	5	42%	3	2	33%	2	0	0%	29%
Organism	3	1	33%	3	0	0%	1	3	100%	33%
Physical object	8	3	38%	3	2	33%	3	3	33%	35%
Environment or geographical location	3	2	67%	1	0	0%	3	3	33%	36%
Staging and scales	3	1	33%	1	0	0%	2	3	50%	36%
Pharmaceutical	42	16	38%	19	12	32%	12	24	67%	45%
Substance	82	39	48%	25	40	80%	15	30	67%	62%
Record artifact	0	0	n/a	2	2	50%	1	3	100%	71%
Situation with explicit content	9	6	67%	3	6	100%	0	0	n/a	80%
Total:	321	93	29%	102	76	37%	65	72	37%	33%

Table 3:

Results of reviewing concepts from small-size groups and control samples ($N_{\text{rev}}(\text{H})$: total number of concepts in groups with one to three concepts in hierarchy H)

Hierarchy (H)	$N_{\text{rev}}(\text{H})$	# Errors in Size-1-to-3 Groups	Size of Control Sample	# Errors in Control Sample	Error % in Size-1 to 3 Groups	Error % in Control Sample
Qualifier value	57	1	40	2	2%	5%
Special concept	35	1	40	5	3%	13%
Observable entity	66	4	60	1	6%	2%
Linkage concept	48	3	40	3	6%	8%
Social context	12	1	20	0	8%	0%
Physical force	8	1	20	1	13%	5%
Specimen	13	2	20	0	15%	0%
Event	17	4	20	2	24%	10%
Body structure	22	6	20	2	27%	10%
Clinic Finding	43	12	40	1	28%	3%
Procedure	24	7	20	0	29%	0%
Organism	12	4	20	0	33%	0%
Physical object	23	8	20	0	35%	0%
Environment or geographical location	14	5	20	0	36%	0%
Staging and scales	11	4	20	0	36%	0%
Pharmaceutical	116	52	80	16	45%	20%
Substance	177	109	120	17	62%	14%
Record artifact	7	5	20	0	71%	0%
Situation with explicit content	15	12	20	1	80%	5%
Total:	720	241	660	51	33%	8%