Published in final edited form as:

Methods Inf Med. 2018 February ; 57(1): 43–53. doi:10.3414/ME17-01-0120.

Validating UMLS Semantic Type Assignments Using SNOMED CT Semantic Tags

Huanying Gu¹, Zhe He², Duo Wei³, Gai Elhanan^{4,5}, and Yan Chen⁶

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

²School of Information, Florida State University, Tallahassee, FL, USA

³Computer Science and Information Systems, Stockton University, Galloway, NJ, USA

⁴Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA

⁵Desert Research Institute, Reno, NV, USA

⁶Department of Computer Information Systems, Borough of Manhattan Community College, City University of New York, New York, NY, USA

Summary

Background: The UMLS assigns semantic types to all its integrated concepts. The semantic types are widely used in various natural language processing tasks in the biomedical domain, such as named entity recognition, semantic disambiguation, and semantic annotation. Due to the size of the UMLS, erroneous semantic type assignments are hard to detect. It is imperative to devise automated techniques to identify errors and inconsistencies in semantic type assignments.

Objectives: Designing a methodology to perform programmatic checks to detect semantic type assignment errors for UMLS concepts with one or more SNOMED CT terms and evaluating concepts in a selected set of SNOMED CT hierarchies to verify our hypothesis that UMLS semantic type assignment errors may exist in concepts residing in semantically inconsistent groups.

Methods: Our methodology is a four-stage process. 1) partitioning concepts in a SNOMED CT hierarchy into semantically uniform groups based on their assigned semantic tags; 2) partitioning concepts in each group from 1) into the disjoint subgroups based on their semantic type assignments; 3) mapping all SNOMED CT semantic tags into one or more semantic types in the UMLS; 4) identifying semantically inconsistent groups that have inconsistent assignments between semantic tags and semantic types according to the mapping from 3) and providing concepts in such groups to the domain experts for reviewing.

Results: We applied our method on the UMLS 2013AA release. Concepts of the semantically inconsistent groups in the PHYSICAL FORCE and RECORD ARTIFACT hierarchies have error

---

Correspondence to: Huanying (Helen) Gu, PhD, New York Institute of Technology, Department of Computer Science, New York, NY 10023, USA, hgu03@nyit.edu.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

rates 33% and 62.5% respectively, which are greatly larger than error rates 0.6% and 1% in semantically consistent groups of the two hierarchies.

Conclusion: Concepts in semantically inconsistent groups are more likely to contain semantic type assignment errors. Our methodology can make auditing more efficient by limiting auditing resources on concepts of semantically inconsistent groups.

Keywords

Controlled medical terminology; quality assurance; SNOMED CT; UMLS

1. Introduction

Biomedical ontologies are widely used in healthcare information systems such as electronic health records, clinical decision support systems, and administrative documents. Notably, the Unified Medical Language System (UMLS), developed and maintained by the United States National Library of Medicine (NLM), is one of the most comprehensive terminological systems in biomedicine [1]. The UMLS has a twolayer structure. Its Metathesaurus is a repository of over three million concept names and over ten million terms with hierarchical and lateral relationships among them [2]. The Semantic Network, which overlays the Metathesaurus of the UMLS, consists of 127 broad categories called semantic types [3]. Each UMLS concept is assigned one or more semantic types. For example, the concept **Myocardial Infarction** is assigned the semantic type Disease or Syndrome. As described in [4], the semantic type assignment task is done semiautomatically. The NLM first carries out the semantic type assignment algorithmically when possible using the semantic information provided by the source terminologies, then subject matter experts are recruited to conduct further reviews of the automated semantic type assignments as well as additional manual assignments. Due to the scale and complexity of the UMLS Metathesaurus, it is unavoidable that errors and inconsistencies of semantic type assignments are introduced to the UMLS during the semiautomated semantic type assignment process. As a matter of fact, in a recent paper of He et al. [5], we found that the editors of the UMLS reintroduced inconsistencies and errors of the semantic type assignments that were corrected in earlier versions. For example, the combination of semantic types *Congenital Abnormality* and *Finding* was assigned to concepts in 2007AB. This combination of semantic types disappeared in 2008AA and reappeared in 2011AA. Due to the size of the UMLS, semantic type assignment errors are unavoidable and hard to detect [5]. It is thus imperative to devise automated and semiautomated techniques to detect the semantic type assignment errors and aid the ongoing quality assurance for the UMLS, an important part of any terminology's development life cycle [6]. It is likely that the UMLS with more accurate semantic type assignments will improve its utility in various healthrelated applications that use the UMLS for named entity recognition and semantic annotation [7, 8, 9, 10].

In this paper, we develop a methodology that programmatically detects inconsistencies between semantic type assignments of a group of concepts in the UMLS and their associated semantic tag in their source terminology SNOMED CT [11]. We hypothesize that UMLS semantic type assignment errors may exist in concepts residing in these inconsistent groups. In this method, we first partition SNOMED CT concepts into groups with the same semantic

tag. Then we partition each group such that the concepts in a subgroup have the same set of semantic types. To enable automated detection of inconsistencies of semantic tags and semantic types, we map every SNOMED CT semantic tag to one or more UMLS semantic types that semantically fit the semantic tag according to the usage notes of the semantic types. We then programmatically inspect semantic consistency between the semantic type assignment in the UMLS and its semantic tag in SNOMED CT. We demonstrate this crossvalidation QA method by applying it to 2013AA release of the UMLS, which contains January 2013 version of SNOMED CT.

The rest of the paper will be organized as follows: Section 2 briefly reviews various auditing methodologies as well as the application of the UMLS semantic types. Section 3 thoroughly describes our method of the application of semantically inconsistent groups to identify semantic type assignment errors. Section 4 presents the results, followed by conclusions and discussions in Section 5 and 6.

2. Background

2.1 SNOMED CT

SNOMED CT is one of the most comprehensive clinical healthcare terminologies in the world [11]. According to January 2017 release of SNOMED CT, it consists of more than 430,000 concepts, more than 2.5 million relationships and more than 1.3 million descriptions [12]. SNOMED CT concepts are organized into 19 hierarchies, with each hierarchy capturing a specific topic and similar types of concepts in the clinical domain, such as CLINICAL FINDING, PROCEDURE, PHARMACEUTICALS AND BIOLOGIC PRODUCTS, and so on. Within each hierarchy, concepts are organized from the general to the more specific with the ISA relations. The ‘Top Level Concepts’, which are usually the root of each hierarchy, are used to name the main branch of the hierarchy. Each concept in SNOMED CT has a fully specified name (FSN) to uniquely describe a concept and clarify its meaning. Each FSN ends with a semantic tag to represent the concept’s semantic category [11]. All the concepts of a hierarchy may share the same semantic tag. For example, all concepts in SPECIMEN hierarchy have the same semantic tag specimen that is the hierarchy’s Top Level Concept name. The majority of hierarchies have one semantic tag. Although each hierarchy stands for one category, concepts from the same hierarchy may have different semantic tags. For example, the concepts **protein diet** and **avian sarcoma** of CLINICAL FINDING hierarchy have semantic tags *finding* and *disorder* respectively. As a matter of fact, semantic tags *finding* and *disorder* are the only two semantic tags that concepts in CLINICAL FINDING hierarchy have. SNOMED CT has been first released in 2002 and then made available to U.S. users within the UMLS since 2003 [13].

2.2 Related Work on UMLS Auditing

The UMLS integrates more than 170 biomedical terminologies, including MeSH, RxNorm, SNOMED CT, CPT, and LOINC. When a concept is integrated into this large terminology system, it is assigned at least one semantic type to specify its categorization at an abstract level. To automatically identify semantic type assignment errors, a compact abstract view of the UMLS was described and a divide-and-conquer approach was proposed to handle the

categorization problem in UMLS [14]. In [15], inconsistencies between UMLS Metathesaurus and Semantic Network were detected by checking the ancestor/descendant relationships in UMLS Metathesaurus. In [5, 16], concepts with similar semantics were grouped together by using an algorithmic method. Suspicious concepts were presented to human auditors, and concepts with wrong or missing semantic type (ST) assignment became noticeable to the auditors. To support auditors to efficiently detect inconsistencies, such as semantic type assignment errors, various auditing tools have been developed and utilized in [17, 18]. The Neighborhood Auditing Tool (NAT) [19] provides a “neighborhood-based” auditing platform, where it displays the focus concept and a variety of its closed related concepts [19]. This hybrid diagram and text interface allows an auditor to present knowledge from both concept level and semantic type level. As a preventive method, a rule-based support system AdviseEditor was developed to specify whether a combination of semantic types is permissible, which can help facilitate the work of UMLS editors and prevent erroneous semantic type assignment [18].

In recent research studies [20, 21], a cross-validation strategy involving SNOMED CT’s hierarchy has been proposed to effectively and efficiently identify UMLS semantic type assignment errors [20, 21]. Gu et al. proposed a UMLS quality assurance (QA) methodology based on the 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. The auditing results confirmed our hypothesis [20].

2.3 Applications of UMLS Semantic Types

The semantic types of the UMLS have been widely used in natural language processing [22, 23, 24], ontology learning [25], and various health informatics applications [7, 8, 9, 10, 26]. The semantic types are often used to capture the semantic information of the terms in specific categories in order to acquire their precise meanings [22]. Albright et al. created annotated clinical narratives with syntactic and semantic labels using UMLS semantic types to facilitate the advancement of clinical natural language processing [23]. Zhang et al. used UMLS semantic types to extract and categorize three types of new information in clinical notes (i.e., problems, medications, and laboratory information) that is not copied and pasted from previous notes and performed a longitudinal analysis of these information in the notes [24]. Hoxha et al. presented an automated method for taxonomy learning using agglomerative hierarchical clustering [25]. Luo et al. introduced a system called SimQ to retrieve similar consumer questions, which reduced the response delay by instantly providing related questions and answers [7]. In the SimQ system, consumer questions are represented with UMLS semantic types and syntactic features (part-of-speech); the similarities among the questions are then computed for retrieving similar questions. Park et al. analyzed the coverage of the UMLS concepts and semantic types in consumers’ generated text on social platforms such as blogs and social Q&As [8]. Yu and He used the semantic types to analyze the semantic coverage of the consumer terms in the online communities of deaf and autism [27]. Weng et al. developed EliXR – a semantic representation for clinical trial eligibility criteria to support the electronic patient eligibility determination [9]. In the EliXR system, 20 UMLS semantic types were used to identify related medical concepts in the freetext eligibility criteria. Frequently occurring semantic

patterns were observed to capture the complex semantic information in the eligibility criteria. He et al. recently developed a tool called *simiTerm* to identify consumer terms in social media free text that were contextually and syntactically similar to existing terms in the Open Access and Collaborative Consumer Health Vocabulary [10]. To capture the similarities among terms, the *simiTerm* tool annotated the semantic information of the given term and the terms in its context using the UMLS semantic types [10]. Due to the wide use of semantic types in various applications, erroneous semantic type assignments of the UMLS concepts can negatively affect these applications such as those mentioned above and therefore detecting and correcting these errors are important.

3. Methods

Our methodology is a fourstage process.

We first partition concepts of a SNOMED CT hierarchy into groups such that concepts in a group have the same semantic tag. We use $\text{Concept}(H)$ to represent the set of all concepts in hierarchy H . $\text{Concept}_{tg}(H)$ is used to denote a subset of $\text{Concept}(H)$, which contains concepts with the same semantic tag tg in hierarchy H . Since each concept has only one semantic tag, all subsets $\text{Concept}_{tg}(H)$ of $\text{Concept}(H)$ are disjoint. For example, the set of all 97,630 concepts of hierarchy CLINICAL FINDING, $\text{Concept}(\text{CLINICAL FINDING})$, is divided into two disjoint subsets, $\text{Concept}_{finding}(\text{CLINICAL FINDING})$ that contains 33,038 concepts with semantic tag *finding*, $\text{Concept}_{disorder}(\text{CLINICAL FINDING})$ with 64,592 concepts that have semantic tag *disorder*. If hierarchy H contains only one semantic tag tg , $\text{Concept}(H)$ will be the same as $\text{Concept}_{tg}(H)$. For example, all concepts of the SPECIMEN hierarchy have the same semantic tag *specimen*. Therefore, $\text{Concept}(\text{SPECIMEN})$ is the same as $\text{Concept}_{specimen}(\text{SPECIMEN})$.

When a SNOMED CT concept is integrated into the UMLS, it is assigned one or more semantic types that represent the concept's semantic categorization. Concepts in $\text{Concept}_{tg}(H)$ can be assigned to the same semantic type or different semantic types. For example, concepts **Animal trials**, **Clinical drug trials**, **Daily bathing**, and **Gene replacement therapy** are all from $\text{Concept}_{procedure}(\text{PROCEDURE})$, but their semantic type assignments are not all the same. **Animal trials** and **Clinical drug trials** are assigned to the same semantic type *Research Activity*, while **Daily bathing**'s semantic type is *Daily or Recreational Activity*, and **Gene replacement therapy** is assigned to both semantic types *Therapeutic or Preventive Procedure* and *Molecular Biology Research Technique* in the UMLS.

In the second stage, all concepts in $\text{Concept}_{tg}(H)$ are further partitioned into the disjoint groups such that all concepts of each group have the same semantic type assignment – the set of semantic types assigned to concepts. We use $\text{Group}_{sty}(\text{Concept}_{tg}(H))$ to represent a subset of $\text{Concept}_{tg}(H)$, in which all concepts have the same semantic type assignment sty . Note that a sty can consist of single semantic type or multiple semantic types, i.e., in the previous example, **Animal trials** and **Clinical drug trials** are members of $\text{Group}_{\{Research Activity\}}(\text{Concept}_{procedure}(\text{PROCEDURE}))$ and **Daily bathing** is a member of $\text{Group}_{\{Daily or Recreational Activity\}}(\text{Concept}_{procedure}(\text{PROCEDURE}))$. Both groups have single

semantic type assignments $\{Research\ Activity\}$ and $\{Daily\ or\ Recreational\ Activity\}$ respectively, while **Gene replacement therapy** is the member of Group $\{Therapeutic\ or\ Preventive\ Procedure,\ Molecular\ Biology\ Research\ Technique\}$ (Concept_{procedure}(PROCEDURE)) whose semantic type assignment consists of two semantic types *Therapeutic or Preventive Procedure* and *Molecular Biology Research Technique*. For example, concepts of Concept_{procedure}(PROCEDURE) are partitioned into 34 disjoint groups. Four of them possess semantic type assignments consisting of at least two semantic types.

In order to enable automatic inconsistency detection, every semantic tag tg is mapped to one or more UMLS Semantic Types that semantically fit the tg in the third stage of the process. The mapping was done through the following three steps: 1) identification of the introduction concept of the semantic tag tg . This can be the root concept of the hierarchy H with semantic tag tg , or a nonroot concept with semantic tag tg which is different from that of its parents (see Figure 1); 2) review of the FSNs of the introduction concept and its child concepts with the same semantic tag tg and identification of appropriate semantic type(s) by reviewing the definitions and usage notes [3] for semantic type(s); and 3) map semantic tag tg to semantic type(s) that are identified in 2). For example, in CLINICAL FINDING hierarchy, concept **clinical finding** is the root concept of hierarchy with the semantic tag *finding*. Thus it is identified as the introduction concept of semantic tag *finding*. After reviewing the FSNs of **clinical finding**, and its child concepts, such as **bleeding**, **calculus finding**, **edema**, and etc. with the same semantic tag *finding*, we mapped the semantic tag *finding* to semantic type *finding*, which was defined as “*That which is discovered by direct observation or measurement of an organism attribute or condition, including the clinical history of the patient. The history of the presence of a disease is a ‘Finding’ and is distinguished from the disease itself.*” [3] For the other semantic tag *disorder* in this hierarchy, **disease** was identified as the introduction concept since its parent concept **clinical finding** has the semantic tag *finding* that is different from *disorder*. After carefully reviewing FSNs of **disease** and its child concepts with the semantic tag *disorder* we conclude semantic tag *disorder* can be mapped to multiple semantic types *Mental or Behavioral Dysfunction*, *Disease or Syndrome*, *Congenital Abnormality*, *Acquired Abnormality*, *Injury or Poisoning*, *Pathologic Function* and *finding*.

The fourth stage is to verify the consistency between semantic tag tg and semantic type assignment sty of Group_{sty}(Concept_{tg}(H)). For each Group_{sty}(Concept_{tg}(H)), we programmatically inspect semantic consistency between its semantic type assignment sty in the UMLS and its semantic tag tg in SNOMED CT. The consistency is determined by whether there exists an agreement between semantic type sty of Group_{sty}(Concept_{tg}(H)) and the mapped semantic type(s) ($msty$) or their descendants of a semantic tag tg . For example, semantic type *bacterium* $\in \{bacterium\}$ of Group_{bacterium}(Concept_{organism}(ORGANISM)) is a child of semantic tag *organism*'s mapped semantic type *organism*. Therefore, the semantic type assignment $\{bacterium\}$ for the concepts of Group_{bacterium}(Concept_{organism}(ORGANISM)) is considered consistent with semantic tag *organism* in the SNOMED CT. However, semantic type *food* $\{food\}$ of Group_{food}(Concept_{cell structure}(BODYSTRUCTURE)) is considered inconsistent, because the semantic

tag *cell structure* was mapped to semantic type *cell component* that doesn't have semantic type *food* as a descendant.

We present our semantic inconsistency detection algorithm as on the left.

Using the above algorithm, we identify a list of concept groups for each SNOMED CT hierarchy with semantic inconsistency between semantic type assignment and semantic tag. These groups will be provided to domain experts for further reviewing. We will evaluate concepts in a selected set of hierarchies to verify our hypothesis that UMLS semantic type assignment errors may exist in concepts residing in semantically inconsistent groups.

Algorithm SemanticInconsistencyDetection

```

let SH be the set of all SNOMED Hierarchies, |SH|= 19;
let T be the set of all semantic tags in the SNOMED CT, |T|=39;
let ST be the set of all semantic types of semantic network of UMLS, |ST|=133;

for every hierarchy H ∈ SH
  let set SuspiciousGroup(H) = ∅;
  generate set Concept(H) = {c|c is the concept in hierarchy H};
  generate set Tag(H) = {tg|tg ∈ T and ∃c ∈ concept(H) with semantic tag tg};
  for every semantic tag tg ∈ Tag(H)
    generate set concepttg(H) = {c|c ∈ Concept(H) whose semantic tag is tg};

    //map semantic tag tg to one or more semantic types that have the same semantics as tg
    generate set MType(tg) = {msty|msty ∈ ST and has the same semantics as tg};

    //find all descendants of semantic types in MTypes(tg)
    generate set ST_Related(tg) = {st|∃msty ∈ MTypes(tg) and st is the descendent of msty};

    //partition Concepttg(H) into a collection of disjoint groups Groupsty(Concepttg(H)) such that all
    //concepts in a group have the same semantic type assignment sty
    generate set Groupsty1(Concepttg(H)) = {Groupsty1(Concepttg(H))};
    Groupsty2(Concepttg(H)), ...,
    Groupstyn(Concepttg(H));
    generate set ST_Assignment(tg) = {sty1, sty2, ..., styn};
  for every semantic type assignment sty ∈ ST_Assignment(tg)
    if ∃s ∈ sty and s ∉ (MType(tg) ∪ ST_Assignment(tg))
      then
        sty is not consistent with tg;
        SuspiciousGroup(H)=SuspiciousGroup(H) ∪ Groupsty(Concepttg(H));

```

4. Results

Our auditing method was applied on the UMLS 2013AA and January 2013 release of the SNOMED CT. For every hierarchy, its semantic tag(s) and their mapped semantic types are listed in Table 1. This mapping process was done semiautomatically by programmatically identifying introduction concepts of semantic tags and manually reviewing (by YC and GE, two of the authors) their FSNs and the definitions of semantic types. Out of 19 hierarchies,

nine hierarchies have only one semantic tag, and other 10 have 2 to 7 semantic tags. Each semantic tag, except *product*, appears in only one hierarchy. There are 12 concepts with semantic tag *product* that belong to both hierarchies PHARMACEUTICAL PRODUCT and PHYSICAL OBJECT. The 39 different SNOMED CT semantic tags were mapped to 47 different semantic types of UMLS Semantic Network. Some of the semantic tags were mapped to single semantic type, e.g., semantic tags *cell* and *event* were mapped to semantic types *Cell* and *Event* respectively, while others were mapped to multiple semantic types, e.g., semantic tag *procedure* was mapped to three semantic types *Diagnostic Procedure*, *Laboratory Procedure*, and *Therapeutic or Preventive Procedure*. Due to the different levels of semantic refinements between semantic tags and semantic types, one or more semantic tags may be mapped to the same semantic type. For example, semantic tags *attribute*, *link assertion*, and *qualifier value* were all mapped to semantic type *Idea or Concept*. There are eight such semantic types that receive mapping from different semantic tags.

For each hierarchy, the concept groups with semantic inconsistencies between semantic type assignment and semantic tag were programmatically identified. For example, seven out of nine concept groups were detected with semantic inconsistencies in the PHYSICAL FORCE hierarchy (Table 2). We also reviewed all concepts in the hierarchy for verification purposes and the results were shown in the column “# of concepts with error” of Table 2. Six out of 18 concepts in seven semantically inconsistent groups have semantic type assignment errors. For example, both concept **Ultraviolet radiation in diagnosis** in Group {*Diagnostic Procedure*} (Concept_{physicalforce}(PHYSICAL FORCE) and concept **Continued movement in** Group {*Qualitative Concept*} (Concept_{physicalforce}(PHYSICAL FORCE) were identified by the program as potential erroneous concept with improper semantic assignments. Upon examination of the concepts, the auditor considered that concept **Ultraviolet radiation in diagnosis** should be assigned semantic type *Natural Phenomenon or Process*, and concept **Continued movement** should have *Phenomenon or Process* as its semantic type. In two semantically consistent groups, concept **high altitude** is the only concept in Group {*Natural Phenomenon or Process*} (Concept_{physicalforce}(PHYSICAL FORCE) which has semantic type assignment error. It should be assigned to the semantic type *Quantitative Concept*, not *Natural Phenomenon or Process*. In Table 3, we showed the results for hierarchy RECORD ARTIFACT.

Table 4 shows the set of semantic tags (Semantic Tags), the number of concept disjoint groups for each semantic tag (# Of Groups), the number of concepts in the disjoint groups for each semantic tag (# Of Concepts), the number of groups with semantic inconsistency between concepts' semantic type assignment and semantic tag (# of Semantically Inconsistent Groups), and the number of concepts in semantically inconsistent groups (# of Concepts with Semantically Inconsistent Groups) for every hierarchy. These detected semantically inconsistent groups will be provided to domain experts for further review.

We assume that concepts in the groups that are considered to be semantically consistent most likely have no semantic type assignment errors. To confirm our hypothesis, we selected two hierarchies PHYSICAL FORCE and RECORD ARTIFACT with manageable size and reviewed all their concepts. Table 5 shows the statistics of semantic type assignment errors. Note that six out of 18 concepts in seven detected groups have semantic type assignment

errors in PHYSICAL FORCE hierarchy, while one out of 157 concepts in two semantically consistent groups was found semantic type assignment error. The methodology achieved the accuracy 33% (6/18), the recall 86% (6/7), and the precision 100% (6/6), which only need to review 18 concepts of identified groups (10% of total concepts in the hierarchy). For hierarchy RECORD ARTIFACT, we got similar results. As shown in Table 6, the concepts in semantically inconsistent groups have statistically significantly higher percentage of erroneous semantic type assignments than the concepts in semantically consistent groups in both PHYSICAL FORCE hierarchy and RECORD ARTIFACT hierarchies ($p < 0.0001$).

Table 7 provides the erroneous semantic type assignments and their corrections in the Inconsistent Semantic groups of the PHYSICAL FORCE and RECORD ARTIFACT hierarchies. The concept **Temperature change** which resides in the PHYSICAL FORCE hierarchy with semantic tag *physical force* was assigned the semantic type *Qualitative Concept*. The auditors deemed it to be erroneous assignment and suggested the correct assignment to be *Natural Phenomenon or Process*.

Furthermore, we reviewed two other semantically consistent groups
Group {Anatomical Abnormality}(Concept_{morphologic abnormality}(BODY STRUCTURES)) and
Group {Diagnostic Procedure, Therapeutic or Preventive Procedure}(Concept_{procedure}(PROCEDURE))
with 153 and 235 concepts respectively. No semantic type assignment errors were found in the first group, and only 12 concepts were found with semantic type assignment errors in the second group. Only 3% of total number of concepts in these two semantically consistent groups had the semantic type assignment errors. The results confirm our hypothesis.

5. Discussion

The UMLS semantic types are a cornerstone for natural language processing and numerous methodologies related to healthcare informatics applications. As such, their correct assignment is of utmost importance for the applications that rely on them. While most of the biomedical ontologies that have been integrated into the UMLS do not provide semantic tagging of their content, SNOMED CT is an exception with its semantic tags. While SNOMED CT's semantic tags are not as granular as the UMLS's semantic types, one can expect a high degree of correlation between SNOMED CT's semantic tags and the UMLS semantic type assignments.

In this study, based on premapping of SNOMED CT's semantic tags to UMLS's semantic types, we formulated and tested a methodology to evaluate the correctness of the UMLS assignment of semantic types to SNOMED CT concepts. The method can validate correctness and highlight inconsistencies that may occur during the semiautomated semantic type assignment process.

We tested our methodology on the PHYSICAL FORCE and the RECORD ARTIFACT hierarchies of SNOMED CT. Our results indicate that seven and four groups (respectively) had semantic type assignments that were inconsistent with the original semantic tags of SNOMED CT affecting four and 5.3 percent of the concepts in the respective hierarchies (Tables 2, 3, and 5). For example, the SNOMED CT concept **CMS supervisory note** with

the semantic tag **record artifact** was assigned the semantic type *Health Care Activity*. However, SNOMED CT's semantic tag, its SNOMED CT ancestor (**record type**) and the relevant SNOMED CT hierarchy (RECORD ARTIFACT) all provide a context that does not support the notion that *Health Care Activity* actually is an activity. Based on our prior mapping of SNOMED CT semantic tags to UMLS semantic type and the contextual information we can conclude that the assigned UMLS semantic type is erroneous and, in fact, this is not an activity but a form type and therefore an *Intellectual Product* UMLS semantic type.

Although 585 out of 1,118 concept groups (52%) were detected with semantic inconsistency, the number of concepts in those groups was less than 8% of the total number of concepts in the UMLS. This result indicates that the detected groups with semantic inconsistency should be small in size, which is in line with our previous research [20]. Our hypothesis is that concepts in those semantic inconsistency groups will have semantic type assignment errors. On the other hand, concepts in the groups that are considered to be semantically consistent most likely have no semantic type assignment errors. Our results confirm our hypothesis.

Our methodology achieved an accuracy of 33% (6/18), recall of 86% (6/7), and precision of 100% (6/6) for the PHYSICAL FORCE hierarchy with similar results for the RECORD ARTIFACT hierarchy. Thus, one only needs to review 18 and 16 concepts within these respective groups. Such reduction of the number of concepts that should be reviewed enhances the effective use of traditionally limited resources that are directed towards retrospective analysis. For any future integration of additional or modified SNOMED CT concepts into the UMLS, our algorithmic approach can be easily integrated into the process to reduce the chance of semantic type assignment errors. Furthermore, the mapping between SNOMED CT's semantic tags and the UMLS's semantic types can be used to enhance the integration effort of any biomedical ontology that overlaps with the domain coverage of SNOMED CT.

Arguably, there is a certain level of subjectivity in the mapping between UMLS semantic types and SNOMED semantic tags. Nevertheless, a small number of arguable mappings should not affect the validity of this method and the overall finding of this study. In this study, our intention was not test the validity of semantic tag assignments within SNOMED CT. As a gold standard, SNOMED CT semantic tag assignments may not be perfect. Ceusters and Bona found frequent changes in semantic tags, the nature of which suggests ambiguity in the concepts, SNOMED CT's underlying ontological model, or both [28]. Accordingly, we observe that changes in SNOMED CT semantic tags can be associated with semantic type assignments for the same concept between different releases of the UMLS. For example the SNOMED CT concept **Electrons** had semantic tag *Physical force* in earlier releases, which may not be appropriate. In the 2017 version, the semantic tag has been changed to *Substance*. Despite this change, our methodological mapping between the SNOMED CT semantic tags and the UMLS semantic types still stands and our mapping between semantic tag *Physical force* and semantic type *Natural Phenomenon or Process* in Table 1 remains in effect. The semantic type assignment of the concept **Electrons** was *Quantitative concept* in the 2013AA version, which was identified as the inconsistency using our methodology. In the 2017AA version, the concept's semantic type has been changed to

Element, ion, or isotope, arguably inappropriate assignment as well. It is of note that in the 2017AA version the concept **Positrons** has the same semantic tag *Physical force* and semantic type assignment *Quantitative concept* as ones in the 2013AA version. As both **Electrons** and **Positrons** should likely have the same semantic tag and semantic type, these inconsistencies within SNOMED CT and the UMLS highlight the fact that neither can offer sufficiently satisfying semantic tags or types.

However, it is not clear to what extent semantic type assignments in UMLS releases may actually reflect changes to the semantic tags of SNOMED CT concepts. For example, while the concept **Ultraviolet radiation in diagnosis** (Table 7) has the semantic tag of *Physical force*, in the UMLS it has the semantic type *Diagnostic Procedure*. This is a clear mismatch and a reasonable indication that SNOMED CT semantic tags are not a major consideration, if any, during the mapping and the semantic type assignment processes of the UMLS. Had the semantic tag been considered during the process, it should have excluded that specific assignment in the UMLS and, hence, highlights the applicability of our methodology.

Admittedly, the cited incorrect semantic type assignments in this paper may not reflect the prevalence of the actual impact of incorrect semantic type assignments on practical performance metrics. Our previous work has shown the frequency of incorrect semantic type assignments [20]. We also reported the decrease in performance of the various kinds of systems that rely on them, which is directly attributable to incorrect semantic type assignments [20, 29].

In this work, we used the 2013AA version of the UMLS. In 2015 (2015AA release), the UMLS deleted six semantic types and the UMLS was left with only 127 semantic types. However, none of the deleted semantic types were involved in our work since none of them were mapped to SNOMED CT semantic tags. Therefore the use of the 2013AA release of the UMLS does not affect the validity of our results.

In the future, we plan to expand our investigation into the semantic type assignments for other hierarchies of SNOMED CT using our methodology with specific focus regarding instances where one SNOMED CT semantic tag may be mapped to multiple UMLS semantic types with hierarchical relationships between them.

6. Conclusions

In this work, we devise a crossvalidation strategy that leverages the semantic tags of SNOMED CT concepts to audit the semantic type assignments of the UMLS concepts. We show that concepts in semantically inconsistent groups are more likely to contain semantic type assignment errors. Our methodology can be utilized to optimize the use of the limited auditing resources by concentrating only on concepts in semantically inconsistent groups.

Acknowledgments

Funding

Research reported in this publication was partially supported by the National Cancer Institute of the National Institutes of Health (NIH) under Award Number R01CA190779. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Bodenreider O The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32(Database issue): D267–270. [PubMed: 14681409]
2. The Statistics of the UMLS 2016AB Release [May 1, 2017] Available from: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html.
3. The UMLS Semantic Network [cited 2012 Dec 5] Available from: <https://semanticnetwork.nlm.nih.gov/>.
4. McCray AT, Hole WT. The scope and structure of the first version of the UMLS Semantic Network. *Proc 14th Annu Symp Comput Appl Med Care*; Los Alamitos, CA 1990 p. 126–130.
5. He Z, Morrey CP, Perl Y, Elhanan G, Chen L, Chen Y, Geller J. Sculpting the UMLS Refined Semantic Network. *Online J Public Health Inform* 2014; 6(2): e181. [PubMed: 25422719]
6. Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as part of the terminology design life cycle. *J Am Med Inform Assoc* 2006; 13(6): 676–690. [PubMed: 16929044]
7. Luo J, Zhang GQ, Wentz S, Cui L, Xu R. SimQ: realtime retrieval of similar consumer health questions. *J Med Internet Res* 2015; 17(2): e43. [PubMed: 25689608]
8. Park MS, He Z, Chen Z, Oh S, Bian J. Consumers' Use of UMLS Concepts on Social Media: DiabetesRelated Textual Data Analysis in Blog and Social Q&A Sites. *JMIR Med Inform* 2016; 4(4): e41. [PubMed: 27884812]
9. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. *J Am Med Inform Assoc* 2011; 18 Suppl 1: i116–124. [PubMed: 21807647]
10. He Z, Chen Z, Oh S, Hou J, Bian J. Enriching consumer health vocabulary through mining a social Q&A site: a similaritybased approach. *J Biomed Inform* 2017; 69: 75–85. [PubMed: 28359728]
11. SNOMED CT User Guide [cited 2013 Apr 2] Available from: http://www.ihtsdo.org/fileadmin/user_upload/doc/en_us/ug.html.
12. Release Notes of SNOMED CT International Edition [May 1, 2017] Available from: <https://www.nlm.nih.gov/healthit/snomedct/international.html>.
13. Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. *J Am Med Inform Assoc* 2005; 12(4): 486–494. [PubMed: 15802483]
14. Gu HH, Perl Y, Elhanan G, Min H, Zhang L, Peng Y. Auditing concept categorizations in the UMLS. *Artif Intell Med* 2004; 31(1): 29–44. [PubMed: 15182845]
15. Cimino JJ, Min H, Perl Y. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. *J Biomed Inform* 2003; 36(6): 450–461. [PubMed: 14759818]
16. Chen Y, Gu HH, Perl Y, Geller J. Structural groupbased auditing of missing hierarchical relationships in UMLS. *J Biomed Inform* 2009; 42(3): 452–467. [PubMed: 18824248]
17. Morrey CP, Geller J, Halper M, Perl Y. The Neighborhood Auditing Tool: a hybrid interface for auditing the UMLS. *J Biomed Inform* 2009; 42(3): 468–489. [PubMed: 19475725]
18. Geller J, He Z, Perl Y, Morrey CP, Xu J. Rulebased support system for multiple UMLS semantic type assignments. *J Biomed Inform* 2013; 46(1): 97–110. [PubMed: 23041716]
19. Morrey CP. Auditing the Unified Medical Language System and Enhancing the Refined Semantic Network: Dissertation in the Department of Computer Science, New Jersey Institute of Technology; 2009.
20. Gu H, Chen Y, He Z, Halper M, Chen L. Quality Assurance of UMLS Semantic Type Assignments Using SNOMED CT Hierarchies. *Methods Inf Med* 2016; 55(2): 158–165. [PubMed: 25925776]
21. Wei D, Halper M, Elhanan G, editors. Using SNOMED semantic concept groupings to enhance semantictype assignments in the UMLS. *Proceedings of the 2nd ACM SIGHT International Health Informatics Symposium*; 2012; Miami, Florida, USA: ACM.

22. Sfakianaki P, Koumakis L, Sfakianakis S, Iatraki G, Zacharioudakis G, Graf N, Marias K, Tsiknakis M. Semantic biomedical resource discovery: a Natural Language Processing framework. *BMC Med Inform Decis Mak* 2015; 15: 77. [PubMed: 26423616]
23. Albright D, Lanfranchi A, Fredriksen A, Styler WF, Warner C, Hwang JD, Choi JD, Dligach D, Nielsen RD, Martin J, Ward W, Palmer M, Savova GK. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *J Am Med Inform Assoc* 2013; 20(5): 922–930. [PubMed: 23355458]
24. Zhang R, Pakhomov S, Melton GB. Longitudinal analysis of new information types in clinical notes. *AMIA Jt Summits Transl Sci Proc* 2014; 2014: 232–237. [PubMed: 25717418]
25. Hoxha J, Jiang G, Weng C. Automated learning of domain taxonomies from text using background knowledge. *J Biomed Inform* 2016; 63: 295–306. [PubMed: 27597572]
26. Fan JW, Li J, Lussier YA. Semantic Modeling for Exposomics with Exploratory Evaluation in Clinical Context. *J Healthc Eng* 2017; 2017: 3818302. [PubMed: 29065591]
27. Yu B, He Z, editors. Exploratory Textual Analysis of Consumer Health Languages for People Who are Deaf/Hard of Hearing. *Proceedings of 2017 IEEE International Conference on Bioinformatics and Biomedicine*; 2017; Kansas City, MO: IEEE.
28. Ceusters W, Bona JP. Analyzing SNOMED CT's Historical Data: Pitfalls and Possibilities. *AMIA Annu Symp Proc* 2016; 2016: 361–370. [PubMed: 28269831]
29. Chen Y, Gu HH, Perl Y, Halper M, Xu J. Expanding the extent of a UMLS semantic type via group neighborhood auditing. *J Am Med Inform Assoc* 2009; 16(5): 746–757. [PubMed: 19567802]

Figure 1.
Illustration of introduction concepts of semantic tags in SNOMED CT hierarchy.

Table 1

Mappings of SNOMED CT semantic tags and UMLS semantic types.

SNOMED CT Hierarchies	SNOMED CT Semantic Tags	Mapped UMLS Semantic Types
BODY STRUCTURE	body structure	Body Space or Junction, Body Location or Region
	cell	Cell
	cell structure	Cell Component
CLINICAL FINDING	morphologic abnormality	Anatomical Abnormality, Neoplastic Process
	disorder	Acquired Abnormality, Congenital Abnormality, Disease or Syndrome, Finding, Injury or Poisoning, Mental or Behavioral Dysfunction, Pathologic Function
	finding	Finding
ENVIRONMENT OR GEOGRAPHICAL LOCATION	environment	Geographic area, Health Care Related Organization, Manufactured Object, Phenomenon or Process, Spatial Concept
	environment / location	Geographic Area, Spatial Concept
	geographic location	Geographic Area
EVENT	event	Event
LINKAGE CONCEPT	attribute	Idea or Concept
	link assertion	Idea or Concept
	linkage concept	Functional Concept, Intellectual Product, Qualitative Concept, Quantitative Concept, Spatial Concept, Temporal Concept
OBSERVABLE ENTITY	observable entity	Clinical Attribute, Finding
ORGANISM	organism	Organism
PHARMACEUTICAL PRODUCT	product	Chemical, Clinical Drug
PHYSICAL FORCE	physical force	Natural Phenomenon or Process, Phenomenon or Process
PHYSICAL OBJECT	physical object	Biomedical or Dental Material, Manufactured objects
	product	Manufactured Objects
PROCEDURE	procedure	Diagnostic Procedure, Laboratory Procedure, Therapeutic or Preventive Procedure
	regime/therapy	Therapeutic or Preventive Procedure
QUALIFIER VALUE	qualifier value	Intellectual Product, Idea or Concept
RECORD ARTIFACT	record artifact	Intellectual Product
SITUATION WITH EXPLICIT CONTENT	situation with explicit content	Finding
SOCIAL CONTEXT	ethnic group	Population Group
	life style	Behavior
	occupation	Occupation or Discipline, Professional or Occupational Group

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

SNOMED CT Hierarchies	SNOMED CT Semantic Tags	Mapped UMLS Semantic Types
	<i>person</i>	<i>Age Group, Family Group, Patient or Disabled Group, Population Group, Professional or Occupational Group</i>
	<i>racial group</i>	<i>Population Group</i>
	<i>religion/philosophy</i>	<i>Findings, Idea or Concept, Population Group</i>
	<i>social concept</i>	<i>Idea or Concept</i>
	<i>inactive concept</i>	<i>Functional Concept</i>
	<i>namespace concept</i>	<i>Functional Concept</i>
	<i>navigational concept</i>	<i>Functional Concept</i>
	<i>special concept</i>	<i>Functional Concept</i>
SPECIAL CONCEPT	<i>specimen</i>	<i>Body Part, Organ, or Organ Component, Body Substance, Substance, Tissue</i>
	<i>assessment scale</i>	<i>Classification, Intellectual Product</i>
SPECIMEN	<i>staging scale</i>	<i>Classification, Intellectual Product</i>
STAGING AND SCALES	<i>tumor staging</i>	<i>Classification, Intellectual product</i>
	<i>Substance</i>	<i>Substance</i>
SUBSTANCE		
Total: 19	39	47

Table 2

Semantic inconsistency groups detected in hierarchy PHYSICAL FORCE.

Semantic Tag	Semantic Type Assignments	Semantically Inconsistent	# Of Concepts	# Of Concepts with Error
<i>physical force</i>	<i>{Natural Phenomenon or Process}</i>	0	126	1
	<i>{Phenomenon or Process}</i>	0	31	0
	<i>{Quantitative Concept}</i>	1	6	2
	<i>{Manufactured Object}</i>	1	4	0
	<i>{Physical Object}</i>	1	3	0
	<i>{Qualitative Concept}</i>	1	2	2
	<i>{Medical Device}</i>	1	1	0
	<i>{Element, Ion, or Isotope}</i>	1	1	1
	<i>{Diagnostic Procedure}</i>	1	1	1
Total	9	7	175	7

Table 3

Semantic inconsistency groups detected in hierarchy RECORD ARTIFACT.

Semantic Tag	Semantic Type Assignments	Semantically Inconsistent	# Of Concepts	# Of Concepts with Error
<i>record artifact</i>	<i>{Intellectual Product}</i>	0	209	2
	<i>{Finding}</i>	1	9	3
	<i>{Functional Concept}</i>	1	3	3
	<i>{Health Care Activity}</i>	1	2	2
	<i>{Manufactured Object, Intellectual Product}</i>	1	2	2
Total	5	4	225	12

Table 4

Statistics of detected semantic inconsistency groups.

Hierarchies	Semantic Tags	# Of Groups	# Of Concepts	# Of Semantically Inconsistent Groups
BODY STRUCTURE	<i>body structure</i>	28	25,261	15
	<i>cell</i>	7	625	6
	<i>cell structure</i>	17	517	16
CLINICAL FINDING	<i>morphologic abnormality</i>	41	4,606	37
	<i>Disorder</i>	39	64,592	24
	<i>finding</i>	81	33,038	78
ENVIRONMENT OR GEOGRAPHICAL LOCATION	<i>environment</i>	17	1,088	8
	<i>environment / location</i>	1	1	0
	<i>geographic location</i>	1	618	0
EVENT	<i>event</i>	22	3,632	2
LINKAGE CONCEPT	<i>attribute</i>	51	1,111	42
	<i>link assertion</i>	1	8	0
	<i>linkage concept</i>	1	1	1
OBSERVABLE ENTITY	<i>observable entity</i>	60	8,221	56
ORGANISM	<i>organism</i>	20	32,319	5
PHARMACEUTICAL PRODUCT	<i>product</i>	148	15,570	9
PHYSICAL FORCE	<i>physical force</i>	9	175	7
PHYSICAL OBJECT	<i>physical object</i>	21	4,489	16
	<i>product</i>	1	12	0
	<i>procedure</i>	34	50,039	30
PROCEDURE	<i>regime/therapy</i>	13	2,455	12
QUALIFIER VALUE	<i>qualifier value</i>	83	8,881	71
RECORD ARTIFACT	<i>record artifact</i>	5	225	4
SITUATION WITH EXPLICIT CONTENT	<i>situation with explicit content</i>	27	3,233	24
SOCIAL CONTEXT	<i>ethnic group</i>	1	259	0
	<i>life style</i>	3	21	1
	<i>occupation</i>	7	3816	4
	<i>person</i>	10	421	5

Hierarchies	Semantic Tags	# Of Groups	# Of Concepts	# Of Semantically Inconsistent Groups
	<i>racial group</i>	2	20	1
	<i>religion/philosophy</i>	6	203	3
	<i>social concept</i>	11	27	7
SPECIAL CONCEPT	<i>inactive concept</i>	2	8	1
	<i>namespace concept</i>	4	153	3
	<i>navigational concept</i>	41	640	40
	<i>special concept</i>	1	1	0
SPECIMEN	<i>specimen</i>	17	1,327	10
STAGING AND SCALES	<i>assessment scale</i>	9	1,056	8
	<i>staging scale</i>	2	16	0
	<i>tumor staging</i>	6	214	4
SUBSTANCE	<i>Substance</i>	268	23,938	35
Total: 19	39	1,118	292,837	585

Table 5

Statistics of semantic type assignment errors in detected semantic inconsistency groups.

Characteristics	Hierarchies	
	PHYSICAL FORCE	RECORD ARTIFACT
# Of Concepts	175	225
# Of Concepts with Error	7	12
# Of Concepts in Semantically Inconsistent Groups	18	16
# Of Concepts in Semantically Inconsistent Groups with Error	6	10
Accuracy	6/18=33%	10/16=63%
Recall	6/7=86%	10/12=83%
Precision	6/6=100%	10/10=100%

Table 6

The comparison of semantic type assignment errors in semantically consistent groups and semantically inconsistent groups within the PHYSICAL FORCE and RECORD ARTIFACT hierarchies.

Characteristics	Hierarchies		
	PHYSICAL FORCE		RECORD ARTIFACT
	# Of Concepts without Errors	# Of Concepts with Errors	# Of Concepts with Errors
# Of Concepts in Semantically Inconsistent Groups	12	6	10
# Of Concepts in Semantically Consistent Groups	156	1	2
Fisher's Exact Test, 2-tail	P < 0.001		P < 0.001

Table 7

Semantic type assignment suggestions for the concepts in Semantically Inconsistent groups of the PHYSICAL FORCE and RECORD ARTIFACT hierarchies.

Hierarchies	Concept Name	Original Inconsistent Semantic Type Assignments	Suggested Semantic Type Assignments
PHYSICAL FORCE	Ultraviolet radiation in diagnosis	<i>Diagnostic Procedure</i>	<i>Natural Phenomenon or Process</i>
	Accelerated particle	<i>Element, Ion, or Isotope</i>	<i>Phenomenon or Process</i>
	Continued movement	<i>Qualitative Concept</i>	<i>Phenomenon or Process</i>
	Temperature change	<i>Qualitative Concept</i>	<i>Natural Phenomenon or Process</i>
	Electrons	<i>Quantitative Concept</i>	<i>Natural Phenomenon or Process</i>
	Positrons	<i>Quantitative Concept</i>	<i>Natural Phenomenon or Process</i>
RECORD ARTIFACT	Summary of needs	<i>Finding</i>	<i>Intellectual Product</i>
	Administrative documentation	<i>Finding</i>	<i>Intellectual Product</i>
	Statement of needs	<i>Finding</i>	<i>Intellectual Product</i>
	Report for Coroner	<i>Functional Concept</i>	<i>Intellectual Product</i>
	Social security claim report	<i>Functional Concept</i>	<i>Intellectual Product</i>
	Disabled drivers orange badge	<i>Functional Concept</i>	<i>Intellectual Product</i>
	Report to benefits agency	<i>Health Care Activity</i>	<i>Intellectual Product</i>
	CMS supervisory note	<i>Health Care Activity</i>	<i>Intellectual Product</i>
	Death Certificates	<i>Manufactured Object Intellectual Product</i>	<i>Intellectual Product</i>
	Intraoperative record	<i>Manufactured Object Intellectual Product</i>	<i>Intellectual Product</i>