Research Paper ■

Expanding the Extent of a UMLS Semantic Type via Group Neighborhood Auditing

Yan Chen, PhD, Huanying Gu, PhD, Yehoshua Perl, PhD, Michael Halper, PhD, Junchuan Xu, MD, MS

Abstract **Objective:** Each Unified Medical Language System (UMLS) concept is assigned one or more semantic types (ST). A dynamic methodology for aiding an auditor in finding concepts that are missing the assignment of a given ST, **S** is presented.

Design: The first part of the methodology exploits the previously introduced *Refined Semantic Network* and accompanying refined semantic types (RST) to help narrow the search space for offending concepts. The auditing is focused in a neighborhood surrounding the extent of an RST, **T** (of **S**) called an *envelope*, consisting of parents and children of concepts in the extent. The audit moves outward as long as missing assignments are discovered. In the second part, concepts not reached previously are processed and reassigned **T** as needed during the processing of S's other RSTs. The set of such concepts is expanded in a similar way to that in the first part.

Measurements: The number of errors discovered is reported. To measure the methodology's efficiency, "error hit rates" (i.e., errors found in concepts examined) are computed.

Results: The methodology was applied to three STs: **Experimental Model of Disease (EMD), Environmental Effect of Humans,** and **Governmental or Regulatory Activity.** The **EMD** experienced the most drastic change. For its RST "**EMD** ∩ **Neoplastic Process**" (RST "**EMD**") with only 33 (31) original concepts, 915 (134) concepts were found by the first (second) part to be missing the EMD assignment. Changes to the other two STs were smaller.

Conclusion: The results show that the proposed auditing methodology can help to effectively and efficiently identify concepts lacking the assignment of a particular semantic type.

■ **J Am Med Inform Assoc.** 2009;16:746–757. DOI 10.1197/jamia.M2951.

Introduction

The concept database of the Unified Medical Language System (UMLS), the Metathesaurus (META), contains about 1.5 million concepts in its 2007AC release.[1–6] Its Semantic Network (SN) overlays a consistent categorization via the assignment of one or more of its 135 semantic types (STs) to each concept.[7,8] However, because the META is so large and inherently complex, ST-assignment errors are all but unavoidable. Furthermore, the UMLS's construction through the integration of many source vocabularies that are not necessarily consistent may contribute to such problems. The differing views of various subject experts who carry out the ST assignments can also be seen as a contributing factor. In fact, ST mis-assignments may reflect a variety of misunderstandings, including inaccurate or incorrect meanings or ambiguities with respect to concepts. An ST mis-assignment may therefore imply the presence of other errors.

In a study involving UMLS users,[9] it was clearly expressed that significant attention should be paid to auditing. The ST mis-assignments were found to be among the leading concerns. Thus, weeding them out should be an important aspect of UMLS maintenance.

Regarding a specific ST **S** (semantic types are written in bold, while concept names are in italics), there are several possibilities for an assignment error: (i) a concept may be assigned **S** incorrectly; (ii) it may be assigned **S** correctly but have errors with respect to its other semantic types; or (iii) it may be missing the assignment of **S**. The first two possibilities were addressed in Chen et al. 2008.[10] We present a methodology for dealing with the third possibility in this paper.

Randomly searching through the META for concepts that warrant an assignment of **S** is certain to be tedious and unlikely to prove fruitful. The challenge is to design an effective algorithmic technique to identify "suspicious" concepts that may be missing the assignment of **S**, using a technique similar to that of Chen et al. 2008[10] for finding other such concepts with respect to different kinds of errors. However, in contrast to Chen et al. 2008,[10] where the basis was the overall extent (i.e., set of assigned concepts) of **S**, there is no obvious set from which to commence the search for omissions of the **S** assignment. This observation makes

Affiliations of the authors: CIS Department, Borough of Manhattan Community College, CUNY (YC), New York, NY; Health Informatics Department, SHRP, University of Medicine and Dentistry of New Jersey (HG), Newark, NJ; Department of Computer Science, New Jersey Institute of Technology (YP, JX), Newark, NJ; Computer Science Department, Kean University (MH), Union, NJ.

This work was partially supported by the NLM under grant R-01-LM008445-01A2.

Correspondence: Yehoshua Perl, Structural Analysis of Biomedical Ontologies Center (SABOC), Department of Computer Science, New Jersey Institute of Technology, University Heights, Newark, NJ 07102-1982; e-mail: <perl@oak.njit.edu>.

Received for review: 08/07/08; accepted for publication: 05/06/09.

the current ST-assignment errors more difficult to uncover than those encountered in Chen et al. 2008.[10]

To overcome this difficulty, we define a guided search for the auditor that emanates outward from the extent of **S**. We proceed from the assumption that concepts requiring the assignment of **S** are in all likelihood already in the vicinity of its extent. Therefore, the search for erroneous concepts is focused in a neighborhood surrounding the extent of a semantic type. Actually, to refine the search space further, we use a refined semantic type (RST)—a subtype of a semantic type—from our previously introduced *Refined Semantic Network*[11,12] as the starting point, since concepts in the same RST extent tend to share an overarching uniform broad meaning, which is not necessarily true for the entire ST. Having a search space of concepts with uniform broad meaning tends to simplify the auditing work, since concepts lacking the expected uniform meaning naturally stand out in a review.

The first part of our methodology concentrates on auditing concepts in a neighborhood surrounding the extent of an RST **T** of **S** that we call an *envelope*. An envelope is defined with respect to the UMLS's parent/child relationships whose origins are in the various UMLS source vocabularies. From there, the search space emanates outward in a concentric progression to encompass more and more distant neighborhoods as type assignment errors continue to be discovered by the auditor. Overall, the methodology allows ancestors and descendants of the concepts in **T** to be systematically examined and possibly brought into **T**'s extent—when a previous ancestor or descendant in the progression is reassigned **T**. These ancestors and descendants are related to a concept (previously assigned an RST) via the parent/child relationships. All RSTs of **S** are, in turn, processed in this manner.

The second part of the methodology constitutes a cross-processing step, where concepts potentially needing an assignment of an RST **T** are identified and reassigned **T** while processing another RST **T**′. Subsequently, those concepts are processed in a manner similar to that of the original extent of **T** in the first part of the methodology, with various tiers of envelopes created and audited.

We demonstrate our methodology by applying it to three semantic types: **Experimental Model of Disease**, **Environmental Effect of Humans**, and **Governmental or Regulatory Activity**. The errors discovered during this effort are reported, and the effectiveness of our approach is discussed.

Background

Refined Semantic Network

We have previously introduced the Refined Semantic Network (RSN), a modified version of the existing Semantic Network, as an enhanced abstraction mechanism for the UMLS.[11,12] It consists of two types: *pure semantic types (PSTs)* and *intersection semantic types (ISTs)*. Collectively, we refer to them as *refined semantic types (RSTs)*. The RSTs are derived automatically from the existing STs in the Semantic Network (SN) and their assignments to concepts.

One PST in the RSN is defined for each ST **S** in the SN. While the PST is given the same name as its corresponding ST **S** in the SN, we will often denote it as $\mathbf{S}^R$ to avoid confusion. The assignments of $\mathbf{S}^R$ in the RSN differ from those of **S**. Specifically, $\mathbf{S}^R$ is assigned strictly to those concepts that originally had **S** as their sole ST assignment. The ISTs serve to provide assignments for the remaining concepts originally in the extent of **S**, denoted E(**S**). Such a concept will have been assigned at least one other semantic type. In fact, let us assume that some concepts were assigned **S** and one other ST, say, **U** simultaneously. This implies the existence of an IST named **S** ∩ **U** that is assigned to exactly those concepts originally assigned both **S** and **U** and no other types. The symbol "∩ " is mathematical intersection, and we use it and "intersection type" because E(**S** ∩ **U**) = E(**S**) ∩ E(**U**). That is, the extent of the IST is the intersection of the extents of the STs from the SN.

Let us note that an empty intersection of E(**S**) and the extent of another type, say, **W** means that **S** ∩ **W** would not appear in the RSN. This avoids any potential combinatorial explosion of ISTs. The ISTs can involve more than two types.

As an example, let us consider the ST **Experimental Model of Disease** (**EMD**). Figure 1 uses a Venn Diagram [13] to show some of the concepts assigned **EMD** and its overlap with **Neoplastic Process** (**NP**). The ellipses represent the respective extents of the STs. As we see, the concepts *Arthritis, Experimental* and *Disease Model*, along with 29 others, are solely assigned **EMD**. Thus, these 31 concepts would be assigned the PST $\mathbf{EMD}^R$ with respect to the RSN (see Figure 2a). By contrast, *Melanoma, Experimental* and *Experimental Hepatoma*, along with 31 more concepts, are assigned both **EMD** and **NP**. In the RSN, these 33 concepts would be assigned the IST, **EMD ∩ NP** (Figure 2a). Figure 2b shows the portion of the RSN involving **EMD**, **NP** and **EMD** ∩ **NP**.

An important characteristic of the RSTs is that they collectively serve to partition the concepts of **S**. (Overall, the RSN's types partition the entire META.) That is, all concepts have unique assignments in the RSN. A concept originally assigned just **S**, will be uniquely assigned $\mathbf{S}^R$. A concept originally assigned **S** and one or more other STs at the same time will now be uniquely assigned the appropriate IST. As a consequence, all concepts in the extent of the same RST have exactly the same ST assignments in the context of the SN. This property, which we call *semantic uniformity*, is one of the main benefits of the RSN. In previous work, we have exploited it in the creation of auditing techniques for the UMLS.[10–12,14,15] Here, again, we make use of it in the attempted expansion of the extent of a given ST.

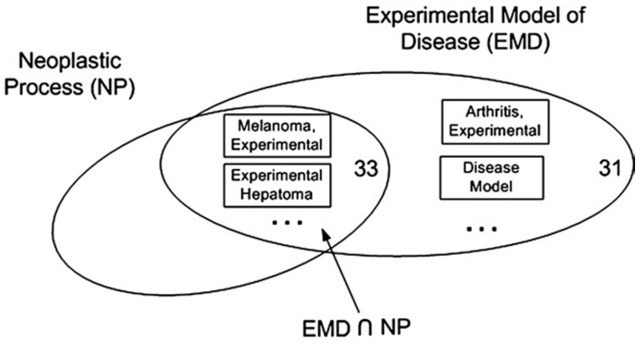

Figure 1. Semantic types **EMD** and **NP** and their intersection.