Journal of Biomedical Informatics 45 (2012) 61–70

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Overcoming an obstacle in expanding a UMLS semantic type extent

Yan Chen [a,*], Huanying Gu [b], Yehoshua Perl [c], James Geller [c]

[a] CIS Department, Borough of Manhattan Community College, CUNY, 199 Chamber Street, New York, NY 10007, United States
[b] Department of Computer Science, New York Institute of Technology, New York, NY 10023, United States
[c] Department of Computer Science, New Jersey Institute of Technology, Newark NJ 07102, United States

ARTICLE INFO

Article history:
Received 28 April 2011
Accepted 31 August 2011
Available online 9 September 2011

Keywords:
UMLS
Semantic type assignment
Auditing
Group auditing
Neighborhood auditing
Refined semantic type

ABSTRACT

This paper strives to overcome a major problem encountered by a previous expansion methodology for discovering concepts highly likely to be missing a specific semantic type assignment in the UMLS. This methodology is the basis for an algorithm that presents the discovered concepts to a human auditor for review and possible correction. We analyzed the problem of the previous expansion methodology and discovered that it was due to an obstacle constituted by one or more concepts assigned the UMLS Semantic Network semantic type **Classification**. A new methodology was designed that bypasses such an obstacle without a combinatorial explosion in the number of concepts presented to the human auditor for review. The new *expansion methodology with obstacle avoidance* was tested with the semantic type **Experimental Model of Disease** and found over 500 concepts missed by the previous methodology that are in need of this semantic type assignment. Furthermore, other semantic types suffering from the same major problem were discovered, indicating that the methodology is of more general applicability. The algorithmic discovery of concepts that are likely missing a semantic type assignment is possible even in the face of obstacles, without an explosion in the number of processed concepts.

1. Introduction

The Unified Medical Language System (ULMS) [3,4,15,16] is a very large and complex terminological system for biomedicine. It consists of two layers, the Metathesaurus (META) [24,25], which is a repository of concepts, and the Semantic Network (SN) [17,18], which is a compact abstraction network consisting of a small number (133) of broad categories called semantic types (STs). The connection between the layers is implemented by assigning each concept one or more semantic types.

The assignments of STs to concepts play a major role in the integration of new terminologies into the UMLS. Due to the extensive size and complexity of the UMLS, errors are inevitable. Auditing is therefore essential to ensure the quality of the UMLS. The ST assignments were proven instrumental in auditing the UMLS for various errors [7,11–14]. ST assignment errors, including incorrect and missing ST assignments, were discovered [6,8,13,14,23]. Redundancy, circularity, omissions and other problems in hierarchical relationships were located [1,2,7,20]. Classification errors were found [9,10,13]. Tools such as the Neighborhood Auditing Tool (NAT) [19] have been developed to facilitate auditing. For an extensive review of auditing of terminologies in general and the UMLS in particular, see [26].

In a study of uses of the UMLS [5], users expressed that incorrect and missing semantic type assignments are errors of greatest concern. Certain structural configurations indicate concepts with a high likelihood of incorrect ST assignments [6,13,14]. However, for the problem of exposing concepts with missing ST assignments there are no structural indicators.

The difficulty of exposing missing ST assignments was demonstrated by the findings of Chen et al. [8], where about thousand concepts of the UMLS that had been correctly assigned **Neoplastic Process**[1] (**NP**)[2] were missing the assignment of the second ST **Experimental Model of Disease** (**EMD**). Those concepts were mainly experimental cancers in mice. They were integrated into the UMLS from the National Cancer Institute thesaurus (NCIt), where they are in the Experimental Organism Diagnoses (EOD) hierarchy. The NCIt maintains its own ST assignments. According to Mougin and Bodenreider [21], these assignments differ from the UMLS ST assignments for some concepts and were proven more accurate. However, the **EMD** assignments were missing for those approximately thousand concepts in the NCIt as well.

In previous research we corrected some **EMD** assignments, but did not detect the concepts missing **EMD** [12]. Furthermore, in

[1] Semantic types are written in bold, while concept names are in italics.

[2] The following abbreviations are used in the paper: **CL** (Classification), **DS** (Disease or Syndrome), **EM/OA** (Expansion methodology with obstacle avoidance), **EMD** (Experimental Model of Disease), **EOD** (Experimental Organism Diagnoses), **HPS** (Hazardous or Poisonous Substance), **NAT** (Neighborhood Auditing Tool), **NCIt** (National Cancer Institute thesaurus), **NP** (Neoplastic Process), **OC** (Organic Chemical), **RST** (Refined Semantic Type), **SN** (Semantic Network), **ST** (Semantic Type), **SV** (Secondary enVelope), **UMLS** (Unified Medical Language System), and **XV** (auXiliary envelope).

* Corresponding author. Fax: +1 212 220 1287.
E-mail address: ychen@bmcc.cuny.edu (Y. Chen).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.08.021

2004, our team audited the Experimental Organism Diagnoses hierarchy of the NCIt for missing relationships and still did not detect the missing ST assignments. The difficulty of detecting concepts with missing ST assignments stems from the lack of a suspicious configuration which indicates their absence, in contrast to the existence of structural indicators for detecting incorrect ST assignments. Without such an indicator, an auditor receives no guidance where to search for missing ST assignments. Searching in an arbitrarily selected part of the UMLS is likely to offer a low yield for an extensive effort.

In the work of Chen et al. [8] we presented a methodology for finding concepts with missing ST assignments. It is based on the assumption that a concept that is in the neighborhood of other concepts that are already assigned a specific ST, but does not have this assignment, very likely *should* have this ST assigned. Furthermore, this process was dynamic; once a concept had been assigned the additional ST, its neighbors were also reviewed [8]. For more details, see Section 2.

In spite of our success in algorithmically discovering many concepts missing **EMD** assignments, confirmed by human auditors [8], not all concepts in the Experimental Organism Diagnoses hierarchy of NCIt missing this assignment were discovered. There are hundreds of experimental diseases (mainly cancers of different kinds) in rats which should be assigned **EMD** and are currently assigned **NP** for cancers or **Disease or Syndrome** (**DS**) for non-cancer experimental diseases. Of course, once this fact has been exposed, one could screen this hierarchy and correct the ST assignments of these concepts, but we would like a methodology for the detection of such cases. In this paper, we are presenting such a methodology for discovering missing ST assignments. When we analyzed what prevented our previous methodology from reaching the missed concepts, we found that an "obstacle" was separating the discovered concepts from those which were not discovered. In this paper, we present a methodology that bypasses such an obstacle and reaches the concepts behind it that are missing the correct ST assignments. The results of applying this methodology for **EMD** are reported. This methodology is applicable to other STs to discover more missing ST assignments.

2. Background

2.1. The Refined Semantic Network for the UMLS

In the UMLS, each concept is assigned at least one semantic type. The set of all concepts that are assigned the same ST is called its *extent*. However, the concepts in the extent of an ST are not necessarily assigned only that ST. For example, the concepts *Arthritis*, *Experimental* and *Experimental Hepatoma* are in the extent of **EMD**. However, *Experimental Hepatoma* is also assigned **Neoplastic Process**. Therefore, these two concepts do not share the same semantics (expressed by the ST assignment) even though they are both in the extent of **EMD**. Hence, the extent of **EMD** is not semantically uniform.

To achieve semantically uniform sets of concepts, each extent needs to be partitioned into subsets to reflect a refinement of this ST. We proposed the Refined Semantic Network for the UMLS, consisting of Refined Semantic Types (RSTs) for this purpose [11,12]. Each RST is either a "Pure Semantic Type" or an "Intersection Semantic Type." Each Pure Semantic Type corresponds to one ST from the SN and is assigned to concepts that were *only* assigned this one ST in the UMLS. All concepts with multiple ST assignments are removed from the extent of the Pure Semantic Type. An Intersection Semantic Type is a combination of two or more STs from the SN and its extent contains concepts assigned exactly such a combination of STs. Hence, in contrast to the extents of the original STs, the extent of each RST contains the concepts that are *only* assigned this RST and have the semantics expressed by it.

Our previous auditing methodology, reported by Chen et al., expands the extent of an ST by separately expanding each of its RSTs [8]. This expansion process identifies any neighboring concepts that have the same semantics as the concepts in the RST's extent and inserts them into the extent. The semantic uniformity of RSTs' extents makes human auditing of the concepts in those extents more effective and efficient.

2.2. Methodology for expanding the extent of a semantic type

In the work of Chen et al. [8], a two-part methodology was introduced for aiding an auditor in discovering missing ST assignments, by narrowing down the set of concepts presented to him. The auditing focused on a neighborhood surrounding the extent of an RST[3] $\mathbf{T^R}$ (E($\mathbf{T^R}$)) called an *envelope* (denoted as V($\mathbf{T^R}$)), consisting of neighbors, i.e. parents and children of the concepts in the extent, which are themselves not in the extent. All concepts in an envelope are audited by a human expert. If a concept with a missing ST assignment is identified then it is corrected and the neighbors of this concept are inserted into the next envelope.

Part 1 of the auditing methodology can be depicted as expanding outward from an extent in a series of concentric circles, as shown for $\mathbf{EMD^R}$ (Fig. 1). For example, *Arthritis*, *Animal Model* and *diencephalic hyperactivity* reside in V($\mathbf{EMD^R}$). An auditor finds that *Animal Model* is lacking the assignment of $\mathbf{EMD^R}$. Thus, its parents, *Animal Study*, *In vivo Model*, *Investigative Techniques* and *Study models* and its children, *Dorsal Skin Fold Window Chamber Model* and *Olfactory Learning*, not already in E($\mathbf{EMD^R}$) or V($\mathbf{EMD^R}$), are included in the second-level envelope $V^2(\mathbf{EMD^R})$ and await auditing after the processing of V($\mathbf{EMD^R}$) has been completed. If any concepts in $V^2(\mathbf{EMD^R})$ are later found to be missing the assignment of $\mathbf{EMD^R}$, then their parents and children not already in E($\mathbf{EMD^R}$), V($\mathbf{EMD^R}$), or $V^2(\mathbf{EMD^R})$ will be entered into the third-level envelope $V^3(\mathbf{EMD^R})$ that is processed after $V^2(\mathbf{EMD^R})$. This process continues until the next envelope remains empty. Due to the auditing process, the concepts that are in dashed-line boxes in Fig. 1 are assigned $\mathbf{EMD^R}$.

This methodology might lead to the assignment of an RST to a concept that is quite far from the concepts in the original extent of this RST. The condition for a concept *c* to be assigned $\mathbf{T^R}$ is that there exists a path of concepts connected by parent or child relationships from a concept *s*, originally assigned $\mathbf{T^R}$, all the way to the concept *c*, such that each intermediate concept on this path is also assigned $\mathbf{T^R}$. The expansion in a sequence of concentric envelopes implements the expansion process in a stepwise manner. Hence a "long distance" expansion is achieved via repeated local expansion steps.

The described process is efficient, since it does not expand in every direction for the longest possible distance. The stepwise expansion happens only for concepts where an ST assignment was made in the previous step. Hence, even if an expansion proceeds along a path of, say, ten concepts, the actual processing done is proportional only to the number of concepts, that were assigned the new RST and their parents and children, but not for all concepts within a distance of ten from the concept originally assigned the RST.

Part 2: As explained in Section 2.1, the extent of a semantic type **T** consists of disjoint subsets of concepts, such that there exists one subset for each RST generated from **T**. While reviewing the envelope of another RST of **T**, say, $\mathbf{T^{R2}}$, the auditor might realize that some of the concepts in the envelope of $\mathbf{T^{R2}}$ should be assigned

[3] $\mathbf{T^R}$ is the **R**efined (in this case "pure") semantic type of the semantic type **T**.