Journal of Biomedical Informatics 46 (2013) 97–110

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

Rule-based support system for multiple UMLS semantic type assignments

James Geller [a], Zhe He [a,*], Yehoshua Perl [a], C. Paul Morrey [b], Julia Xu [c]

[a] New Jersey Institute of Technology, Newark, NJ, United States
[b] Utah Valley University, Orem, UT, United States
[c] NIH Clinical Center, Bethesda, MD, United States

ARTICLE INFO

Article history:
Received 9 May 2012
Accepted 15 September 2012
Available online 3 October 2012

Keywords:
UMLS
Semantic Network
Metathesaurus
UMLS editing
Semantic type definitions
Concept Insertion

ABSTRACT

Background: When new concepts are inserted into the UMLS, they are assigned one or several semantic types from the UMLS Semantic Network by the UMLS editors. However, not every combination of semantic types is permissible. It was observed that many concepts with rare combinations of semantic types have erroneous semantic type assignments or prohibited combinations of semantic types. The correction of such errors is resource-intensive.
Objective: We design a computational system to inform UMLS editors as to whether a specific combination of two, three, four, or five semantic types is permissible or prohibited or questionable.
Methods: We identify a set of inclusion and exclusion instructions in the UMLS Semantic Network documentation and derive corresponding rule-categories as well as rule-categories from the UMLS concept content. We then design an algorithm *adviseEditor* based on these rule-categories. The algorithm specifies rules for an editor how to proceed when considering a tuple (pair, triple, quadruple, quintuple) of semantic types to be assigned to a concept.
Results: Eight rule-categories were identified. A Web-based system was developed to implement the *adviseEditor* algorithm, which returns for an input combination of semantic types whether it is permitted, prohibited or (in a few cases) requires more research. The numbers of semantic type pairs assigned to each rule-category are reported. Interesting examples for each rule-category are illustrated. Cases of semantic type assignments that contradict rules are listed, including recently introduced ones.
Conclusion: The *adviseEditor* system implements explicit and implicit knowledge available in the UMLS in a system that informs UMLS editors about the permissibility of a desired combination of semantic types. Using *adviseEditor* might help accelerate the work of the UMLS editors and prevent erroneous semantic type assignments.

1. Introduction

The Unified Medical Language System (UMLS) [1–4], is derived from about 160 source terminologies. Its Metathesaurus [5,6] contains over two and a half million concepts. The UMLS Semantic Network (SN) [7–10] provides a compact semantic abstraction layer, consisting of 133 high-level, broad categories, called semantic types. One or more semantic types of the Semantic Network are assigned to each Metathesaurus concept, providing it with semantics, in the sense of describing the nature of the concept by identifying its one or more broad categories.

When there are two semantic types assigned to the same concept, a number of problems may occur. In some cases, one semantic type assignment may be redundant, because the other semantic type expresses the meaning of the concept in a more specific way. In other cases, one semantic type assignment may outright contradict another one, indicating an inconsistency in the UMLS semantic type assignments. These problems notwithstanding, multiple assignments are important to express fine shades of semantics. For some cases, e.g. for chemical concepts, multiple assignments are explicitly encouraged in the documentation of the UMLS Semantic Network. There is no public repository that expresses all the different legitimate ways of interplay between the 133 semantic types. Neither is there a complete list of prohibited combinations of semantic types.

When a concept is assigned multiple semantic types, it has compound semantics [11,12], which is the combination of the semantics of the multiple semantic types. Such concepts are complex, due to their compound semantics of being simultaneously "this and that." Our experience shows [11–15] that concepts with rare combinations of semantic types, i.e. there are only a few Metathesaurus concepts assigned exactly this combination, have a high likelihood of erroneous semantic type assignments. Furthermore, some semantic type assignments stand in contradiction to the

* Corresponding author. Address: Computer Science Department, New Jersey Institute of Technology, University Heights, Newark, NJ 07102-1982, United States.
E-mail address: zh5@njit.edu (Z. He).

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.09.007

explicit documentation of the UMLS Semantic Network. This situation suggests that UMLS editors would benefit from a support system, informing them regarding the permissibility of assigning a specific combination of semantic types to a concept.

The objective of this research is to develop a system *adviseEditor* that will inform an editor as to whether a specific tuple (pair, triple, quadruple, quintuple) of semantic types is permitted or prohibited. There is a need for such a system, because UMLS editors have introduced prohibited combinations of semantic types and even reintroduced them after the UMLS was corrected by eliminating those prohibited combinations. (Examples of such reintroduced combinations appear in Section 4.7.) To achieve this objective, we first need to define categories of rules that govern the possible interactions of pairs of semantic types. We will point out examples where concepts in the Metathesaurus violate the identified rules. If the *adviseEditor* system would have been in place when those concepts were originally introduced into the UMLS and assigned semantic types, these errors could have been prevented. We will also provide counts of semantic type pairs belonging to different rule-categories, as determined by the *adviseEditor* system.

2. Background

The Metathesaurus of the UMLS is the result of integrating about 160 source terminologies into one knowledge source. An important conceptual tool for this integration is the UMLS Semantic Network. Every concept in the Metathesaurus is assigned one or more semantic types of the Semantic Network at the time of integration [16,17]. These assignments were performed by many UMLS editors at the National Library of Medicine over a long period of time, and thus are not necessarily done in a consistent manner.

The UMLS Semantic Network is structured as two separate trees, rooted in the semantic types **Entity** and **Event**, respectively. The 133 semantic types of the Semantic Network constitute its nodes and are connected by IS-A links. They are furthermore connected by 53 lateral relationship kinds. Inheritance of lateral relationships along IS-A links is by default a defined operation, except for a few cases where it is explicitly blocked.

When working with semantic types we make use of the following definition.

Definition. The set of all concepts assigned a specific semantic type **T** is called the *extent of* **T**, abbreviated as E(**T**).

Whenever a concept is assigned two semantic types, then it is contained in the extents of both semantic types at the same time. Mathematically this means that the concept is in the set intersection of the two extents. The mathematical symbol ∩, expressing intersection, will occasionally be used when describing sets of concepts that are assigned two semantic types.

In [11,12,16] auditing of the UMLS for inconsistencies was carried out, based on intersections of extents of semantic types. We hypothesized [12] that concepts in small intersections have a high likelihood of wrong semantic type assignments. In a sample of 100 intersections, each containing only a single concept, analyzed by Cimino [12], only 11 concepts were found to have correct semantic type assignments.

Gu et al. showed [17] that concepts assigned pairs of semantic types, such that the intersections of their extents are small, were more likely to have erroneous semantic type assignments than other concepts. In this paper, we make use of this observation for developing an algorithm for classifying pairs of semantic types according to rule-categories.

This research also builds on an algorithm [18] for identifying all redundant semantic type assignments, namely assignments in which a concept is assigned the semantic types **X** and **Y** such that **X** is a child or descendant of **Y**. Such redundant assignments are prohibited by the rules of the Semantic Network [19], and only **X** should be assigned. Assigning the respective pairs of semantic types is not legal, and they should never be assigned to the same concept. However, in the 1998 release we found 8622 concepts with redundant semantic type assignments in 77 prohibited intersections [12].

To help both editors and users of the UMLS, the National Library of Medicine provides a definition for each semantic type in the Semantic Network source data. Usage notes (UNs) are provided for some, but by far not all, semantic types. Note that in the balance of this paper, when we refer to a semantic type definition, *we mean to include any usage notes* attached to this definition. Some usage notes include instructions concerning the combination of two semantic types. These instructions describe situations in which a concept assigned one semantic type may not, may, or should be assigned a specific second semantic type.

3. Methods

3.1. Text-based instructions

Studying the documentation of the Semantic Network, one can distinguish between two kinds of instructions, *inclusion instructions* and *exclusion instructions*. An inclusion instruction expresses the fact that two semantic types *may* be used for the same concept or even *should* be used for the same concept. An exclusion instruction expresses the fact that two semantic types *may not* be used for the same concept.

We will use the semantic type **Anatomical Abnormality** to describe the following possible parts of a usage note: (1) specification, (2) inclusion instruction, and (3) exclusion instruction. Below is the UN provided in the UMLS about this semantic type.

UN: Use this type if the abnormality in question can be either an acquired or congenital abnormality. Neoplasms are not included here. These are given the type '**Neoplastic Process**'. If an anatomical abnormality has a pathologic manifestation, then it will additionally be given the type '**Disease or Syndrome**', e.g., "Diabetic Cataract" will be double-typed for this reason.

3.1.1. Specification

A specification may contain an additional explanation of what a certain semantic type stands for, or a set of requirements to be satisfied by a concept to be assigned this semantic type, or a clarification to distinguish between two semantic types.

In the above usage note of **Anatomical Abnormality** the following part corresponds to a specification. "Use this type if the abnormality in question can be either an acquired or congenital abnormality."

In this case, one needs to realize that, as shown in Fig. 1, **Acquired Abnormality** and **Congenital Abnormality** are the two children of **Anatomical Abnormality** in the Semantic Network.

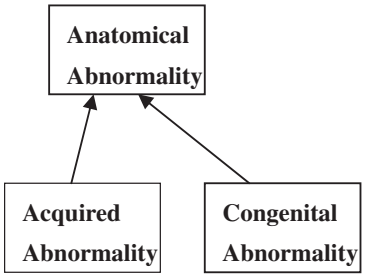

Fig. 1. Anatomical Abnormality subhierarchy of SN.