Journal of Biomedical Informatics 81 (2018) 1–15

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Mismatches between major subhierarchies and semantic tags in SNOMED CT

Jonathan P. Bona[a,*], Werner Ceusters[b]

[a] Department of Biomedical Informatics, College of Medicine, University of Arkansas for Medical Sciences, 4301 W. Markham St., #782, Little Rock, AR 72205-7199, USA

[b] Department of Biomedical Informatics, Jacobs School of Biomedical and Medical Sciences, University at Buffalo, 77 Goodell street, Buffalo, NY 14203, USA

ARTICLE INFO

Keywords:
SNOMED CT
Semantic tags
Quality assurance

ABSTRACT

The *fully specified name* of a concept in SNOMED CT is formed by a *term* to which in the typical case is added a *semantic tag*. The latter is meant to disambiguate homonymous terms and to indicate in which major sub-hierarchy of SNOMED CT that concept fits. We have developed a method to determine whether a concept's tag correctly identifies its place in the hierarchy, and applied this method to an analysis of all active concepts in every SNOMED CT release from January 2003 to January 2017. Our results show (1) that there are concepts in almost every release whose semantic tag does not match their placement in the hierarchy, (2) that it is primarily disorder concepts that are involved, and (3) that the number of such mismatches increase since the July 2012 version. Our analysis determined that it is primarily the absence of a mechanism in the SNOMED CT authoring environment to suggest stated relationships for very similar concepts that is responsible for the mismatches. We argue that the SNOMED CT authoring environment should treat the semantic tags as part of the formal structure so that methods can be implemented to keep the sub-hierarchies in sync with the semantic tags.

1. Introduction

SNOMED CT is a large reference terminology for the clinical domain in which what are called 'concepts', claimed to be representations of 'clinical meanings'[1], are linked to 'descriptions' which contain terms indicating various ways of how these clinical meanings are expressed in natural language.

The January 2017 version of SNOMED CT consists of 326,734 active and 214,969 inactive concepts. Concepts are linked to other concepts by means of relationships some of which are grouped so as to form machine-readable logical definitions that can be used for logical inference [2, p757]. SNOMED CT concepts are organized into a hierarchy of 'Is-a' relationships. The top concept, [138875005 | SNOMED CT Concept (SNOMED RT + CTV3)] directly subsumes 19 high level concepts. Most of these concepts are first-order concepts such as [404684003 | Clinical finding (finding)] and [123037004 | Body structure (body structure)] which serve as the root of sub-hierarchies of concepts about entities directly relevant to and within the domain of healthcare. Some of these concepts are second-order concepts that describe the structure of SNOMED CT rather than the structure of what the first-order concepts are about. SNOMED CT comes with a history mechanism that allows for a detailed analysis of how the system has changed over time [3]. The complete SNOMED CT hierarchy for each release is generated by a description logic classifier applied to "stated" definitions and relationships that are created and edited by human authors or editors of the ontology [2, p757].

Every SNOMED CT concept comes with *descriptions* one of which is selected as the *Fully Specified Name* (FSN). For example, the FSN of the concept with unique identifier '35566002' is 'Hematoma (morphologic abnormality)'. This FSN informs us that 'hematoma' – i.e. the part of the FSN that precedes the part written in parentheses – is an acceptable term by means of which concept 35566002 may be expressed in clinical language. An FSN typically ends with a short text surrounded by parentheses that is called the '*semantic tag*'. One function of this tag is to disambiguate the FSN of this concept from the FSNs of other concepts that may be expressed by the same term [2, p41]. It is thus the semantic tag 'morphologic abnormality' which disambiguates the display name of the concept [35566002 | Hematoma (morphologic abnormality)] from the concept [385494008 | Hematoma (disorder)]. This is useful when the user interface of, for example, an electronic healthcare record system returns in response to a search for 'hematoma' all the FSNs of all concepts in which this term appears in at least one of their descriptions without, however, showing the entire hierarchy: without the semantic tag, it would not be possible to determine what the difference in

* Corresponding author.
E-mail addresses: jpbona@uams.edu (J.P. Bona), ceusters@buffalo.edu (W. Ceusters).

https://doi.org/10.1016/j.jbi.2018.02.009
Received 1 October 2017; Received in revised form 14 February 2018; Accepted 16 February 2018
Available online 17 February 2018
1532-0464/

meaning would be between what would be displayed, for example, as [35566002 | Hematoma] and [385494008 | Hematoma].

The semantic tag (now also called the '*hierarchy tag*') is said to '*identify the hierarchy into which the concept is placed via its Relationships*' [4, p237]. Although the SNOMED CT documentation does not provide more detail on what this exactly means, our understanding of this is that the directed acyclic graph (DAG) formed by SNOMED CT's complete Is-a hierarchy is intended to be composed out of smaller DAGs, one for each semantic tag. Each one of these smaller DAGs, so we assume, is intended to satisfy the following criteria: (1) it is populated by all concepts whose FSNs contain the same semantic tag, and (2) there is only one concept at the root of this DAG: the 'corresponding concept'. Further, these smaller DAGs may be nested so that, for example, the DAG formed by the concepts with the semantic tag 'finding' includes the DAG formed by the concepts with the semantic tag 'disorder'.

Because semantic tags are substrings added to *names* inside FSNs and are not represented separately as part of SNOMED CT's formal model, it is not easy to determine whether there is for each semantic tag indeed a DAG that satisfies the above mentioned criteria. Moreover, there does not appear to be an official published mapping that lists the semantic tag/concept correspondences for SNOMED CT. In many cases this correspondence may seem obvious to a human observer. For many tags there is indeed a single high-level concept whose semantic tag matches exactly the part of the FSN that precedes the tag. For example, one direct sub-concept of the top SNOMED CT Concept is [71388002 | Procedure (procedure)]. This concept has the semantic tag 'procedure' and its name in the FSN is the word 'Procedure'. In other cases, the correspondence is less obvious. For instance, no direct sub-concept of SNOMED CT's top concept is tagged 'morphologic abnormality', nor is there any concept whose name is exactly 'Morphologic abnormality'. The same holds for the semantic tag 'disorder'. The concept [118956008 | Body structure, altered from its original anatomical structure (morphologic abnormality)] is a child of [123037004 | Body structure (body structure)] and appears to be the highest concept (i.e. closest to the top) tagged with 'morphologic abnormality'. If we are correct in our interpretation, then the concept [35566002 | Hematoma (morphologic abnormality)] should be classified in the sub-hierarchy of morphologic abnormalities and be subsumed by [118956008 | Body structure, altered from its original anatomical structure (morphologic abnormality)] while [385494008 | Hematoma (disorder)] should be classified in the sub-hierarchy of diseases, the highest level concept of this sub-hierarchy being [64572001 | Disease (disorder)].

The exact relationship between SNOMED CT's semantic tags and concepts has thus far not been widely researched. In [3] we explored how the semantic tags of concepts changed over time. We found in total 285 patterns according to which SNOMED CT concepts underwent changes in the semantic tags assigned to them in the collection of SNOMED CT versions studied. This included 43 patterns according to which an FSN *without* a semantic tag was changed into one *with* a semantic tag. There were no patterns with more than 3 changes over time. Changes in semantic tags were found to happen for a variety of reasons. One is a change in SNOMED CT's concept model, for example when in the newer version distinctions were made that did not exist in earlier versions, or when different interpretations were introduced (e.g. the product / substance distinction). Such changes have a global impact on large parts of the ontology. Another reason is that concepts were in one or other way erroneous and had to be corrected. While doing these analyses, we were nevertheless hampered by the fact that the SNOMED CT documentation available from the IHTSDO provides insufficient information on what the precise set of semantic tags the SNOMED CT editors are working with might be. The information that a semantic tag is that what appears at the end of an FSN between brackets [2, p41] turned out not to be reliable. Historically, FSNs didn't have a semantic tag at all as this was apparently introduced later as witnessed by the many changes in descriptions to that end. It was also found that parsing anything that terminates an FSN between brackets leads to many false positives in older concepts, thus requiring manual inspection for disambiguation.

Furthermore, some FSNs end with more than one parenthesized substring, which makes it look at first glance as if the concepts with such FSNs might have multiple semantic tags. This in turn further confuses the question of what, exactly, counts as a semantic tag. For example, the string "contextual qualifier" appears surrounded in parentheses in 103 FSNs immediately preceding the official semantic tag "qualifier value", as in the concept: [30207005 | Risk of (contextual qualifier) (qualifier value)] and its children. A similar pattern occurs with the quasi-tag "property", as seen in [118597006 | Quantity rate (property) (qualifier value)] and 92 others. This phenomenon is not limited to qualifier values: [110818007 | Bile duct and stomach (combined site) (body structure)] is one of 296 concepts whose FSN ends with '(combined site) (body structure)'. Other examples include terms that appear to be more parenthetical clarifications rather than indicative of an implicit sub-hierarchy among tags: 'less than 2 years' in [4359001 | Early congenital syphilis (less than 2 years) (disorder)] and 'chemical processes, except Petroleum' in [9101001 | Reactor-converter operator (chemical processes, except Petroleum) (occupation)] are two examples.

Throughout this analysis we treat as semantic tags only those parenthesized substrings that occur last in an FSN. The SNOMED CT Editorial Guide supports this interpretation: '*Each FSN term **ends*** [bold emphasis added] *with a 'semantic tag' in parentheses*'[4, p208].

The work presented here assesses the January 31, 2017 International Release of SNOMED CT, including the history information that it contains starting with the January version of 2003, to determine the extent to which SNOMED CT's use of semantic tags is systematic and consistent with its placement of concepts that use those semantic tags within the concept hierarchy.

2. Material and methods

The research hypotheses driving this work are:

(1) Within a specific release of SNOMED CT, all semantic tags are intended to be related to the concept system through a one-to-one correspondence between the semantic tag and some unique high-level concept which we call the 'corresponding concept' for that tag.
(2) Every concept that uses a particular semantic tag *t* within a specific SNOMED CT version should be subsumed by that semantic tag's corresponding concept *Ct*, where *Ct* is the highest level concept that uses *t*, within that version. This hypothesis is motivated by the apparent change in terminology from '*semantic tag*' in [2] to '*hierarchy tag*' in [4, p227].
(3) The fact that semantic tags, so we assume, are not part of SNOMED CT's formal model may lead to mismatches: we consider a concept to be 'mismatched' in a specific SNOMED CT version if it has the semantic tag *t* but is not subsumed by that tag's corresponding concept *Ct*.
(4) Where such mismatches exist, they are due to errors in either the concept's placement in the SNOMED CT hierarchy or in its semantic tag. Such errors, when discovered by the SNOMED editors, are corrected in later releases.

To test these hypotheses, we implemented automated procedures (1) to find for each semantic tag its corresponding concept in each release, (2) to identify mismatched concepts, and (3) to group these mismatches in categories based on how mismatched concepts relate to other mismatched concepts.

Because the semantic tag 'disorder' contains the most mismatches in the latest release investigated, semi-automated and manual methods were used to identify possible causes. To that end, we retrieved and analyzed the subsumption hierarchy of all mismatched concepts for the semantic tag 'disorder' as well as their other relationships and we