Detecting Underspecification in SNOMED CT Concept Definitions Through Natural Language Processing

Edson Pacheco[1,2], Holger Stenzhorn[3], Percy Nohama[1], Jan Paetzold[3,4], Stefan Schulz[2,3,4]
[1]Federal Technical University of Paraná (CPGEI/UTFPR), Curitiba, Brazil;
[2]Pontifical Catholic University of Paraná (PUCPR), Curitiba, Brazil;
[3]University Medical Center Freiburg, Germany; [4]AVERBIS GmbH, Freiburg, Germany

Abstract

Quality assurance and audit issues play a major role in maintening large biomedical terminology, such as SNOMED CT. Several automatized techniques have been proposed to facilitate the identification of weak spots and suggest adequate improvements.

In this study, we address a well-known issue within SNOMED CT: Albeit the wording of many free-text concept descriptions suggests a connection to other concepts, they are often not referred to in the logical concept definition.

To detect such inconsistencies, we use a semantic indexing approach which maps free text onto a sequence of semantic identifiers. Applied to SNOMED CT concepts without attributes, our technique spots refinable concepts and suggests appropriate attributes, i.e., connections to other concepts. Based on a manual analysis of random samples, we estimate that approximately 18,000 refinable concepts can be found.

INTRODUCTION

SNOMED CT[1] is a large and heterogeneous clinical terminology. This is due to several factors:

- It grew out as a meger of two legacy systems (SNOMED RT and NHS Clinical Terms Version 3) with different, partly contradicting design principles[2];
- It faces constant requests for content inclusion and in the past, this used to be handled quite generously;
- The content maintenance and auditing process seriously lags behind the needs.

Due to SNOMED CT's sheer size, it is impossible to maintain, audit and assure the quality in a completely manual way. Several semi-automated methods have been proposed for detecting defects in terms of content and architecture [8,9,10,11].

A well-known quality problem in SNOMED CT is underspecification. In contrast to "real" errors, content is not false but missing. The advantage is that such problems can be remedied in a monotonous way, i.e., without removing content. Underspecification can be found with numerous SNOMED CT concepts, which – although their textual descriptions exhibit composed meanings – they are not logically related to any other SNOMED CT concept besides their taxonomic parent(s).

For instance, the concept *Cerebral function* is only related to its parent *Nervous system function*, yet the expected relation with the concept *Brain structure* is missing. But as "cerebral" is derived from "cerebrum" (as a synonym of "brain"), a lexicon-based method could infer the missing logical attribution. The inclusion of such an approach in the process of terminology maintenance would help to fill in definitional gaps, thus increasing SNOMED CT's power of providing semantic interoperability.

This study describes our mechanism to detect underspecified SNOMED concepts and to propose possible refinement attributes by natural language processing methods.

MATERIAL AND METHODS

SNOMED CT Sources

We use the descriptions, concepts, and relationships tables from the English 01/2009 release of SNOMED CT[1]. The descriptions table provides several synonymous terms for each concept (named SNOMED CT "descriptions"), among them exactly one, unique FSN (fully specified name), exactly one PT (preferred term) and zero to many synonyms.

For our purpose, only the concept status field from the concept table is relevant, as it allows the distinction between active and inactive concepts.

Finally, the relationships table holds the associations between concepts. For our purpose, only the distinction between defining and qualifying relationships is relevant.

Semantic Indexing

Due to the high diversity of natural language expressions in terms of inflection, derivation, and synonymy we perform a conceptual abstraction of the meaning of each description. More exactly, we map a sequence of text tokens (t_1, t_2, t_3,…,t_m) to a sequence of morphosemantic identifiers (m_1, m_2, m_3,…,m_n), using the MorphoSaurus system[5]. This system uses so-called subwords[5] as lexical units which are defined as the minimum lexical units of meaning-bearing terms in a given domain. Subwords often correspond to word fragments. For example, “hepatitis” is split into the subwords “hepat” and “itis”.

The semantic layer of MorphoSaurus is represented by subword equivalence classes, identified by so-called MIDs (MorphoSaurus identifiers). Each lexical entry is associated with exactly one equivalence class. Equivalence classes group lexical variants, synonyms, and translations. For instance the subwords “hepat” and “liver” are in the same equivalence class, just as “itis” and “inflamm”.

Currently, over 100,000 lexical entries exist in the MorphoSaurus lexicon. This assures a high performance extraction of subwords and their mapping by using finite-state techniques for lexicon-based decomposition, derivation and deflection[5].

Morphosemantic indexing was performed for each of the 837,105 active SNOMED CT descriptions yielding and average 4.95 MIDs per description.

Selection of Underspecified Concept Candidates

We selected the candidates which are possibly underspecified concepts according to the following criteria: Firstly, we used active concepts only. Secondly, we excluded all concepts that had defining relationships other than *is-a* (taxonomic subsumption relationship).

Attribute Harvesting

The attributes of some SNOMED CT concept are all (non-*is-a*) relation – concept pairs that are assigned to this concept in the relationships table. For instance, the concept *Inflammatory disease of liver* has the attribute *Finding site*: *Liver structure*.

In contrast, the concept *Hepatitis notification* has no attribute at all, although one would expect a link to the concept *Inflammatory disease of liver*.

In the latter case we want our system to propose suited attributes. However, we ignore the nature of the relationship and focus on the target concept only. The reason for this decision is that it is often unclear which existing relationship should be used or whether a new one should be introduced into SNOMED CT.

We developed the following approach:

Let C be a non-attributed concept and FSN $_C$ its fully specified name and P_C = {P_1(C), P_2(C), …, P_k(C)} the set of the concept’s direct parents. For any parent P_i(C), again, the FSN is used: $FSN_{Pi(C)}$ = FSN(P_i(C)).

So we compare the MID sequences of each element in FSN_C with the MID sequences of each element in $FSN_{Pi(C)}$ as follows: each MID occurring in both sequences is eliminated from the sequence of the former. For the remaining MID sequence it is checked whether it exactly matches the MID sequence of any other description across the whole set of SNOMED CT descriptions (here not only FSNs). In this case, the concept belonging to that description is suggested as a candidate for refining the original concept.

Evaluation Methodology

For the evaluation of each semantic type (as given by the bracketed expression in the FSN, e.g. *Organism, Substance, Body Structure*) a random sample of twenty underspecified concepts is extracted and listed together with all the attribute refinement candidate the system proposed. For each of the sample concept a domain expert verified (i) whether this concept should be refined, and (ii) whether one of the suggested refinement candidates can be plausibly used for refinement.

In order to measure the inter-rater agreement, a second domain expert performed the same verifications for half the sample (ten concepts in each hierarchy).

RESULTS

Nearly half (45.2%) of the SNOMED CT concepts (132,125) have no attributes. Our system identified 48,552 (16.6%) as refinable, i.e. suggested on average 2.8 potential target concepts (which, together with a suitable relation, would refine the logical description of the concept under scrutiny).

Table 1 provides the exact figures classified by the main SNOMED CT hierarchies. According to the estimations based on the sample analysis, approx. 18,500 concepts are refinable and for over 12,000 the system suggests the right target concept.

		Underspecified Concepts		Refinement candidates		Analysis of samples(n=20)		Sample based estimation	
SNOMED hierarchies	Active Concepts	n	%	n	%	justified refinement	correct suggestion	refinable concepts	with correct suggestions
Organism	31840	31840	100.0	4973	15.6	0%	0%	0	0
Substance	23554	23554	100.0	8627	36.6	55%	35%	4700	3000
body structure	25637	22386	87.3	15076	58.8	5%	0%	800	0
qualifier value	8823	8823	100.0	3533	40.0	0%	0%	0	0
observable entity	7885	7885	100.0	3647	46.3	70%	50%	2600	1800
Finding	32780	5356	16.3	2253	6.9	90%	75%	2000	1700
physical object	4408	4408	100.0	1339	30.4	85%	80%	1100	1100
morphologic abnormality	4297	4289	99.8	2164	50.4	80%	60%	1700	1300
Occupation	3843	3843	100.0	1330	34.6	75%	10%	1000	100
Product	19310	3541	18.3	686	3.6	100%	60%	700	400
Event	3578	3529	98.6	447	12.5	85%	45%	400	200
Disorder	63874	2812	4.4	1080	1.7	90%	60%	1000	600
Procedure	47764	2256	4.7	1001	2.1	85%	65%	900	700
Others	14511	7603	52.4	2396	16.5	75%	60%	1800	1400
TOTAL	292104	132125	45.2	48552	16.6			18700	12300

Table 1. Analysis of underspecified SNOMED CT concepts by subhierarchies. Underspecified concepts: concepts that have no attributes. Refinement candidates: concepts for which missing attributes were suggested by the system. Justified refinement: the concept under scrutiny is underspecified and its formal definition should be refined. Correct suggestion: For the concept to be refined one of the suggested attributes is correct.

A closer look on the distribution reveals that in some of the hierarchies not a single concept is provided with any attribute. This is the case with *Organism*, *Substance*, *Qualifier Value*, *Observable Entity*, *Physical Object*, and *Occupation*. This is consistent with the SNOMED CT editing guidelines as applied so far. However, our system also suggested refinement concepts for these hierarchies, e.g. *Macaroni* for *Macaroni maker* (occupation), *Canada* for *Salmonella canada* (organism), *Metal* for *Metal device*, or *Acyl carnitine* for *Acylcarnitine hydrolase (substance)*. Whereas we rejected all suggestions in the Organism and the Qualifier value branch, we accepted some in the others, as they seemed plausible. However, it must be discussed under a clinical point of view, whether the material a profession uses or the substrate of an enzyme should be specified by SNOMED CT.

Body Structure is another interesting case, as we rejected all offered target concept suggestions and only accepted one of twenty refinability judgment. The reason is SNOMED CT's idiosyncratic way to emulate part-of hierarchies by taxonomies of so-called "structure" or "part" concepts according to the SEP triplet model[6]. So the part-of relations were already there (albeit masked by the SEP constructs): *Cardiac wall structure* isa *Heart Part*. The proposed target concepts proved useless. We also rejected the suggested refinement of certain body parts by ordinal numbers, such as *Fifth metatarsal structure* by *Five*.

A quite common reason to reject the system's classification of a concept as refinable is that is already sufficiently defined by its parents, such as *Female first cousin* by the intersection of *First cousin* and *Female cousin*. A final analysis tackles the semantic types of the concepts found. *Qualifier value* accounted for one third, followed by *Substance*, *Body structure*, *Observable entity*, *Physical object*, *Finding*, and *Person*.

Kappa provides a measure of the degree to which two judges, A and B. A 'judge' in this context is a domain expert. The interrater agreement analysis yielded only a fair agreement on which sample concepts should be refined (Kohen's Kappa 0.55). The agreement on whether the correct target concept was proposed was better, with Kohen's Kappa equalling 0.74.

DISCUSSION

Several authors addressed error detection in SNOMED CT: Wang *et al.*[11] performed a structural analysis and split SNOMED CT into partitions that contain structurally and semantically related concepts. Two different taxonomies were extracted from SNOMED CT based on the stated relationships between concepts thus allowing the concept hierarchy to be viewed at different levels of granularity.

Whereas the "area taxonomy" (an area contains all concepts with the exact same structure of relationships) highlights structural irregularities, the "p-area taxonomy" presents a finer structure as well as semantic information. Based on those taxonomies, one audit methodology shows errors which appear as irregularities at the structural level in the first taxonomy and highlights structural irregularities found in the second taxonomy. Finally, the p-area taxonomy is reviewed for sets of related concepts based on structural similarity. The main goal of this approach to present high-level (better apprehendable) views of the terminology allowing better navigation and orientation into the content and structure of a terminology together with direct display of structural issues.

Wei *et al.*[12] hypothesized that such errors contribute to the structural disorder and therefore investigated if their correction simplifies the hierarchical structure. The complexity assessment was carried out by using the area and p-area taxonomies. It was then asserted that concepts with one relationship are simpler than ones with more relationships. Also, since p-areas are seen to represent sets of semantically-related concepts, an area with fewer p-areas for the same concept number is considered to have fewer different meanings. Experiments showed that indeed the complexity the more errors are fixed: The number of partial areas became much less when errors were fixed and when erroneous relationships were deleted, the mean number of relationships per class decreased as well.

Campbell *et al.*[13] introduce the "lexically-suggested logical closure" for evaluating the maturity and quality of terminologies and apply this metric to SNOMED-RT's development progress. They correlate within the terminology the number of omission errors that can be algorithmically detected though analyzing the language structure among the terms. For example, if important relationships are omitted this can lead to incomplete class retrieval, such as in the case of "retinal vasulitis" which was defined as "eye disease" but had no relation to "vasculitis". A longest common substring algorithm or similarity scoring approach can identify and suggest the latter class as a superclass of the former. The proposed metric is the coefficient of proposed relationships accepted vs. rejected by experts and thus shows the quality of the proposals.

Cornet and Abu-Hanna[14] introduced a method for auditing medical terminologies based on detecting (non-primitive) concepts with equivalent logical definitions for higlighting cases where concepts 1) are redundantly defined more than once (but by different terms) or 2) have the same definition but are supposed to be different (i.e., they are underspecified and lack additional information). A description logic reasoner is used first to retrieve the sets of logically equivalent concepts and then those sets are analyzed manually towards detecting the two scenarios. This evaluation method has been applied to the DICE terminology by the authors where four double-defined and 300 underspecified concepts were found. Since SNOMED-CT is based on DL too, this methodology could be directly applied to this terminology as well.

The approach described by Jiang & Chute[10] uses Formal Concept Analysis (FCA). Here, SNOMED CT's normal forms, are reformulated in form of lattice theory. This permits to visualize partial or incomplete orders, such as the SNOMED CT structure, in an information lattice and its consequences and can thus represent the complete (decomposed) semantics underlying concept definitions. Thus anonymous (non-labeled) concepts that appear in several concept definitions are detected and are propose as new (labeled) concepts for inclusion into SNOMED CT. Experiments showed that the more anonymous nodes existed, the smaller was the number of fully defined concepts, which might indicate that SNOMED CT contents are quantifiably semantically incomplete.

Bodenreider *et al.* [8] proposed proper ontology design principles for SNOMED CT auditing: So should each class have at least one parent, non-leaf classes must have minimum two children and class must be different from any other class in its definition. It was shown that almost a third of all classes with children broke the second rule. On the other hand there exist also classes which have hundreds or even thousands of direct children, hinting that some intermediate classification level(s) is/are missing. Another finding was that the last rule was broken often as well, namely that more than half of all parent/child relations have no differentiae between the parent description and their own.

These approaches seem highly valuable for improving the quality of SNOMED CT, as each of them pinpoints addressed different classes of defects. But in contrast to the methodology we propose, none of them includes any analysis based on the natural language descriptions. From our results above, we conclude that our method is supposed to detect gaps the other presented approaches are unable to identify.

However, the reported methodology still has several drawbacks: The MorphoSaurus indexer occasionally creates artifacts due to lexical underspecification. For *Struck by falling lumber (event)* our system suggests the missing concept *lumbar*. Furthermore, the subtraction criterion (which compares the MID(s) between the child and the parent concept) sometimes seems too strong: So is *Vitamin A overdose* a child of *Vitamin overdose*, but the remainder (*A*) gives to hint to the associated concept. Too strong may also be the assumption that only those concepts are underspecified that have no attributes at all.

CONCLUSION

We have proposed a method that supports the audit of SNOMED CT by pinpointing specification gaps in logical concept definitions though exploring free-text descriptions. It targets, first of all, concepts that have no attributes, which currently constitute nearly half of all concepts. By comparing a simplified semantic representation of the meaning of the concepts' fully specified names with those of their parents, our system generates hypotheses regarding possible attribute candidates. Based on a manual analysis of random samples we estimate that approximately 18,000 SNOMED CT concepts can be refined. A literature survey suggests that the presented approach highlights issues which cannot be found by other existing approaches and thus effectively complements them.

Acknowledgments

This work was supported by CNPq, National Council of Scientific and Technological Development (Brazil, 550830/2005-7), and the International Bureau of the Federal Ministry of Education and Research (Germany, BRA05/022).

REFERENCES

1. International Health Terminology Standards Development Organisation (IHTSDO). Systematised Nomenclature of Medicine – Clinical Terms (SNOMED CT). http://www.ihtsdo.org
2. Cornet R, de Keizer N. Forty years of SNOMED: a literature review. BMC Med Inform Decis Mak. 2008 Oct 27;8 Suppl 1:S2.
3. The Unified Medical Language System Knowledge Source Server. http://umlsks.nlm.nih.gov
4. Honeck M, Hahn U, Klar R, Schulz S. Text Retrieval Based on Medical Subwords. Stud Health Technol Inform. 2002;90:241-245.
5. Markó K, Schulz S, Hahn U: MorphoSaurus - Design and Evaluation of an Interlingua-based, Cross-language Document Retrieval Engine for the Medical Domain. Meth Inf Med 4/2005(44): 537-545.
6. Schulz S, Romacker M, Hahn U. Part-whole reasoning in medical ontologies revisited – introducing SEP triplets into classification-based description logics. Proc AMIA Symp. 1998:830-834.
7. Markó K, Daumke P, Schulz S, Klar R, Hahn U: Large-Scale Evaluation of a Medical Cross-Language Information Retrieval System. MEDINFO. 2007: 392-396.
8. Bodenreider O, Smith B, Kumar A, Burgun A. Investigating Subsumption in SNOMED CT: An Exploration into Large Description Logic-Based Biomedical Terminologies, Artif Intell Med. 2007 Mar;39(3):183-195.
9. Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT. Stud Health Technol Inform. 2004;107 (Pt1):482-486.
10. Jiang G, Chute CG. Auditing the semantic completeness of SNOMED CT using formal concept analysis. J Am Med Inform Assoc. 2009 Jan-Feb;16(1):89-102.
11. Wang Y, Halper M, Min H, Perl Y, Chen Y, Spackman KA. Structural methodologies for auditing SNOMED. J Biomed Inform. 2007 Oct;40(5):561-581.
12. Wei D, Wang Y, Perl Y, Xu J, Halper M, Spackman K. Complexity Measures to Track the Evolution of a SNOMED Hierarchy. AMIA Annu Symp Proc. 2008 Nov 6:778-782
13. Campbell, K.E., M.S. Tuttle, and K.A. Spackman. A "lexically-suggested logical losure" metric for medical terminology maturity in Proceedings of the 1998 AMIA Annual Fall Symposium. 1998. Orlando, FL, USA
14. Cornet R, Abu-Hanna A. Auditing Description-Logic-based Medical Terminological Systems by Detecting Equivalent Concept Definitions. International Journal of Medical Informatics, 2008; 77(5): 336-345