Lexical ambiguity in SNOMED CT

Stefan Schulz[1], Catalina Martínez-Costa, Jose Antonio Miñarro-Giménez
Institute of Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria

Abstract. Terminology systems that represent the language as used in human communication have to deal with the problem of lexical ambiguity; i.e. the same natural language term is assigned to two or more codes. A scrutiny of the large international terminology standard SNOMED CT focused on concepts that are linked by the same term and exhibit problems especially when using the terminology in a Natural Language Processing context. We found 8,338 ambiguous terms from about 700k terms in SNOMED CT and provide recommendations in order to improve its quality by curators.

Keywords. SNOMED CT, Biomedical terminologies, Biomedical ontologies, Lexical ambiguity

1. Introduction

Lexical ambiguity is the capacity of a term to have multiple meanings. Humans constantly produce ambiguous utterances, due to our capacity of intuitively inferring the right sense guided by linguistic context and domain knowledge. For machine processing, the processing of ambiguities is a known problem, and WSD (word sense disambiguation) is a classical Natural Language Processing (NLP) task.

Reference ontologies and terminologies tend to avoid ambiguities in their choice of preferred terms or labels. They are ideally self-explaining (e.g. "Transplantation of liver (procedure)"), but often far away from clinicians' jargon ("Liver transplant" or "LT"). However, as soon as the terminology is enriched by (quasi)synonyms or close-to-user entry terms, the ambiguity problem arises (e.g. "LT" for "Leishmania tropica", "Low testosterone" and others).

Besides classical cases of lexical ambiguity (e.g. "bank" for riverside vs. financial institution), a notorious source of lexical ambiguity in medical language is the overuse of short forms [1]. Among several types of word shortenings, acronyms are the most common ones like CT, ECG, CA and LT. They are single tokens derived from the initial (mostly capitalised) components of words in a phrase and/or syllables in a word. In particular, short acronyms are known to have dozens of expansions. AcronymFinder [2], provides 61 expansions of the acronym "CT" ("Clinical Trial", "Cognitive Therapy,

[1] Corresponding author: Stefan Schulz, Medical University of Graz, Auenbruggerplatz 2/V, 8036 Graz (Austria), E-mail: stefan.schulz@medunigraz.at.

“Computed Tomography”, “Connective Tissue”,…) in the context of science and medicine.

In the following we will scrutinise the problem of lexical ambiguity in the large ontology-based clinical terminology standard SNOMED CT [3] and will provide naming recommendations for SNOMED CT curators in a similar way as existing naming conventions [4] for ontologies in OBO Foundry [5]. SNOMED CT’s January 2017 international release counts about 300k concepts and 700k terms. The main purpose of this – still preliminary – work is to identify and to discuss the relevance of lexical ambiguity in SNOMED CT.

Ambiguous terms in a terminology are those that map to more than one code. Lexical ambiguity matters in terminologies particularly when they are used as dictionaries in natural language processing systems, because the choice of the right code becomes a matter of chance, unless contextual information is used. SNOMED CT distinguishes between fully specified names (FSNs) and other terms, so-called synonyms. All FSNs end with a hierarchy tag (i.e. semantic type) in parentheses identifying the hierarchy into which the concept is placed, e.g. “B-cell lymphoma (disorder)”. This guarantees a bijective function between the set of SNOMED CT codes and the set of SNOMED CT FSNs. Like in all terminology systems, the set of synonyms in SNOMED CT never fully covers the linguistic diversity of a domain. This is the reason for advocating so-called interface terminologies as containers for synonyms and close-to-user expressions in general, which should be constructed bottom-up and linked to reference terminologies [6, 7].

2. Material and Methods

The international release of SNOMED CT is a hybrid between a reference terminology and a user interface terminology, according to [6]. Ambiguous terms are especially frequent when stripping the hierarchy tag from FSNs: “B-cell lymphoma” is therefore a synonym both of the SNOMED CT concept “B-cell lymphoma (disorder)” and “B-cell lymphoma (morphologic abnormality)”. Acronyms rarely appear as synonyms in SNOMED CT, because naming conventions [8] require that acronyms be followed by their expansion, such as in “PIN - Prostatic intraepithelial neoplasia”, with a dash enclosed in white space characters as delimiting sequence. For retrieving acronyms in text, however, the expanded form needs to be suppressed in the matching procedure; in this example this means the match is done with “PIN”. These are cases where lexical ambiguity becomes a serious issue, e.g. where the system has to choose between “Prostatic intraepithelial neoplasia” and “Pressure-induced nystagmus” when matching “PIN”. In the following we consider this a special case of lexical ambiguity.

Our analysis of the ambiguity of terms in SNOMED CT is based on all active concepts and terms from the January 2017 release. Lexical ambiguity is investigated at two different levels, viz. (i) full terms as obtained from the SNOMED CT description table, and (ii) acronym extracts that correspond to our definition (see below).

To this end, two dictionaries D_1 and D_2 are built. D_1 collects all SNOMED CT concept IDs to which an ambiguous term was assigned. D_2 does the same for acronyms, by matching the abovementioned pattern and ignoring the expansion section. The selection of acronyms was done according to a simple rule of thumb, which proved highly selective in medical terminologies: only tokens between two and seven characters, in which at least the second or third character is capitalised, are considered acronyms.

Both D_1 and D_2 are then analysed according to the following criteria:

- Combinations of SNOMED CT hierarchy tags, in order to better delineate where ambiguities occur.
- Cases where concepts that belong to ambiguous terms are semantically related by direct non-taxonomic links like **Associated morphology** or **Has active Ingredient**.
- Cases where concepts that belong to ambiguous terms are semantically related by direct taxonomic (***is-a***) links.

3. Results

Table 1 characterises either set. The existence of outliers is explained by the fact that a few acronyms are not followed by their expansions. For example, the acronym "O/E" (which means "on examination") occurs in hundreds of terms like "O/E - toe" or "O/E - eye". This shows that SNOMED CT's acronym – expansion pattern is not specific. It is not very sensitive either, because there are occurrences of acronyms that do not comply with the naming pattern at all. For example, "ENT" (Ear - nose - throat) is never introduced according to the naming pattern. It occurs not only in terms like "O/E - ENT" but also in isolation (just "ENT") as synonym of "Ear, nose and throat surgery".

Table 1: Frequency and distribution of ambiguous readings of SNOMED CT terms.

Dictionary	Count	Cardinality		Maximum
		Mean	Median	
D_1 (non-acronym terms)	7,439	2.02	2	6
D_2 (acronyms)	899	5.54	2	1678

Regarding the five most frequent hierarchy tag patterns, Table 2 and 3 show very different results comparing SNOMED CT full terms (D_1) and acronyms extracted according to the SNOMED CT acronym / definition pattern (D_2).

Table 2: Leading patterns of concept tuples connected by the same SNOMED CT (non-acronym) term

Hierarchy tag combination patterns	Pattern count	Rate of non-taxonomic links	Rate of taxonomic links
| product | substance |	4,064	0.888	0.000
| disorder | morphologic abnormality |	1,047	0.707	0.000
| organism | organism |	221	0.000	0.452
| procedure | substance |	213	0.911	0.000
| procedure | procedure |	200	0.000	0.465
Other n-tuples ($2 \leq n \leq 6$)	1,694		

Regarding D_1, we see a high aggregation of ambiguous terms with two combinations, viz. "| product | substance |" and "| disorder | morphologic abnormality |".

These two distributions also exhibit a high degree of ontological connection, which is also true for the combination "| procedure | substance |". Taxonomic links between concepts that share a term are quite frequent in all cases in which the ambiguity occurs within the same hierarchy. This also applies to many for the less frequent patterns not distinguished in Table 2.

In D_2, the distribution between patterns is more balanced, and the degree of connection between concepts that share the same acronym is lower.

Table 3: Leading patterns of concept tuples linked by the same acronym extracted from SNOMED CT terms

Hierarchy tag combination Patterns	Pattern count	Rate of non-taxonomic links	Rate of taxonomic links
| disorder | disorder |	66	0.015	0.167
| disorder | procedure |	59	0.034	0.000
| procedure | procedure |	38	0.000	0.263
| procedure | substance |	33	0.333	0.000
| disorder | substance |	28	0.000	0.000
Other n-tuples (2 ≤ n ≤ 1678)	675		

4. Discussion and recommendation

The way acronyms are introduced in SNOMED CT is neither specific nor sensitive. The pattern recommended by SNOMED CT to characterise acronym/definition pairs (e.g. "DNA - Did not attend") is also found in as acronym/specialisation pairs (e.g. "DNA - appointment mix-up"), which explains extreme cardinality outliers. Besides, the overall number of acronyms in SNOMED CT is not high, compared to the size of the terminology.

More than half of the term-level ambiguities are explained by concept pairs that are also ontologically connected. It concerns the combination of product concepts with substance concepts via the relation **Has active ingredient**, which relates, "Folinic acid (product)" with "Folinic acid (substance)". This is quite similar with disorder and morphology concepts connected via **Associated morphology**, relating, e.g. "Solar keratosis (disorder)" with "Solar keratosis (morphologic abnormality)", as well as substance concepts connected, e.g., via **Component**, such as "Curcumin stain (procedure)" and "Curcumin stain (substance)".

These frequent types ambiguities occur in a rather systematic way. Especially in the case of | disorder | morphology |, these parings can be considered as dot categories [9], i.e. complex categories that classify tightly connected concepts. Dot categories are often not really discerned by language and common sense. A commonly cited example for this is "book" as being both an information object and a physical object depending on the context (e.g. "this thick <physical object> book" is an incomprehensible "<information object> book"). Dot categories are well defined and easy to handle and comply, in their majority, with the SNOMED CT concept model. More problematic are lexical ambiguities in which the two competing concepts represent children and parents in the taxonomy, which is most likely to be found in the procedure and the organism branch. For instance, "Blepharotomy (procedure)" is a child of "Incision of eyelid (procedure)" and has as synonym "Incision of eyelid".

Nevertheless, the phenomenon of ambiguous terms in SNOMED CT has to be seen in the context of the size of the terminology. We found 7,439 ambiguous SNOMED CT terms and 899 ambiguous acronyms, which represents just 8,338 ambiguous terms. Especially the diversity of acronyms found in SNOMED seems small compared to their real occurrence in medical texts.

For the SNOMED CT curators we give the following recommendations:

- Create awareness that the lexical coverage of SNOMED CT with synonyms is limited.
- Concentrate on SNOMED CT as *reference ontology*, leaving the maintenance of collections of close-to-user terms (*user interface terminologies*) to user groups and release centres, according to the ASSESS-CT recommendations.
- Eliminate synonyms from parent concepts if there is already the same term in a child concept.
- Reconsider naming conventions.
- If possible, complete missing relations between those concepts that are linked by truly polysemous terms.

References

[1] Grange B, Bloom DA. Acronyms, abbreviations and initialisms. BJU Int. 2000 Jul;86(1):1-6.

[2] Acronym Finder. http://www.acronymfinder.com/

[3] SNOMED International (January 2017). http://www.snomed.org/snomed-ct

[4] Schober D, Smith B, Lewis SE, Kusnierczyk W, Lomax J, Mungall C, et al. Survey-based naming conventions for use in OBO Foundry ontology development. BMC Bioinformatics. 2009 Apr 27;10:125

[5] Smith B, Ashburner M, Rosse C et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007, 25: 1251–1255

[6] ASSESS-CT. Assessing SNOMED CT for Large Scale eHealth Deployments in the EU. http://assess-ct.eu

[7] Schulz S, Rodrigues JM, Rector A, Chute CG. Interface Terminologies, Reference Terminologies and Aggregation Terminologies: A Strategy for Better Integration. Accepted for MEDINFO 2017

[8] SNOMED CT Editorial Guide (January 2017). https://confluence.ihtsdotools.org/display/DOCEG

[9] Arapinis A,Vieu L (2015). A plea for complex categories in ontologies. *Applied Ontology*, *10*(3-4), 285-296.