Qualitative assessment of annotations using SNOMED CT

Jose Antonio Miñarro-Giménez[1,*], Catalina Martínez-Costa[1] and Stefan Schulz[1]

[1] Institute for Medical Informatics, Statistics and Documentation. Medical University of Graz, Austria.

ABSTRACT

Motivation: SNOMED CT provides about 300,000 codes with fine-grained definitions to support interoperability of health data. However, even experienced human coders tend to disagree about which codes to choose for expressing clinical content.

Results: 20 short clinical text fragments were independently annotated with SNOMED CT codes by two terminology experts. We analysed each disagreement instance and classified disagreements into eight categories, for which representative examples are presented.

Conclusion: For each disagreement category measures to improve the terminology and to support guidelines for human and machine annotation are proposed and discussed.

*** Contact:** jose.minarro-gimenez@medunigraz.at

1 INTRODUCTION

SNOMED CT (SNOMED CT, 2016) is the world's most comprehensive multilingual healthcare terminology. It enables machine-processable representation of clinical content in electronic health records (EHRs). SNOMED CT (SCT) provides terms in different languages, which are linked to language-independent concepts. Concepts are arranged in subclass hierarchies. Many concepts are described by formal axioms, for which SCT can be regarded as a formal ontology. The use of SCT in EHRs is expected to improve communication and semantic retrieval, real-time decision support, cross-border interoperability of health data, and retrospective reporting for research and management (Bodenreider, 2008). The ASSESS CT project (ASSESS CT, 2016) has collected empirical evidence for the fitness for purpose of SCT, compared to other terminologies. One of the project experiments focused on measuring manual annotation of samples of clinical text.

2 MATERIALS AND METHODS

Two terminology settings, viz. subsets of SCT and of the UMLS Metathesaurus (UMLS, 2009) without SCT were used as a testbed for manual annotations. To this end, annotation guidelines had been formulated for the two settings. Among others, they establish preference criteria and determine how to deal with concept composition: existing pre-coordinated concepts should be preferred. 20 English medical text samples (around 500 characters each) were annotated by independent medical experts. These samples had been collected from different sources, clinical disciplines and text types, partly translated to English from other European languages. We will describe the results of a qualitative assessment of annotations, for which types of disagreement and errors were identified and submitted to an in-depth analysis.

3 RESULTS AND DISCUSSION

The annotations of the corpus were split into 232 fragments, mostly noun phrases, according to the judgment of authors and annotators. The group of annotations consisted on average of 2.5 different codes. The SNOMED CT coordination syntax was not used, in order to ensure compatibility with the alternative scenario. The agreement between experts was astonishingly low: only in 20% of the fragments the same set of codes was found. Our analysis showed that the most common type of disagreement was due to sheer human error. E.g., content was missing, wrong codes were selected, guidelines were ignored, or existing content was not retrieved. Since more than one cause of disagreement can affect the same fragment, more errors than fragments were identified.

We categorize the reasons for disagreement into eight categories, based on the analysis of the annotated corpus. They are described in detail in the next subsections from more to less frequent.

3.1 Ambiguity in the interpretation of medical text

Parts of our texts were highly elliptical and therefore difficult to parse. This is a known issue in clinical narratives, which affects understanding of the content even by clinicians from a different specialty. As a result, diverging interpretations were found, e.g. in "the mitral valve liquid was removed by tapping", "tapping" was encoded as *Drainage procedure (procedure)* by one and as *Aspiration (procedure)* by the other annotator. Also in the text "The examination cannot be completed", "examination" was interpreted as *Clinical examination (procedure)* by one annotator and *Autopsy examination (procedure)* by the other due to the text is an autopsy report.

3.2 Ambiguous interface terms

Interface terms provide close-to-human expressions, including synonyms and abbreviations. Our analysis showed that different SCT concepts had similar interface terms, which complicated the choice: *Worried (finding)* and *Anxiety (finding*) are different concepts with the first one having "Anxious cognitions" as interface term. Exact definitions are missing.

In contrast to this example, most cases with similar interface terms belong to different hierarchies. In-depth knowledge of SNOMED CT and guideline adherence would avoid these errors. E.g. *Finding of measures of palpebral fissure (finding)* vs. *Measure of palpebral fissure (observable entity)* belong to the *Clinical Finding* and *Observable entity* hierarchies, respectively.

This type of disagreement is even more convoluted due to the existence of concepts from one hierarchy with interface terms suggesting the name of another one: *Pain (finding)* has "Pain observation" as an interface term, but belongs to the *Clinical finding* hierarchy.

3.3 Different concept granularity

The selection of the most appropriate concept for a text fragment by different annotators is biased by the ambiguity of the text and the level of annotators' domain knowledge. An example of this disagreement type is the annotation of the following text: "*The alcohol test in the vitreous body revealed an ethanol content of 2.7%*". Alcohol test can be annotated by *Alcohol measurement (procedure)* but the use of the descendant *Ethanol measurement (procedure)* is preferable because the substance measured is ethanol (albeit being commonly denoted by "alcohol").

3.4 Overlooking of pre-coordinated concepts

This type of disagreement occurs when the existence of a pre-coordinated concept is ignored and a combination of simpler concepts is taken instead. E.g. "Diabetes monitoring" is annotated by *Diabetes mellitus (disorder)* together with *Monitoring - action (qualifier value)* instead of *Diabetic monitoring (regime/therapy)*. In cases where the complex concept is fully defined, description logics reasoning might detect equivalence, however not in this case.

3.5 Qualifier value ambiguity

The SCT *Qualifier value* hierarchy offers many options for encoding what is normally expressed as adjectives. These concepts lack formal and textual definitions, which leads to boundary problems. The scope of use of qualifiers can often guessed from the ancestors, e.g. *Slight (qualifier value)* represents a magnitude whereas *Mild (qualifier value)* represents a severity degree. This often clashes with the use of words like in "*The examination revealed slight bleeding in the area of the mitral valve*", interpreted as (quantifiable) amount of bleeding by one annotator and as (qualitative) bleeding severity by the other one.

3.6 Ambiguity due to the use of laterality qualifiers

The ambiguity due to laterality is caused by the inheritance of laterality qualifiers such as *Left*, *Right*, *Bilateral*, *Unilateral* (Van Berkum et al, 2002). In contrast to such ambiguity, our annotation experiment the ambiguity found was slightly different. As we did not allow post-coordination, we formed annotation groups (i.e. unordered set of codes wherever more than one concept was needed for a fragment). Thus laterality qualifiers within a group caused ambiguity like in "Wound on the dorsum of the right hand", with *Right (qualifier value)* instead of *Structure of right hand (body structure)* because it could mean "the right side of the dorsum of the hand".

3.7 Ambiguous or incomplete guideline specification

In order to reduce annotation disagreements, we had provided the annotators with annotation guidelines with recommendations like "Use concepts from the *Substance* hierarchy instead of from the *Pharmaceutical product* one whenever possible". In several cases, non-compliance with guidelines led to disagreements: E.g. *Clinical finding* concepts were used instead of *Observable entity* concepts in cases where a value could be post-coordinated or the use of navigational nodes (i.e. internal hierarchy grouping nodes) as *Clinical Findings*, such as *Finding of muscle tone*.

3.8 Weaknesses in SNOMED CT

Some weaknesses of SCT became obvious in the study, such as lack of formal or textual definitions, or incomplete taxonomies. An example of the former is the sub-hierarchy of the concept *Hodgkin's disease (disorder)*, where "Classic Hodgkin disease" was not among the leaf concepts, although this was intuitively assumed. Thus the annotators had the dilemma of choosing the parent concept or guessing the most likely one of the subtypes, cf. Figure 1.

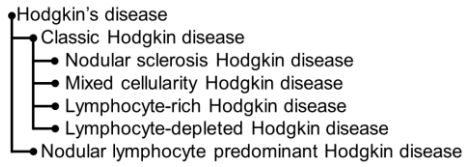

Figure 1. Hodgkin's disease hierarchy according to PDQ (2016)

4 CONCLUSION

The analysis of expert annotations of clinical text with SCT highlighted typical ambiguities and errors, which produced annotation disagreements. Some of them could be contributed by the annotators themselves. Which could be mitigated by more sophisticated tools supported by rich interface terms. Others were caused by ambiguity of medical text and typical clinical language phenomena. Weaknesses in the terminology proper and/or in the annotation guidelines accounted for other errors. Table 1 shows the types of disagreements, together with their frequency.

Table 1. Typology of disagreements with frequency

Reason of disagreement	Count
Sheer human error	150
3.1. Ambiguity in the interpretation of medical text	42
3.2. Ambiguous interface terms	41
3.3. Different concept granularity	39
3.4. Overlooking of pre-coordinated concepts	19
3.5. Qualifier value ambiguity	14
3.6. Ambiguity due to the use of laterality qualifiers	11
3.7. Ambiguous or incomplete guideline specification	10
3.8. Weakness in SNOMED CT	10

We conclude that reaching a complete agreement of all annotations is difficult. A route towards more consistent annotations includes the improvement of the quality of the terminology, appropriate tools, and more precise definitions of preferences among possible codes in the guidelines. It should finally be mentioned that the low inter-annotator agreement for SCT was paralleled by an equally low agreement for the alternative, UMLS-based scenario.

Acknowledgements. This paper has been written as part of ASSESS CT which is funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 643818.

REFERENCES

ASSESS CT (2016) http://assess-ct.eu

Bodenreider, O. (2008). Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. *Yearb Med Inform.*, 67-79.

PDQ® Adult Treatment Editorial Board (2016). PDQ Adult Hodgkin Lymphoma Treatment. Bethesda, MD: National Cancer Institute. Available at: www.cancer.gov/types/lymphoma/patient/adult-hodgkin-treatment-pdq. [PMID: 26389245]

SNOMED CT (2016). http://www.ihtsdo.org/snomed-ct

UMLS® Reference Manual [Internet]. Bethesda (MD): National Library of Medicine (US); 2009 Sep-. 2, Metathesaurus. Available from: http://www.ncbi.nlm.nih.gov/books/NBK9684/

Van Berkum, M. M., Rallins, M. C., & Spackman, K. A. (2002). Choosing Sides. Assigning Laterality as an Attribute in SNOMED® CT. *Proceedings of the AMIA Symposium*, 1184