A comparative analysis of the density of the SNOMED CT conceptual content for semantic harmonization

Zhe He, James Geller, Yan Chen

Medical terminologies vary in the amount of concept information (the “density”) represented, even in the same sub-domains. This causes problems in terminology mapping, semantic harmonization and terminology integration. Moreover, complex clinical scenarios need to be encoded by a medical terminology with comprehensive content. SNOMED Clinical Terms (SNOMED CT), a leading clinical terminology, was reported to lack concepts and synonyms, problems that cannot be fully alleviated by using post-coordination. Therefore, a scalable solution is needed to enrich the conceptual content of SNOMED CT. We are developing a structure-based, algorithmic method to identify potential concepts for enriching the conceptual content of SNOMED CT and to support semantic harmonization of SNOMED CT with selected other Unified Medical Language System (UMLS) terminologies.

We first identified a subset of English terminologies in the UMLS that have ‘PAR’ relationship labeled with ‘IS_A’ and over 10% overlap with one or more of the 19 hierarchies of SNOMED CT. We call these “reference terminologies” and we note that our use of this name is different from the standard use. Next, we defined a set of topological patterns across pairs of terminologies, with SNOMED CT being one terminology in each pair and the other being one of the reference terminologies. We then explored how often these topological patterns appear between SNOMED CT and each reference terminology, and how to interpret them.

Four viable reference terminologies were identified. Large density differences between terminologies were found. Expected interpretations of these differences were indeed observed, as follows. A random sample of 299 instances of special topological patterns (“2:3 and 3:2 trapezoids”) showed that 39.1% and 59.5% of analyzed concepts in SNOMED CT and in a reference terminology, respectively, were deemed to be alternative classifications of the same conceptual content. In 30.5% and 17.6% of the cases, it was found that intermediate concepts could be imported into SNOMED CT or into the reference terminology, respectively, to enhance their conceptual content, if approved by a human curator. Other cases included synonymy and errors in one of the terminologies.

These results show that structure-based algorithmic methods can be used to identify potential concepts to enrich SNOMED CT and the four reference terminologies. The comparative analysis has the future potential of supporting terminology authoring by suggesting new content to improve content coverage and semantic harmonization between terminologies.