AMIA Annu Symp Proc 2017:163-164

Identifying Potentially Missing Hierarchical Relations in SNOMED CT based on Lexical Features – Impact of Synonyms and Lexico-syntactic Constraints

Satyajeet Raje, PhD, Olivier Bodenreider, MD, PhD
U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Introduction

The quality assurance of large bio-ontologies is extremely critical for their effective and continued use and is an active area of research[1]. For example, recent investigations highlighted issues in the hierarchical structure of SNOMED CT and its detrimental effects on biomedical applications[2]. Previous work by one of the authors[3] established a method to identify potentially missing hierarchical relations leveraging lexical features in SNOMED CT. It used the preferred term for each concept in SNOMED CT to create logical definitions for concepts. These definitions were used to identify missing hierarchical relations. The method was tested on a subset of SNOMED CT concepts and showed limited precision (20% of false positives). In this paper, we propose two improvements on our original method: 1) by adding lexico-syntactic constraints based on shallow parsing, we expect to increase precision; 2) by processing all synonyms, we expect to increase recall. This work is a contribution to quality assurance in SNOMED CT.

Methods

We used the Sept. 2016 release of SNOMED CT (US Edition) in this study. Our methodology can be summarized as follows. We create logical definitions for concepts leveraging all concept labels (preferred terms and synonyms) provided by SNOMED CT and their lexico-syntactic analysis. We infer hierarchical relations from the logical definitions and filter out those already present in the original hierarchy of SNOMED CT. Finally, we review the potentially missing relations. We compare the results of the enhanced method to those of the original (baseline) method.

A. Creating logical definitions. In the original method, each concept in SNOMED CT is represented as a set of words using description logics (see example below in Table 1). In the enhanced method:

1) We distinguish between head nouns and modifiers through lexico-syntactic analysis using the minimal commitment parser provided by SemRep[4]. E.g. "Photogenic epilepsy" has head "epilepsy" and has modifier "photogenic". We then create logical definitions representing the head and modifier roles of words in the term. Complex terms are ignored. (In the definition below, "some" is the existential qualifier in the OWL syntax.)

2) We create logical definitions for each synonym of the concept in addition to the preferred term. For example, the concept *Photogenic epilepsy* has a synonym, "Television epilepsy" in addition to its preferred term, "Photogenic epilepsy". Definitions created for terms are later mapped to the corresponding concept.

Table 1. Example logical definitions from the original and enhanced methods.

Original	"Photogenic epilepsy" ≡ *disorder* AND (has_word some *photogenic*) AND (has_word some *epilepsy*)
Enhanced	"Photogenic epilepsy" ≡ *disorder* AND (has_head some *epilepsy*) AND (has_mod some *photogenic*) "Television epilepsy" ≡ *disorder* AND (has_head some *epilepsy*) AND (has_mod some *television*)

We apply this enhanced method to the same hierarchies as the original study, namely the *Disorder of head (disorder) (118934005)* and *Operative procedure on head (procedure) (89901005).*

B. Inferring and filtering hierarchical relations. We use the ELK reasoner[5] to infer a hierarchy of terms from the logical definitions (expressed in OWL 2 EL profile[6]). For instance, ELK infers that *disorder* AND (has_head some *epilepsy*) AND (has_mod some *photogenic*) is a subclass of *disorder* AND (has_head some *epilepsy*). We further derive a hierarchy of concepts from the hierarchy of terms. For example, *Photogenic epilepsy (disorder)* is a subclass of *Epilepsy (disorder)*. We only keep those subclass relations (between concepts) that are not already present in SNOMED CT (transitively closed).

C. Comparing results from original and enhanced methods. We compare the performance of the original and enhanced methods. To validate the results, we manually review all potentially missing subclass relations generated by either method. The results were pooled during review in order to mask the source of the relations.

Results

A. Creating logical definitions. We created logical definitions for the 12,088 concepts of the subhierarchy rooted with the concept *Disorder of head (disorder)* and for the 3798 concepts from *Operative procedure on head (proce-*

dure) by the original method. With the enhanced method, we could create definitions for 15,757 terms (9687 concepts) and 3129 terms (2171 concepts), respectively.

B. Inferring and filtering hierarchical relations. A total of 612 potentially missing relations were found by the original method, and 525 by the enhanced method. Common to both methods were 225 relations, giving a combined total of 912 relations.

C. Comparing results from original and enhanced methods. Of the 912 combined relations, 657 (72%) were judged valid. Table 2 gives a breakdown of the results per method. Contrary to our expectation, the addition of lexico-syntactic constraints and synonyms does not provide the expected gain in performance. However, there is significant gain in precision (~10 points) for the 225 relations retrieved by both methods (intersection).

Table 2. Performance of the original and enhanced methods.

	Original (A)	Enhanced (B)	Both (A ∩ B)	Total (A ∪ B)
Valid	488	370	201	657
Invalid	124	155	24	255
Total relations identified	612	525	225	912
Precision	*0.797*	*0.705*	*0.893*	*0.72*
*Recall**	*0.742*	*0.563*	*0.304*	-
F1 measure	*0.769*	*0.626*	*0.456*	-

**The combined 912 relations are considered as reference. Recall is based on total 657 "valid" relations in this set.*

Discussion

A. Adding lexico-syntactic constraints increases precision. The gain in precision for the 225 relations retrieved by both methods (intersection) indicates that the lexico-syntactic constraints reduce false positives (adding precision). For example, *Removal of fixation of mandible (procedure)* is no longer recognized as subclass of *Fixation of mandible (procedure)*, because the two terms have different head nouns. This level of precision has potential application to quality assurance. However, it also causes a significant drop in recall, because our method excludes complex syntactic patterns.

B. Enhancements degrade performance. We observed that adding synonyms created additional valid relations that cannot be inferred by the original method. For example, *Enlarged parietal foramina (disorder)* is recognized as subclass of *Craniolacunia (disorder)*. However, it also adds many false positives. For example, *Acquired keratoglobus (disorder)* is inferred as subclass of *Congenital keratoglobus (disorder)*, because it has a synonym "Keratoglobus", reflecting the predominant congenitality of this condition.

Conclusion

We have evaluated the effects of adding synonyms and lexico-syntactic constraints on the identification of potentially missing hierarchical relations in a subset of SNOMED CT. In the future, we would like to apply our method to all of SNOMED CT and study the relative contribution of each enhancement (adding synonyms vs. lexico-syntactic constraints). Importantly, the missing hierarchical relations identified by our methods may only indicate underlying issues in SNOMED CT's concept definitions, which will require domain expertise to address.

Acknowledgements

This work is supported by the Intramural Research Program of the U.S. National Library of Medicine (NLM) and an appointment to the NLM Research Participation Program administered by ORISE through an interagency agreement between the U.S Dept. of Energy and the NLM.

References

1. Geller J, Perl Y, Halper M, Cornet R. Special Issue on Auditing of Terminologies. J Biomed Inform. 2009;42(3):407-11.
2. Rector AL, Brandt S, Schneider T. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. J Am Med Inform Assoc. 2011;18(4):432-40.
3. Bodenreider O. Identifying missing hierarchical relations in SNOMED CT from logical definitions based on the lexical features of concept names. Proceedings of the 6th International Conference on Biomedical Ontology (ICBO 2016). 2016:(electronic proceedings: http://ceur-ws.org/Vol-1747/IT601_ICBO2016.pdf).
4. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003;36(6):462-77.
5. Kazakov Y, Krötzsch M, Simančík F. ELK: a reasoner for OWL EL ontologies. System Description. 2012.
6. W3C. OWL 2 web ontology language document overview. 2012.