Journal of Biomedical Informatics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Auditing SNOMED CT hierarchical relations based on lexical features of concepts in non-lattice subgraphs

Licong Cui[a,b,*], Olivier Bodenreider[d], Jay Shi[c], Guo-Qiang Zhang[a,b,c]

[a] Department of Computer Science, University of Kentucky, Lexington, KY, USA
[b] Institute for Biomedical Informatics, University of Kentucky, Lexington, KY, USA
[c] Department of Internal Medicine, University of Kentucky, Lexington, KY, USA
[d] National Library of Medicine, Bethesda, MD, USA

ARTICLE INFO

Keywords:
Biomedical ontologies
SNOMED CT
Quality assurance
Non-lattice subgraphs
Lexical attributes

ABSTRACT

Objective: We introduce a structural-lexical approach for auditing SNOMED CT using a combination of non-lattice subgraphs of the underlying hierarchical relations and enriched lexical attributes of fully specified concept names. Our goal is to develop a scalable and effective approach that automatically identifies missing hierarchical IS-A relations.

Methods: Our approach involves 3 stages. In stage 1, all non-lattice subgraphs of SNOMED CT's IS-A hierarchical relations are extracted. In stage 2, lexical attributes of fully-specified concept names in such non-lattice subgraphs are extracted. For each concept in a non-lattice subgraph, we enrich its set of attributes with attributes from its ancestor concepts within the non-lattice subgraph. In stage 3, subset inclusion relations between the lexical attribute sets of each pair of concepts in each non-lattice subgraph are compared to existing IS-A relations in SNOMED CT. For concept pairs within each non-lattice subgraph, if a subset relation is identified but an IS-A relation is not present in SNOMED CT IS-A transitive closure, then a missing IS-A relation is reported. The September 2017 release of SNOMED CT (US edition) was used in this investigation.

Results: A total of 14,380 non-lattice subgraphs were extracted, from which we suggested a total of 41,357 missing IS-A relations. For evaluation purposes, 200 non-lattice subgraphs were randomly selected from 996 smaller subgraphs (of size 4, 5, or 6) within the "Clinical Finding" and "Procedure" sub-hierarchies. Two domain experts confirmed 185 (among 223) suggested missing IS-A relations, a precision of 82.96%.

Conclusions: Our results demonstrate that analyzing the lexical features of concepts in non-lattice subgraphs is an effective approach for auditing SNOMED CT.

1. Introduction

Biomedical ontologies and standardized terminologies such as SNOMED CT play an important role in healthcare information management, biomedical information extraction, and data integration [1]. SNOMED CT [2], the primary focus of this paper, is the largest clinical terminology used worldwide. Managed by the SNOMED International, SNOMED CT has been used in electronic health records (EHRs) and for clinical decision support, information retrieval, and semantic interoperability. Under the Health Information Technology for Economic and Clinical Health (HITECH) Act [3], SNOMED CT has been required in the United States for encoding relevant clinical information to ensure meaningful use of EHRs. The use of SNOMED CT in EHRs supports cost-effective delivery of care.

The quality of SNOMED CT impacts the quality of EHR and patient safety. For example, an increasing variety of value sets (consisting of subsets of SNOMED CT concepts) have been specified for EHR-based decision support, quality reporting, and cohort selection. Value sets can be intensionally defined, i.e., as the list of concepts sharing some common features, e.g., all descendants of "Malignant epithelial neoplasm of skin" in the Clinical Finding sub-hierarchy. However, "Squamous cell carcinoma of skin" is currently not listed as one of its descendants, and would thus be missing from the corresponding value set. As a consequence, patients with "Squamous cell carcinoma of skin" would not be selected for a cohort of patients with "Malignant epithelial neoplasm of skin."

Due to the large size and complexity of SNOMED CT (over 300,000 concepts and over 1.5 million relations), quality issues such as wrong hierarchical classifications, missing hierarchical relations, and missing concepts are inevitable, and the root cause of these problems can

[*] Corresponding author at: Department of Computer Science, University of Kentucky, 301 Rose Street, Lexington, KY 40506, USA.
E-mail address: licong.cui@uky.edu (L. Cui).

https://doi.org/10.1016/j.jbi.2017.12.010
Received 16 October 2017; Received in revised form 18 December 2017; Accepted 19 December 2017

sometimes be traced back to incomplete or inaccurate logical definitions. Most existing approaches to quality assurance of SNOMED CT merely indicate the presence of possible quality issues and do not precisely identify the location or nature of the problem. Arduous manual review by domain experts or ontology auditors is then required to validate the potential errors and, more importantly, fix these errors in future versions.

We introduce a structural-lexical approach for auditing SNOMED CT using a combination of non-lattice subgraphs of the underlying hierarchical relations and enriched lexical attributes of fully specified concept names. Our goal is to develop a scalable and effective approach that automatically identifies missing IS-A relations with high precision. A secondary goal is to uncover related incorrect IS-A relations in the subgraphs. Our approach involves three stages. In stage 1, all non-lattice subgraphs of SNOMED CT's IS-A hierarchical relations are extracted. In stage 2, lexical attributes of fully-specified concept names in such non-lattice subgraphs are extracted. For each concept in a non-lattice subgraph, we enrich its set of attributes with attributes from its ancestor concepts within the non-lattice subgraph. In stage 3, subset inclusion relations between the lexical attribute sets of each pair of concepts in each non-lattice subgraph are compared to existing IS-A relations in SNOMED CT. For concept pairs within each non-lattice subgraph, if a subset relation is identified but an IS-A relation is not present in SNOMED CT IS-A transitive closure, then a missing IS-A relation is reported.

2. Background

2.1. SNOMED CT

SNOMED CT, owned and distributed by SNOMED International, is the most comprehensive clinical health terminology worldwide [2]. It contains over 300,000 concepts that are hierarchically organized in a Directed Acyclic Graph (DAG) of IS-A relations. SNOMED CT has 19 top-level sub-hierarchies including "Clinical finding," "Procedure," and "Body Structure." Each concept in SNOMED CT has a fully specified name, which is in the form of the preferred term followed by a semantic tag in parentheses, e.g., "Congenital sacral meningocele (disorder)."

2.2. Non-lattice subgraphs

From the point of view of the hierarchical structure (i.e., DAG of IS-A relations), lattice is a desirable property for a well-formed ontology or terminology [12]. A lattice is a specific type of DAG such that any two nodes (or concepts) have a unique maximal shared descendant and a unique minimal shared ancestor. A pair of concepts is called a *non-lattice pair*, if the two concepts have more than one maximal shared common descendant [13–15]. For example, in Fig. 1, the concept pair (1, 2) is a non-lattice pair, since they have two maximal shared common descendants: 5 and 6. In previous work [12–14], we have developed various computational approaches to systematically extract all the non-lattice pairs in SNOMED CT for further auditing.

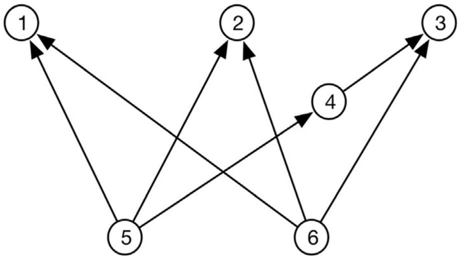

Fig. 1. An example of a non-lattice subgraph of size 6. Here nodes represent concepts, and edges represent subconcept-superconcept relations. For instance, the edge from 5 to 1 means 5 is a subclass of 1.

Since there may exist multiple non-lattice pairs having the same maximal shared descendants (such as (1, 2), (1, 3), and (2, 3) in Fig. 1), separately analyzing each such non-lattice pair would be redundant. Therefore, a notion of *non-lattice subgraph* is further introduced to avoid redundant analysis [15]. Given a non-lattice pair $p = (c_1,c_2)$ and its maximal common descendants $mcd(p)$, the corresponding non-lattice subgraph can be obtained by first computing the minimal common ancestors of the maximal common descendants, $mca(mcd(p))$, and then aggregating the concepts and the IS-A edges between (including) any concept in $mca(mcd(p))$ and any concept in $mcd(p)$. For instance, given the non-lattice pair (1, 2) in Fig. 1 and its maximal common descendants {5, 6}. Computing the minimal common ancestors of {5, 6} yields {1, 2, 3}. Then aggregating all the concepts and edges between {1, 2, 3} and {5, 6} yields a non-lattice subgraph consisting of the concepts {1, 2, 3, 4, 5, 6} and IS-A edges {(5, 1), (6, 1), (5, 2), (6, 2), (4, 3), (6, 3), (5, 4)}. The size of a non-lattice subgraph is defined as the number of concepts it contains.

2.3. Related work and specific contribution

Auditing or quality assurance of biomedical terminologies (including SNOMED CT) has been an active research area given its importance. The three main approaches to auditing terminologies are based on lexical, structural and semantic features (see [4] for a review of auditing techniques). Structural auditing methods include abstraction networks (AbNs), which have been extensively investigated as a means to help identify SNOMED CT subdomains that may need more attention for quality assurance work [5–9]. AbNs group concepts based on shared outgoing attribute relationships. AbNs-based approaches only identify areas of SNOMED CT where errors may be concentrated, with limited precision. In contrast, our approach identifies errors with high precision and pinpoints their location. Based on this information, SNOMED CT editors can focus on correcting the logical definitions.

Somewhat similar to our approach, Agrawal et al. used a combination of lexical and structural indicators to identify inconsistency issues in the logical definitions of SNOMED CT concepts [10,11]. They first identify lexically similar concepts (i.e., with terms of the same length, but differing by one word) and then compare the concepts' logical definitions in attribute relationships (structural part) to detect inconsistently modeled concepts. However, Agrawal's method relies on lexically similar concepts and has limited applicability, as well as limited precision. In contrast, our approach first identifies non-lattice subgraphs and then utilizes enriched lexical attributes of concepts in such non-lattice subgraphs to suggest missing IS-A relations. Therefore, our approach is widely applicable to biomedical ontologies and achieves a higher precision.

In previous work [15], we introduced a hybrid structural-lexical approach based on the lexical patterns of concept names in non-lattice subgraphs to automatically suggest missing hierarchical relations and concepts in SNOMED CT. However, the predefined lexical patterns only covered 4% of non-lattice subgraphs in SNOMED CT. In this work, we expand on this work and enrich the lexical attributes of each concept in non-lattice subgraphs to facilitate the identification of missing IS-A relations. This approach takes advantage of the rich lexical information contained in the ancestors of each concept in non-lattice subgraphs to facilitate the auditing process. The structural-lexical approach introduced in this work is more general. It supports the analysis of a larger proportion (7.4%) of the non-lattice subgraphs and identifies previously undiscovered missing hierarchical relations.

3. Material and methods

We use the September 2017 release of SNOMED CT (US edition) in this work. We extract all the non-lattice subgraphs in SNOMED CT. We