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*

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

*Corresponding author. Email address: licong.cui@uky.edu (Licong Cui). Address: Department of Computer Science, University of Kentucky, 301 Rose Street, Lexington, KY 40506, USA.

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.

Keywords: Biomedical ontologies, SNOMED CT, quality assurance, non-lattice subgraph, lexical atrributes

---

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 decision support, quality reporting, and cohort selection. Value sets can be intensionally defined, i.e., as the list of concepts sharing some common feature, e.g., all descendants of "Malignant epithelial neoplasm of skin" in the disease subhierarchy. 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 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 subhierarchies 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, 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, 14, 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, 13, 14], we have developed various computational approaches to systematically extract all the non-lattice pairs in SNOMED CT for further auditing.

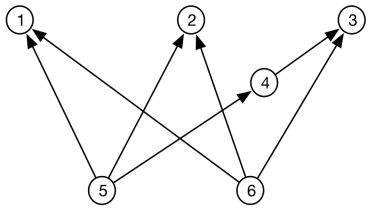

Figure 1: An example of a non-lattice subgraph of size 6. Here nodes represent concepts, and edges represent subconcept-superconcept relations. For instance, the edge from 5 to 1 means 5 is a subclass of 1.

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, $mca(mcd(p))$; 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, 6, 7, 8, 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 enrich the lexical attributes of concepts in non-lattice subgraphs, identify missing hierarchical IS-A relations between concepts based on the enriched lexical attributes. Clinical experts evaluate a random sample of suggested missing IS-A relations to verify missing IS-A relations and incorrect IS-A relations.

Algorithm 1 presents the pseudocode for identifying missing IS-A relations for a given non-lattice subgraph based on enriched lexical attributes. The algorithm mainly consists of three steps: detection of stop words and antonyms (lines 1 – 5), construction of enriched lexical attributes (lines 6 – 12), and identification of missing IS-A relations (lines 13 – 19). We describe these steps in detail and provide illustrative examples.

3.1. Detection of stop words and antonyms

Since the lexical attributes of the concept "Fetal hypertrophic cardiomyopathy due to maternal diabetes mellitus" contain that of the concept "Diabetes mellitus," the relation "Fetal hypertrophic cardiomyopathy due to maternal diabetes mellitus" IS-A "Diabetes mellitus" would be incorrectly generated. Similarly, "Periostitis without osteomyelitis" IS-A "Osteomyelitis" would be incorrectly generated. Along the same lines, the concept "Open reduction of closed sacral fracture" contains antonyms "open" and "closed," and the concept "Acute on chronic endometritis disorder" contain antonyms "acute" and "chronic." Ignoring terms that contains antonyms prevents us from suggesting wrong relations, for example, between "Open reduction of

Input: A non-lattice subgraph G consisting of concepts and IS-A relations
Output: Reclassified IS-A relations

if *the fully specified name of a concept in* G *contains stop word(s) or antonyms* **then**
    stop here;
else
    continue;
end
Compute the transitive closure of the IS-A relations in G;
Derive term pairs based on the transitive closure and fully specified names of the concepts in G;
for *each concept* c *in* G **do**
    Initialize a set L_c of lexical attributes for c using its fully specified name;
    Enrich L_c by leveraging the lexical attributes of c's ancestors;
    Enrich L_c by the derived term pairs;
end
for *each concept* c_1 *in* G **do**
    for *each concept* c_2 *in* G **do**
        if $c_1 \neq c_2$ *and* $L_{c_1} \subseteq L_{c_2}$ **then**
            Suggest c_2 IS-A c_1;
    end
end
Reduce the resulted IS-A relations to direct IS-A relations;

Algorithm 1: Pseudocode for identifying missing IS-A relations for a non-lattice subgraph based on enriched lexical attributes.

closed sacral fracture" and "Open reduction of open sacral fracture." To prevent such issues, we exclude from processing those terms containing words, such as "due to" and "without." More generally, we extend this measure to a list of stop words and antonyms.

We consider the following as stop words: "and," "or," "and/or," "no," "not," "without," "due to," "secondary to," "except," "by," "after," "co-occurrent," "bilateral," "examination," "able," "amputation," "removal," "replacement," "resection," "excision." For antonyms, we rely on a list of pairs of antonyms from WordNet [17, 18], including ("anterior," "posterior"), ("chronic," "acute"), ("open," "closed"), ("positive," "negative"), ("high," "low"), ("benign," "malignant"), ("right," "left"), ("simple," "compound").

Given a non-lattice subgraph G, we detect if any concept in G contains stop word(s) and antonyms, which are prone to generate incorrect IS-A relations using lexical attributes in practice. If stop word(s) or antonyms are detected, we discontinue the investigation of the non-lattice subgraph (i.e., stop the process of identifying missing IS-A relations for G).

3.2. Construction of enriched lexical attributes

Given a non-lattice subgraph G, we construct an enriched set of lexical attributes for each concept in G by leveraging three sources. The first source is the fully specified name of the concept itself, i.e., its own lexical attributes; the second source is the fully specified names of the concept's ancestors within the subgraph, i.e., often more generic words compared to the attributes of the concept itself; and the third source is a set of derived term pairs, intended to capture hypernymy relations between individual words from hierarchically related concepts in the non-lattice subgraph.

To obtain the second source, we compute the transitive closure of the IS-A relations in G, denoted by $T = \{(d, a)|$ concept a is an ancestor of concept d and $a \in G\}$. To obtain the third source, for each concept pair (d, a) in T, assuming W_d and W_a represent the sets of words contained in the concepts d and a, respectively; if $W_d \cap W_a \neq \emptyset$, $W_d - (W_d \cap W_a) \neq \emptyset$, and $W_a - (W_d \cap W_a) \neq \emptyset$, we obtain a derived term pair $\Big(W_d - (W_d \cap W_a), W_a - (W_d \cap W_a)\Big)$. Take the concept pair ("Fracture subluxation of perilunate joint," "Fracture dislocation of perilunate joint") as an example. We have W_d={fracture, subluxation, of, perilunate, joint} and W_a={fracture, dislocation, of, perilunate, joint}, and thus $W_d \cap W_a$={fracture, of, perilunate, joint}, from which we derive the term pair ("subluxation," "dislocation"). This derived term pair captures the fact that dislocation is a hypernym of (i.e., is more generic than) subluxation.

Leveraging the three sources, we build an enriched set of lexical attributes (in lowercase) for each concept c in G as follows.

1. We initialize a set L_c of lexical attributes using the set of words contained in the fully specified name of c.
2. For each ancestor a of c within G, we enrich L_c by adding the set of words contained in the fully specified name of a.
3. For any derived term pair (p_1, p_2), if the term p_1 is contained in the fully specified name of c, then we further enrich L_c by adding the set of words in the term p_2.

We illustrate the process of constructing enriched lexical attributes using the non-lattice subgraph shown in Fig. 2A. This non-lattice subgraph consists of 6 concepts (numbered in circles). The initialized sets of lexical attributes

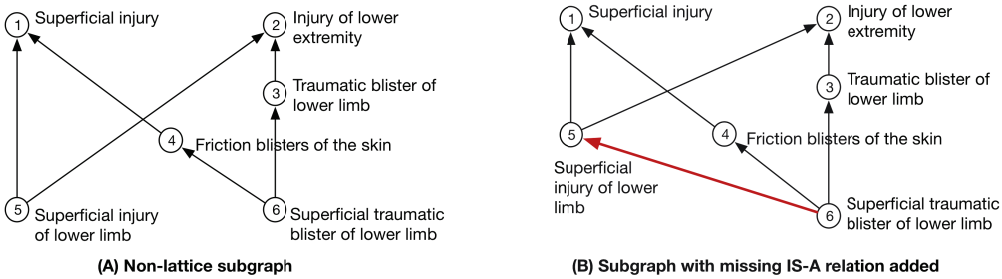

Figure 2: An example of a non-lattice subgraph of size 6 in the "Clinical finding" sub-hierarchy, as well as the resulted subgraph after adding a missing IS-A relation (red link): "Superficial traumatic blister of lower limb" IS-A "Superficial injury of lower limb."

using the fully specified names of the six concepts are:

$L_1 = \{\text{superficial, injury}\}$,

$L_2 = \{\text{injury, of, lower, extremity}\}$,

$L_3 = \{\text{traumatic, blister, of, lower, limb}\}$,

$L_4 = \{\text{friction, blisters, of, the, skin}\}$,

$L_5 = \{\text{superficial, injury, of, lower, limb}\}$,

$L_6 = \{\text{superficial, traumatic, blister, of, lower, limb}\}$.

Leveraging the ancestors' lexical attributes results in the following enriched sets (with newly added lexical attributes *italicized*):

$L_1 = \{\text{superficial, injury}\}$,

$L_2 = \{\text{injury, of, lower, extremity}\}$,

$L_3 = \{\text{traumatic, blister, of, lower, limb, } \textit{injury, extremity}\}$,

$L_4 = \{\text{friction, blisters, of, the, skin, } \textit{superficial, injury}\}$,

$L_5 = \{\text{superficial, injury, of, lower, limb, } \textit{extremity}\}$,

$L_6 = \{\text{superficial, traumatic, blister, of, lower, limb, } \textit{injury, extremity, friction, blisters, the, skin}\}$.

Leveraging the derived term pairs results in the same sets of lexical attributes

(i.e., no additional lexical attributes are added for the concepts).

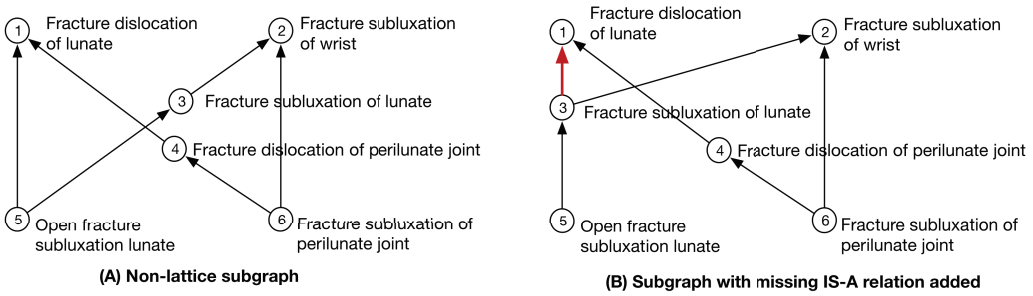

Figure 3: An example of a non-lattice subgraph of size 6 in the "Clinical finding" sub-hierarchy, as well as the resulted subgraph after adding a missing IS-A relation (red link): "Fracture subluxation of lunate" IS-A "Fracture dislocation of lunate."

Fig. 3A shows another example of non-lattice subgraph. The initial sets of lexical attributes using the fully specified names of the six concepts are:

L_1 = {fracture, dislocation, of, lunate},

L_2 = {fracture, subluxation, of, wrist},

L_3 = {fracture, subluxation, of, lunate},

L_4 = {fracture, dislocation, of, perilunate, joint},

L_5 = {open, fracture, subluxation, lunate},

L_6 = {fracture, subluxation, of, perilunate, joint},

Leveraging the ancestors' lexical attributes results in the following enriched sets (with newly added lexical attributes *italicized*):

L_1 = {fracture, dislocation, of, lunate},

L_2 = {fracture, subluxation, of, wrist},

L_3 = {fracture, subluxation, of, lunate, *wrist*},

L_4 = {fracture, dislocation, of, perilunate, joint, *lunate*},

L_5 = {open, fracture, subluxation, lunate, *dislocation, of, wrist*},

L_6 = {fracture, subluxation, of, perilunate, joint, *dislocation, lunate, wrist*}.

Leveraging the derived term pairs results in the following final sets of lexical attributes (with newly added lexical attributes *italicized*):

L_1 = {fracture, dislocation, of, lunate},

L_2 = {fracture, subluxation, of, wrist, *dislocation*},

L_3 = {fracture, subluxation, of, lunate, wrist, *dislocation*},

L_4 = {fracture, dislocation, of, perilunate, joint, lunate},

L_5 = {open, fracture, subluxation, lunate, dislocation, of, wrist},

L_6 = {fracture, subluxation, of, perilunate, joint, dislocation, lunate, wrist}.

Note that the enrichment of L_2 and L_3 is due to the derived term pair ("subluxation", "dislocation"), which is obtained by the concept pair (6, 4) in the transitive closure, that is, ("Fracture subluxation of perilunate joint", "Fracture dislocation of perilunate joint").

3.3. Identification of missing IS-A relations

We compute all possible IS-A relations between concepts in a given non-lattice subgraph G using the enriched lexical attributes for each concept (L_{c_i}). For any two concepts c_1 and c_2, if L_{c_1} is a proper subset of L_{c_2}, then we suggest c_2 is more specific than c_1 (or c_2 IS-A c_1). Then we further reduce the computed IS-A relations to direct IS-A relations to eliminate relations that can be inferred from other relations. We compare the set of relations obtained from enriched lexical attributes of concepts in non-lattice subgraphs to the IS-A relations present in the inferred hierarchy of SNOMED CT. The

relations obtained through our approach, but not present in SNOMED CT, are considered missing relations.

For example, for the concepts numbered 5 and 6 in Fig. 2A, $L_5 = \{$superficial, injury, of, lower, limb, extremity$\}$ is a proper subset of $L_6 = \{$superficial, traumatic, blister, of, lower, limb, injury, extremity, friction, blisters, the, skin$\}$, thus we suggest concept 6 is more specific than concept 5, that is, "Superficial traumatic blister of lower limb" IS-A "Superficial injury of lower limb" (see the red link in Fig. 2B). Computing all IS-A relations in the graph in Fig. 2A results in the following set of IS-A relations: {(4, 1), (5, 1), (6, 1), (3, 2), (5, 2), (6, 2), (6, 3), (6, 4), (6, 5)}, which can be further reduced to direct relations: {(4, 1), (5, 1), (3, 2), (5, 2), (6, 3), (6, 4), (6, 5)}. Here (6, 5) is the newly identified relation, because all the others already exist in the original non-lattice subgraph.

For the concepts 1 and 3 in Fig. 3A, $L_1 = \{$fracture, dislocation, of, lunate$\}$ is a proper subset of $L_3 = \{$fracture, subluxation, of, lunate, wrist, dislocation$\}$, thus we suggest concept 3 is more specific than concept 1, that is, "Fracture subluxation of lunate" IS-A "Fracture dislocation of lunate" (see the red link in Fig. 3B). Here (3, 1) is a newly identified relation. Among the existing relations, our method also identifies (3, 2), since $L_2 = \{$fracture, subluxation, of, wrist, dislocation$\}$ is a proper subset of $L_3 = \{$fracture, subluxation, of, lunate, wrist, dislocation$\}$.

3.4. Evaluation

We focus on small non-lattice subgraphs (of size 4, 5, and 6) to evaluate the effectiveness of our approach to suggesting missing IS-A relations and revealing incorrect IS-A relations in SNOMED CT. The rationale for focusing

on non-lattice subgraphs of smaller size is twofold: one is that it is easier for experts to review these subgraphs, the other is that the errors found in small subgraphs are often also contained in larger subgraphs [15].

We selected a random sample of 200 non-lattice subgraphs from "Clinical finding" and "Procedure," the two largest sub-hierarchies of SNOMED CT. The 200 subgraphs (223 IS-A instances) were split into two sample sets (125 subgraphs each), with a shared common subset of 50 subgraphs (56 IS-A instances). Two clinical experts (authors OB and JS, two physicians familiar with SNOMED CT, who were not involved in the development of the method) independently reviewed the two sample sets with suggested missing IS-A relations. For the commonly evaluated 50 subgraphs, differences in evaluation results were reconciled by discussion.

For the suggestions that were found incorrect by a clinical expert, we further reviewed the existing IS-A relations in the original non-lattice subgraphs that were used to generate the suggestions. This is because the identification of a missing IS-A relation can be due to the presence of an erroneous IS-A relation in the subgraph. If the clinical expert also disagrees with the existing IS-A relation, then this relation is identified as an incorrect IS-A relation in SNOMED CT (source error). For instance, from the non-lattice subgraph in Fig. 3A, we also suggest that concept 6 is more specific than concept 3, that is, "Fracture subluxation of perilunate joint" IS-A "Fracture subluxation of lunate." However, this invalid suggestion is derived in part from the existing relation (4, 1): "Fracture dislocation of perilunate joint" IS-A "Fracture dislocation of lunate." Since perilunate dislocation is distinct from lunate dislocation, the existing relation is invalid in the first place. Therefore, al-

though the missing IS-A relation we identified is a false positive, our analysis of the non-lattice subgraph in Fig. 3A reveals an incorrect IS-A relation (4, 1).

4. Results

4.1. Non-lattice subgraphs

A total of 195,121 non-lattice subgraphs were extracted. Among these, our approach based on enriched lexical attributes of the non-lattice subgraphs identified 14,380 subgraphs containing missing IS-A relations. Table 1 shows the distribution of such non-lattice subgraphs by the SNOMED CT sub-hierarchies. There were a total of 1,474 small non-lattice subgraphs (size of 4, 5, and 6). The distribution of such small non-lattice subgraphs within each sub-hierarchy is also given in Table 1. The "Clinical finding" sub-hierarchy accounted for the largest number of non-lattice subgraphs (6,612 any-size and 692 small-size).

It is worth noting that a non-lattice subgraph may contain more than one missing IS-A relations. For instance, the non-lattice subgraph shown in Fig. 4A contains two missing IS-A relations: "Congenital sacral meningocele" IS-A "Congenital meningocele," and "Cervical spinal hydromeningocele" IS-A "Congenital meningocele" (see the red links). Therefore, the number of missing IS-A relations suggested was larger than the number of non-lattice subgraphs. Overall, the 14,380 non-lattice subgraphs contain a total of 41,357 missing IS-A relations. The 1,474 small non-lattice subgraphs contain a total of 1,629 missing IS-A relations.

Table 1: Numbers of non-lattice subgraphs and small non-lattice subgraphs (of size 4, 5, and 6) that suggested missing IS-A relations, according to the SNOMED CT sub-hierarchies.

Sub-hierarchy	No. of non-lattice subgraphs	No. of small non-lattice subgraphs
Clinical finding	6,612	692
Body structure	3,634	245
Procedure	3,004	304
Substance	401	56
Pharmaceutical / biologic product	264	60
Physical object	216	53
Social context	66	14
Specimen	53	18
Qualifier value	41	10
Organism	46	7
Observable entity	30	8
Situation with explicit context	10	6
Event	1	0
Record artifact	1	0
Physical force	1	1
Total	14,380	1,474

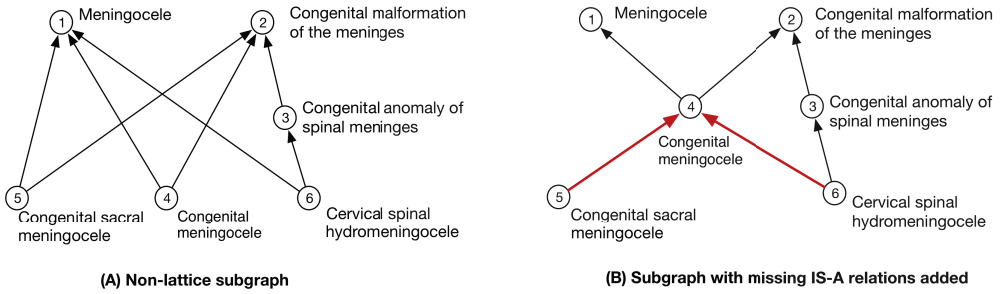

Figure 4: A non-lattice subgraph of size 6 in the "Clinical finding" sub-hierarchy, as well as the resulting subgraph after adding two missing IS-A relations (red links): "Congenital sacral meningocele" IS-A "Congenital meningocele," and "Cervical spinal hydromeningocele" IS-A "Congenital meningocele."

4.2. Evaluation

Of the 200 subgraphs randomly selected from 937 small non-lattice subgraphs in the two largest sub-hierarchies, 139 were in the "Clinical finding" sub-hierarchy, and 61 in the "Procedure" sub-hierarchy. Of the 200 sub-

graphs, 32 were of size 4, 86 of size 5, and 82 of size 6.

The 200 subgraphs contain a total of 223 missing IS-A relations. Upon review, two clinical experts concluded that 185 (82.96%) missing IS-A relations are valid. For the invalid suggestions (false positives for suggested missing IS-A relations), the experts further examined the existing IS-A relations in SNOMED CT which were used for generating the suggestions, and identified 22 existing IS-A relations to be incorrect (confirmed source errors), beyond those that were evaluated.

Table 2 summarizes the evaluation results by the two domain experts.

Table 2: The precisions of our approach in terms of evaluators.

Evaluator	No. of subgraphs	No. of suggestions	True Positive	False Positive	Precision
1	125	139	115	24	82.73%
2	125	138	115	23	83.33%

A total of 56 missing IS-A relations within the 50 non-lattice subgraphs were evaluated by both evaluators. The two evaluators initially had agreement on 46 out of 56 (82.14%) of the cases. After reconciliation, all the discrepancies were resolved except 1 case (no agreement was reached for this case). In addition, 3 cases were flagged as potentially contentious although agreement was reached. The invalid suggestions further revealed 4 incorrect IS-A relations in SNOMED CT as the source of error.

Table 3 lists 15 examples of valid missing IS-A relations in SNOMED CT verified by clinical experts, including "Renal angle tenderness" IS-A "Renal pain" suggested from the non-lattice subgraph shown in Fig. 5, and "Transient neonatal hyperglycemia" IS-A "Acute hyperglycemia" suggested from the non-lattice subgraph shown in Fig. 6.

Table 3: Examples of missing IS-A relations in SNOMED CT identified by our approach.

Child	Parent
Renal angle tenderness (finding)	Renal pain (finding)
Congenital alveolar hyperplasia of maxilla (disorder)	Congenital maxillary hyperplasia (disorder)
Revision of prosthesis of abdominal aorta (procedure)	Revision of abdominal vascular prosthesis (procedure)
Revision of prosthesis of bifurcation of aorta (procedure)	Revision of prosthesis of abdominal aorta (procedure)
Longitudinal deficiency of femur (disorder)	Deformity of femur (disorder)
Suture of periosteum of vertebra (procedure)	Operation on vertebra (procedure)
Transient neonatal hyperglycemia (disorder)	Acute hyperglycemia (disorder)
Superficial traumatic blister of lower limb (disorder)	Superficial injury of lower limb (disorder)
Acute lymphangitis of finger (disorder)	Acute lymphangitis of hand (disorder)
Syphilitic parkinsonism (disorder)	Late syphilitic encephalitis (disorder)
Angioplasty of external iliac artery (procedure)	Repair of iliac artery (procedure)
Burn of conjuctival sac (disorder)	Burn of conjunctiva (disorder)
Computed tomography of salivary gland with contrast (procedure)	Computed tomography sialogram (procedure)
Neoplasm of peripheral nerves of hip (disorder)	Neoplasm of peripheral nerves of lower limb (disorder)
Esophageal atresia with tracheoesophageal fistula (disorder)	Congenital esophageal fistula (disorder)

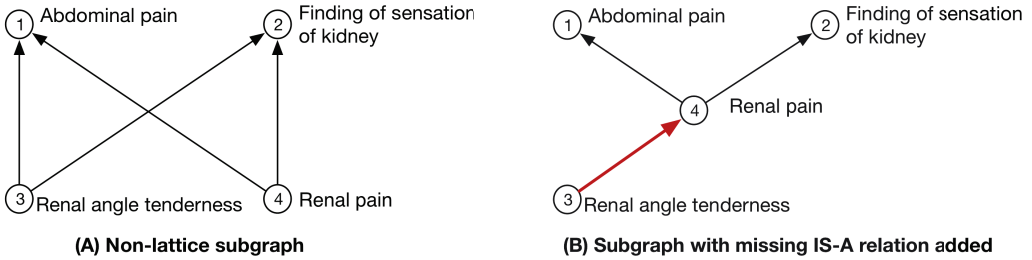

Figure 5: A non-lattice subgraph of size 4 and the resulted subgraph after adding a missing IS-A relations (red link): "Renal angle tenderness" IS-A "Renal pain."

Table 4 lists 4 examples of incorrect IS-A relations in SNOMED CT (source errors) verified by clinical experts. Fig. 7 shows the non-lattice subgraph exhibiting the incorrect IS-A relation: "Congenital cyst of posterior segment of eye" IS-A "Disorder of anterior segment of eye" (see the red cross), which leads to the incorrect suggestion "Congenital cyst of posterior segment of eye" IS-A "Congenital anomaly of anterior segment of eye" using our approach.

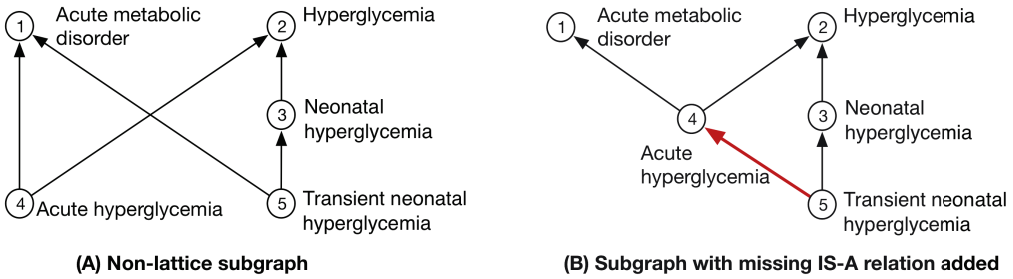

Figure 6: A non-lattice subgraph of size 5 and the resulted subgraph after adding a missing IS-A relations (red link): "Transient neonatal hyperglycemia" IS-A "Acute hyperglycemia."

Table 4: Examples of incorrect IS-A relations (source errors) in SNOMED CT identified by our approach.

Child	Parent
Congenital cyst of posterior segment of eye (disorder)	Disorder of anterior segment of eye (disorder)
Mobile cecum (disorder)	Congenital malrotation of intestine (disorder)
Division of mitral valve chordae tendineae (procedure)	Commissurotomy of heart valve (procedure)
Stripping of cranial suture (procedure)	Operation on bone (procedure)

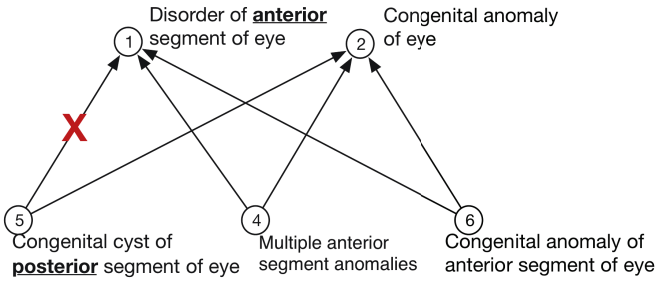

Figure 7: Non-lattice subgraph exhibiting an **incorrect** IS-A relation (source error) in SNOMED CT (red cross): "Congenital cyst of posterior segment of eye" IS-A "Disorder of anterior segment of eye."

We will submit the verified suggestions to SNOMED International for review as part of its ongoing internal quality improvement activities.

5. Discussion

5.1. False positives and intricate cases

Even though our hybrid approach was aimed at identifying missing hierarchical relations with high precision, false positives could not be completely eliminated. In some cases, the concepts contain implicit knowledge and have misleading surface forms. For example, our method suggests that "Infection of toe web" IS-A "Infection of toe," which is not correct. Toe web refers to the interdigital space of foot, and is not a part of toe. Another example is that our method recommends "Humerus head juvenile osteochondritis" as a subclass of "Humerus juvenile osteochondritis," based on the observation that both concepts denote a form of humerus osteochondritis, and that one of them is further specified as juvenile. However, the juvenile form affects the humerus head, while the more general form affects the epicondyle, at the other extremity of the bone. In both cases, lexical similarity between the two terms is responsible for the false positives.

Similarly, our method suggests an IS-A relation between the disorder concepts "Budd-Chiari syndrome" and "Hepatic vein thrombosis." However, this relation does not always stand, since Budd-Chiari syndrome can also be due to compression (not thrombosis) of the hepatic veins. Here, the presence of an erroneous IS-A relation in SNOMED CT between "Budd-Chiari syndrome" and "Thrombosis of vein of trunk" contributed to the wrongful suggestion.

Sometimes part-whole relationships may give opposite conclusions in different contexts. For example, one of the false positives is the suggestion that "Does use the elements of language" IS-A "Does use language." Since using

elements of a language is not the same as the ability to use the language, this IS-A relation is incorrect. However, if a subject has "Difficulty using the elements of language," then the subject must have "Difficulty using [the] language." This would result in a true positive for our method.

Of note, during the evaluation, we observed a few cases for which it was difficult to determine whether the suggested missing IS-A relation was correct or not. However, in the vast majority of cases, the experts had no difficulty agreeing on whether the suggested IS-A relations were true or false positives.

5.2. Precision and recall

This paper focused on the evaluation of precision. Unlike traditional information retrieval tasks but similar to finding software bugs, standard reference data sets for the evaluation of "recall" for ontology quality assurance methods are virtually impossible to construct, except in very restricted settings.

Despite the unavailability of ground truth on ontological errors, one can use cumulative SNOMED CT changes as a surrogate reference set for evaluating recall. In [19], it was demonstrated that small-sized (≤ 15) non-lattice fragments captured more than 60% of SNOMED CT's relational changes. Coupled with the precision demonstrated in this paper using lexical attributes, our approach strikes a balance between precision and "recall," while also maintaining consistency with SNOMED CT's logically inferred statements.

5.3. Enhanced coverage of non-lattice subgraphs

This work builds on our previous work reported in [15], in that both leverage non-lattice graph substructures. The distinction is the substantially larger number of non-lattice subgraphs that were covered by the approach presented in this paper. Applying the approach reported in [15], to the same SNOMED CT version (September 2017 US edition), only 2,124 of 14,380 non-lattice subgraphs identified in this work can be detected using the previous approach. This represents 85.23% increase in coverage. Among non-lattice subgraphs of size 4, 5 and 6, 77.61% were newly identified (1,144 out of 1,474). For example, none of the missing IS-A relations in Table 3 or incorrect IS-A relations in Table 4 would be detectable using the approach in [15]. However, the approach in [15] addressed missing concepts in addition to missing relations. Therefore, our recommendation for ontology quality assurance would be to use both approaches.

5.4. Limitations

Despite the substantially increased coverage of non-lattice subgraphs, we are only able to cover 7.4% of all non-lattice subgraphs. Identifying new lexical patterns among the non-lattice subgraphs remains an active topic for research.

Automatic change suggestion for identified errors is a unique feature of our approach. However, the change suggestions pertain to the inferred hierarchy. Since this hierarchy is inferred by a description logic classifier based on the logical definitions of concepts, the only meaningful remediation would be to find the root cause and modify the logical definitions so that the appropriate

hierarchy can be inferred. Identifying erroneous and missing axioms in logical definitions will be the object of future work.

6. Conclusions

This paper introduced a novel approach to predicting missing IS-A relations in SNOMED CT by combining non-lattice subgraphs and enriched lexical attributes of concepts. Our result of a 82.96% precision on the predicted missing relations demonstrates that leveraging enriched lexical attributes within non-lattice subgraphs is an effective approach for auditing SNOMED CT. Since a hierarchical substructure and lexical attributes of concepts are present in almost all biomedical ontologies, our method is generally applicable for ontology quality assurance purposes.

Acknowledgements

This work was supported by the National Science Foundation through grants IIS-1657306 and ACI-1626364, and the National Institutes of Health (NIH) National Center for Advancing Translational Sciences through grant UL1TR001998. This work was also supported by the Intramural Research Program of the NIH, National Library of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] O. Bodenreider, Biomedical ontologies in action: role in knowledge management, data integration and decision support,Yearbook of medical informatics (2008) 67-79.

[2] SNOMED CT. http://www.snomed.org/snomed-ct, 2017 (accessed 26 September 2017).

[3] Health Information Technology for Economic and Clinical Health (HITECH) Act. http://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf, 2009 (accessed 26 September 2017).

[4] X. Zhu, J.W. Fan, D.M. Baorto, C. Weng, J.J. Cimino, A review of auditing methods applied to the content of controlled biomedical terminologies, J. Biomed. Inform. 42(3) (2009) 413-425.

[5] Y. Wang, M. Halper, H. Min, Y. Perl, Y. Chen, K.A. Spackman, Structural methodologies for auditing SNOMED, J. Biomed. Inform. 40(5) (2007) 561-581.

[6] Y. Wang, M. Halper, D. Wei, Y. Perl, J. Geller, Abstraction of complex concepts with a refined partial-area taxonomy of SNOMED. J. Biomed. Inform. 45(1) (2012) 15-29.

[7] Y. Wang, M. Halper, D. Wei, H. Gu, Y. Perl, J. Xu, G. Elhanan, Y. Chen, K.A. Spackman, J.T. Case, G. Hripcsak, Auditing complex concepts of SNOMED using a refined hierarchical abstraction network, J. Biomed. Inform. 45(1) (2012) 1-14.

[8] C. Ochs, J. Geller, Y. Perl, Y. Chen, J. Xu, H. Min, J.T. Case, Z. Wei, Scalable quality assurance for large SNOMED CT hierarchies using subject-based subtaxonomies, J. Am. Med. Inform. Assoc. 22(3) (2015) 507-518.

[9] C. Ochs, J. Geller, Y. Perl, Y. Chen, A. Agrawal, J.T. Case, G. Hripcsak, A tribal abstraction network for SNOMED CT target hierarchies without attribute relationships, J. Am. Med. Inform. Assoc. 22(3) (2014) 628-639.

[10] A. Agrawal, G. Elhanan, Contrasting lexical similarity and formal definitions in SNOMED CT: Consistency and implications, J. Biomed. Inform. 47 (2014):192-8.

[11] A. Agrawal, Y. Perl, C. Ochs, G. Elhanan, Algorithmic detection of inconsistent modeling among SNOMED CT concepts by combining lexical and structural indicators, In IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2015) 476-483.

[12] G.Q. Zhang, O. Bodenreider, Large-scale, exhaustive lattice-based structural auditing of SNOMED CT, In AMIA Annual Symposium Proceedings (2010) 922-926.

[13] G.Q. Zhang, W. Zhu, M. Sun, S. Tao, O. Bodenreider, L. Cui, October. MaPLE: a MapReduce pipeline for lattice-based evaluation and its application to SNOMED CT, In IEEE International Conference on Big Data (2014) 754-759.

[14] L. Cui, S. Tao, G.Q. Zhang, Biomedical ontology quality assurance using a big data approach, ACM T. Knowl. Discov. D. 10(4) (2016) 41.

[15] L. Cui, W. Zhu, S. Tao, J.T. Case, O. Bodenreider, G.Q. Zhang, Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT, J. Am. Med. Inform. Assoc. 24(4) (2017) 788-798.

[16] D.A. Cruse, Lexical semantics, Cambridge University Press, Vancouver, 1986.

[17] G.A. Miller. WordNet: a lexical database for English, Communications of the ACM 38(11) (1995) 39-41.

[18] WordNet. https://wordnet.princeton.edu/, 2017 (accessed 26 September 2017).

[19] G.Q. Zhang, Y. Huang, L. Cui, Can SNOMED CT Changes Be Used as a Surrogate Standard for Evaluating the Performance of Its Auditing Methods?, In AMIA Annual Symposium Proceedings (2017) 1886-1895.

[20] Logical Model of SNOMED CT Components - Relationships and Concept Definitions. `https://confluence.ihtsdotools.org/display/DOCRELFMT/1.3.+Relationships+and+Concept+Definitions`, 2017 (accessed 29 November 2017).