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

Licong Cui\textsuperscript{a,b,*}, Olivier Bodenreider\textsuperscript{d}, Jay Shi\textsuperscript{c}, Guo-Qiang Zhang\textsuperscript{a,b,c}

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

Abstract

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. Our approach involves 3 stages. In stage 1, all non-lattice subgraphs of SNOMED CT’s IS-A relations are extracted. In stage 2, lexical attributes of fully specified concept names in such non-lattice subgraphs are extracted. The attribute set of each concept is enlarged (enriched) with those of its parent concepts that are present in their local non-lattice subgraphs. In stage 3, subset inclusion relations between the attribute sets of each pair of concepts in each non-lattice subgraph are compared. For concept pairs within each non-lattice subgraph, if a subset relation between respective attribute sets is identified but an IS-A relation is not present, then a missing IS-A relation is reported. The September 2017 release of SNOMED CT (US edition) was used for our evaluation. A total of 14,380 non-lattice subgraphs were extracted, which suggested a total of 41,357 missing IS-A relations. For evaluation, 200 non-lattice subgraphs were randomly selected from 996 smaller subgraphs (of sizes 4, 5, or 6) within the “Clinical Finding” and “Procedure” sub-hierarchies. Two domain experts confirmed 185 (among 223) missing IS-A relations, a precision of 82.96%.

\textsuperscript{*}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.

Preprint submitted to Journal of Biomedical Informatics October 15, 2017
Our result demonstrates that enriching lexical attributes within non-lattice subgraphs is an effective approach for auditing SNOMED CT.

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

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 increase the adoption and ensure meaningful use of EHRs. The use of SNOMED CT in EHRs brings multitudes of benefits in the caring of patients and enabling cost-effective delivery of care to the population.

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 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 exist, with their root cause sometimes 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 with limited precision, but do not automatically suggest correction measures to systematically fix
the issues. Arduous manual review by domain experts or ontology auditors are 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 for the true positive cases, and uncovers related incorrect IS-A relations in the subgraphs for the false positive cases. 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. The lexical attribute set of each concept is enlarged (enriched) with those of its ancestor concepts that are present within each 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. 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

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 [19, 20, 21]. For example, in Fig. 1, the concept pair (1, 2) is a non-lattice pair, since they
share two maximal shared common descendants 5 and 6. In our previous work [18, 19, 20], we have developed various computational approaches to systematically extract all the non-lattice pairs in SNOMED CT for further auditing.

\[ \text{Figure 1: An example of a non-lattice subgraph of size 6.} \]

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 of such non-lattice pairs would be redundant. Therefore, a notion of non-lattice subgraph is further introduced to avoid redundant analysis [21]. 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 \( \{5, 6\} \) obtains \( \{1, 2, 3\} \), then aggregating all the concepts and edges between \( \{1, 2, 3\} \) and \( \{5, 6\} \) obtains 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. From the point of view of the hierarchical structure, lattice is a desirable property for a well-formed ontology or terminology [18].

2.3. Related work on auditing techniques for SNOMED CT

Auditing or quality assurance of SNOMED CT has been an active research area given its importance. Zhu et al. performed a literature survey on early auditing methods applied to biomedical terminologies [4], including SNOMED CT. Jiang and Chute utilized formal concept analysis to audit the semantic completeness of SNOMED CT to identify missing concepts [5]. Rector and Iannone examined the use of common qualifiers in SNOMED
CT definitions by leveraging both lexical and semantic techniques [6]. Wang et al. proposed structural approaches based on abstraction networks to detect erroneous concepts in SNOMED [7, 8, 9]. Ochs et al. investigated two new types of abstraction networks (subject-based and tribal-based methods) to reveal erroneous concepts in SNOMED CT [10, 11]. Mortensen et al. performed a crowdsourcing study to identify errors in SNOMED CT [12]. Agrawal et al. used a combination of lexical and structural indicators to identify inconsistency issues in modeling SNOMED CT concepts [13]. He et al. used a comparative, structural method to enrich SNOMED CT concepts using Unified Medical Language System [14]. Bodenreider introduced a method to identify missing hierarchical relations in SNOMED CT from logical definitions based on the lexical features of concept names using a Description Logics reasoner [15]. Ceusters et al. [16] and Tao et al. [17] analyzed evolutional changes of SNOMED CT to uncover possible quality issues. Zhang and Bodenreider proposed a lattice-based structural approach to exhaustively extracting non-lattice fragments to audit SNOMED CT [18]. Cui et al. introduced a hybrid structural-lexical approach based on the lexical patterns of concept names in non-lattice subgraphs, so as to automatically suggest missing hierarchical relations and concepts in SNOMED CT [21].

2.4. Specific contribution

Analyzing non-lattice subgraphs has shown potential to uncover quality issues in SNOMED CT. In previous work [21], we explored a structural-lexical approach to mine lexical patterns in non-lattice subgraphs in SNOMED CT to identify missing IS-A relations and concepts. However the predefined lexical patterns covered only 4% of non-lattice subgraphs in SNOMED CT.

In this work, we introduce a novel structural-lexical approach which leverages more existing knowledge in SNOMED CT by enriching 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. Most importantly, the structural-lexical approach introduced in this work is capable of identifying previously undiscovered missing IS-A relations and incorrect IS-A relations, as well as reaching a higher coverage of non-lattice subgraphs (7.4% coverage rate).
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.

\begin{algorithm}
\textbf{Input}: A non-lattice subgraph }G\text{ consisting of concepts and IS-A relations  \\
\textbf{Output}: Reclassified IS-A relations  \\
1 \textbf{if} the fully specified name of a concept in }G\text{ contains stop word(s) or antonyms \textbf{then}  \\
2 \hspace{1em} \text{stop here;}  \\
3 \textbf{else}  \\
4 \hspace{1em} \text{continue;}  \\
5 \textbf{end}  \\
6 \text{Compute the transitive closure of the IS-A relations in }G;  \\
7 \text{Derive term pairs based on the transitive closure and fully specified names of the concepts in }G;  \\
8 \textbf{for each concept }c\text{ in }G\text{ do}  \\
9 \hspace{1em} \text{Initialize a set }L_c\text{ of lexical attributes for }c\text{ using its fully specified name;}  \\
10 \hspace{1em} \text{Enrich }L_c\text{ by leveraging the lexical attributes of }c\text{’s ancestors;}  \\
11 \hspace{1em} \text{Enrich }L_c\text{ by the derived term pairs;}  \\
12 \textbf{end}  \\
13 \textbf{for each concept }c_1\text{ in }G\text{ do}  \\
14 \hspace{1em} \textbf{for each concept }c_2\text{ in }G\text{ do}  \\
15 \hspace{2em} \textbf{if }c_1 \neq c_2\text{ and }L_{c_1} \subseteq L_{c_2}\text{ then}  \\
16 \hspace{2em} \text{Suggest }c_2\text{ IS-A }c_1;  \\
17 \hspace{1em} \textbf{end}  \\
18 \textbf{end}  \\
19 \text{Reduce the resulted IS-A relations to direct IS-A relations;}  \\
\end{algorithm}

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

Algorithm 1 presents the pseudocode for identifying missing IS-A rela-
tions 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 next with illustrative examples provided.

3.1. Detection of stop words and antonyms

Given a non-lattice subgraph $G$, we detect if any concept in $G$ contains stop word(s) and antonyms, which are prone to generate wrong 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. We consider the followings as the stop words: “and,” “or,” “and/or,” “no,” “not,” “without,” “due,” “secondary,” “except,” “by,” “after,” “co-occurrence,” “bilateral,” “examination,” “able,” “amputation,” “removal,” “replacement,” “resection,” “excision.” For example, the concept “Excision of vessel of head and/or neck” contain two stop words.

For antonyms, we leverage a list of pairs of antonyms from WordNet [23, 24], including (“anterior,” “posterior”), (“chronic,” “acute”), (“open,” “closed”), (“positive,” “negative”), (“high,” “low”), (“benign,” “malignant”), (“right,” “left”), (“simple,” “compound”). For example, the concept “open reduction of closed radial shaft fracture procedure” contains antonyms “open” and “closed,” and the concept “acute on chronic endometritis disorder” contain antonyms “acute” and “chronic.”

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; the second source is the fully specified names of the concept’s ancestors within the subgraph; and the third source is a set of derived term pairs. The rationale for leveraging derived term pairs is to include additional lexical attributes that may not be covered by the ancestors’ lexical attributes.

To obtain the second source, we compute the transitive closure of the IS-A relations in $G$, denoted by $T = \{(d, a) | \text{concept } a \text{ is an ancestor of concept } d \text{ and } a \in G\}$. To obtain the third source, for each concept pair $(d, a)$ in $T$, assume $W_d$ 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 $W_a - (W_d \cap W_a) \neq \emptyset$, then we obtain a derived term pair $\left(W_d - (W_d \cap W_a), W_a - (W_d \cap W_a)\right)$.
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 $p_2$.

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.”

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 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):

- $L_1 = \{\text{superficial, injury}\}$,
- $L_2 = \{\text{injury, of, lower, extremity}\}$,
- $L_3 = \{\text{traumatic, blister, of, lower, limb, injury, extremity}\}$,
\[ L_4 = \{ \text{friction, blisters, of, the, skin, superficial, injury} \}, \]
\[ L_5 = \{ \text{superficial, injury, of, lower, limb, extremity} \}, \]
\[ L_6 = \{ \text{superficial, traumatic, blister, of, lower, limb, injury, extremity, friction, blisters, the, skin} \}. \]

Leveraging the derived term pairs result in the same sets of lexical attributes (i.e., no additional lexical attributes are added for the concepts). The derived term pairs are ("traumatic blister of lower limb," "injury"), ("superficial limb," "extremity"), ("traumatic blister limb," "injury extremity"), ("superficial traumatic blister limb," "injury extremity"), and ("superficial traumatic lower limb," "friction the skin"). For example, the term pair ("traumatic blister of lower limb," "injury") is obtained by the concept pair (6, 1) in the transitive closure, that is, ("Superficial traumatic blister of lower limb," "Superficial injury"). Since the ancestors’ lexical attributes already included the lexical attributes obtained using the derived term pairs, there is no new lexical attributes added.

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:
\[ L_1 = \{ \text{fracture, dislocation, of, lunate} \}, \]
\[ L_2 = \{ \text{fracture, subluxation, of, wrist} \}, \]
\[ L_3 = \{ \text{fracture, subluxation, of, lunate} \}, \]
\[ L_4 = \{ \text{fracture, dislocation, of, perilunate, joint} \}, \]
\[ L_5 = \{ \text{open, fracture, subluxation, lunate} \}, \]
\[ L_6 = \{ \text{fracture, subluxation, of, perilunate, joint} \}. \]

Leveraging the ancestors’ lexical attributes results in the following enriched sets (with newly added lexical attributes italicized):
$L_1 = \{\text{fracture, dislocation, of, lunate}\}$,
$L_2 = \{\text{fracture, subluxation, of, wrist}\}$,
$L_3 = \{\text{fracture, subluxation, of, lunate, wrist}\}$,
$L_4 = \{\text{fracture, dislocation, of, perilunate, joint, lunate}\}$,
$L_5 = \{\text{open, fracture, subluxation, lunate, dislocation, of, wrist}\}$,
$L_6 = \{\text{fracture, subluxation, of, perilunate, joint, dislocation, lunate, wrist}\}$.

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

$L_1 = \{\text{fracture, dislocation, of, lunate}\}$,
$L_2 = \{\text{fracture, subluxation, of, wrist, dislocation}\}$,
$L_3 = \{\text{fracture, subluxation, of, lunate, wrist, dislocation}\}$,
$L_4 = \{\text{fracture, dislocation, of, perilunate, joint, lunate}\}$,
$L_5 = \{\text{open, fracture, subluxation, lunate, dislocation, of, wrist}\}$,
$L_6 = \{\text{fracture, subluxation, of, perilunate, joint, dislocation, lunate, wrist}\}$.

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. 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.

For example, for the concepts numbered 5 and 6 in Fig. 2A, $L_5 = \{\text{superficial, injury, of, lower, limb, extremity}\}$ is a proper subset of $L_6 = \{\text{superficial, traumatic, blister, of, lower, limb, injury, extremity, friction, blisters, the, skin}\}$, 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, while all the others already exist in the original non-lattice subgraph.
For the concepts 1 and 3 in Fig. 3A, $L_1 = \{\text{fracture, dislocation, of, lunate}\}$ is a proper subset of $L_3 = \{\text{fracture, subluxation, of, lunate, wrist, dislocation}\}$, thus we suggest concept No. 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).

3.4. Evaluation

We focus on small non-lattice subgraphs (in size of 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 it is easier for experts to review these subgraphs, the other is the errors found in small subgraphs are often also contained in larger subgraphs [21].

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 a 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 disagreed 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 (source error) in SNOMED CT. For instance, the non-lattice subgraph in Fig. 3A also suggests 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 also invalid. Therefore, the non-lattice subgraph in Fig. 3A further 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 which 14,380 suggested missing IS-A relations using our approach based on enriched lexical attributes of the non-lattice subgraphs. 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).

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.

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

It is worth noting that a non-lattice subgraph may suggest one or more missing IS-A relations. For instance, the non-lattice subgraph shown in Fig. 4A suggested 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. The 14,380 non-lattice subgraphs suggested a total of 41,357 missing IS-A relations. The 1,474 small non-lattice subgraphs suggested a total of 1,629 missing IS-A relations.
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 subgraphs, 32 were of size 4, 86 of size 5, and 82 of size 6.

The 200 subgraphs suggested a total of 223 missing IS-A relations. Two clinical experts verified 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>
<thead>
<tr>
<th>Evaluator</th>
<th>No. of subgraphs</th>
<th>No. of suggestions</th>
<th>True Positive</th>
<th>False Positive</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>125</td>
<td>139</td>
<td>115</td>
<td>24</td>
<td>82.73%</td>
</tr>
<tr>
<td>2</td>
<td>125</td>
<td>138</td>
<td>115</td>
<td>23</td>
<td>83.33%</td>
</tr>
</tbody>
</table>

56 identified 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 source of error.

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

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

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

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 the incorrect IS-A relations (source errors) in SNOMED CT verified by clinical experts. Fig. 7 shows the non-lattice
Figure 6: A non-lattice subgraph of size 5 and the resulted subgraph after adding a missing IS-A relation (red link): “Transient neonatal hyperglycemia” IS-A “Acute hyperglycemia.”

subgraph revealing the incorrect IS-A relation: “Congenital cyst of posterior segment of eye” IS-A “Disorder of anterior segment of eye” (see the red cross).

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

<table>
<thead>
<tr>
<th>Child</th>
<th>Parent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Congenital cyst of posterior segment of eye (disorder)</td>
<td>Disorder of anterior segment of eye (disorder)</td>
</tr>
<tr>
<td>Mobile cecum (disorder)</td>
<td>Congenital malrotation of intestine (disorder)</td>
</tr>
<tr>
<td>Division of mitral valve chordae tendineae (procedure)</td>
<td>Commissurotomy of heart valve (procedure)</td>
</tr>
<tr>
<td>Stripping of cranial suture (procedure)</td>
<td>Operation on bone (procedure)</td>
</tr>
</tbody>
</table>

Figure 7: Non-lattice subgraph revealing 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
5. Discussion

5.1. False positives

Even though our use of non-lattice subgraphs was aimed at reducing false positives, they cannot be completely eliminated. For example, our method led to the suggestion of “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.

Sometimes part-whole relationship 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.

Another collection of false positives have to do with syndromes that may have many manifestations, such as “Ehlers-Danlos syndrome” and “Budd-Chiari syndrome.” In such cases the lexical-based attributes become less effective in identifying missing IS-A relations.

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 is virtually impossible to construct, unless 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 [25], 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. Distinction with related work

Our structural-lexical approach is distinct from Agrawal’s work [13] on combining lexical and structural indicators to detect inconsistent concept modeling in SNOMED CT. Agrawal’s method first identifies lexically similar concepts (i.e., concepts with the same length but names differing at a specific position) and then compares their concept modeling in attribute relationships to detect inconsistencies. Our approach first identifies non-lattice subgraphs and then utilize enriched lexical attributes of concepts in non-lattice subgraphs to suggest missing IS-A relations.

Our contribution builds on the work reported in [21], in that both leverage non-lattice graph substructures in order to address precision while providing automatic change suggestion. 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 [21], to the same SNOMED CT version (September 2017 US edition), only 2,124 out 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 sizes 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 [21]. However, the approach in [21] 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% with our method. Identifying de novo 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 are still in the realm of the inferred hierarchy. It remains highly desirable to develop methods that can trace errors all the way to SNOMED CT’s logical definitions. Our informal review of the logical definitions has revealed 2 main issues: (1) the presence of many primitive concepts, for which Description Logics classifiers cannot infer subclasses; (2) the presence of incomplete logical definitions among concepts. Finding the root cause of erroneous ramifications remains the “Holy Grail” for ontology quality assurance.
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 enriching lexical attributes within non-lattice subgraphs is an effective approach for auditing SNOMED CT. Since our approach is based on the hierarchical substructure and lexical attributes of concept classes which are existent in almost all biomedical ontologies, it 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


