Quality Assurance of NCI Thesaurus by Mining Structural-Lexical Patterns

Rashmie Abeysinghe[1], Michael A. Brooks, MD[2], Jeffery Talbert, PhD[3], Licong Cui, PhD[1,⋆]
[1]Department of Computer Science, University of Kentucky, Lexington, KY
[2]Departments of Radiology and Medicine, University of Kentucky, Lexington, KY
[3]Institute for Pharmaceutical Outcomes and Policy, University of Kentucky, Lexington, KY

Abstract

Quality assurance of biomedical terminologies such as the National Cancer Institute (NCI) Thesaurus is an essential part of the terminology management lifecycle. We investigate a structural-lexical approach based on non-lattice subgraphs to automatically identify missing hierarchical relations and missing concepts in the NCI Thesaurus. We mine six structural-lexical patterns exhibiting in non-lattice subgraphs: containment, union, intersection, union-intersection, inference-contradiction, and inference union. Each pattern indicates a potential specific type of error and suggests a potential type of remediation. We found 809 non-lattice subgraphs with these patterns in the NCI Thesaurus (version 16.12d). Domain experts evaluated a random sample of 50 small non-lattice subgraphs, of which 33 were confirmed to contain errors and make correct suggestions ($33/50 = 66\%$). Of the 25 evaluated subgraphs revealing multiple patterns, 22 were verified correct ($22/25 = 88\%$). This shows the effectiveness of our structural-lexical-pattern-based approach in detecting errors and suggesting remediations in the NCI Thesaurus.

Introduction

Biomedical terminologies and ontologies serve as a knowledge source for many biomedical applications, including natural language processing applications and decision support systems[1]. Quality issues in terminologies, if not addressed, can affect the quality of all downstream information systems relying on them as a knowledge source[2]. Terminology Quality Assurance (TQA) strives to estimate and enhance the quality of terminologies by improving consistency, coverage and completeness, non-redundancy and clarity[3]. However, it is labor-intensive and time-consuming to discover errors or inconsistencies by manual review of large biomedical terminologies. Therefore, automating TQA has been an active area of research[4].

Developed and maintained by the National Cancer Institute (NCI), the NCI Thesaurus (NCIt) is a reference terminology used in an increasing number of NCI and other systems[5,6]. It contains over 100,000 concepts related to cancer research, including cancer-related diseases, findings and abnormalities; anatomy; agents, drugs and chemicals; genes and so on[7]. Given the sheer size of the NCIt, it is unavoidable that errors may be introduced in its development, update, and maintenance phases. Moreover, it is impractical for human experts to manually review the terminology to discover quality issues such as missing concepts, concept redundancies, and missing hierarchical relations. Automatic approaches to quality assurance of the NCIt are highly desirable to provide human experts with error candidates for review and verification.

In this paper, we develop an automatic approach to detecting missing hierarchical IS-A relations and missing concepts in the NCIt based on non-lattice subgraphs, which were initially introduced for auditing SNOMED CT[8]. We investigate six structural-lexical patterns exhibiting in non-lattice subgraphs in the NCIt, with each pattern indicating a certain type of potential error and suggesting a potential correction. Human experts reviewed a random subset of non-lattice subgraphs automatically detected using this approach to confirm the uncovered errors and suggested corrections.

1 Background

1.1 NCI Thesaurus (NCIt)

The NCIt is a biomedical terminology for cancer research, covering vocabulary for clinical care, translational and basic research, and public information and administrative activities[5,6]. It was first published in 2000 with the intention to facilitate data sharing and interoperability by different NCI components. Concepts in NCIt are hierarchically organized in 19 domains, including *Abnormal Cell*; *Anatomic Structure, System, or Substance*; *Biological Process*; *Disease,*

⋆Corresponding author: licong.cui@uky.edu

Disorder or Finding; *Drug, Food, Chemical or Biomedical Material*, *Gene*, *Gene Product*, *Molecular Abnormality*, and *Organism*. The version 16.12d of NCIt contains over 118,000 concepts.

The NCIt was built using Ontylog, a description logic explicitly for building large complex terminologies[7]. It is published in several formats including Ontylog XML, Web Ontology Language (OWL), and flat files. The NCIt also has defined and inferred versions. The defined version contains the assertions about each concept by the terminology editors. The inferred version includes additional assertions and classifications inferred by DL classifiers. In this paper, we used the inferred version of the NCIt to perform our pattern-based error detection and correction.

1.2 Quality Assurance of Biomedical Terminologies

Researchers have proposed various approaches[9] to auditing biomedical terminologies, such as the NCIt and SNOMED CT. Min et al.[10] proposed the abstraction networks (AbNs) approach to audit NCIt based on area taxonomies and partial-area taxonomies, where area taxonomies are groups of concepts that have exactly same roles, and partial-area taxonomies are further divisions of areas so that they are structurally uniform and singly-rooted. Ochs et al.[11,12] introduced subject-based AbN methods that summarize a subhierarchy rooted at a subject concept within a large hierarchy, and tribal-based AbN methods that are based on a subhierarchy rooted at a child of a hierarchy root to audit SNOMED CT. Verspoor et al.[13] introduced an automated method for identifying univocality violations in Gene Ontology. Zhang et al.[14,15] proposed a lattice-based structural auditing method to exhaustively detect non-lattice pairs in SNOMED CT. Cui et al.[3] presented a big data approach to perform lattice-based terminology quality assurance on SNOMED CT. Agrawal et al.[16] used a combination of positional similarity sets and structural indicators to identify modeling inconsistencies in SNOMED CT. Bodenreider[17] introduced a method to identify missing hierarchical relations in SNOMED CT from logical definitions based on the lexical features of concept names. Ceusters et al.[18] assessed the conformity of NCIt to widely accepted principles in terminology construction and ontology building. Mougin et al.[19] presented a semantic web technology method for quality assurance of NCIt. Zhe et al.[20,21] introduced a topological pattern-based method to recommend new concepts to include to NCIt.

More recently, Cui et al.[8] have introduced a hybrid structural-lexical method based on non-lattice subgraphs for scalable and systematic discovery of missing hierarchical relations and concepts in SNOMED CT. Four lexical patterns in non-lattice subgraphs were proposed for error detection and correction in SNOMED CT. In this paper, we apply these four patterns to NCIt and introduce two new patterns to identify errors and suggest corrections. To provide better readability, we review the definitions of the four patterns proposed for SNOMED CT in the Methods section and illustrate the patterns with examples in the NCIt.

2 Methods

Our approach leverages both structural and lexical information in the NCIt to systematically detect potential errors and automatically suggest remediations. Firstly, we identify all non-lattice subgraphs in NCIt. Secondly, we mine structural and lexical patterns in the non-lattice subgraphs, where each pattern indicates a potential missing hierarchical relation or missing concept in the NCIt. Finally, human domain experts evaluate a randomly selected sample of the potential errors detected, as well as the proposed remediation. We used the 16.12d version of the NCIt in this work.

2.1 Detecting Non-lattice Subgraphs

Non-lattice pairs. From a structural point of view, lattice is a desirable property for a well-formed terminology[14,15]. A terminology is a lattice if any two concepts in the terminology have a unique maximal shared descendant, as well as a unique minimal shared ancestor. A pair of concepts is known as a non-lattice pair, if the two concepts have more than one maximal shared descendant (alternatively minimal shared ancestor). A non-lattice pair generates a graph fragment with the nodes (or concepts) between the concept pair and the maximal shared descendants. There could be multiple non-lattice pairs which possess the same maximal shared descendants. In this case, it is not economical to examine each of these separately. If non-lattice pairs possessing the same maximal shared descendants are added together, this is also not economical since there might be concepts with ancestor-descendant relationship, which cause redundant analysis. Therefore the notion of non-lattice subgraphs has been introduced to facilitate effective analysis[8].

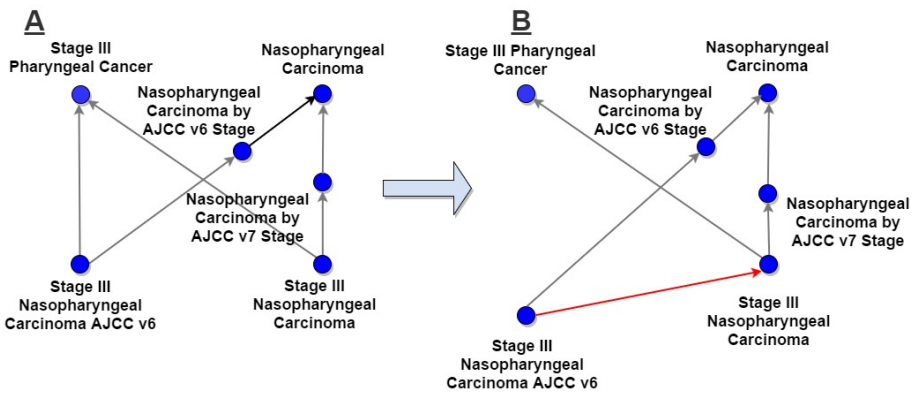

Figure 1: (**A**) An example of a non-lattice subgraph in the NCIt. (**B**) The suggested remediation of the non-lattice subgraph in (A): *Stage III Nasopharyngeal Carcinoma AJCC v6* is a subclass of *Stage III Nasopharyngeal Carcinoma*.

Non-lattice subgraphs. A non-lattice subgraph[8] can be obtained by a given non-lattice pair $p = (c_1, c_2)$ and its maximal common descendants $mcd(p)$ by first reversely computing the minimal common ancestors of the maximal common descendants, $mca(mcd(p))$; then accumulating the concepts and the edges between (including) any concept in $mca(mcd(p))$ and any concept in $mcd(p)$. The reverse computation obtains the minimal concepts sharing the same maximal common descendants to avoid redundant analysis. The minimal concepts $mca(mcd(p))$ are called the upper bounds of the non-lattice subgraph, and $mcd(p)$ are called the lower bounds. The size of a non-lattice subgraph is the number of concepts it contains. For instance, Figure 1A shows a non-lattice subgraph of size 6 in the NCIt, where *Stage III Pharyngeal Cancer* and *Nasopharyngeal Carcinoma* are the concepts in the upper bounds, and *Stage III Nasopharyngeal Carcinoma AJCC v6* and *Stage III Nasopharyngeal Carcinoma* are the concepts in the lower bounds.

In this work, we first parse the NCIt distribution file "*ThesaurusInferred.owl*" to extract all the concepts and their labels, as well as hierarchical IS-A relations (i.e., Child-Parent relations). Then we leverage the computational pipeline implemented in previous work[8] to exhaustively detect non-lattice subgraphs in the NCIt. Each resulting non-lattice subgraph consists of five components: concepts in the lower bounds, concepts in the upper bounds, concepts in the non-lattice subgraph, IS-A relationships in the non-lattice subgraph, and the size of the non-lattice subgraph.

2.2 Mining Structural and Lexical patterns in Non-lattice Subgraphs

Since manual review of non-lattice subgraphs to discover potential errors is labor-intensive and time-consuming, we further take into account of the lexical information (concept labels) to automatically identify structural and lexical patterns in non-lattice subgraphs. Each pattern indicates certain type of errors and suggests a potential remediation.

For lexical information, we consider the label of a concept as a set of words in lower case. For example, the concept label *Stage III Pharyngeal Cancer* is considered as a set of words {*stage, iii, pharyngeal, cancer*}. For structural information, given a non-lattice subgraph, we use U_i to denote the set of words for a certain concept in the upper bounds, and L_j to denote the set of words for a certain concept in the lower bounds.

We define six patterns taking into account of such lexical and structural information in the NCIt: Containment, Union, Intersection, Union-Intersection, Inference-Contradiction, and Inference-Union. The Containment, Union, Intersection, and Union-Intersection patterns were initially proposed in previous work[8] for SNOMED CT. The Inference-Contradiction and Inference-Union patterns are newly proposed in this work, incorporating inference into the structural and lexical information.

2.2.1 Containment

A non-lattice subgraph is defined as exhibiting a containment pattern[8], if the set of words for one concept U_i in the upper bounds is contained in the set of words for another concept U_j in the upper bounds, or the set of words for one concept L_i in the lower bounds is contained in the set of words for another concept L_j in the lower bounds. That

is, $U_i \subset U_j$, or $L_i \subset L_j$. This pattern may suggest a missing IS-A relation between the two concepts in the upper bounds (or lower bounds), that is, U_j IS-A U_i (or L_j IS-A L_i). Consider the example in Figure 1A, $L_1 = \{$*stage, iii, nasopharyngeal, carcinoma*$\}$ in the lower bounds is contained in $L_2 = \{$*stage, iii, nasopharyngeal, carcinoma, ajcc, v6*$\}$ in the lower bounds. This indicates a potential missing hierarchical relation: L_2 IS-A L_1, with L_2 more specific than L_1. The suggested correction is to add the relation *Stage III Nasopharyngeal Carcinoma AJCC v6* is a subclass of *Stage III Nasopharyngeal Carcinoma* (highlighted as a red edge in Figure 1B).

For the containment pattern, we do not consider non-lattice subgraphs with concepts involving negation words such as *no, not, without, absence*, since that would incorrectly suggest a missing hierarchical relation between a concept with negation and a concept without negation.

2.2.2 Union

A non-lattice subgraph is defined as exhibiting a union pattern[8], if the union of the sets of words for two concepts U_i and U_j in the upper bounds is equal to the set of words for some concept L_k in the lower bounds, that is, $U_i \cup U_j = L_k$. This pattern may suggest a missing IS-A relation between other concepts in the lower bounds and L_k. For instance, in Figure 2A, the union of $U_1 = \{$*testicular, non-seminomatous, germ, cell, tumor*$\}$ and $U_2 = \{$*malignant, testicular, germ, cell, tumor*$\}$ in the upper bounds is equal to $L_1 = \{$*malignant, testicular, non-seminomatous, germ, cell, tumor*$\}$ in the lower bound. This indicates a potential missing IS-A relation between the other concept *Childhood Testicular Yolk Sac Tumor* in the lower bounds and L_1. That is, *Childhood Testicular Yolk Sac Tumor* IS-A *Malignant Testicular Non-Seminomatous Germ Cell Tumor* (highlighted as a red edge in Figure 2B).

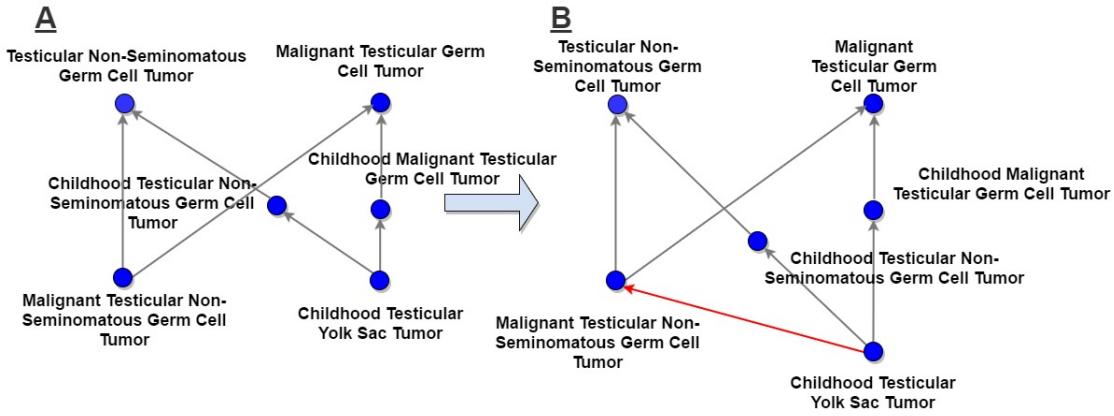

Figure 2: (**A**) A non-lattice subgraph exhibiting the Union pattern. (**B**) The suggested remediation of the non-lattice subgraph in (A): *Childhood Testicular Yolk Sac Tumor* IS-A *Malignant Testicular Non-Seminomatous Germ Cell Tumor*.

2.2.3 Intersection

A non-lattice subgraph is defined as exhibiting an intersection pattern[8], if the intersection of the set of words for two concepts L_i and L_j in the lower bounds is equal to the set of words for some concept U_k in the upper bound, that is, $L_i \cap L_j = U_k$. This pattern may suggest a missing IS-A relation between U_k and other concepts in the upper bounds. For instance, in Figure 3A, the intersection of $L_1 = \{$*splenic, t, lymphoblastic, lymphoma*$\}$ and $L_2 = \{$*splenic, b, lymphoblastic, lymphoma*$\}$ in the lower bounds is equal to $U_1 = \{$*splenic, lymphoblastic, lymphoma*$\}$ in the upper bound. This indicates a potential missing IS-A relation between U_1 and the other concept *Aggressive Non-Hodgkin Lymphoma* in the upper bound. That is, *Splenic Lymphoblastic Lymphoma* IS-A *Aggressive Non-Hodgkin Lymphoma* (the red edge in Figure 3B).

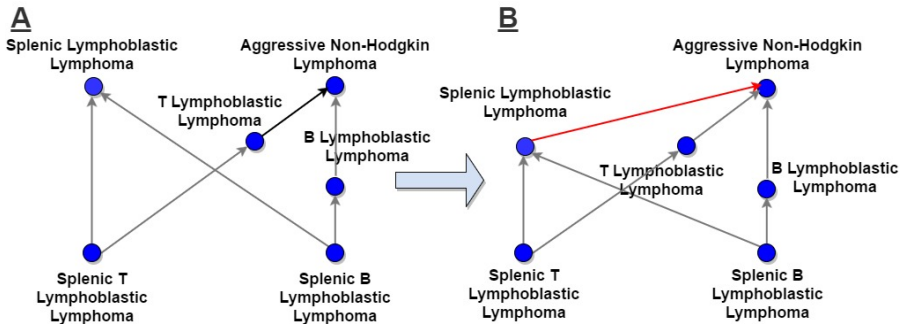

Figure 3: (**A**) A non-lattice subgraph exhibiting the Intersection pattern. (**B**) The suggested remediation of the non-lattice subgraph in (A): *Splenic Lymphoblastic Lymphoma* IS-A *Aggressive Non-Hodgkin Lymphoma*.

2.2.4 Union-Intersection

A non-lattice subgraph is defined as exhibiting an union-intersection pattern[8], if the union of the set of words for two concepts U_i and U_j in the upper bounds is equal to the intersection of the set of words for two concepts L_s and L_t in the lower bounds, that is, $U_i \cup U_j = L_s \cap L_t$. This pattern may suggest a missing intermediary concept between the two concepts (U_i and U_j) in upper bounds and the two concepts (L_s and L_t) in the lower bounds. For example, in Figure 4A, the union of $U_1 = \{$*localized, carcinoma*$\}$ and $U_2 = \{$*adult, liver, carcinoma*$\}$ is equal to the intersection of $L_1 = \{$*localized, non-resectable, adult, liver, carcinoma*$\}$ and $L_2 = \{$*localized, resectable, adult, liver, carcinoma*$\}$, that is, $U_i \cup U_j = L_s \cap L_t = \{$*localized, adult, liver, carcinoma*$\}$. This indicates a potential missing concept *Localized Adult Liver Carcinoma* (green node in Figure 4B), which represents the features that are common to L_s and L_t in the lower bounds and inherited from U_i and U_j in the upper bounds.

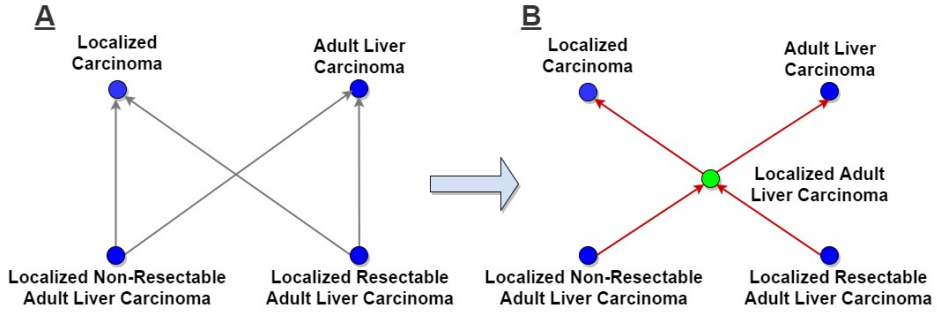

Figure 4: (**A**) A non-lattice subgraph exhibiting the Union-Intersection pattern. (**B**) The suggested remediation of the non-lattice subgraph in (A): adding a missing concept *Localized Adult Liver Carcinoma*.

It is worth noting that if $U_i \cup U_j = L_s \cap L_t$ happens to be equal to L_s or L_t, then the non-lattice subgraph falls into the union pattern as well; if it happens to be equal to U_i or U_j, then the non-lattice subgraph falls into the intersection pattern as well. In such cases, the suggestion for union pattern or intersection pattern is adopted, since no intermediary concept is needed.

2.2.5 Inference-Contradiction

Given a non-lattice subgraph G, we define two types of concept pairs appearing in G: related and unrelated. A pair of concepts (C_i, C_j) in G is called related if C_i is a subclass or descendant of C_j; otherwise, (C_i, C_j) is called unrelated. For instance, in Figure 5A, the concept pair (*Anaplastic Cell*, *Neoplastic Large Cell*) is related; while the concept pair (*Anaplastic T-Lymphocyte*, *Neoplastic Large T-Lymphocyte*) is unrelated.

Suppose R is the set of all related concept pairs in G, and $\overline{R}$ is the set of all unrelated concept pairs in G. We perform a set-difference-based inference to derive contradiction in the following way. For each related concept pair (B_d, B_a) in

R, if $B_d - (B_d \cap B_a) \neq \emptyset$ and $B_a - (B_d \cap B_a) \neq \emptyset$, an inferred term pair $\Big(B_d - (B_d \cap B_a), B_a - (B_d \cap B_a)\Big)$ can be derived. Similarly, for each unrelated concept pair (N_i, N_j) in $\overline{R}$, if $N_i - (N_i \cap N_j) \neq \emptyset$ and $N_j - (N_i \cap N_j) \neq \emptyset$, an inferred term pair $\Big(N_i - (N_i \cap N_j), N_j - (N_i \cap N_j)\Big)$ can be derived. If there exists some common term pair that can be derived from both a related pair in R and an unrelated pair in $\overline{R}$, we say that the non-lattice subgraph is exhibiting an inference-contradiction pattern. For instance, in Figure 5A, the related concept pair (*Anaplastic Cell*, *Neoplastic Large Cell*) derives a term pair (*Anaplastic*, *Neoplastic Large*); while the unrelated concept pair (*Anaplastic T-Lymphocyte*, *Neoplastic Large T-Lymphocyte*) derives the same term pair (*Anaplastic*, *Neoplastic Large*). This pattern may suggest a potential missing IS-A relation between the unrelated concept pair: *Anaplastic T-Lymphocyte* IS-A *Neoplastic Large T-Lymphocyte* (the red edge in Figure 5B).

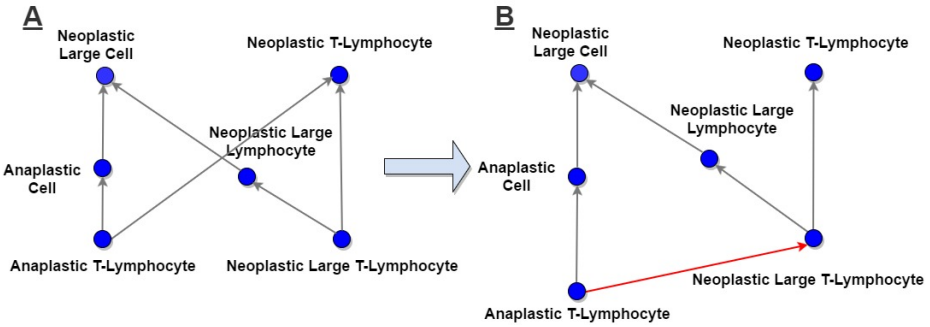

Figure 5: (**A**) A non-lattice subgraph exhibiting the Inference-Contradiction pattern. (**B**) The suggested remediation of the non-lattice subgraph in (A): *Anaplastic T-Lymphocyte* IS-A *Neoplastic Large T-Lymphocyte*.

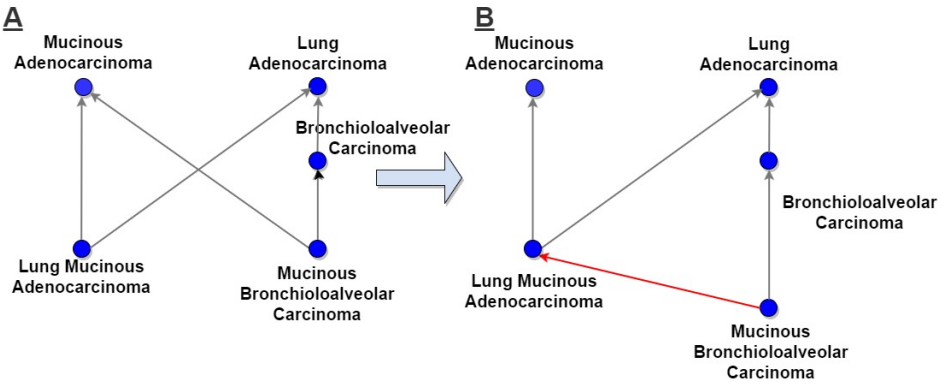

Figure 6: (**A**) A non-lattice subgraph exhibiting the Union, Inference-Contradiction, and Inference-Union patterns. (**B**) The suggested remediation of (A): *Mucinous Bronchioloalveolar Carcinoma* IS-A *Lung Mucinous Adenocarcinoma*.

2.2.6 Inference-Union

A non-lattice subgraph is defined as exhibiting an inference-union pattern, if the union of the set of words for some concept U_s in the upper bounds and the intersection of the set of words for two concepts L_i and L_j in the lower bounds is equal to the set of words for some concept L_k in the lower bounds, that is, $U_s \cup (L_i \cap L_j) = L_k$. This may suggest a missing IS-A relation between other concepts in the lower bounds and L_k.

For instance, in Figure 6A, the intersection of $L_1 = \{$*lung, mucinous, adenocarcinoma*$\}$ and $L_2 = \{$*mucinous, bronchioloalveolar, carcinoma*$\}$ in the lower bounds is $\{$*mucinous*$\}$, whose union with $U_1 = \{$*lung, adenocarcinoma*$\}$ is equal to $L_1 = \{$*lung, mucinous, adenocarcinoma*$\}$. This indicates a potential missing IS-A relation between the other concept L_2 in the lower bounds and L_1. That is, *Mucinous Bronchioloalveolar Carcinoma* IS-A *Lung Mucinous Adenocarcinoma* (the red edge in Figure 6B).

2.2.7 Non-lattice Subgraphs with Multiple Patterns

We also investigate non-lattice subgraphs exhibiting multiple patterns among the above-mentioned six patterns. For instance, the non-lattice subgraph in Figure 1A exhibits both containment and inference-union patterns, and both patterns suggest a missing IS-A relation: *Stage III Nasopharyngeal Carcinoma AJCC v6* IS-A *Stage III Nasopharyngeal Carcinoma*. The non-lattice subgraph in Figure 6A is following three patterns: union, inference-contradiction, and inference-union, and all these patterns suggest a missing relation between *Mucinous Bronchioloalveolar Carcinoma* and *Lung Mucinous Adenocarcinoma*.

2.3 Evaluation

For evaluation, we focus on small non-lattice subgraphs (size of 4, 5, and 6) due to two reasons. One is that small ones are relatively easy to visually inspect by domain experts. The other reason is that small non-lattice subgraphs may be contained in larger ones, and fixing errors in small ones will automatically eliminate the same errors propagated in the larger ones (although there might be other errors in the larger ones).

To evaluate the performance of applying different patterns in small non-lattice subgraphs to automatically detect real errors in NCIt and suggest corrections, we randomly selected 25 non-lattice subgraphs with a single pattern, and 25 ones with multiple patterns, respectively. These 50 sample non-lattice subgraphs as well as their suggested remediations were rendered in scalable vector graphics and provided to experts (authors MAB and JT) for evaluation. MAB evaluated cancer-related samples, and JT evaluated drug-related samples.

Table 1: Number of non-lattice subgraphs exhibiting each of the 24 patterns.

Pattern	**No. of non-lattice subgraphs**	**No. of small non-lattice subgraphs (size of 4-6)**
Containment	159	84
Union	7	3
Intersection	430	166
Union-Intersection	24	2
Inference-Contradiction	37	3
Inference-Union	21	12
Inference-Contradiction, Containment	3	1
Inference-Union, Containment	19	13
Inference-Contradiction, Inference-Union	12	9
Intersection, Containment	2	1
Intersection, Inference-Contradiction	33	9
Union, Inference-Union	1	0
Inference-Contradiction, Union-Intersection	1	0
Intersection, Inference-Union	3	0
Inference-Union, Inference-Contradiction, Containment	14	7
Intersection, Inference-Union, Containment	2	1
Union, Inference-Union, Inference-Contradiction	7	4
Union, Intersection, Inference-Union	13	12
Intersection, Inference-Contradiction, Containment	2	0
Union, Union-Intersection, Inference-Union, Containment	6	4
Union, Intersection, Inference-Union, Inference-Contradiction	5	3
Intersection, Inference-Contradiction, Containment, Union-Intersection	1	0
Intersection, Inference-Union, Inference-Contradiction, Containment	2	0
Union, Union-Intersection, Inference-Union, Inference-Contradiction, Containment	5	3
Total	809	337

3 Results

3.1 Non-lattice Subgraphs Exhibiting Structural and Lexical Patterns

A total of 8,143 non-lattice subgraphs were identified in the NCIt (version 16.12d), among which 809 exhibits a single pattern or multiple patterns. Of these 809 non-lattice subgraphs, 678 were found exhibiting a single lexical pattern, 131 exhibiting multiple patterns. Of the 809 non-lattice subgraphs, 337 were small ones (size of 4, 5, and 6), among which 270 exhibited a single pattern, 67 exhibited multiple patterns. Table 1 shows the numbers of both non-lattice

subgraphs and small non-lattice subgraphs exhibiting different combinations of patterns (six single pattern, eighteen multiple patterns). For instance, there were 159 non-lattice subgraphs exhibiting a single containment pattern (the first row in Table 1), and 5 non-lattice subgraphs exhibiting multiple patterns: union, union-intersection, inference-union, inference-contradiction, and containment (the last row in Table 1). Figure 7 shows an example of non-lattice subgraph with these five patterns. For the 678 non-lattice subgraphs with a single pattern, the intersection pattern accounted for the largest proportion (430 non-lattice subgraphs). For the 131 non-lattice subgraphs with multiple patterns, the intersection and inference-contradiction patterns accounted for the largest proportion (33 non-lattice subgraphs).

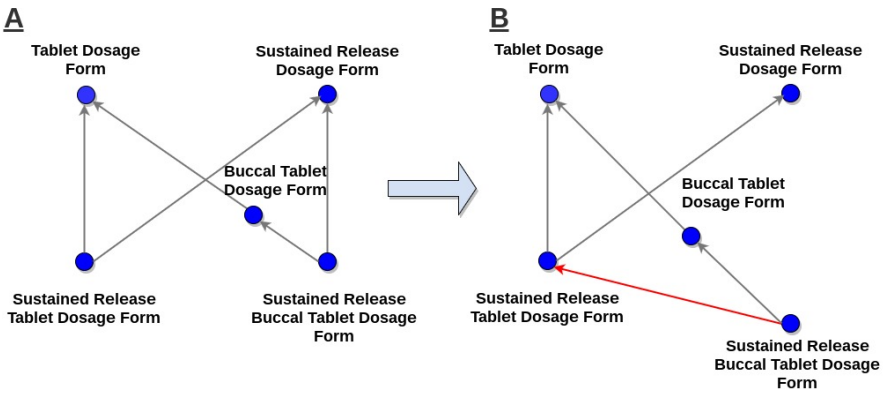

Figure 7: (**A**) A non-lattice subgraph exhibiting five patterns: Union, Union-Intersection, Inference-Union, Inference-Contradiction, and Containment. (**B**) The suggested remediation of the non-lattice subgraph in (A): *Sustained Release Buccal Tablet Dosage Form* IS-A *Sustained Release Tablet Dosage Form*.

Table 2: Numbers of small non-lattice subgraphs evaluated by domain experts in terms of patterns, as well as numbers of correct suggestions verified by experts.

Pattern	**No. of non-lattice subgraphs**	**No. of non-lattice subgraphs with correct suggestions**	**Correction rate**
Containment	7	6	85.7%
Union	1	1	100%
Intersection	14	2	14.3%
Union-Intersection	1	1	100%
Inference-Contradiction	1	1	100%
Inference-Union	1	0	0%
Inference-Contradiction, Containment	1	1	100%
Inference-Union, Containment	4	3	75%
Inference-Contradiction, Inference-Union	3	3	100%
Intersection, Containment	1	1	100%
Intersection, Inference-Contradiction	3	2	66.7%
Inference-Union, Inference-Contradiction, Containment	2	2	100%
Intersection, Inference-Union, Containment	1	0	0%
Union, Inference-Union, Inference-Contradiction	2	2	100%
Union, Intersection, Inference-Union	4	4	100%
Union, Union-Intersection, Inference-Union, Containment	2	2	100%
Union, Intersection, Inference-Union, Inference-Contradiction	1	1	100%
Union, Union-intersection, Inference-Union, Inference-Contradiction, Containment	1	1	100%
Total	50	33	66%

3.2 Evaluation

Of the 50 sample non-lattice subgraphs evaluated by domain experts, 33 were verified to contain errors and make correct suggestions ($33/50 = 66\%$). Among these 33 correct cases, 32 were missing hierarchical relations and one was a missing intermediary concept. Table 2 presents the numbers of evaluated non-lattice subgraphs exhibiting each combination of patterns, and the numbers of correct suggestions confirmed by domain experts. Of the 25 evaluated non-lattice subgraphs with a single pattern, 11 were verified correct ($11/25 = 44\%$). Of the 25 evaluated non-lattice

subgraphs with multiple patterns, 22 were verified correct ($22/25 = 88\%$). This illustrates that non-lattice subgraphs with multiple patterns achieved a better performance than those with a single pattern in terms of the correction rate.

4 Discussion

In this paper, we investigated non-lattice subgraphs in NCIt based on six structural and lexical patterns, with each pattern automatically suggesting a potential missing hierarchical relation or missing concept. Our pattern-based approach leveraging both structural and lexical information is scalable and applicable to other terminologies for quality assurance work, since it generally takes concepts (as well as concept labels) and hierarchical relations of a terminology as the input, and generates erroneous non-lattice subgraphs and potential corrections as the output.

Analysis of failure cases. For the single-pattern non-lattice subgraphs evaluated in Table 2, the suggestions made by the intersection pattern turned out to have a low correction rate ($2/14 = 14.3\%$). Figure 8A shows a non-lattice subgraph exhibiting the intersection pattern: $\{$*gestational, choriocarcinoma*$\} \cap \{$*ovarian, choriocarcinoma*$\} = \{$*choriocarcinoma*$\}$. However, its automatic suggestion in Figure 8B is not correct. That is, *Choriocarcinoma* is NOT a subclass of *Malignant Female Reproductive System Neoplasm*, since Choriocarcinoma can be a malignant female reproductive system neoplasm, but it can also arise in the male testis. Another example of wrongly suggested case by the containment pattern is: $\{$*osteoma*$\} \subset \{$*osteoid, osteoma*$\}$. However, despite the similarity in names, *Osteoid Osteoma* and *Osteoma* are two completely different types of tumor, and *Osteoid Osteoma* is thus NOT a subclass of *Osteoma*.

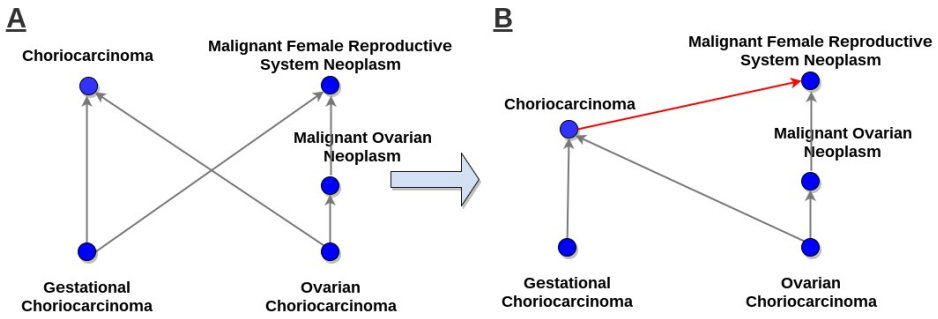

Figure 8: (**A**) A non-lattice subgraph exhibiting an Intersection pattern. (**B**) The **wrongly** suggested remediation of (A).

Comparison with previous work. The hybrid approach to mining structural-lexical patterns in non-lattice subgraphs were initially proposed in previous work[8] for quality assurance of SNOMED CT, where four patterns were studied: containment, union, intersection, and union-intersection. In this paper, we applied these four patterns to NCIt, and further proposed two new patterns with implicit inference: inference-contradiction and inference-union. In addition, only single-pattern non-lattice subgraphs were investigated in previous work[8], while in this paper, we not only studied non-lattice subgraphs with a single pattern, but also those with multiple patterns. Non-lattice subgraphs in NCIt with multiple patterns turned out to have a higher error detection and correction rate than those with a single pattern (see Table 2). For SNOMED CT[8], the overall correction rate of the four patterns (by single pattern) was 59%. For the NCIt in this paper, the overall correction rate of the six patterns (by both single pattern and mixed patterns) is 66%.

Limitations and future work. A limitation of this work is that we only evaluated small non-lattice subgraphs (size of 4, 5, 6) for experts' ease to review and validate. It would be interesting to further examine larger-size non-lattice subgraphs for evaluation. In addition, our evaluation was limited on the number of samples and only one domain expert was involved. We plan to evaluate more samples by multiple experts in the future. Another limitation of this work is that the list of negation words used in detecting the containment pattern was manually constructed based on our observation and previous experience. In the future, we expect to use a resource like NegEx for this purpose. The followings are a couple of directions for additional future work. When defining different patterns, we only used the concept labels for lexical information. We plan to take into account of the concept synonyms to complement concept labels, which may obtain more non-lattice subgraphs with patterns. Note that there are non-lattice subgraphs that may be erroneous but are not exhibiting any of the six patterns studied in this paper. New patterns or approaches are needed to uncover potential errors in such non-lattice subgraphs.

5 Conclusions

In this paper, we investigated a hybrid approach to identifying potential errors in the NCI Thesaurus and automatically suggesting remediations, by mining structural and lexical patterns in non-lattice subgraphs. This approach proved an effective way for error detection and correction in the NCI Thesaurus, and is applicable to other biomedical terminologies for quality assurance purposes.

Acknowledgement

This work was supported by the National Institutes of Health (NIH) National Center for Advancing Translational Sciences through grant UL1TR001998 and National Science Foundation through grant IIS-1657306.

References

1. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearbook of Medical Informatics. 2008;67-79.
2. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. Journal of the American Medical Informatics Association. 2014;21(e1):e11-9.
3. Cui L, Tao S, Zhang GQ. Biomedical ontology quality assurance using a big data approach. ACM Transactions on Knowledge Discovery from Data. 2016;10(4):1-28.
4. Geller J, Perl Y, Halper M, Cornet R. Special issue on auditing of terminologies. Journal of Biomedical Informatics. 2009;42(3): 407-411.
5. NCI Thesaurus (NCIt). https://wiki.nci.nih.gov/pages/viewpage.action?pageId=7472532. Accessed February 27, 2017.
6. de Coronado S, Wright LW, Fragoso G, et al. The NCI thesaurus quality assurance life cycle. Journal of Biomedical Informatics. 2009;42(3):530-9.
7. Hartel FW, de Coronado S, Dionne R, Fragoso G, Golbeck J. Modeling a description logic vocabulary for cancer research. Journal of Biomedical Informatics. 2005 Apr 30;38(2):114-29.
8. Cui L, Zhu W, Tao S, Case JT, Bodenreider O, Zhang GQ. Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT. Journal of the American Medical Informatics Association 2017 ocw175. doi: 10.1093/jamia/ocw175
9. Zhu X, Fan JW, Baorto DM, Weng C, Cimino JJ. A review of auditing methods applied to the content of controlled biomedical terminologies. Journal of Biomedical Informatics. 2009;42(3):413-25.
10. Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as part of the terminology design life cycle. Journal of the American Medical Informatics Association. 2006;13(6):676-90.
11. Ochs C, Geller J, Perl Y, et al. Scalable quality assurance for large SNOMED CT hierarchies using subject-based subtaxonomies. Journal of the American Medical Informatics Association. 2015;22(3):507-518.
12. Ochs C, Geller J, Perl Y, et.al. A tribal abstraction network for SNOMED CT target hierarchies without attribute relationships. Journal of the American Medical Informatics Association. 2015;22(3):628-39.
13. Verspoor K, Dvorkin D, Cohen K, Hunter L. Ontology quality assurance through analysis of term transformations. Bioinformatics. 2009;25(12):i77-i84.
14. Zhang GQ, Bodenreider O. Large-scale, exhaustive lattice-based structural auditing of SNOMED CT. AMIA Annu Symp Proc. 2010;922-26.
15. Zhang GQ, Bodenreider O. Using SPARQL to test for lattices: application to quality assurance in biomedical ontologies. The Semantic Web-ISWC. 2010;273-288.
16. Agrawal A, Perl Y, Ochs C, Elhanan G. Algorithmic detection of inconsistent modeling among SNOMED CT concepts by combining lexical and structural indicators. IEEE International Conference on Bioinformatics and Biomedicine, 2015;476-483.
17. Bodenreider O. Identifying missing hierarchical relations in SNOMED CT from logical definitions based on the lexical features of concept names. International Conference on Biomedical Ontology and BioCreative. 2016;1747.
18. Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods of Information in Medicine 2005;44:498-507.
19. Mougin F, Bodenreider O. Auditing the NCI thesaurus with semantic web technologies. AMIA Annu Symp Proc. 2008;500-504.
20. He Z, Geller J. Preliminary analysis of difficulty of importing pattern-based concepts into the National Cancer Institute thesaurus. Studies in Health Technology and Informatics. 2016;228:389-93.
21. Zhe He, Yan Chen, Sherri de Coronado, Katrina Piskorski, and James Geller. Topological-pattern-based recommendation of UMLS concepts for National Cancer Institute thesaurus. AMIA Annu Symp Proc. 2016;618-627.