A Lexical Approach to Identifying Subtype Inconsistencies in Biomedical Terminologies

Rashmie Abeysinghe*, Fengbo Zheng*, Eugene W. Hinderer III†, Hunter N.B. Moseley†‡§¶, Licong Cui*‡
*Department of Computer Science
†Department of Molecular and Cellular Biochemistry
‡Institute for Biomedical Informatics
§Markey Cancer Center
¶Center for Environmental and Systems Biochemistry
University of Kentucky, Lexington, Kentucky, USA

***Abstract*—We introduce a lexical-based inference approach for identifying subtype (or *is_a* relation) inconsistencies in biomedical terminologies. Given a terminology, we first represent the name of each concept in the terminology as a sequence of words. We then generate hierarchically-linked and -unlinked pairs of concepts, such that the two concepts in a pair have the same number of words, and contain at least one word in common and a fixed number n of different words ($n = 1, 2, 3, 4, 5$). The linked and unlinked concept-pairs further infer corresponding linked and unlinked term-pairs, respectively. If a linked concept-pair and an unlinked concept-pair infer the same term-pair, we consider this as a potential subtype inconsistency, which may indicate a missing subtype relation or an incorrect subtype relation. We applied this approach to Gene Ontology (GO), National Cancer Institute thesaurus (NCIt) and SNOMED CT. A total of 4,841 potential subtype inconsistencies were found in GO, 2,677 in NCIt, and 53,782 in SNOMED CT. Domain experts evaluated a random sample of 211 potential inconsistencies in GO, and verified that 124 of them are valid (i.e., a precision of 58.77% for detecting subtype inconsistencies in GO). We also performed a preliminary study on the extent to which external knowledge in the Unified Medical Language System (UMLS) can provide supporting evidence for validating the detected potential inconsistencies: 0.54% (=26/4841) for GO, 11.43% (=306/2677) for NCIt, and 3.61% (=1940/53782) for SNOMED CT. Results indicate that our lexical-based inference approach is a promising way to identify subtype inconsistencies and facilitates the quality improvement of biomedical terminologies.**

***Index Terms*—Terminology quality assurance, Gene Ontology, National Cancer Institute thesaurus, SNOMED CT, Unified Medical Language System, Subtype inconsistencies, Missing subtype relations, Incorrect subtype relations**

I. Introduction

Biomedical terminologies and ontologies play important roles in knowledge management; data integration, exchange and semantic interoperability; and decision support and reasoning in biomedicine [1], [2], [3], [4], [5]. For example, Gene Ontology (GO) provides a dynamic, controlled vocabulary that can be applied to all branches of life, and has been widely used for modeling and codifying biological knowledge [6], [7], [8]. The National Cancer Institute thesaurus (NCIt) has been designed to provide structured and principled representation of key cancer-related concepts, covering topics including cancers, drugs, therapies, anatomy, genes, pathways, cellular and subcellular processes, and proteins [9]. SNOMED CT is the most comprehensive clinical health terminology in the world and supports development of high-quality electronic health records [1], [10], [11].

Since biomedical terminologies are constantly evolving, inconsistencies or errors may be introduced during the terminology evolution and modeling process. Therefore, quality assurance has been an integral part of the terminology management lifecycle. However, quality assurance becomes increasingly challenging due to the ever-growing size and structural complexity of biomedical terminologies. It is time-consuming and labor-intensive to manually review these terminologies and uncover potential quality issues. There is a pressing need to develop effective, semi-automated approaches to detect and fix potential quality issues in terminologies in a manner that minimizes manual review.

In this paper, we introduce a lexical approach to systematically detect potential *subtype* (or *is_a* relation) inconsistencies that can be generally applied to biomedical terminologies. This approach leverages the names of pairs of concepts that are hierarchically linked and unlinked in a given terminology to derive potential inconsistencies, which may be indicative of missing subtype relations or incorrect existing subtype relations. We applied this approach to three terminologies: GO, NCIt, and SNOMED CT. To evaluate the effectiveness of this approach, we select a random sample of potential inconsistencies detected in GO, which were manually reviewed and validated by domain experts. We also performed a preliminary study on utilizing external knowledge from the Unified Medical Language System (UMLS) to automatically identify supporting evidence for the detected potential inconsistencies. This study shows the degree to which cross-terminology evaluation can help with the validation of the detected subtype inconsistencies.

This work was supported by the National Science Foundation through grants 1816805, 1657306 and 1252893, and by the National Institutes of Health through grants UL1TR001998 and R21CA231904. Correspondence: licong.cui@uky.edu

II. Background

A. Gene Ontology

Maintained by the Gene Ontology Consortium, GO provides computer-readable knowledge regarding the functions, organization, and localization of genes and gene products (GO concepts or terms) and how these functions relate to each other (relations) [7], [8], [12]. GO covers three subdomains (or subontologies): biological process (the broad biological system in which a gene product is involved), molecular function (the specific role a gene product has or potentially has within a biological process), and cellular component (the location or organized unit in a cell where the gene product performs its molecular function) [12], [13]. The 03/28/2017 release of GO contains over 40,000 concepts.

B. National Cancer Institute Thesaurus

NCIt is the National Cancer Institute (NCI)'s reference terminology that includes broad coverage of the cancer domain. It covers vocabulary for clinical care, translational and basic research, and public information and administrative activities [14], [15]. NCIt concepts are hierarchically organized into 19 domains, including abnormal cell; anatomic structure, system or substance; biological process; disease, disorder or finding [16]. The 07/2018 release of NCIt contains more than 135,000 concepts.

C. SNOMED CT

Maintained and distributed by SNOMED International, SNOMED CT is the largest clinical terminology in the world [10], [11]. It is a multilingual and multinational terminology with comprehensive, scientifically validated content [17]. SNOMED CT content covers clinical medicine which includes findings, diseases, and procedures for use in electronic medical records [1]. The 03/01/2018 release of the SNOMED CT United States (US) edition contains more than 300,000 concepts.

D. Unified Medical Language System (UMLS)

UMLS is an integrated repository of biomedical vocabularies provided by the National Library of Medicine (NLM) [18]. It contains over 200 biomedical terminologies including GO, NCIt, SNOMED CT, Medical Dictionary for Regulatory Activities (MedDRA) and Human Phenotype Ontology, to enable interoperability between computer systems. Term variants from source vocabularies are mapped to UMLS concepts, and each concept is assigned a unique concept identifier (CUI). For example, terms *Heart attack, Myocardial infarction* and *Cardiovascular Stroke* from different sources represent the same meaning and are mapped to a UMLS concept (CUI: *C0027051*). The 2018AA release of UMLS contains more than 3.6 million concepts [19].

E. Quality Assurance of Biomedical Terminologies

Various approaches have been investigated for quality assurance or auditing of biomedical terminologies [20], [21]. For example, Ochs et al. [22] have developed two kinds of abstraction networks: area taxonomy and partial-area taxonomy, for auditing GO. Here, area taxonomies are groups of concepts that have exactly the same roles, and partial-area taxonomies are further divisions of areas, which are structurally uniform and singly-rooted. They identified groups of anomalous terms that are expected to have a higher error rate when compared to other terms.

Agrawal et al. [23] have proposed positional similarity sets of concepts, which are concepts with the same length of names but differing by one word in a single position, to uncover inconsistently modeled concepts in SNOMED CT.

Verspoor et al. [24] have introduced a quality assurance method for GO based on univocality (similar concepts being expressed consistently). They have developed a transformation-based clustering methodology to identify terms which express similar semantics, but use different linguistic conventions.

Zhang et al. [25] have proposed a lattice-based approach to auditing biomedical terminologies. They extracted non-lattice pairs, which are pairs of concepts which do not satisfy the lattice property, a desirable property for a well-formed terminology, to audit SNOMED CT. Cui et al. [26] have introduced a big data approach (using Hadoop MapReduce) to exhaustively detect non-lattice pairs in SNOMED CT, achieving several orders of magnitude in speed-up in comparison with [25]. Recently, Cui et al. [27] have mined lexical patterns of concept names in non-lattice subgraphs to detect missing hierarchical relations and concepts. Abeysinghe et al. [28] have applied that approach to NCIt and further introduced additional lexical patterns in non-lattice subgraphs.

In previous work [29], we performed a preliminary study on representing each GO concept name as a set of words and deriving subtype inconsistencies from hierarchically linked and unlinked pairs of GO concepts, which have the same number of words, containing common words as well as a single different word.

These existing auditing approaches are sometimes limited in precision, lack applicability, or focused on analyzing substructures of a terminology. In this work, we expand on our previous work [29] and present a general, lexical-based inference approach to identify subtype inconsistencies in a given terminology. This approach is widely applicable to biomedical terminologies and not limited to any substructures.

III. Methods

Our lexical-based inference approach, implemented in Java programming language, aims at identifying potential subtype inconsistencies among concept-pairs in a given terminology. This approach leverages three intrinsic aspects of knowledge in the terminology: the names of concepts, the existing subtype relations, and the absent subtype relations. Firstly, we represent each concept name as a sequence of words. Then, we generate hierarchically-linked and -unlinked partial matching pairs of concepts. Such concept-pairs further derive linked and unlinked term-pairs. Then, we identify potential subtype

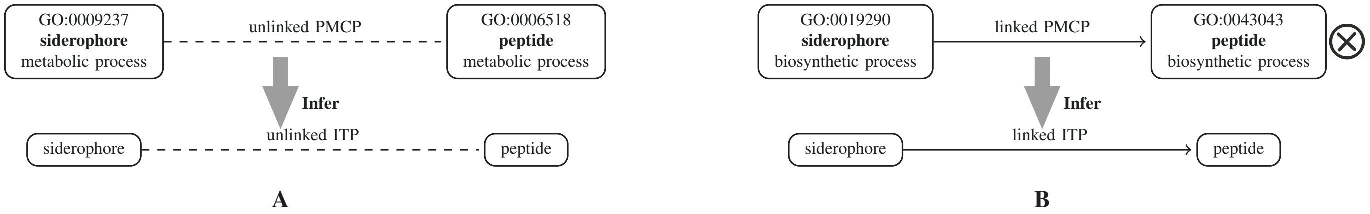

Fig. 1. **A**: Unlinked PMCP with diff 1 in GO and its unlinked ITP derived; **B**: Linked PMCP with diff 1 in GO and its linked ITP derived. This example reveals a potentially **incorrect existing subtype relation** in **B**, that is, GO:0019290 (*siderophore biosynthetic process*) is not a subtype of GO:0043043 (*peptide biosynthetic process*).

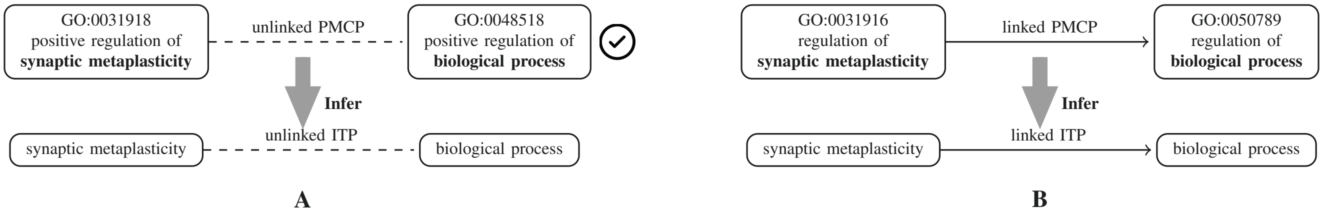

Fig. 2. **A**: An unlinked PMCP with diff 2 in GO and its unlinked ITP derived; **B**: A linked PMCP with diff 2 in GO and its linked ITP derived. This example reveals a potentially **missing subtype relation** in **A**, that is, GO:0031918 (*positive regulation of synaptic metaplasticity*) is-a GO:0048518 (*positive regulation of biological process*).

inconsistencies through linked and unlinked concept-pairs that derive the same term-pair. We apply this approach to GO, NCIt, and SNOMED CT, respectively. For evaluation, domain experts manually review a random sample of potential subtype inconsistencies detected in GO. In addition, we perform an automated cross-terminology evaluation by leveraging external knowledge in UMLS to find supporting evidence for the detected potential subtype inconsistencies.

A. Representation of Concept Names

The name of a concept in a terminology usually represents the semantic meaning of a concept unambiguously. Given a terminology, we represent the name of each concept C as an ordered sequence of words $w_1 w_2 \ldots w_m$. For example, the name of a GO concept GO:0042317 (the unique identifier) is *penicillin catabolic process*, and its sequence-of-words representation is [*penicillin, catabolic, process*]. Note that this is different from the set-of-words model in [29], which considers *penicillin catabolic process* and *catabolic process penicillin* as the same.

B. Generation of Linked and Unlinked Concept-Pairs

A pair of concepts belonging to the same sub-hierarchy of a terminology, is defined as a *partial matching concept pair (PMCP)* with diff n, if the names of the two concepts have the same number of words and contain at least one word in common and n different words. We study $n = 1, 2, 3, 4, 5$ in this paper. For instance, GO:0042317 (*penicillin catabolic process*) and GO:0009310 (*amine catabolic process*) is a PMCP with diff 1, because both of them are from the *biological process* sub-hierarchy of GO, contain two common words [*catabolic, process*], and differ in a single word – *penicillin* versus *amine*.

We classify PMCPs into two categories (linked and unlinked) as follows. If the two concepts in a PMCP have a subtype relation (either direct or indirect), then the PMCP is called a *linked PMCP*. If the the two concepts in a PMCP does not have a subtype relation (neither direct nor indirect), then the PMCP is called an *unlinked PMCP*. Note that we pre-compute transitive closure of the subtype relation (i.e., direct and indirect *is-a* relations) to decide whether a PMCP is linked or unlinked. That is, if two concept of a PMCP are in the transitive closure, then the PMCP is linked; otherwise, it is unlinked. In other words, for a linked PMCP (C_1, C_2), the concept C_1 is either a direct subtype of the concept C_2 or an indirect (transitive) subtype of C_2.

Fig. 1A presents an example of an unlinked PMCP with diff 1 in GO, where the two concepts GO:0009237 (*siderophore metabolic process*) and GO:0006518 (*peptide metabolic process*) differ in a single word – *siderophore* versus *peptide*.

Fig. 2B presents an example of a linked PMCP with diff 2 in GO, where the two concepts GO:0031916 (*regulation of synaptic metaplasticity*) and GO:0050789 (*regulation of biological process*) differ in two words – *synaptic metaplasticity* versus *biological process*.

C. Generation of Linked and Unlinked Term-Pairs

For each PMCP (C_1, C_2), an *Inferred Term Pair (ITP)* can be derived as follows. Assume that $C_1 = w_{11} w_{12} \ldots w_{1m}$ and $C_2 = w_{21} w_{22} \ldots w_{2m}$ and there are n different words between C_1 and C_2: $w_{1i_1} w_{1i_2} \ldots w_{1i_n}$ versus $w_{2i_1} w_{2i_2} \ldots w_{2i_n}$, where $1 \leq i_j \leq m$ and $1 \leq j \leq n$. Then an ITP ($w_{1i_1} w_{1i_2} \ldots w_{1i_n}$, $w_{2i_1} w_{2i_2} \ldots w_{2i_n}$) is derived. In other words, the different words between the names of C_1 and C_2 derives an ITP. Note that we also require that the terms in difference cannot contain only numerals when generating ITPs.

We also classify ITPs into two categories (linked and unlinked) based on the PMCPs from which they are derived. An ITP derived from a linked PMCP is called a *linked ITP*. An ITP derived from an unlinked PMCP is called an *unlinked ITP*.

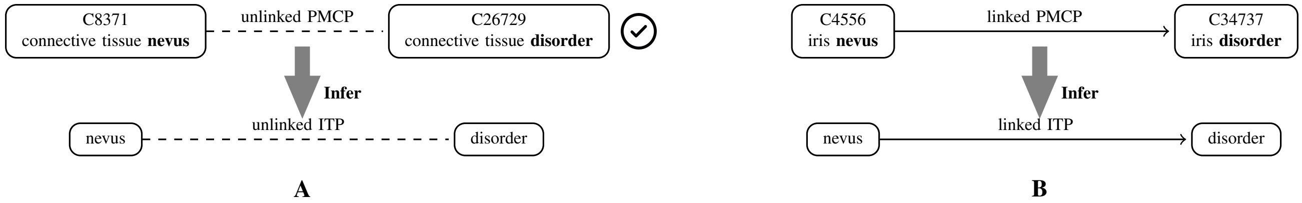

Fig. 3. **A**: An unlinked PMCP with diff 3 in NCIt and its unlinked ITP derived; **B**: A linked PMCP with diff 3 in NCIt and its linked ITP derived. This example reveals a potentially **missing subtype relation** in **A**, that is, C8371 (*connective tissue nevus*) is-a C26729 (*connective tissue disorder*).

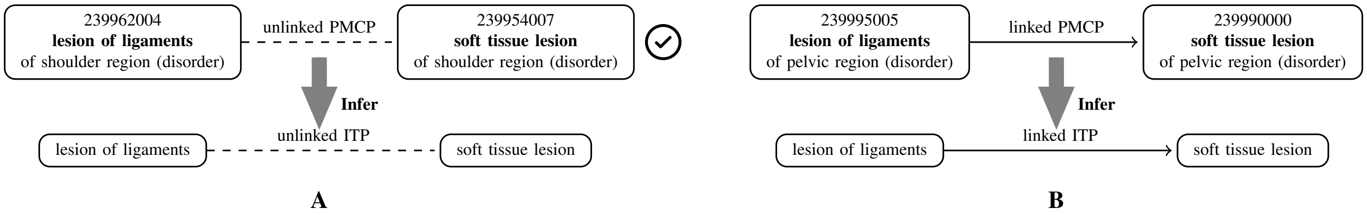

Fig. 4. **A**: An unlinked PMCP with diff 3 in SNOMED CT and its unlinked ITP derived; **B**: A linked PMCP with diff 3 in SNOMED CT and its linked ITP derived. This example reveals a potentially **missing subtype relation** in **A**, that is, 239962004 (*lesion of ligaments of shoulder region (disorder)*) is-a 239954007(*soft tissue lesion of shoulder region (disorder)*).

Take Fig. 1A as an example, the unlinked concepts GO:0009237 (*siderophore metabolic process*) and GO:0006518 (*peptide metabolic process*) differ in the first word and derive an unlinked ITP ([*siderophore*], [*peptide*]). In Fig. 2B, the linked concepts GO:0031916 (*regulation of synaptic metaplasticity*) and GO:0050789 (*regulation of biological process*) differ in the third and fourth words and derive a linked ITP ([*synaptic, metaplasticity*], [*biological, process*]).

D. Detection of Potential Inconsistencies

If the unlinked ITP derived from an unlinked PMCP and the linked ITP derived from a linked PMCP are the same, we consider the two PMCPs as a potential subtype inconsistency. For instance, the unlinked PMCP (GO:0009237, GO:0006518) in Fig. 1A and the linked PMCP (GO:0019290, GO:0043043) in Fig. 1B is considered a potential subtype inconsistency, since they derive the same ITP ([*siderophore*], [*peptide*]).

The unlinked PMCP (GO:0031918, GO:0048518) in Fig. 2A and the linked PMCP (GO:0031916, GO:0050789) in Fig. 2B are considered as a potential subtype inconsistency, since they derive the same ITP ([*synaptic, metaplasticity*], [*biological, process*]). Similarly, Fig. 3 and Fig. 4 give examples of potential subtype inconsistencies in NCIt and SNOMED CT, respectively.

E. Evaluation of Detected Potential Inconsistencies

1) Evaluation by Domain Experts: A random sample of potential subtype inconsistencies detected in GO was selected and evaluated by two domain experts (authors EWH and HNBM), to evaluate the effectiveness of our approach in detecting inconsistencies. The two domain experts reviewed and discussed the samples together.

We classify the detected potential inconsistencies into three categories during the evaluation: missing subtype relations, incorrect existing subtype relations, and false positives. Given an inconsistency I consisting of an unlinked PMCP (u_1, u_2) and a linked PMCP (l_1, l_2), we describe each of the three categories in detail as follows.

- Missing Subtype Relations: If the concepts in the unlinked PMCP (u_1, u_2) form a valid subtype relation, then it is regarded as a missing subtype (i.e., u_1 should be a subtype of u_2). For instance, in Fig. 2A, the concepts in the unlinked PMCP (GO:0031918, GO:0048518) indeed form a valid subtype relation; thus there is a missing subtype relation – GO:0031918 (*positive regulation of synaptic metaplasticity*) should be a subtype of GO:0048518 (*positive regulation of biological process*).
- Incorrect Existing Subtype Relations: If the concepts in the linked PMCP (l_1, l_2) are found to be an invalid subtype relation, then it is regarded as an incorrect existing subtype relation (i.e., l_1 should not be a subtype of l_2). For example, in Fig. 1B, the concepts in the linked PMCP (GO:0019290, GO:0043043) are found to form an invalid subtype relation, because the definition of siderophores clearly indicates that some are not peptides, for example quinolbactin produced by Pseudomonas fluorescens [30]. That is, GO:0019290 (*siderophore biosynthetic process*) should not be a subtype of GO:0043043 (*peptide biosynthetic process*).
- False Positives: If the concepts in the linked PMCP (l_1, l_2) indeed form a valid subtype relation and the concepts in the unlinked PMCP (u_1, u_2) are found to be an invalid subtype relation, then I is regarded as a false positive. For example, the concepts in the linked PMCP (GO:0002728, GO:0002716) in Fig. 5B indeed forms a valid subtype relation, and the unlinked PMCP (GO:0061082, GO:0002444) in Fig. 5A does not form a valid subtype relation. Therefore, the inconsistency shown in Fig. 5 is a false positive.

2) Cross-terminology Evaluation based on UMLS: We also leveraged external knowledge in UMLS (i.e., other terminolo-

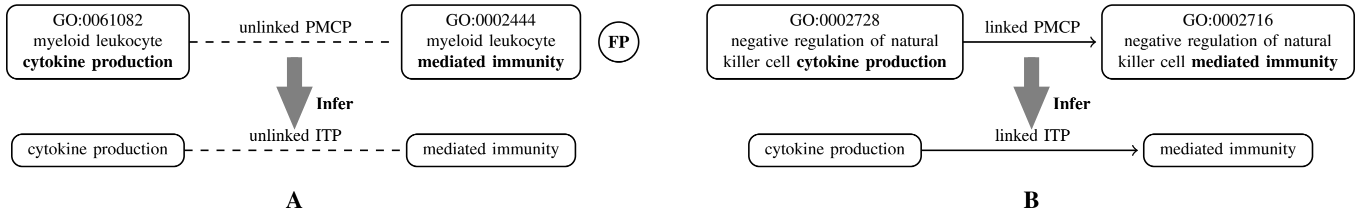

Fig. 5. **A**: An unlinked PMCP with diff 2 in GO and its unlinked ITP derived; **B**: A linked PMCP with diff 2 in GO and its linked ITP derived. Evaluated by the domain experts, the unlinked PMCP in **A** is an invalid subtype relation, the linked PMCP in **B** is a valid subtype relation, and therefore the potential inconsistency in this example is a **false positive** (**FP**).

gies in UMLS) to identify supporting evidence for detected potential subtype inconsistencies, which indicates the extent to which cross-terminology can help with validating whether a detected subtype inconsistency is a missing subtype relation. We performed such automated cross-terminology evaluation for GO, NCIt and SNOMED CT, respectively.

Given a terminology, we perform a systematic check for each detected potential subtype inconsistency I. Assume that (u_1, u_2) is the unlinked PMCP involved in the inconsistency I. Then we map concepts u_1 and u_2 to the corresponding UMLS concepts m_1 and m_2. If there exists a path p from m_1 to m_2 in UMLS such that $p = m_1, m_{i_1}, m_{i_2}, \ldots, m_{i_k}, m_2$ where m_1 *is_a* m_{i_1}, m_{i_1} *is_a* m_{i_2}, …, and m_{i_k} *is_a* m_2, then we say that there is an evidence in UMLS supporting that u_1 is a subtype of u_2. Note that the subtype relations along the path may be from different terminologies. For instance, in Fig. 3, the path from C8371 *(connective tissues nevus)* to C26729 *(connective tissue disorder)* in UMLS was found through terminologies SNOMED CT and MEDCIN.

IV. Results

A. Summary Results

A total of 4,841 potential inconsistencies were found in the 03/28/2017 release of GO, 2,677 in the 07/2018 release of NCIt, and 53,782 in the 01/03/2018 release of SNOMED CT US edition, respectively (see Table I). The distribution of inconsistencies with respect to the number of word differences between concepts (diff) is also given in Table I. The majority of inconsistencies were obtained by a diff of 1.

TABLE I
Number of potential inconsistencies derived from GO, NCIt, and SNOMED CT with respect to the number of word differences between concepts ($n = 1, 2, 3, 4, 5$).

Terminology	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	Total
GO	3,527	998	243	64	9	4,841
NCIt	2,256	317	86	4	14	2,677
SNOMED CT	32,954	13,518	4,852	2,092	366	53,782

B. Evaluation

1) Evaluation by Domain Experts: Each detected inconsistency indicates a potentially missing subtype relation or an incorrect existing subtype relation (a valid inconsistency), or is a falsely identified inconsistency (an invalid inconsistency).

A random sample of 211 detected inconsistencies was reviewed by the domain experts, and 124 were found to be valid. Among the valid inconsistencies, 94 were missing subtype relations and 30 were incorrect existing subtype relations. The overall precision of the method is 58.77% (124/211).

Table II shows the distribution of the valid inconsistencies in terms of the number of word differences. For instance, the samples with 1 difference achieved a precision of 60.27% (88/146), while those with 2 differences got less precision 55.81% (24/43). The highest precision is 83.33% (5/6) for the samples with 4 differences. Table III lists 10 examples of valid inconsistencies confirmed by the domain experts.

TABLE II
Valid inconsistencies found during manual evaluation for $n = 1, 2, 3, 4, 5$ in GO.

n	Evaluation sample size	Inconsistencies (valid)	Precision
1	146	88	60.27%
2	43	24	55.81%
3	13	7	53.85%
4	6	5	83.33%
5	3	0	0%
Overall	211	124	58.77%

2) Cross-terminology Evaluation based on UMLS: The UMLS-based evaluation identified supporting evidence for missing subtype relations involved in 26 detected inconsistencies in GO, 306 in NCIt, and 1,940 in SNOMED CT, respectively. Table IV shows the distribution of the missing subtype relations identified in terms of the number of word differences. Tables V, VI and VII present terminologies in UMLS and their corresponding path contributions (PC) to identify supporting evidence for the detected potential subtype inconsistencies in GO, NCIt, and SNOMED CT, respectively. These tables contain the top 10 terminologies with the maximum path contributions. For GO, Medical Subject Headings contributed the most. For NCIt, SNOMED CT contributed the most. For SNOMED CT, Read Thesaurus contributed the most.

V. Discussion

A. Analysis of Failure Cases

The invalid inconsistencies confirmed by the domain experts are considered false positives. Fig. 5 shows an example of

TABLE III
EXAMPLES OF THE SUBTYPE INCONSISTENCIES (I: INCORRECT RELATION, M: MISSING RELATION) DETECTED IN GO VERIFIED BY DOMAIN EXPERTS.

ITP	**Unlinked PMCP**	**Linked PMCP**	**Type**
(cephalosporin, amine)	GO:0043646: cephalosporin biosynthetic process GO:0009309: amine biosynthetic process	GO:0043645: cephalosporin metabolic process GO:0009308: amine metabolic process	M
(gamma-tubulin, tubulin)	GO:1902481: gamma-tubulin complex assembly GO:0007021: tubulin complex assembly	GO:0043015: gamma-tubulin binding GO:0015631: tubulin binding	M
(fusion, morphogenesis)	GO:0046528: imaginal disc fusion GO:0007560: imaginal disc morphogenesis	GO:0035146: tube fusion GO:0035239: tube morphogenesis	I
(rRNA, RNA)	GO:1901259: chloroplast rRNA processing GO:0031425: chloroplast RNA processing	GO:0031167: rRNA methylation GO:0001510: RNA methylation	M
(nickel, inorganic)	GO:0090509: nickel cation import into cell GO:0098659: inorganic cation import into cell	GO:0035444: nickel cation transmembrane transport GO:0098662: inorganic cation transmembrane transport	M
(galactosylceramide, phospholipid)	GO:0006683: galactosylceramide catabolic process GO:0009395: phospholipid catabolic process	GO:0061591: calcium activated galactosylceramide scrambling GO:0061588: calcium activated phospholipid scrambling	M
([activin, receptor], [protein, kinase])	GO:0070697: activin receptor binding GO:0019901: protein kinase binding	GO:0048179: activin receptor complex GO:1902911: protein kinase complex	M
([dimethyl, sulfoxide], [organic, substance])	GO:1904620: cellular response to dimethyl sulfoxide GO:0071310: cellular response to organic substance	GO:0018907: dimethyl sulfoxide metabolic process GO:0071704: organic substance metabolic process	M
([systemic, acquired, resistance], [innate, immune, response])	GO:0052160: modulation by symbiont of host systemic acquired resistance GO:0052167: modulation by symbiont of host innate immune response	GO:1901672: positive regulation of systemic acquired resistance GO:0045089: positive regulation of innate immune response	I
([complement, activation, classical, pathway], [response, to, external, stimulus])	GO:0045959: negative regulation of complement activation, classical pathway GO:0032102: negative regulation of response to external stimulus	GO:0030450: regulation of complement, activation classical pathway GO:0032101: regulation of response to external stimulus	M

TABLE IV
SUPPORTING EVIDENCE OF MISSING SUBTYPE RELATIONS IDENTIFIED BY THE UMLS-BASED EVALUATION IN TERMS OF THE NUMBER OF WORD DIFFERENCES ($n = 1, 2, 3, 4, 5$).

	GO	NCIt	SNOMED CT
$n = 1$	22	249	1,502
$n = 2$	4	57	358
$n = 3$	0	0	66
$n = 4$	0	0	13
$n = 5$	0	0	1
Totals	26	306	1,940

TABLE V
TERMINOLOGIES AND CORRESPONDING PATH CONTRIBUTIONS (PC) FOR THE UMLS-BASED EVALUATION OF DETECTED SUBTYPE INCONSISTENCIES IN GO.

$n = 1$		$n = 2$	
Terminology	**PC**	Terminology	**PC**
Medical Subject Headings	15	NCIt	3
NCIt	11	CRISP Thesaurus	2
Crisp Thesaurus	11	Alcohol and Other Drug Thesaurus	1
Alcohol and Other Drug Thesaurus	7		
Foundation Model of Anatomy Ontology	6		
LOINC	5		
Thesaurus of Psychological Index Terms	5		
SNOMED CT	5		
University of Washington Digital Anatomist	5		
Read Thesaurus	5		

false positives, where the linked PMCP is correct, and the unlinked PMCP is incorrect. This is due to the existing relation in GO in Fig. 5B being a regulation of a complex pathway of two concepts which could be hierarchically related while the suggested relation in Fig. 5A being the concepts themselves which cannot be related. In scenarios such as these, the suitable relationship is *part_of* instead of *is_a*. An analogy could be made to the two concepts *Engine* and *Cylinder block*. The regulation of the *Cylinder block* may be a subclass of regulation of the *Engine*, but deriving that *Cylinder block is-a Engine* is incorrect. However, it is correct that *Cylinder block* is *part_of Engine*.

Another scenario of false positives is that the ITPs involve general terms such as (*senescence*, *development*), which may not be suitable to serve as a good candidate to detect subtype inconsistencies. An example of unlinked PMCPs is GO:0080187 (*floral organ senescence*) and GO:0048437 (*floral organ development*). Senescence is not a specific type of development, rather it is a state within the process of development and would more accurately be considered a component of development. Therefore, there should be a *part-of* relation between GO:0080187 (*floral organ senescence*) and GO:0048437 (*floral organ development*), which is already existent in the current GO.

TABLE VI
TERMINOLOGIES AND CORRESPONDING PATH CONTRIBUTIONS (PC) FOR THE UMLS-BASED EVALUATION OF DETECTED SUBTYPE INCONSISTENCIES IN NCIT.

$n = 1$		$n = 2$	
Terminology	**PC**	**Terminology**	**PC**
SNOMED CT	184	SNOMED CT	48
Read Thesaurus	86	Read Thesaurus	31
Medical Subject Headings	60	MedDRA	15
MEDCIN	57	International Classification of Diseases Related Health Problems	13
MedDRA	44	Medical Subject Headings	13
National Drug File-Reference Terminology	33	MEDCIN	10
CRISP Thesaurus	32	National Drug File-Reference Terminology	9
Alcohol and Other Drug Thesaurus	24	CRISP Thesaurus	9
COSTART	22	COSTART	9
International Classification of Diseases and Related Health Problems	17	Human Phenotype Ontology	8

B. Distinction with Related Work

In [23], Agrawal et al. leveraged lexically similar concepts in SNOMED CT with only one different word at the same position of their names to identify concept modeling inconsistencies (from the point of view of concepts). Our work is focused on detecting subtype defects in biomedical terminologies by leveraging the inconsistent ITPs derived across linked and unlinked PMCPs (from the perspective of relations). In addition, our approach does not limit the number of different words between concepts to one.

In [28], we investigated a structural-lexical approach to auditing the NCI Thesaurus, where six lexical patterns were applied to substructures called non-lattice subgraphs. Here, one of the lexical patterns leveraged inferred terms in non-lattice subgraphs to suggest potentially missing *is-a* relations in the NCI Thesaurus. In this work, we exhaustively consider all the linked and unlinked PMCPs for investigating potential inconsistencies in a given terminology without limiting to any substructure, although we employ a similar idea of lexical-based inference to [28] (concept names were represented using the set-of-words model in [28]). Moreover, this work identifies potentially incorrect existing *is-a* relations in addition to missing *is-a* relations.

We performed a preliminary study [29] on detecting potential subtype inconsistencies in GO representing concept names using the set-of-words model, which motivated this work of using the sequence-of-words model to take into consideration of orders of words. In [29], PMCPs were derived by concept-pairs having the same number of words and at least a word in common and a single different word (i.e., *diff* = 1), while in this work we allow *diff* to be any of $\{1, 2, 3, 4, 5\}$. Also, in [29], the evaluation was only performed by domain experts, while in this work in addition to the domain experts' manual evaluation, we also performed an automated evaluation based on UMLS to measure the degree to which it can help with reducing the manual evaluation effort needed. Additionally, the approach discussed in [29] was only applied to GO, while in this work, we generally apply our approach to GO, NICt and SNOMED CT. It should also be noted that the precision for $n = 1$ in this work (60.27%) is further improved than that of [29] (56.33%).

C. Limitations and Future Work

In this work, we limited the definition of PMCPs to concept-pairs having the same number of words. However, it should be noted that ITPs could be derived by any pair of concepts without such a restriction. We plan to perform such an analysis to study whether disregarding the restriction will affect the overall performance of our approach.

Another limitation of this work is that we did not take into account the granularity of the inferred term pairs when generating potential inconsistencies. It would be useful to leverage some weight functions to inferred term pairs so that more general terms are given a less weight and more specific terms are given a higher weight when generating potential inconsistencies. The weight function may also consider the frequency of occurrence of the ITP in the current terminology (as a linked ITP) where a higher frequency would give it more prominence. We expect such a strategy would enable us to reduce the number of false positives.

Additionally, we only performed automated evaluation by leveraging external knowledge in UMLS which showed limited supporting evidence (Table IV): 0.54% (=26/4841) for GO, 11.43% (=306/2677) for NCIt, and 3.61% (=1940/53782) for SNOMED CT. It would be interesting to further investigate methods to leverage other external knowledge such as biomedical literature to automatically identify supporting evidence for detected potential inconsistencies and reduce domain experts' manual effort.

VI. CONCLUSION

In this paper, we investigated a lexical-based inference approach to audit biomedical terminologies based on the inconsistencies of inferred term-pairs derived from linked and unlinked concept-pairs. We applied this approach to GO, NCIt and SNOMED CT respectively to detect potential subtype inconsistencies. From the evaluation performed by domain experts, our approach achieved an overall precision of 58.77% in detecting valid subtype inconsistencies in GO. This is a large enrichment of inconsistencies in comparison to the low rate of inconsistencies expected across the ontology. We also performed a preliminary study on a UMLS-based method to automatically identify supporting evidence of missing subtype relations to understand to what extent the external knowledge in UMLS can help with reducing the manual evaluation effort required from domain experts. The results demonstrated that the lexical-based inference approach is a promising way to detect potential subtype inconsistencies, which indicate missing subtype relations as well as incorrect subtype relations. This approach is also applicable to other biomedical terminologies for quality assurance analysis.

REFERENCES

[1] D. Lee, N. de Keizer, F. Lau, and R. Cornet, "Literature review of SNOMED CT use," *Journal of the American Medical Informatics Association*, vol. 21, no. e1, pp. e11–e19, 2013.

TABLE VII

TERMINOLOGIES AND CORRESPONDING PATH CONTRIBUTIONS (PC) FOR THE UMLS-BASED EVALUATION OF DETECTED SUBTYPE INCONSISTENCIES IN SNOMED CT.

$n = 1$		$n = 2$		$n = 3$		$n = 4$		$n = 5$	
Terminology	**PC**	**Terminology**	**PC**	**Terminology**	**PC**	**Terminology**	**PC**	**Terminology**	**PC**
Read Thesaurus	954	Read Thesaurus	283	Read Thesaurus	49	Read Thesaurus	13	Read Thesaurus	1
MEDCIN	323	MEDCIN	69	MEDCIN	19	Medical Subject Headings	4		
NCIt	291	Medical Subject Headings	63	Foundational Model of Anatomy Ontology	17	NCIt	3		
Medical Subject Headings	257	NCIt	61	University of Washington Digital Anatomist	17	National Drug File - Reference Terminology	3		
CRISP Thesaurus	164	Alcohol and Other Drug Thesaurus	43	NCIt	13	CRISP Thesaurus	2		
Alcohol and Other Drug Thesaurus	138	CRISP Thesaurus	40	International Classification of Diseases and Related Health Problems	10	University of Washington Digital Anatomist	1		
National Drug File - Reference Terminology	109	National Drug File - Reference Terminology	30	Medical Subject Headings	10	MEDCIN	1		
International Classification of Diseases and Related Health Problems	85	University of Washington Digital Anatomist	29	Human Phenotype Ontology	4				
Foundational Model of Anatomy Ontology	84	Foundational Model of Anatomy Ontology	27	National Drug File - Reference Terminology	3				
MedDRA	77	International Classification of Diseases and Related Health Problems	21	MedlinePlus Health Topics	2				

[2] O. Bodenreider, "Biomedical ontologies in action: role in knowledge management, data integration and decision support," *Yearbook of medical informatics*, p. 67, 2008.

[3] G. L. Holliday, R. Davidson, E. Akiva, and P. C. Babbitt, "Evaluating functional annotations of enzymes using the gene ontology," *The Gene Ontology Handbook*, pp. 111–132, 2017.

[4] G.-Q. Zhang, L. Cui, S. Lhatoo, S. U. Schuele, and S. S. Sahoo, "MEDCIS: multi-modality epilepsy data capture and integration system," in *AMIA Annual Symposium Proceedings*, vol. 2014. American Medical Informatics Association, 2014, p. 1248.

[5] L. Cui, A. Bozorgi, S. D. Lhatoo, G.-Q. Zhang, and S. S. Sahoo, "EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification," in *AMIA Annual Symposium Proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 1191.

[6] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene Ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, p. 25, 2000.

[7] G. O. Consortium *et al.*, "The gene ontology (GO) project in 2006," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D322–D326, 2006.

[8] G. O. Consortium, "Expansion of the gene ontology knowledgebase and resources," *Nucleic acids research*, vol. 45, no. D1, pp. D331–D338, 2016.

[9] N. Sioutos, S. de Coronado, M. W. Haber, F. W. Hartel, W.-L. Shaiu, and L. W. Wright, "Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information," *Journal of biomedical informatics*, vol. 40, no. 1, pp. 30–43, 2007.

[10] K. Donnelly, "SNOMED-CT: The advanced terminology and coding system for eHealth," *Studies in health technology and informatics*, vol. 121, p. 279, 2006.

[11] (2018, Sep.) Snomed international. [Online]. Available: https://www.snomed.org/snomed-ct

[12] (2017, Dec.) Gene ontology consortium - documentation. [Online]. Available: http://www.geneontology.org/page/documentation

[13] R. W. Francis, "Golink: finding cooccurring terms across gene ontology namespaces," *International journal of genomics*, vol. 2013, 2013.

[14] (2018, Sep.) Nci thesaurus (ncit). [Online]. Available: https://wiki.nci.nih.gov/pages/viewpage.action?pageId=7472532

[15] S. De Coronado, M. W. Haber, N. Sioutos, M. S. Tuttle, L. W. Wright *et al.*, "Nci thesaurus: using science-based terminology to integrate cancer research results." in *Medinfo*, 2004, pp. 33–37.

[16] S. de Coronado, L. W. Wright, G. Fragoso, M. W. Haber, E. A. Hahn-Dantona, F. W. Hartel, S. L. Quan, T. Safran, N. Thomas, and L. Whiteman, "The nci thesaurus quality assurance life cycle," *Journal of biomedical informatics*, vol. 42, no. 3, pp. 530–539, 2009.

[17] (2018, Sep.) Snomed ct basics. [Online]. Available: https://confluence.ihtsdotools.org/display/DOCSTART/4.+SNOMED+CT+Basics

[18] Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl1, pp. D267–D270, 2004.

[19] (2018, Sep.) Statistics - 2018aa release. [Online]. Available: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html

[20] X. Zhu, J.-W. Fan, D. M. Baorto, C. Weng, and J. J. Cimino, "A review of auditing methods applied to the content of controlled biomedical terminologies," *Journal of biomedical informatics*, vol. 42, no. 3, pp. 413–425, 2009.

[21] M. F. Amith, Z. He, J. Bian, J. A. Lossio-Ventura, and C. Tao, "Assessing the practice of biomedical ontology evaluation: Gaps and opportunities," *Journal of biomedical informatics*, 2018.

[22] C. Ochs, Y. Perl, M. Halper, J. Geller, and J. Lomax, "Quality assurance of the gene ontology using abstraction networks," *Journal of bioinformatics and computational biology*, vol. 14, no. 03, p. 1642001, 2016.

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

[24] K. Verspoor, D. Dvorkin, K. B. Cohen, and L. Hunter, "Ontology quality assurance through analysis of term transformations," *Bioinformatics*, vol. 25, no. 12, pp. i77–i84, 2009.

[25] G.-Q. Zhang and O. Bodenreider, "Large-scale, exhaustive lattice-based structural auditing of SNOMED CT," in *AMIA Annual Symposium Proceedings*, vol. 2010. American Medical Informatics Association, 2010, p. 922.

[26] L. Cui, S. Tao, and G.-Q. Zhang, "Biomedical ontology quality assurance using a big data approach," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 10, no. 4, p. 41, 2016.

[27] L. Cui, W. Zhu, S. Tao, J. T. Case, O. Bodenreider, and G.-Q. Zhang, "Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT," *Journal of the American Medical Informatics Association*, vol. 24, no. 4, pp. 788–798, 2017.

[28] R. Abeysinghe, M. A. Brooks, J. Talbert, and C. Licong, "Quality assurance of NCI Thesaurus by mining structural-lexical patterns," in *AMIA Annual Symposium Proceedings*, vol. 2017. American Medical Informatics Association, 2017, p. 364.

[29] R. Abeysinghe, E. W. Hinderer, H. N. Moseley, and L. Cui, "Auditing subtype inconsistencies among gene ontology concepts," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017, pp. 1242–1245.

[30] D. Mossialos, J.-M. Meyer, H. Budzikiewicz, U. Wolff, N. Koedam, C. Baysse, V. Anjaiah, and P. Cornelis, "Quinolobactin, a new siderophore of pseudomonas fluorescens atcc 17400, the production of which is repressed by the cognate pyoverdine," *Applied and environmental Microbiology*, vol. 66, no. 2, pp. 487–492, 2000.