

Leveraging Logical Definitions and Lexical Features to Detect Missing IS-A Relations in Biomedical Terminologies

Rashmie Abeysinghe^{1†}, Fengbo Zheng^{2†}, Jay Shi³,
Samden Lhatoo¹, Licong Cui^{2*}

¹Department of Neurology, The University of Texas Health Science
Center at Houston, Houston, TX, USA.

²D. Bradley McWilliams School of Biomedical Informatics, The
University of Texas Health Science Center at Houston, Houston, TX,
USA.

³Intermountain Healthcare, Denver, CO, USA.

*Corresponding author(s). E-mail(s): licong.cui@uth.tmc.edu;

Contributing authors: rashmie.abeyasinghe@uth.tmc.edu;
fzh229@gmail.com; jayshi88@gmail.com; samden.d.lhatoo@uth.tmc.edu;

[†]These authors contributed equally to this work.

Abstract

Biomedical terminologies play a vital role in managing biomedical data. Missing IS-A relations in a biomedical terminology could be detrimental to its downstream usages. In this paper, we investigate an approach combining logical definitions and lexical features to discover missing IS-A relations in two biomedical terminologies: SNOMED CT and the National Cancer Institute (NCI) thesaurus. The method is applied to unrelated concept-pairs within non-lattice subgraphs: graph fragments within a terminology likely to contain various inconsistencies. Our approach first compares whether the logical definitions of a concept are more general than the other concept. Then, we check whether the lexical features of the concept are contained in the other concept. If both constraints are satisfied, we suggest a potentially missing IS-A relation between the two concepts. The method identified 982 potential missing IS-A relations for SNOMED CT and

100 for NCI thesaurus. In order to assess the efficacy of our approach, a random sample of results belonging to the “Clinical Findings” and “Procedure” subhierarchies of SNOMED CT and results belonging to the “Drug, Food, Chemical or Biomedical Material” subhierarchy of the NCI thesaurus were evaluated by domain experts. The evaluation results revealed that 118 out of 150 suggestions are valid for SNOMED CT and 17 out of 20 are valid for NCI thesaurus.

Keywords: SNOMED CT, NCI thesaurus, Terminology Quality Assurance

1 Introduction

Throughout the years, biomedical terminologies have played a significant role in biomedical research and applications, especially in facilitating data management. Two such leading biomedical terminologies are SNOMED CT and National Cancer Institute thesaurus (NCIt). SNOMED CT is the world’s largest clinical terminology, which is a standard for facilitating the exchange of clinical health information [1]. NCI thesaurus is a reference terminology that facilitates translational research in cancers [2].

Many modern biomedical terminologies including SNOMED CT and NCIt have been formally represented using description logics (DL), a family of formal knowledge representation languages. A key reasoning service provided by DL is ontology classification, achieved by DL reasoners (e.g., ELK [3], Snorocket [4]), which can check the consistency of definitions across the whole ontology and automatically infer a hierarchy of concepts (i.e., infer IS-A hierarchical relations among concepts) based on the stated facts.

In both SNOMED CT and NCIt, concepts are logically defined with hierarchical and attribute relations [5, 6]. The curators associate each concept with a stated definition consisting of description logic axioms based on the current knowledge about that concept. Then, a description logic classifier is applied to the stated definitions to generate inferred logical axioms [7]. For instance, Figure 1 shows the inferred logical definitions of two SNOMED CT concepts: “*Malignant neoplasm of peripheral nerve of*

abdomen (disorder)” and “*Neoplasm of peripheral nerves of abdomen (disorder)*”. In SNOMED CT, some relations are grouped into relation groups if they are associated with each other [8].

In SNOMED CT and NCIt, a concept is considered to be fully defined if its definition is sufficient to distinguish its meaning from other similar concepts [6, 9]. Otherwise, its definition status is primitive. If concept A is fully defined, the DL reasoners will identify concepts whose definitions satisfy A ’s defining relations (i.e., whose definitions are more detailed/specific) to be the subtypes of concept A . On the other hand, if a concept is primitive, the DL reasoners will not infer any subtypes for it.

The definition status (i.e. fully defined or primitive) of individual concepts is usually decided by the curators of the terminology. Therefore, valid hierarchical relations among concepts may not be captured by the DL reasoners due to the primitive definition status of the potential supertypes. For instance, in the March 2020 release of the SNOMED CT (US Edition), the concept “*Neoplasm of peripheral nerves of abdomen (disorder)*” is a primitive concept. As shown in Figure 2, the definition of the concept “*Benign ganglioneuroma of abdomen (disorder)*” is more specific than this concept. This is because the corresponding attribute-value pairs (that are shown on the same level) are either the same or more specific in “*Benign ganglioneuroma of abdomen (disorder)*”. Similarly, as shown in Figure 1, the concept “*Neoplasm of peripheral nerves of abdomen (disorder)*” is a primitive concept, and the definition of the concept “*Malignant neoplasm of peripheral nerve of abdomen (disorder)*” is more specific than this. However, still, DL reasoners will not establish a hierarchical relation between these two concepts (i.e., a missing hierarchical relation) as “*Neoplasm of peripheral nerves of abdomen (disorder)*” is primitive. Note that in the March 2021 Release of the SNOMED CT (US Edition) this hierarchical relation exists as “*Neoplasm of peripheral nerves of abdomen (disorder)*” became fully defined and hence the relation became derivable by DL reasoners.

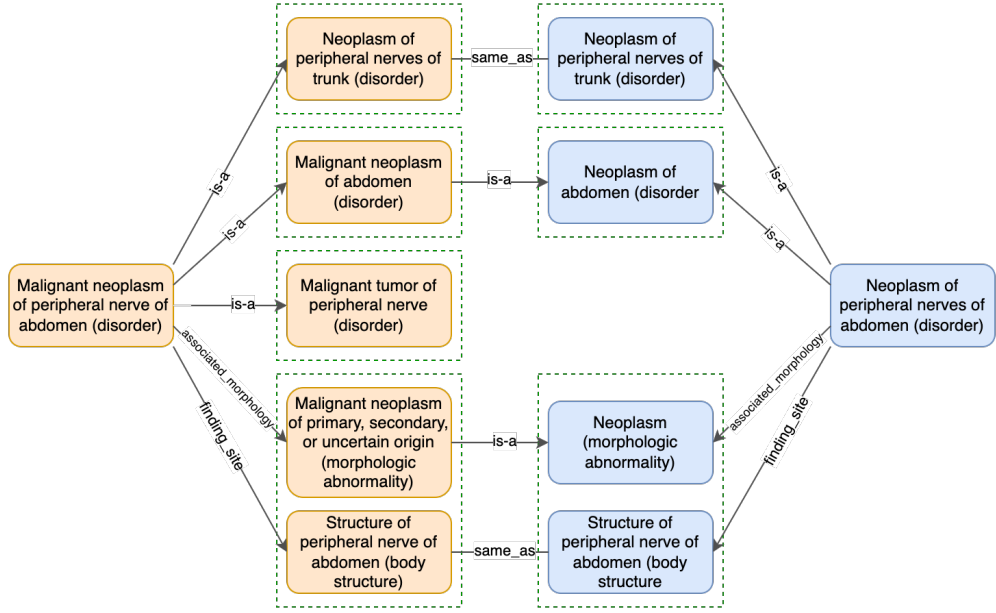


Fig. 1 Comparison of inferred definitions of “*Malignant neoplasm of peripheral nerve of abdomen (disorder)*” (188326001) and “*Neoplasm of peripheral nerves of abdomen (disorder)*” (126992002) in the March 2020 Release of the SNOMED CT (US Edition) [10]. Relationship groups are indicated with dashed lines in green.

Our goal in this paper is to identify such potentially missing hierarchical relations that the DL reasoners missed (i.e., in which the potential subconcepts are more specific than the superconcepts in terms of logical definitions, but the superconcepts are primitive). To achieve this, we first identify candidate pairs of concepts from non-lattice subgraphs which often contain quality issues including missing hierarchical relations. Then, given a candidate pair, we check if the inferred logical definitions of one concept are more specific than those of the other. If so, the potential superconcept should be a primitive concept (otherwise, these hierarchical relations should have been inferred by the DL reasoners) and, there may be a missing hierarchical relation between these two concepts. Since the superconcept is of primitive definition status (i.e., the logical definitions may be insufficient to express its semantic meanings), purely relying on logical definitions may lead to erroneous missing hierarchical relations being suggested.

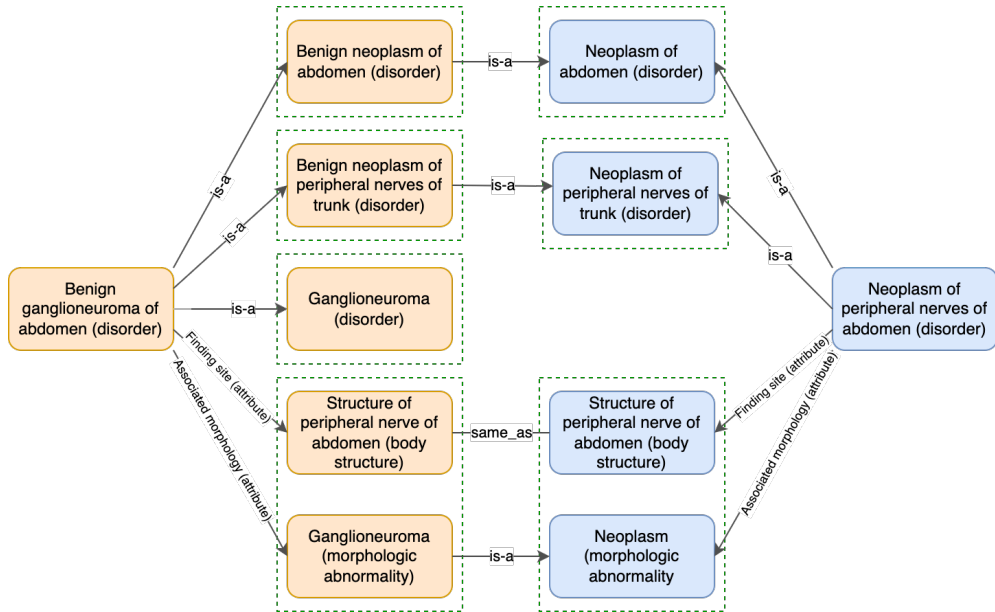


Fig. 2 Comparison of inferred definitions of “*Benign ganglioneuroma of abdomen (disorder)*” (426134002) and “*Neoplasm of peripheral nerves of abdomen (disorder)*” (126992002) in the March 2020 Release of the SNOMED CT (US Edition) [10]. Relationship groups are indicated with dashed lines in green.

Therefore, in this paper, we also utilize lexical features of concepts as supplementary to determine the subsumption relations among concepts.

Throughout the years, there has been considerable exploration of various approaches to identify and address different quality issues including missing hierarchical (IS-A) relations within biomedical terminologies [11]. For instance, Bodenreider has come up with an approach to generate logical definitions of SNOMED CT concepts by lexical features in concept labels. Reasoning on these logical definitions has revealed missing hierarchical relations in SNOMED CT [12]. Graph summarization techniques (called abstraction networks) have been extensively utilized to uncover various modeling issues within biomedical terminologies [13–16]. Abstraction networks summarize the terminology structure and various characteristics of such networks have been investigated to address different quality issues. Agrawal et al. have explored

different approaches to identify concepts that are lexically similar and should be modeled in a similar manner. Inconsistent modeling among such groups of concepts has led to the identification of errors [17–19]. Liu et al. have explored deep learning to suggest missing IS-A relations in NCIIt [20, 21]. Their strategy involves training a Convolutional Neural Network with existing relations as positive samples and uncle-nephew pairs as negative samples. Concept features to train the model are obtained through documents containing concept lexical and hierarchical information. In previous work, we investigated training a Graph Neural Network to predict missing IS-A relations within the Clinical findings subhierarchy of SNOMED CT [22]. We utilized four types of features to train the model: (1) concept name features; (2) hierarchical features; (3) enriched lexical attribute features; and (4) logical definition features. A cross-validation-inspired approach was used to apply the model to all hierarchically unrelated concept pairs. In previous work, we have also proposed several approaches that uncover missing IS-A relations purely utilizing lexical features of concepts [23–29], and approaches that combine lexical and structural features [23, 24, 30]. A more detailed comparison with such approaches that are related to this work is provided later in the paper in the Discussion section.

2 Methods

There are mainly four steps in our method: (1) pre-compute non-lattice subgraphs and identify candidate pairs of concepts that are currently not linked by hierarchical relations; (2) given a candidate pair, check if the inferred definitions of one concept are more specific than the other’s; (3) compute lexical features for concepts and perform lexical-based subsumption checking; (4) remove redundant and cycle-causing potentially missing hierarchical relations.

2.1 Pre-computing non-lattice subgraphs and generating	277
candidate pairs	278
	279
	280
In our previous work [23, 30–32], we found that non-lattice subgraphs often reveal	281
quality issues such as missing hierarchical relations or missing concepts. Non-lattice	282
subgraphs are graph fragments obtained from hierarchical (or IS-A) relations of an	283
ontology. A pair of concepts is known as a non-lattice pair if they share more than one	284
maximal common descendant. A non-lattice subgraph can be obtained from a non-	285
lattice pair by first reversely computing the minimal common ancestors of the maximal	286
common descendants of the non-lattice pair and then aggregating all the concepts and	287
hierarchical relations between them [30]. Figure 3 shows a non-lattice subgraph in the	288
March 2020 Release of the SNOMED CT (US Edition) obtained from non-lattice pair:	289
(<i>“Neoplasm of peripheral nerves of trunk (disorder),” “Neoplasm of abdomen (disor-</i>	290
<i>der)”</i>) with three maximal common descendants <i>“Malignant neoplasm of peripheral</i>	291
<i>nerve of abdomen (disorder),” “Benigh ganglioneuroma of abdomen (disorder),”</i> and	292
<i>“Neoplasm of peripheral nerves of abdomen (disorder).”</i> Similarly, Figure 4 shows a	293
non-lattice subgraph in the 23.05e release of NCIt that contains the non-lattice pair:	294
(<i>“EGFR-targeting Agent,” “Bispecific Monoclonal Antibody”</i>) and five of its maximal	295
common descendants.	296
	297
	298
	299
	300
	301
	302
	303
	304
	305
	306
	307
In this work, we first compute all the non-lattice subgraphs using an efficient non-	308
lattice extraction algorithm [33]. Then we generate a list of candidate concept pairs	309
which are concepts that are currently not linked by hierarchical relations in non-lattice	310
subgraphs. Consider the SNOMED CT non-lattice subgraph shown in Figure 3. Two	311
example candidate pairs are (<i>“Malignant neoplasm of peripheral nerve of abdomen</i>	312
<i>(disorder),” “Neoplasm of peripheral nerves of abdomen (disorder)”</i>) and (<i>“Benigh</i>	313
<i>ganglioneuroma of abdomen (disorder),” “Neoplasm of peripheral nerves of abdomen</i>	314
<i>(disorder)”</i>). In the NCIt non-lattice subgraph in Figure 4, two example candidate	315
	316
	317
	318
	319
	320
	321
	322

323 pairs are (“*Amivantamab and Recombinant Human Hyaluronidase*,” “*Amivantamab*”)
 324 and (“*EGFR-targeting Agent*,” “*Bispecific Monoclonal Antibody*”).
 325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

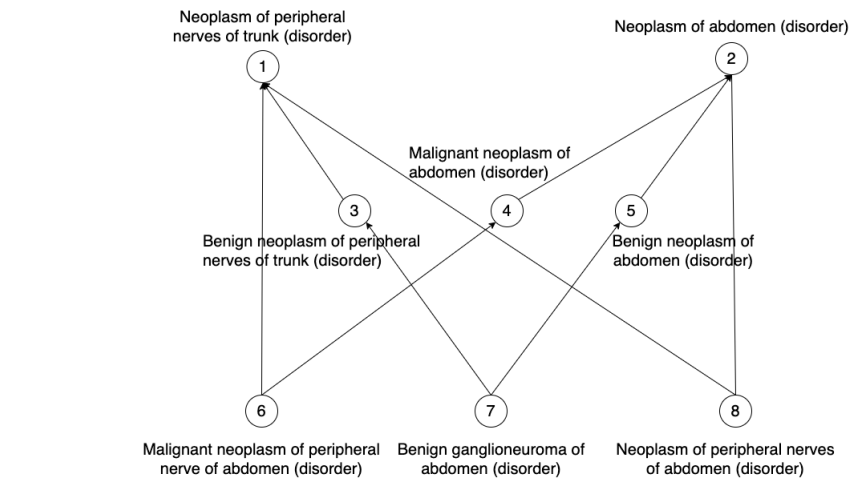


Fig. 3 An example of non-lattice subgraphs in the March 2020 Release of the SNOMED CT (US Edition). Concepts are connected by hierarchical relations.

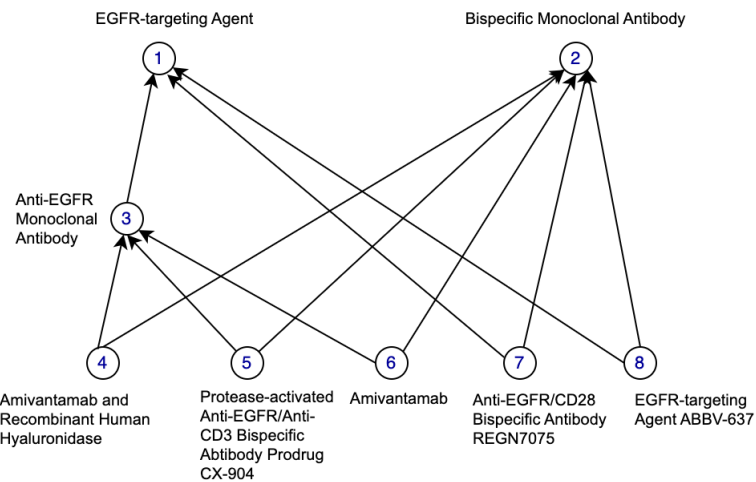


Fig. 4 An example of non-lattice subgraphs in the 23.05e Release of NCIt. Concepts are connected by hierarchical relations.

2.2 Logical definition-based subsumption checking

In this step, given a candidate pair, we check whether the logical definitions of one concept are more general than the other. We perform this comparison at the relation group level. Note that some relations such as IS-A relations in Figure 1, can be ungrouped in SNOMED CT. We consider each of these relations to be in a separate group. In addition, NCIt does not group relations as SNOMED CT does. Therefore, we also consider each relation in NCIt concepts to be in its own relation group to generalize the method's implementation.

Based on relation groups, given a concept X , we consider its logical definition (inferred) as a set of groups of defining relations, $I_X = \{X_n \mid n = 1, \dots, i\}$, where X_n is a group of relations in the form of attribute-value pair(s), i.e., $X_n = \{(k_{nm} : v_{nm}) \mid m = 1, \dots, j\}$. For example, the logical definition of the SNOMED CT concept “*Neoplasm of peripheral nerves of abdomen (disorder)*” in Figure 1 consists of three relation groups $\{X_1, X_2, X_3\}$, where $X_1 = \{(Is\ a: Neoplasm\ of\ peripheral\ nerves\ of\ trunk\ (disorder))\}$, $X_2 = \{(Is\ a: Neoplasm\ of\ abdomen\ (disorder))\}$, and $X_3 = \{(Associated\ morphology: Neoplasm\ (morphologic\ abnormality)), (Finding\ site: Structure\ of\ peripheral\ nerve\ of\ abdomen\ (body\ structure))\}$. Note that X_3 contains two relations while X_1 and X_2 contain one relation each.

Given a candidate pair (X, Y) , I_X is considered to be more specific than I_Y in logical definitions if, for each relation group Y_m in I_Y , there exists a corresponding group X_n in I_X such that X_n is more specific than Y_m . Given two relation groups, X_n is considered to be more specific than Y_m , if for each defining relation (k_Y, v_Y) in Y_m , there exists a corresponding defining relation (k_X, v_X) in X_n such that (k_X, v_X) is more specific than (k_Y, v_Y) . The following two rules are followed to determine whether a defining relation is more specific than another.

The first rule is the inclusion rule which covers most cases. Given two defining relations (k_X, v_X) and (k_Y, v_Y) , (k_X, v_X) is more specific than (k_Y, v_Y) if k_X is the same

415 as or a subtype (i.e., descendant) of k_Y , and v_X is the same as or a subtype (i.e., descen-
 416 dant) of v_Y . Consider the candidate pair in Figure 1. For each relation group in the
 417 inferred definition of concept “*Neoplasm of peripheral nerves of abdomen (disorder)*,”
 418 we could find a corresponding group in the inferred definition of “*Malignant neoplasm*
 419 *of peripheral nerve of abdomen (disorder)*” which is more specific. For example, the
 420 relation groups at the bottom of Figure 1 both contain two relations. The relation
 421 (*Finding site: Structure of peripheral nerve of abdomen (body structure)*) exists under
 422 both concepts. In the other relation, the attribute type “*Associated morphology*” is the
 423 same for both the concepts while the value concept “*Malignant neoplasm of primary,*
 424 *secondary, or uncertain origin (morphologic abnormality)*” is a subtype of “*Neoplasm*
 425 *(morphologic abnormality)*”. As a result, based on their logical definitions “*Malignant*
 426 *neoplasm of peripheral nerve of abdomen (disorder)*” is considered to be more specific
 427 than “*Neoplasm of peripheral nerves of abdomen (disorder)*”.

437 The second rule is the property chains, which include transitive properties. Given
 438 attribute types k_a , k_b and k_Y with a property chain $k_a \circ k_b$ is a sub-property of k_Y ,
 439 defining relation (k_X, v_X) is more specific than (k_Y, v_Y) if attribute type k_X is the
 440 same as or a subtype of k_a , and v_X has a relation to v_Y via attribute type k_b . Con-
 441 sider the SNOMED CT defining relations (*Causative agent: Sodium calcium edetate*
 442 *(substance)*) from concept “*Sodium calcium edetate adverse reaction (disorder)*” and
 443 (*Causative agent: Edetate (substance)*) from concept “*Edetate adverse reaction (disor-*
 444 *der)*.” Here, the value concept “*Sodium calcium edetate (substance)*” is not a subtype of
 445 “*Edetate (substance)*”. However, “*Sodium calcium edetate (substance)*” has a relation
 446 whose attribute type is “*Is modification of*” to “*Edetate (substance)*,” and property
 447 chain of *Causative agent* $\circ$ *Is modification of* is a sub-property of *Causative agent*.
 448 Substituting to the second rule, k_a and k_Y equal to “*Causative agent*,” k_b equals
 449 to “*Is modification of*.” In this case, k_X equals to k_a (i.e., “*Causative agent*”), and
 450
 451
 452
 453
 454
 455
 456
 457
 458
 459
 460

value v_X “*Sodium calcium edetate (substance)*” has a relation to v_Y “*Edetate (substance)*” via k_b “*Is modification of*”. As a result, defining relation (*Causative agent: Sodium calcium edetate (substance)*) is more specific than relation (*Causative agent: Edetate (substance)*) even though they do not comply with the first inclusion rule. In the September 2021 Release of the SNOMED CT (US Edition), all the property chains have attribute type “*Is modification of*” as intermediate property (i.e., $k_b =$ “*Is modification of*”).

In some concepts, the inferred definitions may not contain any attribute relations only containing hierarchical relations. In such cases, we only have limited definitions for the potential supertype, and it could be meaningless to find its potential subtypes considering logical definitions. To improve the quality of suggested missing hierarchical relations, we only consider those candidate pairs where the potential supertype contains at least one attribute relation.

2.3 Supplementary lexical-based subsumption checking

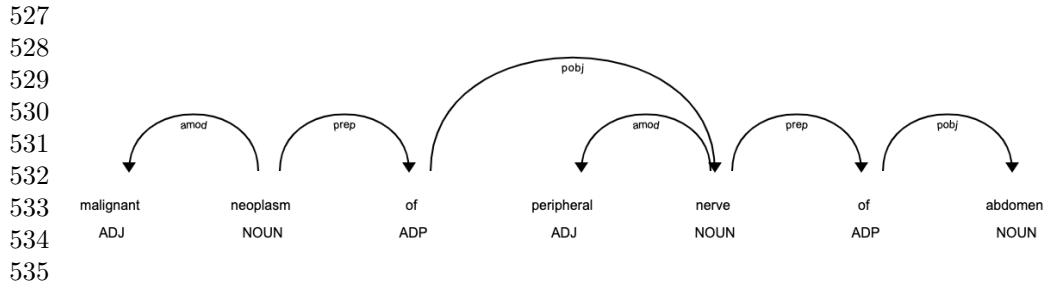
In our previous work [23, 25, 32], we found that lexical features (e.g., words and noun phrases appearing in the concept names) can be used to represent the semantic meaning of concepts. These lexical features may include information that is not conveyed through logical definitions and can be taken as supplementary features in representing the semantic meaning of concepts. In this work, we aggregate three types of lexical features from a concept name to form a lexical feature set for each concept: (1) dependency pairs of two dependencies: object of a preposition “*pobj*” and direct object “*dobj*”; (2) base noun phrases; and (3) single words that were not in dependency pairs.

Given a concept name, we first use Spacy [34], a Natural Language Processing (NLP) library, to perform dependency parsing. Figure 5 shows the dependency parse of the SNOMED CT concept “*Malignant neoplasm of peripheral nerve of abdomen (disorder)*.” As shown, the first occurrence of the word “of” and the word

507 “nerve” has “*pobj*” dependency. Also, the second occurrence of the word “of” and
 508 the word “abdomen” also has “*pobj*” dependency. Therefore, we include of nerve and
 509 of abdomen as dependency pairs in the lexical feature set.
 510

512 Afterward, using Spacy, all the base noun phrases existing in a concept name are
 513 identified and aggregated to the lexical feature set. For instance, the SNOMED CT
 514 concept “*Malignant neoplasm of peripheral nerve of abdomen (disorder)*” contains base
 515 noun phrases: malignant neoplasm, peripheral nerve, and abdomen.
 516

518 Finally, the rest of the words that are not part of the dependency pairs are
 519 aggregated into the lexical feature set. For instance, in the SNOMED CT concept
 520 “*Malignant neoplasm of peripheral nerve of abdomen (disorder)*”, the words ‘malig-
 521 nant,’ ‘neoplasm,’ and ‘peripheral’ are not part of the dependency pairs of nerve and
 522 of abdomen. Therefore, these words are aggregated to the lexical feature set.
 523
 524
 525
 526



536 **Fig. 5** Dependency parsing result for concpet name “*Malignant neoplasm of peripheral nerve of*
 537 *abdomen (disorder)*.” The semantic tag “(disorder)” is not parsed and will not be included in the
 538 lexical feature set of this concept.
 539

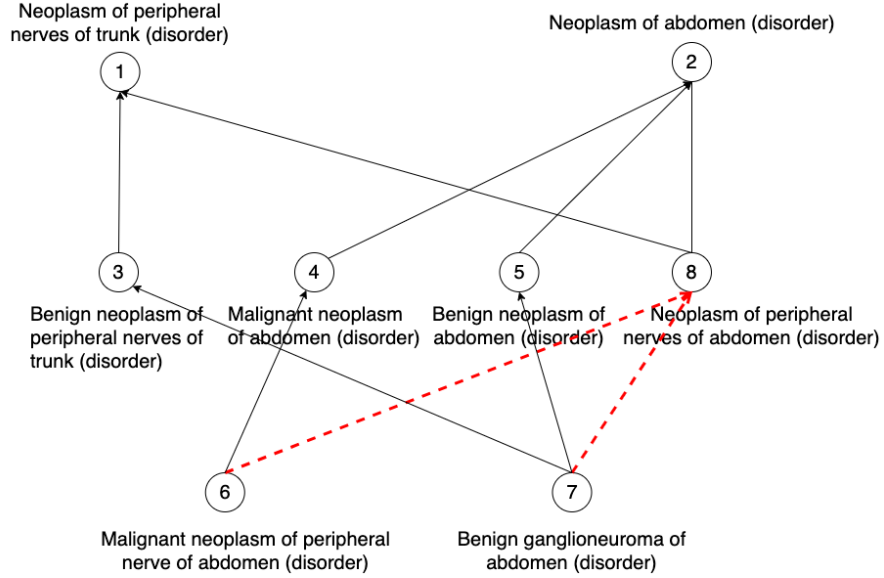
540
 541 To obtain a broader view of the semantics of a concept, we further construct an
 542 enriched set of lexical features by leveraging its ancestors. The lexical features for each
 543 ancestor is computed and aggregated to the concept’s lexical feature set to generate
 544 the enriched lexical feature set. Table 1 shows the initial lexical feature set and the
 545 enriched lexical feature set for the SNOMED CT concept “*Malignant neoplasm of*
 546 *peripheral nerve of abdomen (disorder)*.”
 547
 548
 549
 550
 551
 552

Table 1 The initial and enriched sets of lexical features of concept “*Malignant neoplasm of peripheral nerve of abdomen (disorder)*.” Noun phrases and dependency pairs are underlined.

Concept Name	Malignant neoplasm of peripheral nerve of abdomen (disorder)
Initial lexical feature set	{ <u>of nerve</u> , <u>of abdomen</u> , <u>malignant</u> , <u>neoplasm</u> , <u>peripheral</u> , <u>malignant neoplasm</u> , <u>peripheral nerve</u> }
Enriched lexical feature set	{ <u>clinical finding</u> , <u>soft tissue lesion</u> , <u>malignant neoplasm</u> , <u>trunk</u> , <u>abdominopelvic segment</u> , <u>peripheral nerve disease</u> , <u>body</u> , <u>neoplasm</u> , <u>snomed concept</u> , <u>ct concept</u> , <u>mass</u> , <u>peripheral nerve</u> , <u>of region</u> , <u>of system</u> , <u>abdominal mass</u> , <u>of abdomen</u> , <u>of tissue</u> , <u>general</u> , <u>of nerves</u> , <u>body region</u> , <u>of trunk</u> , <u>nervous</u> , <u>peripheral nerves</u> , <u>nerve</u> , <u>peripheral nerve finding</u> , <u>malignant neoplastic disease</u> , <u>the peripheral nervous system</u> , <u>soft</u> , <u>ct</u> , <u>of segment</u> , <u>neoplastic disease</u> , <u>trunk structure</u> , <u>neoplastic</u> , <u>soft tissue</u> , <u>disorder</u> , <u>tumor</u> , <u>neurological</u> , <u>of structure</u> , <u>body site</u> , <u>clinical</u> , <u>neuropathy</u> , <u>malignant tumor</u> , <u>nervous system</u> , <u>abdominopelvic</u> , <u>peripheral</u> , <u>of nerve</u> , <u>hamartoma</u> , <u>neurological lesion</u> , <u>tissue</u> , <u>body system</u> , <u>trunk nerve lesion</u> , <u>malignant</u> , <u>abdominal</u> , <u>the</u> , <u>and/or</u> , <u>lesion</u> , <u>general finding</u> , <u>disease</u> , <u>body structure</u> , <u>space-occupying lesion</u> , <u>by site</u> , <u>finding</u> }

Given a candidate pair (X, Y) , if X is more specific than Y in terms of logical definitions, we further check whether the enriched lexical feature set of X is a superset of Y 's (i.e. if concept X is also lexical-wise more specific than Y). If so, a potentially missing hierarchical relation X IS-A Y is discovered. Consider the candidate pair (“*Malignant neoplasm of peripheral nerve of abdomen (disorder)*,” “*Neoplasm of peripheral nerves of abdomen (disorder)*”) in the SNOMED CT non-lattice subgraph in Figure 3 as an example. “*Malignant neoplasm of peripheral nerve of abdomen (disorder)*” is more specific than “*Neoplasm of peripheral nerve of abdomen (disorder)*” both in logical definitions and lexical features, and therefore, a potentially missing hierarchical relation “*Malignant neoplasm of peripheral nerve of abdomen (disorder)*” IS-A “*Neoplasm of peripheral nerve of abdomen (disorder)*” is suggested by our method. Note that our approach also found another missing IS-A relation in this particular non-lattice subgraph: “*Benign ganglioneuroma of abdomen (disorder)* IS-A *Neoplasm of peripheral nerves of abdomen (disorder)*.” Both the missing IS-A relations are shown in Figure 6. Similarly, the NCIt candidate-pair (“*Amivantamab and Recombinant Human Hyaluronidase*,” “*Amivantamab*”) in the NCIt non-lattice subgraph in Figure 4 satisfies both these logical and lexical conditions. Therefore, a potential

599 missing IS-A relation “*Amivantamab and Recombinant Human Hyaluronidase*” IS-A
600 “*Amivantamab*” is suggested between these two concepts. This missing IS-A relation
601 is shown in Figure 7.
602
603



623 **Fig. 6** Two potentially missing hierarchical relations identified (marked red) by our methods in
624 the SNOMED non-lattice subgraph shown in Figure 3. Note that the original direct hierarchical
625 relation between “*Malignant neoplasm of peripheral nerve of abdomen (disorder)*” and “*Neoplasm of*
626 *peripheral nerves of trunk (disorder)*” is removed because it can now be transitively inferred by the
627 potential missing hierarchical relation and the existing hierarchical relation.

630 2.4 Redundancy and cycle removal

631
632 Some of the potential missing IS-A suggested by our method might be implied by
633 other potential missing IS-A relations and existing IS-A relations. For example, our
634 approach may suggest two potentially missing hierarchical relations A IS-A B and A
635 IS-A C . If C is an ancestor of B in the original concept hierarchy of SNOMED CT, A
636 IS-A C will be considered redundant as it can be implied transitively by potentially
637 missing hierarchical relation A IS-A B and existing IS-A relation B IS-A C . Such
638 redundant potential missing IS-A relations are removed from the list of discovered
639 potential missing IS-A relations. For each potential missing IS-A relation, we combine
640
641
642
643
644

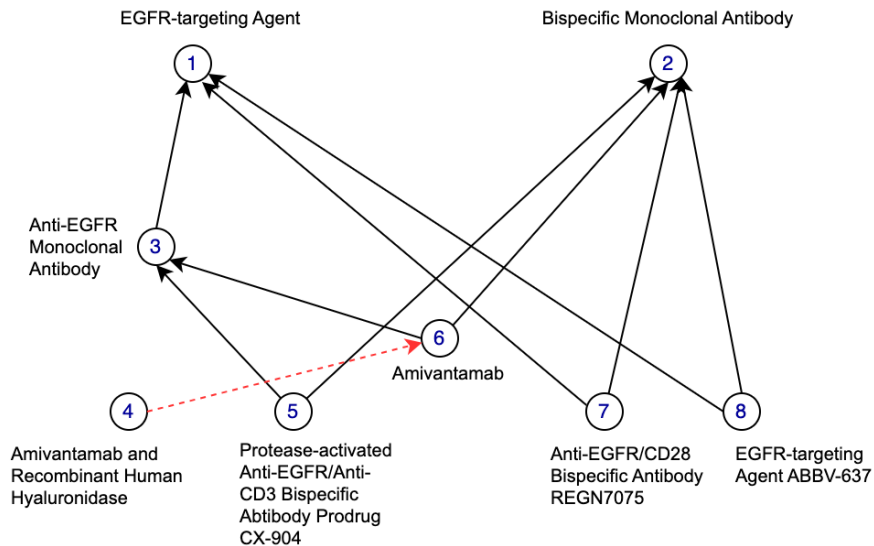


Fig. 7 A potentially missing hierarchical relations identified (marked red) by our methods in the NCI non-lattice subgraph shown in Figure 4. Note that the two original direct hierarchical relations from the concept “*Amivantamab and Recombinant Human Hyaluronidase*,” to the concepts “*Bispecific Monoclonal Antibody*” and “*Anti-EGFR Monoclonal Antibody*” are removed because they can now be transitively inferred by the potential missing hierarchical relation and the existing hierarchical relations.

the rest of the potential missing IS-A relations together with all the existing IS-A relations to check whether it can be inferred.

In addition, we further remove any potential missing IS-A relations that may cause cycles in the ontology. For instance, if our method suggests two potentially missing IS-A relations X IS-A Y and Y IS-A X , then both of these would be removed as they cause a cycle. A potential missing IS-A relation could cause a cycle together with existing IS-A relations in the ontology. For example, if the method suggests X IS-A Y , while Y IS-A X already exists in the ontology, then, X IS-A Y will be removed.

2.5 Evaluation

To evaluate the efficacy of our method in identifying accurate missing IS-A relations, we leveraged the support of domain experts (authors JS and SL) to review a sample of potential missing IS-A discovered by the method. The experts evaluated potential

691 missing IS-A relations in terms of their validity and provided comments where nec-
692 essary indicating why a certain case is valid or not. For SNOMED CT, we randomly
693 picked potential missing IS-A relations from ‘*Clinical Findings*’ and ‘*Procedure*’ sub-
694 hierarchies, and both the domain experts individually reviewed each case. We consider
695 a particular potential missing IS-A relation to be valid if both reviewers agree with
696 it. For NCIIt, we picked all the potential missing IS-A relations from the ‘*Drug, Food,*
697 *Chemical or Biomedical Material*’ subhierarchy which were each manually reviewed
698 by the author JS.
699
700
701
702
703
704
705

706 **3 Results**

707
708 We applied our method to all the active concepts and relations in the inferred versions
709 of the September 2021 Release of the US Edition of SNOMED CT which contained
710 358,356 concepts and the 23.05e release of NCIIt which contained 180,065 concepts. The
711 non-lattice detection algorithm identified 234,963 non-lattice subgraphs in SNOMED
712 CT and 14,529 in NCIIt. Among these non-lattice subgraphs, our approach identified
713 982 non-redundant potentially missing IS-A relations for SNOMED CT and 100 for
714 NCIIt.
715
716
717
718
719
720

721 **3.1 Evaluation results**

722
723 From 982 potential missing IS-A relations discovered in the SNOMED CT, 577 were
724 in the ‘*Clinical Finding*’ subhierarchy and 247 were in the ‘*Procedure*’ subhierarchy.
725 For the evaluation, we randomly picked 150 potential missing IS-A relations from
726 these two subhierarchies. The evaluation findings showed that 118 of them (78.67%)
727 are valid IS-A relations.
728
729
730

731 Among the 100 potential missing IS-A relations detected within the NCIIt, 20
732 were found in the ‘*Drug, Food, Chemical or Biomedical Material*’ subhierarchy. The
733
734
735
736

evaluation by the domain expert revealed that 17 of them (85%) are valid IS-A relations.

Tables 2 and 3 display five valid IS-A relations each identified within SNOMED CT and NCIt respectively.

Table 2 Domain expert confirmed five missing IS-A relations discovered in the SNOMED CT.

Child	Parent
Folliculitis cruris pustulosa atrophicans (disorder)	Degenerative disorder of extremity (disorder)
Accidental nitrous oxide poisoning (disorder)	Accidental poisoning caused by gaseous anesthetic (disorder)
Primary squamous cell carcinoma of tonsillar pillar (disorder)	Primary squamous cell carcinoma of oropharynx (disorder)
Spinal ganglionectomy (procedure)	Ganglionectomy of peripheral nerve (procedure)
Excision of finger joint synovium (procedure)	Arthrectomy of finger (procedure)

Table 3 Domain expert confirmed five missing IS-A relations discovered in NCIt.

Child	Parent
Palbociclib Isethionate	Palbociclib
Estramustine Phosphate Sodium Anhydrous	Estramustine
Rituximab and Hyaluronidase Human	Rituximab
Vinorelbine Tartrate Emulsion	Vinorelbine Tartrate
Liposomal Vinorelbine	Vinorelbine

4 Discussion

In this paper, we introduced a method to identify IS-A relations within a terminology that were not captured during classification by Description Logic reasoners. The approach identifies unrelated concept-pairs within non-lattice subgraphs that are both logically and lexically likely to form IS-A relations.

The number of missing IS-A relations within a terminology is unknown. However, it can be assumed that the number would be much less than the number of existing

IS-A relations within a well-formed terminology. Due to the discovery nature of terminology Quality Assurance, no approach is able to capture all missing IS-A relations. Different approaches usually capture different subsets of missing IS-A relations. The same approach could also capture different numbers of potential missing IS-A in different terminologies based on the characteristics of each terminology. Our approach captures a significantly higher number of potential missing IS-A relations in SNOMED CT than NCIt (982 vs 100). The major reason for this is the number of non-lattice subgraphs in each terminology. While SNOMED CT is twice the size of NCIt in terms of the number of concepts it has (358,356 versus 180,065), it contains around 16 times more non-lattice subgraphs (234,963 versus 14,529). Since our method is applied within non-lattice subgraphs, the method is able to discover many more missing IS-A relations in SNOMED CT than NCIt. It must be mentioned that any number of inconsistencies discovered is immensely valuable to the quality improvement process of these biomedical terminologies and can make a large impact on the downstream applications that use these terminologies.

Although our method uncovers missing IS-A relations between concepts, it is important to mention that rectifying such issues may not be as straightforward as directly adding the missing relations into respective terminologies as there might be other underlying issues within a terminology that cause these missing relations. For instance, the fix may rather involve modifying the logical definitions of the concepts so that the missing relation becomes inferable by a DL reasoner. For instance, Figure 6 shows the missing IS-A relations that were suggested by our methods to the non-lattice subgraph in Figure 3. Figure 8 presents the corresponding concept hierarchy in the March 2021 US Edition of SNOMED CT. Note that the missing IS-A relation we identified: “*Benign ganglioneuroma of abdomen (disorder)*” IS-A “*Neoplasm of peripheral nerves of abdomen (disorder)*” does not exist directly in this new version. The hierarchical relation “*Benign ganglioneuroma of abdomen (disorder)*” IS-A “*Benign*

neoplasm of peripheral nerves of abdomen (disorder)” has been added in this version which together with the existing hierarchical relation “Benign neoplasm of peripheral nerves of abdomen (disorder)” IS-A “Neoplasm of peripheral nerves of abdomen (disorder)” infers the missing IS-A relation our method suggested.

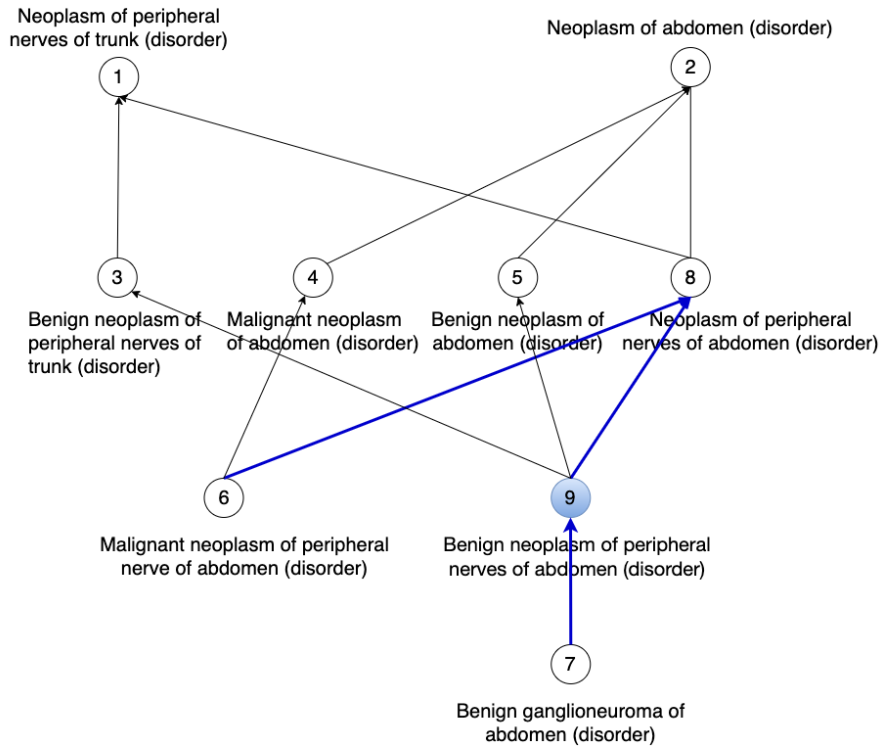


Fig. 8 Concept hierarchy in the March 2021 Release of the SNOMED CT (US Edition) consisting of concepts in the non-lattice subgraph shown in Figure 3. There is a new concept “Benign neoplasm of peripheral nerves of abdomen (disorder)” that is marked blue. The hierarchical relations marked blue are newly added ones which indicate that the missing hierarchical relations we identified were correct.

It should also be mentioned that while a vast majority of missing IS-A suggestions made by our method are correct, it also makes some invalid suggestions. For example, the method suggests an IS-A relation between the SNOMED CT concepts “Accidental fenoprofen poisoning (disorder)” and “Accidental poisoning caused by antirheumatic (disorder).” This is not correct since Fenoprofen is an NSAID medication, not an anti-rheumatic medication. The origin of this invalid suggestion stems from the fact

875 that “*Poisoning caused by antirheumatic (disorder)*” is an ancestor of “*Accidental*
876 *fenoprofen poisoning (disorder)*”. This relationship affects both the logical definition-
877 based and lexical-based subsumption checks.
878

879 Similarly, our approach suggests a missing IS-A between NCI concepts “*Radicicol*
880 *Derivative KF58333*” and “*Radicicol*”. This is also incorrect as the derivative KF58333
881 is a different molecule from Radicol. This suggestion is made because the approach is
882 only checking whether the potential child’s lexical feature set is a superset of that of the
883 potential parent, without further looking into the semantics indicated by the additional
884 lexical features (e.g. “derivative” in this instance) the potential child contains.
885
886
887
888
889

890 4.1 Comparison with related work

891 Logical definitions and lexical features have not often been explored together for
892 quality assurance of relations in biomedical terminologies. In one instance, Quesada-
893 Martínez et al. have investigated natural language content in concept labels and the
894 logical definitions to identify missing relations in SNOMED CT [35]. Their approach
895 identifies lexical regularities from concept labels through natural language processing
896 techniques and they propose relations between classes exhibiting these regularities.
897 Our approach is different from this as it directly compares logical definitions across
898 two concepts to suggest a missing relation. Quesada-Martínez et al.’s approach has
899 identified 585 cases of potential missing relations in SNOMED CT of which they have
900 analyzed one case which was found to be valid. Bodenreider’s approach in identifying
901 missing hierarchical relations in SNOMED CT relies on constructing logical defini-
902 tions from concept labels and running a description logic reasoning on them [12]. In
903 contrast, our approach is applied directly to existing logical definitions of SNOMED
904 CT and since it is applied to primitive parent terms, it captures relationships that
905 cannot be identified through reasoning. Bodenreider’s approach has been applied to
906 disorder and procedure concepts of SNOMED CT. The approach has uncovered 559
907
908
909
910
911
912
913
914
915
916
917
918
919
920

potential missing IS-A relations and an evaluation on a random sample with 100 cases has revealed 78% are valid. Note that the precision of our approach is slightly higher with 78.67% (118 out of 150). However, it should be mentioned that a direct comparison of precision is not appropriate to measure the effectiveness of different quality assurance approaches. This is because different approaches address different kinds of problems and may uncover different types of relational defects. Ontology quality assurance approaches are meant to discover ontological defects that have not been uncovered before. As there is no gold standard, it is difficult to compute recall for such approaches.

Recently, Chen et al. have introduced a deep learning-based IS-A relation prediction method for OWL ontologies [36]. Their method utilizes the pre-trained language model BERT to generate contextual embeddings for a given class with customized templates to incorporate the class context. We experimented with this approach on NCIt and found that out of the 100 potential missing IS-A relations identified by our method, 86 were also found by Chen et al.’s approach. However, it is worth noting that even when the child and parent are switched in 100 potential missing IS-A relations, Chen et al.’s approach still predicts 69 cases as IS-A relations. Therefore, further investigations (particularly by means of a manual evaluation of the predictions) are needed for IS-A relation prediction approaches such as [36] to ensure their effectiveness in identifying missing IS-A relations.

In previous work, we have leveraged different variations of purely enriched lexical attributes to identify missing IS-A relations [23–25]. For example, in [23] and [24], the enriched lexical attributes generated were all at word-level. However, [25] introduced noun phrases in concept names to the enriched lexical attributes in addition to the words.

In [32] we investigated an approach combining enriched lexical attributes and logical definitions of concepts in NCIt to identify missing IS-A relations. However, the

967 enriched lexical attributes generated were based on words and roots of noun chunks
968 distinct from the method used to generate lexical attributes in this paper.
969

970

971 **4.2 Future directions** 972

973

974 In this work, we obtained the enriched lexical attributes leveraging dependency-pairs,
975 base noun phrases, and words of concepts and their ancestors. However, we did not
976 take into account the different variations of words such as singular or plural versions
977 as well as synonymous words and phrases. In the future, we would like to explore
978 a comprehensive normalization strategy to normalize the lexical features leveraging
979 lemmatization and synonym replacement approaches.
980

981

982 As previously stated, in certain scenarios, the underlying reason for the missing IS-
983 A relations might be attributed to issues with logical definitions of concepts. Though
984 we leverage logical definitions to discover missing IS-A relations in this work, we are
985 yet unable to identify changes in logical definitions that may be needed to address the
986 root causes. In the future, we aim to explore approaches that can tackle this important
987 problem.
988

989

990 Recent advancements in Large Language Models (LLMs) have revolutionized Nat-
991 ural Language Processing. It would be interesting to explore how these LLMs could
992 be effectively leveraged to make accurate predictions for missing IS-A relations.
993

994

995

1000 **5 Conclusion** 1001

1002

1003 In this paper, we introduced an approach to discover missing IS-A relations that
1004 would not be captured during internal terminology consistency checking methods such
1005 as classification by description logic reasoners. Given a candidate concept-pair, our
1006 approach first compares whether logical definitions of one concept are more general
1007 than the other. Then, we further check whether the enriched lexical attributes of the
1008 earlier concept are a subset of the latter. If both conditions are satisfied, we suggest a
1009
1010
1011
1012

missing IS-A relation between the two concepts. Then, we remove redundant potential missing IS-A suggestions that can be inferred through other potential missing relations and potential missing IS-A relations that can cause cycles. Applying our method to the September 2021 US Edition of SNOMED CT and 23.05e release of the National Cancer Institute thesaurus (NCIt), we identified 982 and 100 potential missing IS-A relations respectively. To analyze the efficacy of our approach, we performed an evaluation of a sample of cases by domain experts. The evaluation showed that out of the 150 SNOMED CT cases, 118 are valid IS-A relations and 17 out of 20 are valid IS-A relations in NCIt. As a vast majority of cases identified by the method are accurate, this method can be deemed as an effective approach in identifying missing IS-A relations and can readily be adopted by any biomedical terminology that has concept names and logical definitions.

Supplementary information. The results for identified missing IS-A relations are available at: https://github.com/rashmie/OQA_logical_definitions_lexical_features/tree/main/results.

List of abbreviations

NCI	National Cancer Institute
NCIt	National Cancer Institute thesaurus
DL	Description Logic
NLP	Natural Language Processing
LLMs	Large Language Models

Declarations

- Funding: This work was supported by the National Science Foundation (NSF) through grant 2047001, and National Institutes of Health (NIH) through grants R01LM013335 and R01NS116287. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or NIH.
- Competing interests: The authors declare that they have no competing interests.
- Ethics approval: Not applicable.

- 1059 • Consent to participate: Not applicable.
- 1060 • Consent for publication: Not applicable.
- 1061 • Availability of data and materials: Not applicable.
- 1062 • Code availability: [https://github.com/rashmie/OQA_logical_definitions_lexical_](https://github.com/rashmie/OQA_logical_definitions_lexical_features)
- 1063 [features](https://github.com/rashmie/OQA_logical_definitions_lexical_features)
- 1064 • Authors' contributions: LC conceptualized and designed the study. FZ and RA
- 1065 designed and implemented the algorithms, generated the results, and prepared the
- 1066 evaluation sample. JS and SL performed the evaluation. RA, FZ, and LC analyzed
- 1067 the evaluation results. RA, FZ, and LC wrote the manuscript.

1075 References

- 1076 [1] National Library of Medicine.: SNOMED CT. [Accessed 14-July-2023]. <https://www.nlm.nih.gov/healthit/snomedct/index.html>.
- 1077 [2] Fragoso G, de Coronado S, Haber M, Hartel F, Wright L. Overview and utilization
- 1078 of the NCI thesaurus. Comparative and functional genomics. 2004;5(8):648–654.
- 1079 [3] Kazakov Y, Krötzsch M, Simancik F. ELK Reasoner: Architecture and Evalua-
- 1080 tion. In: ORE; 2012. .
- 1081 [4] Lawley MJ, Bousquet C. Fast classification in Protégé: Snorocket as an OWL 2 EL
- 1082 reasoner. In: Proc. 6th Australasian Ontology Workshop (IAOA'10). Conferences
- 1083 in Research and Practice in Information Technology. vol. 122; 2010. p. 45–49.
- 1084 [5] SNOMED International.: SNOMED CT Logical Model. [Accessed 12-July-
- 1085 2023]. [https://confluence.ihtsdotools.org/display/DOCSTART/5.+SNOMED+](https://confluence.ihtsdotools.org/display/DOCSTART/5.+SNOMED+CT+Logical+Model)
- 1086 [CT+Logical+Model](https://confluence.ihtsdotools.org/display/DOCSTART/5.+SNOMED+CT+Logical+Model).
- 1087 [6] National Cancer Institute.: NCI Thesaurus Semantics. [Accessed 12-July-
- 1088 2023]. [https://evs.nci.nih.gov/ftp1/ThesaurusSemantics/NCI%20Thesaurus%](https://evs.nci.nih.gov/ftp1/ThesaurusSemantics/NCI%20Thesaurus%20Semantics)
- 1089
- 1090
- 1091
- 1092
- 1093
- 1094
- 1095
- 1096
- 1097
- 1098
- 1099
- 1100
- 1101
- 1102
- 1103
- 1104

20Semantics.pdf. 1105

1106

[7] SNOMED International.: High Level Logical Model of SNOMED CT. [Accessed 1107
14-Dec-2023]. [https://confluence.ihtsdotools.org/display/DOCRELFMT/2.1+](https://confluence.ihtsdotools.org/display/DOCRELFMT/2.1+High+Level+Logical+Model+of+SNOMED+CT) 1108
[High+Level+Logical+Model+of+SNOMED+CT](https://confluence.ihtsdotools.org/display/DOCRELFMT/2.1+High+Level+Logical+Model+of+SNOMED+CT). 1109
1110
1111
1112

[8] SNOMED International.: attribute group. [Accessed 12-July-2023]. [https://](https://confluence.ihtsdotools.org/display/DOCGLOSS/attribute+group) 1113
confluence.ihtsdotools.org/display/DOCGLOSS/attribute+group. 1114
1115
1116
1117

[9] SNOMED International.: SNOMED CT Starter Guide. [Accessed 1-June- 1118
2021]. [https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+](https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide) 1119
[Starter+Guide](https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide). 1120
1121
1122

[10] SNOMED International.: SNOMED International Release Management Home. 1123
[Accessed 1-June-2021]. <https://confluence.ihtsdotools.org/display/RMT/>. 1124
1125
1126
1127

[11] Zhu X, Fan JW, Baorto DM, Weng C, Cimino JJ. A review of auditing meth- 1128
ods applied to the content of controlled biomedical terminologies. Journal of 1129
biomedical informatics. 2009;42(3):413–425. 1130
1131
1132

[12] Bodenreider O. Identifying missing hierarchical relations in SNOMED CT from 1133
logical definitions based on the lexical features of concept names. In: CEUR 1134
workshop proceedings. vol. 1747. NIH Public Access; 2016. . 1135
1136
1137
1138

[13] Ochs C, Geller J, Perl Y, Chen Y, Agrawal A, Case JT, et al. A tribal abstrac- 1139
tion network for SNOMED CT target hierarchies without attribute relationships. 1140
Journal of the American Medical Informatics Association. 2015;22(3):628–639. 1141
1142
1143
1144

[14] Wei D, Bodenreider O. Using the abstraction network in complement to descrip- 1145
tion logics for quality assurance in biomedical terminologies-a case study in 1146
SNOMED CT. Studies in health technology and informatics. 2010;160(0 2):1070. 1147
1148
1149
1150

1151 [15] Ochs C, Geller J, Perl Y, Chen Y, Xu J, Min H, et al. Scalable quality assurance
1152 for large SNOMED CT hierarchies using subject-based subtaxonomies. *Journal*
1153 *of the American Medical Informatics Association*. 2015;22(3):507–518.
1154
1155
1156 [16] Ochs C, Perl Y, Geller J, Halper M, Gu H, Chen Y, et al. Scalability of
1157 abstraction-network-based quality assurance to large SNOMED hierarchies. In:
1158 *AMIA Annual Symposium Proceedings*. vol. 2013. American Medical Informatics
1159 Association; 2013. p. 1071.
1160
1161
1162 [17] Agrawal A, Perl Y, Ochs C, Elhanan G. Algorithmic detection of inconsistent
1163 modeling among SNOMED CT concepts by combining lexical and structural indi-
1164 cators. In: *2015 IEEE international conference on bioinformatics and biomedicine*
1165 *(BIBM)*. IEEE; 2015. p. 476–483.
1166
1167
1168 [18] Agrawal A, Qazi K. Detecting modeling inconsistencies in SNOMED CT using
1169 a machine learning technique. *Methods*. 2020;179:111–118.
1170
1171
1172 [19] Agrawal A. Evaluating lexical similarity and modeling discrepancies in the proce-
1173 dure hierarchy of SNOMED CT. *BMC Medical Informatics and Decision Making*.
1174 2018;18:27–33.
1175
1176 [20] Liu H, Zheng L, Perl Y, Geller J, Elhanan G. Can a convolutional neural network
1177 support auditing of nci thesaurus neoplasm concepts? In: *ICBO*; 2018. .
1178
1179
1180 [21] Liu H, Perl Y, Geller J. Concept placement using BERT trained by trans-
1181 forming and summarizing biomedical ontology structure. *Journal of Biomedical*
1182 *Informatics*. 2020;112:103607.
1183
1184
1185 [22] Abeysinghe R, Zheng F, Bernstam EV, Shi J, Bodenreider O, Cui L. A deep
1186 learning approach to identify missing is-a relations in SNOMED CT. *Journal of*
1187 *the American Medical Informatics Association*. 2023;30(3):475–484.
1188
1189
1190
1191
1192
1193
1194
1195
1196

[23] Cui L, Bodenreider O, Shi J, Zhang GQ. Auditing SNOMED CT hierarchical relations based on lexical features of concepts in non-lattice subgraphs. Journal of biomedical informatics. 2018;78:177–184.	1197 1198 1199 1200 1201 1202
[24] Abeysinghe R, Brooks MA, Cui L. Leveraging non-lattice subgraphs to audit hierarchical relations in NCI Thesaurus. In: AMIA annual symposium proceedings. vol. 2019. American Medical Informatics Association; 2019. p. 982.	1203 1204 1205 1206 1207 1208
[25] Zheng F, Shi J, Cui L. A lexical-based approach for exhaustive detection of missing hierarchical IS-A relations in SNOMED CT. In: AMIA Annual Symposium Proceedings. vol. 2020. American Medical Informatics Association; 2020. p. 1392.	1209 1210 1211 1212 1213
[26] Manuel W, Abeysinghe R, He Y, Tao C, Cui L. Identification of missing hierarchical relations in the vaccine ontology using acquired term pairs. Journal of Biomedical Semantics. 2022;13(1):22.	1214 1215 1216 1217 1218 1219
[27] Hao X, Abeysinghe R, Shi J, Cui L. A substring replacement approach for identifying missing IS-A relations in SNOMED CT. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2022. p. 2611–2618.	1220 1221 1222 1223 1224 1225 1226 1227
[28] Mohtashamian M, Abeysinghe R, Hao X, Cui L. Identifying Missing IS-A Relations in Orphanet Rare Disease Ontology. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2022. p. 3274–3279.	1228 1229 1230 1231 1232 1233
[29] Abeysinghe R, Yang Y, Bartels M, Zheng WJ, Cui L. An evidence-based lexical pattern approach for quality assurance of Gene Ontology relations. Briefings in Bioinformatics. 2022;23(3):bbac122.	1234 1235 1236 1237 1238
[30] 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	1239 1240 1241 1242

1243 CT. Journal of the American Medical Informatics Association. 2017;24(4):788–
1244 798.
1245
1246
1247 [31] Abeysinghe R, Brooks MA, Talbert J, Licong C. Quality assurance of NCI
1248 Thesaurus by mining structural-lexical patterns. In: AMIA Annual Symposium
1249 Proceedings. vol. 2017. American Medical Informatics Association; 2017. p. 364.
1250
1251
1252
1253 [32] Zheng F, Abeysinghe R, Sioutos N, Whiteman L, Remennik L, Cui L. Detecting
1254 missing IS-A relations in the NCI Thesaurus using an enhanced hybrid approach.
1255 BMC Medical Informatics and Decision Making. 2020;20(10):1–11.
1256
1257
1258
1259 [33] Zhang GQ, Xing G, Cui L. An efficient, large-scale, non-lattice-detection algo-
1260 rithm for exhaustive structural auditing of biomedical ontologies. Journal of
1261 biomedical informatics. 2018;80:106–119.
1262
1263
1264
1265 [34] Explosion.: NCI Thesaurus Semantics. [Accessed 12-July-2023]. [https://evs.nci.](https://evs.nci.nih.gov/ftp1/ThesaurusSemantics/NCI%20Thesaurus%20Semantics.pdf)
1266 [nih.gov/ftp1/ThesaurusSemantics/NCI%20Thesaurus%20Semantics.pdf](https://evs.nci.nih.gov/ftp1/ThesaurusSemantics/NCI%20Thesaurus%20Semantics.pdf).
1267
1268
1269 [35] Quesada-Martínez M, Fernández-Breis JT, Karlsson D. Suggesting Missing Rela-
1270 tions in Biomedical Ontologies Based on Lexical Regularities. In: MIE; 2016. p.
1271 384–388.
1272
1273
1274
1275 [36] Chen J, He Y, Geng Y, Jiménez-Ruiz E, Dong H, Horrocks I. Contextual
1276 semantic embeddings for ontology subsumption prediction. World Wide Web.
1277 2023;26(5):2569–2591.
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288