Leveraging non-lattice subgraphs for suggestion of new concepts for SNOMED CT

Xubing Hao*, Rashmie Abeysinghe†, Fengbo Zheng*, Licong Cui*
*School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA
†Department of Neurology, The University of Texas Health Science Center at Houston, Houston, Texas, USA

***Abstract*—Missing hierarchical *is-a* relations and missing concepts are common quality issues in biomedical ontologies. Non-lattice subgraphs have been extensively studied for automatically identifying missing *is-a* relations in biomedical ontologies like SNOMED CT. However, little is known about non-lattice subgraphs' capability to uncover new or missing concepts in biomedical ontologies. In this work, we investigate a lexical-based intersection approach based on non-lattice subgraphs to identify potential missing concepts in SNOMED CT. We first construct lexical features of concepts using their fully specified names. Then we generate hierarchically unrelated concept pairs in non-lattice subgraphs as the candidates to derive new concepts. For each candidate pair of concepts, we conduct an order-preserving intersection based on the two concepts' lexical features, with the intersection result serving as the potential new concept name suggested. We further perform automatic validation through terminologies in the Unified Medical Language System (UMLS) and literature in PubMed. Applying this approach to the March 2021 release of SNOMED CT US Edition, we obtained 7,702 potential missing concepts, among which 1,288 were validated through UMLS and 1,309 were validated through PubMed. The results showed that non-lattice subgraphs have the potential to facilitate suggestion of new concepts for SNOMED CT.**

***Index Terms*—Ontologies and Terminologies, Non-lattice subgraphs, Ontology Enrichment, UMLS**

I. Introduction

An ontology formally represents knowledge in a domain of interest by a set of concepts and relations between those concepts. Ontologies have been widely used in biomedical research and applications such as knowledge representation, knowledge management, data integration, data sharing, natural language processing, information retrieval, and decision support [1]. This has been accentuated especially due to increasingly large amounts of heterogeneous health related data produced [2].

Biomedical ontologies are continuously updated with new concepts to reflect update-to-date knowledge. Development of automatic approaches to suggest potential new or missing concepts to be added to biomedical ontologies is an active research area for ontology enrichment. Mining lexical patterns in non-lattice subgraphs have shown the potential to automatically suggest missing *is-a* (or subtype) relations and missing concepts in biomedical ontologies such as SNOMED CT and NCI Thesaurus [3], [4]. However, the main focuses of these previous works were to identify missing *is-a* relations in non-lattice subgraphs, and only a small amount of missing concepts were suggested using a specific lexical pattern called *Union-Intersection*.

To unravel the full potential of non-lattice subgraphs for suggesting missing concepts, in this work we exploit a general lexical-based approach to derive potential new concepts from hierarchically unrelated concept pairs contained in non-lattice subgraphs of SNOMED CT. Given a pair of unrelated concepts in a non-lattice subgraph, we perform order-preserving intersection of the two concepts' names to generate potential missing concept. We further leverage biomedical terminologies in the Unified Medical Language System (UMLS) and literature in PubMed to automatically validate suggested missing concepts.

II. Background

A. SNOMED CT

SNOMED CT is the largest clinical terminology system in the world. It provides a common terminology that supports effective communication between different specialties and sites of care. SNOMED CT plays an important role in indexing, storing, retrieving, and aggregating clinical data [5]. Specifically, the United States (US) Edition of SNOMED CT is the official source for use in US healthcare systems, combining the content of both the US Extension and the International releases of SNOMED CT [6].

In SNOMED CT, each concept has a fully specified name (FSN) that represents a unique, unambiguous description of the meaning of the concept [7]. Additionally, each concept is specified as either fully defined or primitive. A concept is fully defined if it contains one or more sufficient definitions that can distinguish itself and its subtypes from all other concepts [8]. On the other hand, a concept is primitive if it's definition is not sufficient to computably distinguish it from other concepts [9].

B. Non-lattice subgraphs

Non-lattice subgraphs in an ontology are derived by non-lattice pairs, where a non-lattice pair is a pair of concepts having more than one maximal shared common descendant [3], [10]. For each non-lattice pair (A, B), its non-lattice subgraph is a graph fragment including concepts and *is-a* relations between the maximal common descendants of A and B, denoted as $mcd(A, B)$ (also called "lower bound"),

This work was supported by the US National Science Foundation (NSF) under grant 1931134 and National Institutes of Health (NIH) under grant R01LM013335. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or NIH. Correspondence: licong.cui@uth.tmc.edu

and the minimal common ancestors of $mcd(A, B)$, denoted as $mca(mcd(A, B))$ (called "upper bound") [3].

Fig. 1 shows a non-lattice subgraph generated from the non-lattice pair *("Disorder of skin appendage (disorder)", "Secondary malignant neoplasm of skin (disorder)")* in the March 2021 Release of the SNOMED CT (US Edition). The size of this non-lattice subgraph is 8, which is the number of concepts it contains.

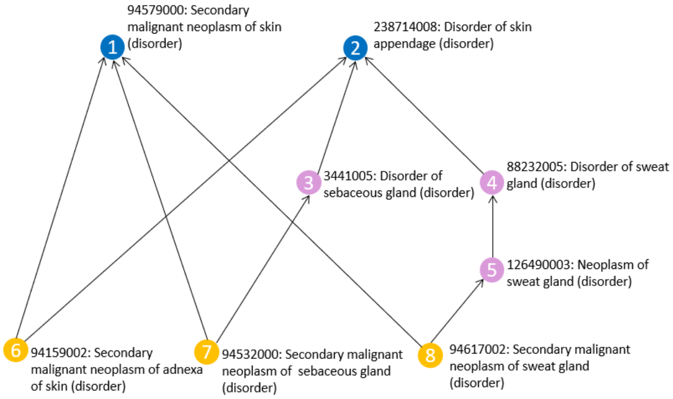
1 94579000: Secondary malignant neoplasm of skin (disorder)
2 238714008: Disorder of skin appendage (disorder)
3 3441005: Disorder of sebaceous gland (disorder)
4 88232005: Disorder of sweat gland (disorder)
5 126490003: Neoplasm of sweat gland (disorder)
6 94159002: Secondary malignant neoplasm of adnexa of skin (disorder)
7 94532000: Secondary malignant neoplasm of sebaceous gland (disorder)
8 94617002: Secondary malignant neoplasm of sweat gland (disorder)

Fig. 1. A non-lattice subgraph of size 8 in SNOMED CT.

C. Identification of missing concepts

A number of ontology quality assurance approaches have focused on identifying missing concepts in biomedical ontologies. In one such approach, He at al. have investigated vertical topological patterns to identify missing concepts in ontologies including SNOMED CT. These are cases where concept pairs existing in two ontologies having different intermediate concepts between the *is-a* relation path of the two concepts. They leverage UMLS to map terms across ontologies and identify these topological patterns. These cases may indicate the possibility to import concepts from one ontology to another [11]–[13].

In previous work, we investigated six lexical patterns in non-lattice subgraphs, of which one pattern named *Union-Intersection* uncovered missing concepts in SNOMED CT [3]. Here, it was checked whether the union of the set-of-words in the upper-bound concepts is equal to the intersection of the set-of-words of the lower bound concepts. If so, a missing concept in between the upper and lower bounds was suggested with a name containing the set-of-words in the union of the upper bound concepts (or intersection of the lower bound concepts). For instance, Fig. 2 shows an example of a non-lattice subgraph exhibiting the Union-Intersection pattern. The union of the set-of-words of upper-bound concepts and the intersection of the set-of-words of lower-bound concepts both results in the set-of-words {"arthritis", "of", "knee", "seropositive", "rheumatoid"}, which indicates a potential missing concept with the resulting set-of-words.

In two recent studies, we proposed a formal concept analysis (FCA) approach based on lexical features of concept names to identify potential missing concepts in the National Cancer Institute thesaurus and SNOMED CT [14], [15]. A formal concept was constructed by extracting lexical features of concept names. Then, multistage intersection was performed to formalize new concepts and detect potential missing concepts. The UMLS and PubMed were leveraged to automatically validate the missing concepts identified.

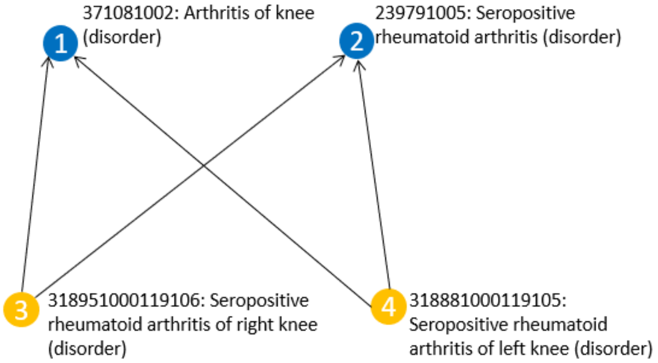
1 371081002: Arthritis of knee (disorder)
2 239791005: Seropositive rheumatoid arthritis (disorder)
3 318951000119106: Seropositive rheumatoid arthritis of right knee (disorder)
4 318881000119105: Seropositive rheumatoid arthritis of left knee (disorder)

Fig. 2. A non-lattice subgraph exhibiting the Union-Intersection pattern.

III. Method

In this work, we use the March 2021 release of SNOMED CT US Edition. We first construct lexical features for each concept using its FSN. We then compute all non-lattice subgraphs and generate unrelated concept pairs in non-lattice subgraphs as candidates to derive new concepts. For each candidate concept pair, we conduct intersection of the lexical features of the two concepts to suggest a potential new concept. We further post-process the intersection results to formalize the names of potential new concepts. Finally, we perform automatic validation of the suggested concepts by leveraging UMLS and PubMed.

A. Constructing lexical features of concepts

For each concept in SNOMED CT, we perform a lexical normalization based on its FSN to formulate lexical features as follows. First the semantic tag of the FSN is removed and the remainder is tokenized to words resulting in a sequence of words. Then we replace words which have synonyms with their preferred terms [14]. Preferred terms are identified through SNOMED CT as well. If a SNOMED CT concept has a single-word preferred term, it will be considered as the preferred term of all its single-word synonyms. For instance, the concept "*Embololalia (finding)*" has a preferred term "*Embololalia*" as well as synonyms "*Embolalia*" and "*Embolophrasia*". Therefore, "*Embololalia*" is considered as the preferred term for both "*Embolalia*" and "*Embolophrasia*". Hence, whenever lexical features of a concept contains "*Embolalia*" or "*Embolophrasia*", we replace it with the preferred term "*Embololalia*".

In addition, we convert all words to lowercase and lemmatize the words using the open source natural language processing library NLTK [16]. This is because the same word may appear in different variations in different concepts (e.g. singular vs plural). For example, the resulting lexical features for concept "*Incision and drainage of deep abscess of soft tissues of neck (procedure)*" after normalization is ["incision", "and", "discharge", "of", "deep", "abscess", "of", "soft", "tissue", "of", "neck"]. Note that word "drainage" has been replaced by its preferred term "discharge", and word "tissues" have been lemmatized to its singular form "tissue".

B. Generating candidate concept pairs

In this step, we generate concept pairs in non-lattice subgraphs serving as candidates for deriving new concepts. We first compute all non-lattice subgraphs in SNOMED CT leveraging an efficient non-lattice-detection algorithm [17]. Since larger non-lattice subgraphs may contain smaller ones, in this work we focus on non-lattice subgraphs with size of less than or equal to 10. A pair of concepts in a non-lattice subgraph is considered as a candidate concept-pair if the following three conditions are met:

- The two concepts are unrelated, that is, they do not have an *is-a* relation (either direct or indirect);
- The two concepts are fully defined. We set this condition because a primitive concept does not have a sufficient definition that can computably distinguish it from other concepts;
- The depths of both concepts are at least 10 hops from the SNOMED CT root concept considering the longest path. This condition is intended to exclude too general concepts whose FSNs are relatively short and may not generate meaningful concept names when performing intersection later.

For example, concepts "*Neoplasm of sweat gland (disorder)*" and "*Secondary malignant neoplasm of sebaceous gland (disorder)*" in the non-lattice subgraph shown in Fig. 1 satisfy all these conditions and form a candidate concept pair. Note that non-lattice subgraphs may overlap with each other. Therefore, a candidate concept-pair may exists in more than one non-lattice subgraph. We remove such duplicate cases from the final list of candidate concept-pairs.

C. Suggesting missing concepts

Potential missing concepts are uncovered through an intersecting procedure of the lexical features of candidate concept pairs. For each candidate pair of concepts, we perform an order-preserving intersection of the two concepts' lexical features, which results in common lexical features shared by both concepts while keeping the order of their appearance in the concept names. Such order-preserving intersection would result in a new concept name that is consistent with the original concepts' semantic organization. For example, intersecting lexical features of concept "*Neoplasm of sweat gland*" and concept "*Secondary malignant neoplasm of sebaceous gland*" in Fig. 1 obtains "*neoplasm of gland*", a potential new concept name.

We further post-process the obtained concept names as follows. In certain circumstances, the obtained names may start with preposition(s) or end with adjective(s) or preposition(s). We discard such adjectives and prepositions from the obtained names so that the names are semantically and grammatically sound. Here prepositions and adjectives are identified through part-of-speech tagging. For instance, intersecting concept "*Secondary malignant neoplasm of junctional region of epiglottis*" and concept "*Secondary malignant neoplasm of soft tissue of head*" results in "*secondary malignant neoplasm of*". By removing the preposition "of" at the end, we have "*secondary malignant neoplasm*" as the potential new concept name.

Also, intersection may result in a concept name with consecutive prepositions in the middle, in which cases we only keep the last preposition. For example, intersecting concept"*Excision of nodule of vocal cord with laser*" and concept "*Excision of reinke's edema with laser*" results in "*excision of with laser*", which contains two consecutive prepositions "of" and "with". By keeping the last preposition "with", we obtain "*excision with laser*" as the name of the potential new concept.

In addition, the same concept name may be obtained by intersecting different concept pairs. For instance, "*cell carcinoma of ear*" can be generated from concept pairs ("*Basal cell carcinoma of ear (disorder)*", "*Squamous cell carcinoma of auricle of ear (disorder)*"); ("*Squamous cell carcinoma of skin of ear (disorder)*", "*Basal cell carcinoma of auricle of ear (disorder)*"); and ("*Basal cell carcinoma of ear (disorder)*", "*Squamous cell carcinoma of skin of ear (disorder)*"). Therefore, after intersecting all candidate concept pairs and obtaining all the potential new concept names, we remove such duplicates from the list of potential new concepts. Furthermore, the new concept names obtained by intersection may already be existing concept names in SNOMED CT. After removing such cases, we suggest the remainder as potential missing concepts.

D. Validating suggested concepts

We validate the identified potential missing concepts by leveraging two external sources: (1) external terminologies from UMLS; and (2) biomedical literature in PubMed.

1) UMLS-based validation: The UMLS integrates many biomedical terminologies including SNOMED CT, Gene Ontology, Medical Subject Headings (MeSH), OMIM and Digital Anatomist Symbolic Knowledge Base [18]. UMLS contains over 16 million concept names from 218 source vocabularies which are gathered through more than 4 million UMLS concepts [19].

In this work, we use the 2021-AA-full version of the UMLS and only leverage the concepts that are in English. We first normalize all the UMLS atoms (concept names from different source vocabularies) as follows. We first tokenize the atoms to words. A word would be replaced with a preferred term (if applicable) as we have performed earlier while normalizing concept names to obtain lexical features. All the words are converted to lowercase and lemmatized. Then we remove stop words from potential new concept names (which are already normalized during lexical feature extraction) and see whether a match can be obtained with a normalized UMLS atom name. If a match is found, the potential new concept is said to be validated through UMLS.

2) PubMed-based validation: We leverage PubMed to perform a literature-based validation of the potential new concepts suggested by our approach [20]. If a potential new concept appears as a base noun phrase in the title or abstract of a biomedical publication, then we say it is validated through PubMed. The requirement of base noun phrase is to make sure that the potentail new concept does not appear as a substring of another concept. For instance, a potential new concept "*thoracic artery*" may exists in an abstract as a substring of "*fetal thoracic artery*", in which case it is more appropriate to say that concept "*fetal thoracic artery*" is validated via the abstract rather than concept "*thoracic artery*".

PubMed contains about 32 million citations and abstracts of biomedical literature. We use the 2021 baseline release of PubMed and its daily update files up to September 13th, 2021. The title and abstract for each publication are extracted from the release files, and parsed with Spacy to identify base noun phrases [21]. Each base noun phrase is then normalized similarly as how the UMLS atoms were normalized.

Due to the enormity of the search space, a sequential search for potential new concepts among these base noun phrases would be time consuming. Therefore, we index the normalized noun phrases using open-source search library Apache Lucene [22]. Then, we search the index for the normalized potential new concepts, which is significantly faster than directly performing a sequential search on the base noun phrases.

IV. Result

We extracted a total of 236,291 non-lattice subgraphs from the March 2021 US Edition of SNOMED CT, among which 43,923 were with a size of less than or equal to 10. From these we identified 92,099 candidate concept pairs (i.e., unrelated fully defined concept pairs with a depth of at least 10 hops from the root). Intersecting candidate concept pairs, we obtained 12,622 potential new concepts, out of which 4,920 already existed as SNOMED CT concepts. The remaining 7,702 are the potential missing concepts uncovered by our approach.

A. UMLS-based validation

UMLS-based validation leveraging external terminologies resulted in 1,288 of the 7,702 concepts being validated. Table I contains 10 examples of such validated missing concepts suggested by our approach, the external UMLS terminologies that they were validated from, and the original concepts in SNOMED CT that resulted in the missing concepts.

For example, in the non-lattice subgraph shown in Fig. 3, concept "*Magnetic resonance imaging arthrography of facet joint (procedure)*" and concept "*Magnetic resonance imaging of lumbar spine and sacroiliac joint (procedure)*" form a candidate concept pair. The intersection of their lexical features suggested a potential missing concept "*magnetic resonance imaging of joint*", which was found in the Medical Dictionary for Regulatory Activities (MedDRA) with atom identifier *A31206611*, corresponding to the UMLS concept "*Magnetic resonance imaging joint*" with CUI *C5208278*.

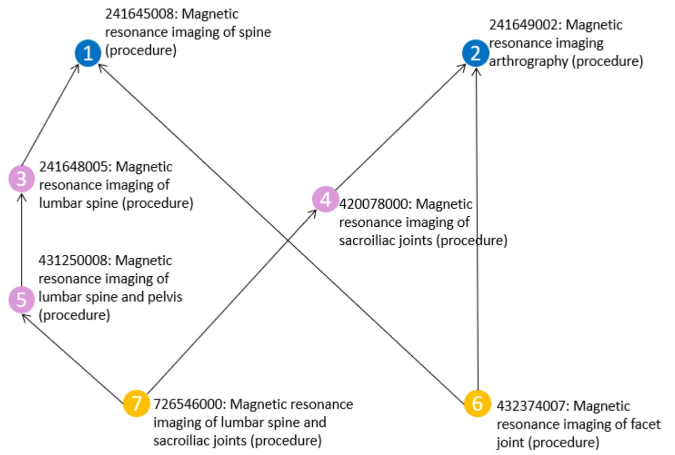
1
241645008: Magnetic resonance imaging of spine (procedure)
2
241649002: Magnetic resonance imaging arthrography (procedure)
3
241648005: Magnetic resonance imaging of lumbar spine (procedure)
4
420078000: Magnetic resonance imaging of sacroiliac joints (procedure)
5
431250008: Magnetic resonance imaging of lumbar spine and pelvis (procedure)
7
726546000: Magnetic resonance imaging of lumbar spine and sacroiliac joints (procedure)
6
432374007: Magnetic resonance imaging of facet joint (procedure)

Fig. 3. Non-lattice subgraph in SNOMED CT.

The 1,288 validated missing concepts were mapped to 3,879 different atoms in UMLS. Note that one concept can be mapped to multiple atoms from different source vocabularies in the UMLS. Table II contains the top 10 UMLS source vocabularies that were mostly used to validate the potential missing concepts and the number of concepts each terminology validated.

B. PubMed-based validation

Literature-based validation leveraging PubMed abstracts resulted in 1,309 potential missing concepts being validated. Out of those, 208 were validated by a single abstract while 1,101 were validated by multiple abstracts. Table III contains 10 examples of missing concepts validated through PubMed. For instance, the missing concept "*principal malignant neuroendocrine neoplasm*" was validated by literature [23]. Note that there were 562 missing concepts that were validated through both UMLS and PubMed.

V. Discussion

A. Supporting evidence from a new SNOMED CT release

In this work, we used the March 2021 release of SNOMED CT US Edition. Since then, the September 2021 SNOMED CT US edition has been released. This new release enables us to find supporting evidence for missing concepts identified by our approach, which is an alternative way to leverage retrospective ground truth (i.e., version difference) to assist in the validation [42]. Comparing the new concepts suggested by our approach against the new version of SNOMED CT, it was seen that 18 new concepts have been already included in the SNOMED CT new version. For example, the new concept "*malignant neoplasm of vertebra*" generated by the original concepts "*Malignant neoplasm of sacral vertebra (disorder)*" and "*Malignant neoplasm of coccygeal vertebra (disorder)*"

TABLE I
TEN EXAMPLES OF VALIDATED MISSING CONCEPTS WITH THE EXTERNAL TERMINOLOGIES IN THE UMLS.

Missing concept	**UMLS source vocabulary**	**Concept pair resulting in the missing concept**
magnetic resonance imaging of joint	MDR	Magnetic resonance imaging arthrography of facet joint (procedure) Magnetic resonance imaging of lumbar spine and sacroiliac joint (procedure)
insertion of suprapubic catheter	CHV, RCD	Cystostomy and insertion of suprapubic catheter (procedure) Insertion of suprapubic catheter using ultrasound guidance (procedure)
lithotripsy of bladder	MEDCIN	Transurethral endoscopic lithotripsy of bladder calculus (procedure) Cystoscopy and electrohydraulic lithotripsy of calculus of bladder (procedure)
fusion of atlantoaxial joint	RCD	Transoral fusion of atlantoaxial joint (procedure) Posterior fusion of atlantoaxial joint (procedure)
chronic nephritis	OMIM, MDR, MTH, ICD10CM, CHV	Chronic radiation nephritis (disorder) Chronic tubulointerstitial nephritis (disorder)
neoplasm of gland	CHV	Neoplasm of sweat gland (disorder) Secondary malignant neoplasm of sebaceous gland (disorder)
adenocarcinoma of skin	MEDCIN	Apocrine adenocarcinoma of skin (disorder) Ceruminous gland adenocarcinoma of skin (disorder)
herpes meningitis	CHV	Herpes zoster with meningitis (disorder) Herpes simplex meningitis (disorder)
congenital obstruction	CHV, SNMI, CHV	Congenital ureterovesical obstruction (disorder) Congenital obstruction of urethra (disorder)
cyst of canal of nuck	CHV, MTH, ICD9	Congenital cyst of canal of Nuck (disorder) Acquired cyst of canal of Nuck (disorder)

TABLE II
TOP TEN TERMINOLOGIES IN THE UMLS THAT VALIDATED THE MOST NUMBER OF CONCEPTS.

External Terminology	**Number of Concepts Validated**
MEDCIN	493
CHV	38
NCI	198
MSH	135
MTH	184
MDR	126
ICD10CM	93
SNMI	166
FMA	146
RCD	146

has been added in the September version as "*Malignant neoplasm of vertebra (disorder)*" with the SNOMED CT identifier *1157067003*. Note that since we used two consecutive releases of SNOMED CT for this comparison, the number of missing concepts seen in the new version is relatively small.

B. Comparison with Union-Intersection approach

We compared the results of our approach to that of Union-Intersection pattern introduced in [3]. Applying the Union-Intersection pattern to the same non-lattice subgraphs used in this work, we obtained 443 potential missing concepts, which is significantly less than what we obtained in this work. Note that in the Union-Intersection approach, the missing concepts are in the form of set of words (without proper ordering of the words provided). Therefore, to compare these two approaches, we did not perform the step to order the words of potential missing concept names in this work.

Results showed that 95 potential missing concepts were identified by both approaches. The Union-Intersection approach alone identified 348 potential missing concepts while our approach in this paper alone identified 7,607 potential missing concepts. Due to the difference between the two approaches (notably in this work we perform normalization while the Union-Intersection pattern did not), it is possible that the name of a potential missing concept is different across the two approaches even if the same concept pairs are intersected. Comparing the concept-pairs intersected, it was seen that both approaches intersected 222 concept pairs in common.

C. Comparison with FCA-based approach

We also compared this work with the previous sequence-based FCA approach [15] (see Table IV). We leveraged the March 2020 release of SNOMED CT US Edition for this comparison as it was the version used in the previous work. For example, in the *Neoplasm and/or hamartoma (disorder)* subhierarchy, the previous work identified 916 potential missing concepts while this approach identified 750; and the two

TABLE III
TEN EXAMPLES OF MISSING CONCEPTS VALIDATED THROUGH BIOMEDICAL LITERATURE IN PUBMED.

Missing concept	**Literature containing the missing concept**	**Concept pair resulting in the missing concept**
interosseous nerve injury	[24]	Anterior interosseous nerve injury (disorder) Posterior interosseous nerve injury (disorder)
allergic contact gingivitis	[25], [26]	Allergic contact gingivitis caused by acrylic dental material (disorder) Allergic contact gingivitis caused by mercury (disorder)
principal malignant neuroendocrine neoplasm	[23]	Primary malignant neuroendocrine neoplasm of appendix (disorder) Primary malignant neuroendocrine neoplasm of rectum (disorder)
percutaneous transluminal valvotomy	[27]	Percutaneous transluminal aortic valvotomy (procedure) Percutaneous transluminal pulmonary valvotomy (procedure)
superficial of groin with infection	[28]–[30]	Superficial injury of groin with infection (disorder) Superficial foreign body of groin with infection (disorder)
rheumatic valve stenosis	[31], [32]	Rheumatic heart valve stenosis with insufficiency (disorder) Rheumatic mitral valve insufficiency and aortic valve stenosis (disorder)
acute perichondritis	[33]–[35]	Acute perichondritis of left external ear (disorder) Acute perichondritis of pinna (disorder)
traumatic sternoclavicular joint	[36]	Open traumatic dislocation, sternoclavicular joint, anterior (disorder) Closed traumatic subluxation sternoclavicular joint (disorder)
intractable lobe epilepsy	[37]	Refractory frontal lobe epilepsy (disorder) Intractable partial parietal lobe epilepsy with impairment of consciousness (disorder)
osteoarthritis of knee joint	[38]–[41]	Osteoarthritis of left knee joint (disorder) Osteoarthritis of right knee joint (disorder)

approaches identified 224 potential missing concepts in common. Note that the previous work was exhaustive not limiting to non-lattice subgraphs and hence, more number of potential missing concepts were uncovered. However, this approach did suggest 526 potential missing concepts in the *Neoplasm and/or hamartoma (disorder)* subhierarchy that were not identified by the previous FCA approach.

In addition, the previous work also limited itself to five subhierarchies of SNOMED CT while this approach was applied to non-lattice subgraphs of all subhierarchies of SNOMED CT. For instance, this work identified 1,804 potential missing concepts in the *Procedure on body region (procedure)* subhierarchy which the previous work did not consider. In Table IV, we only show five such additional subhierarchies that we considered in this work, although this approach identifies potential missing concepts in all 19 subhierarchies of SNOMED CT. Also, it should be noted that a majority of potential missing concepts uncovered in this approach was not identified by the previous work [15].

D. Limitations and future work

One limitation of our lexical-based intersection approach is that the resulting new concept may not have the same semantic type as the original concept pairs. For instance, original concepts "*Opacity of cornea of right eye (disorder)*" and "*Disorder of cornea of bilateral eyes (disorder)*" can generate a potential missing concept "*cornea of eye*", which is not a disorder but a body structure). In future work, we plan to also consider the semantic tags of concepts when suggesting potential missing concept names.

In this work, we only considered non-lattice subgraphs with size of less than or equal to 10. In the future, we would extend this work to all non-lattice subgraphs. We would also like to investigate into placing the missing concepts in the existing hierarchy of SNOMED CT by identifying relations between the new concepts and existing concepts.

In addition, although automatic validation via UMLS and PubMed were performed, whether the validated missing concepts should be added into the SNOMED CT still needs manual evaluation of terminology curators.

VI. CONCLUSION

In this paper, we presented a lexical-based approach that leverages unrelated concept pairs in non-lattice subgraphs to suggest potential missing concepts in SNOMED CT. The automatic validation through external terminologies in the UMLS and literature in PubMed provided encouraging supporting evidence of suggested missing concepts, indicating that non-lattice subgraphs have the potential to facilitate suggestion of new concepts for SNOMED CT.

REFERENCES

[1] O. Bodenreider, "Biomedical ontologies in action: role in knowledge management, data integration and decision support," *Yearbook of medical informatics*, vol. 17, no. 01, pp. 67–79, 2008.

TABLE IV
COMPARISON OF THE RESULT OF THE TWO APPROACHES:SEQUENCE-BASED FCA APPROACH (PREVIOUS) VS. NON-LATTICE SUBGRAPH-BASED APPROACH (CURRENT).

Sub-hierarchy	**Existing concepts**	**Missing concepts (Previous/Current)**	**Overlap**	**Validated by UMLS (Previous/Current)**	**Validated by literature (Previous/Current)**
Neoplasm and/or hamartoma (disorder)	8,758	916/750	224	268/222	138/110
Traumatic AND/OR non-traumatic injury (disorder)	13,689	2,002/1,806	371	130/259	103/242
Degenerative disorder (disorder)	3,428	310/173	24	30/37	103/29
Surgical procedure (procedure)	20,307	3,064/1,245	355	86/193	607/256
Removal (procedure)	10,031	1,983/570	169	107/82	455/77
Procedure on body region (procedure)	31,006	NA/1,804	NA	NA/275	NA/330
Musculoskeletal finding (finding)	14,706	NA/1,712	NA	NA/308	NA/208
Procedure on organ (procedure)	25,643	NA/1,642	NA	NA/247	NA/317
Mass of body structure (finding)	10,719	NA/865	NA	NA/264	NA/127
Wound (disorder)	3,813	NA/567	NA	NA/80	NA/52

[2] M. 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*, vol. 80, pp. 1–13, 2018.

[3] 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.

[4] 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.

[5] "Snomed ct managed service - us edition release notes - march 2021," https://confluence.ihtsdotools.org/display/RMT/SNOMED+CT+Managed+Service+-+US+Edition+Release+Notes+-+March+2021, (Online; accessed June, 2021).

[6] "Snomed ct united states edition," https://www.nlm.nih.gov/healthit/snomedct/us_edition.html, (Online; accessed June, 2021).

[7] "Fully specified name," https://confluence.ihtsdotools.org/display/DOCEG/Fully+Specified+Name, (Online; accessed June, 2021).

[8] "Snomed ct fully defined concept," https://confluence.ihtsdotools.org/display/DOCGLOSS/fully+defined+concept, (Online; accessed June, 2021).

[9] "Snomed ct primitive concept," https://confluence.ihtsdotools.org/display/DOCGLOSS/primitive+concept, (Online; accessed June, 2021).

[10] 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.

[11] Z. He, J. Geller, and Y. Chen, "A comparative analysis of the density of the snomed ct conceptual content for semantic harmonization," *Artificial intelligence in medicine*, vol. 64, no. 1, pp. 29–40, 2015.

[12] Z. He, J. Geller, and G. Elhanan, "Categorizing the relationships between structurally congruent concepts from pairs of terminologies for semantic harmonization," *AMIA Summits on Translational Science Proceedings*, vol. 2014, p. 48, 2014.

[13] Z. He, Y. Chen, S. de Coronado, K. Piskorski, and J. Geller, "Topological-pattern-based recommendation of umls concepts for national cancer institute thesaurus," in *AMIA Annual Symposium Proceedings*, vol. 2016. American Medical Informatics Association, 2016, p. 618.

[14] F. Zheng and L. Cui, "A lexical-based formal concept analysis method to identify missing concepts in the nci thesaurus," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 1757–1760.

[15] F. Zheng, R. Abeysinghe, and L. Cui, "Identification of missing concepts in biomedical terminologies using sequence-based formal concept analysis," *BMC Medical Informatics and Decision Making*, 2021 [in press].

[16] "Text preprocessing in python: Steps, tools, and examples," https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908, (Online; accessed August, 2021).

[17] G.-Q. Zhang, G. Xing, and L. Cui, "An efficient, large-scale, non-lattice-detection algorithm for exhaustive structural auditing of biomedical ontologies," *Journal of biomedical informatics*, vol. 80, pp. 106–119, 2018.

[18] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[19] "Statistics - 2021aa release," https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html, (Online; accessed Aug, 2021).

[20] National Library of Medicine, "Pubmed," https://pubmed.ncbi.nlm.nih.gov/, (Online; accessed September, 2021).

[21] Explosion, "Spacy," https://spacy.io/, (Online; accessed September, 2021).

[22] The Apache Software Foundation, "Apache lucene," https://lucene.apache.org/, (Online; accessed September, 2021).

[23] A. Das and A. Pratap, "Primary malignant neuroendocrine tumour of pleura: First case report," *Case reports in oncological medicine*, vol. 2016, 2016.

[24] C. Oswalt, "Median nerve injuries and their management." *Southern medical journal*, vol. 70, no. 6, pp. 725–727, 1977.

[25] J. Silvestre, M. Albares, M. Blanes, J. Pascual, and N. Pastor, "Allergic contact gingivitis due to eugenol present in a restorative dental material," *Contact dermatitis*, vol. 52, no. 6, pp. 341–341, 2005.

[26] A. K. Izumi, "Allergic contact gingivostomatitis due to gold," *Archives of dermatological research*, vol. 272, no. 3, pp. 387–391, 1982.

[27] J. Král, J. Hradec, J. Petrásek, P. Jebavỳ, and P. Niederle, "Incidence and changes in mitral regurgitation in balloon valvotomy of the mitral valve. a color doppler study," *Casopis lekaru ceskych*, vol. 131, no. 23, pp. 696–699, 1992.

[28] W. Paaske and J. Laustsen, "Early results of 132 aortic or aorto-iliac arterial reconstructions with the new stretch eptfe vascular prosthesis." *International angiology: a journal of the International Union of Angiology*, vol. 13, no. 4, pp. 296–299, 1994.

[29] E. Skagius, M. Bosnjak, M. Björck, J. Steuer, R. Nyman, and A. Wanhainen, "Percutaneous closure of large femoral artery access with prostar xl in thoracic endovascular aortic repair," *European Journal of Vascular and Endovascular Surgery*, vol. 46, no. 5, pp. 558–563, 2013.

[30] E. Prats, J. Banzo, M. D. Abós, F. Garcia-Lopez, T. Escalera, M. Garcia-Miralles, R. Gaston, and M. J. Asenjo, "Diagnosis of prosthetic vascular graft infection by technetium-99m-hmpao-labeled leukocytes," *Journal of Nuclear Medicine*, vol. 35, no. 8, pp. 1303–1307, 1994.

[31] Y. Bi, S. Li, H. Song, G. Zhang, and Y. Zhang, "Direct vision mitral valve reconstruction: a long-term follow-up of 46 cases," *Zhonghua wai ke za zhi [Chinese Journal of Surgery]*, vol. 36, no. 8, pp. 466–468, 1998.

[32] T. Telila, E. Mohamed, and K. M. Jacobson, “Endovascular therapy for rheumatic mitral and aortic valve disease,” *Current treatment options in cardiovascular medicine*, vol. 20, no. 7, pp. 1–10, 2018.

[33] T. Ito, “Recurrent auricular inflammation caused by kimuras disease: reminiscent of the early phase of relapsing polychondritis?” *Oxford medical case reports*, vol. 2019, no. 9, p. omz091, 2019.

[34] M. D. Rivera-Morales, J. L. Rodríguez-Belén, A. Vera, and L. Ganti, “Perichondritis: not all ear pain is otitis,” *Cureus*, vol. 12, no. 10, 2020.

[35] A. Usoro and M. R. Ehmann, “Acute auricular perichondritis with an effusion,” *Clinical practice and cases in emergency medicine*, vol. 3, no. 4, 2019.

[36] S. Wang, Z. Chen, L. Lin, Q. Pan, B. Wang, F. Liu, and C. Zhang, “Long-term results for traumatic sternoclavicular joint dislocation treated with a sternoclavicular joint-specific plate,” *ANZ Journal of Surgery*, vol. 91, no. 4, pp. 653–657, 2021.

[37] E. Semina, K. Rubina, V. Stepanova, and V. Tkachuk, “Participation of urokinase receptor and its endogenous ligands in brain development and formation of cognitive functions,” *Rossiiskii fiziologicheskii zhurnal imeni IM Sechenova*, vol. 102, no. 8, pp. 881–903, 2016.

[38] S. P. Karande and S. Kini, “Osteoarthritis: Clinical and radiological correlation.” *The Journal of the Association of Physicians of India*, vol. 66, no. 7, pp. 37–39, 2018.

[39] G. Weseloh, L. Lenz, and W. Weidlich, “Enzyme studies of synovial fluid from osteoarthritis knee joints (author’s transl),” *Archiv fur Orthopadische und Unfall-chirurgie*, vol. 88, no. 2, pp. 217–223, 1977.

[40] D. Witoński, M. Wagrowska-Danilewicz, and G. Raczyńska-Witońska, “Distribution of substance p nerve fibers in osteoarthritis knee joint.” *Polish journal of pathology: official journal of the Polish Society of Pathologists*, vol. 56, no. 4, pp. 203–206, 2005.

[41] S. Bokhari, “Tendonitis: the major cause of pain in osteoarthritis knee joint,” *J Ayub Med Coll Abbottabad*, vol. 24, pp. 3–4, 2012.

[42] G.-Q. Zhang, Y. Huang, and L. Cui, “Can snomed ct changes be used as a surrogate standard for evaluating the performance of its auditing methods?” in *AMIA Annual Symposium Proceedings*, vol. 2017. American Medical Informatics Association, 2017, p. 1903.