A Comparison of Exhaustive and Non-lattice-based Methods for Auditing Hierarchical Relations in Gene Ontology

Rashmie Abeysinghe, PhD[1], Fengbo Zheng, PhD[2], Licong Cui, PhD[2,⋆]
[1]Department of Neurology, University of Texas Health Science Center at Houston, Houston, TX
[2]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX

Abstract

Uncovering and fixing errors in biomedical terminologies is essential so that they provide accurate knowledge to downstream applications that rely on them. Non-lattice-based methods have been applied to identify various kinds of inconsistencies in different biomedical terminologies. In previous work, we have introduced two inference-based approaches that were applied in an exhaustive manner to audit hierarchical relations in the Gene Ontology: (1) Lexical-based inference framework, and (2) Subsumption-based sub-term inference framework. However, it is unclear how effective these exhaustive approaches perform compared with their corresponding non-lattice-based approaches. Therefore, in this paper, we implement the non-lattice versions of these two exhaustive approaches, and perform a comprehensive comparison between non-lattice-based and exhaustive approaches to audit the Gene Ontology. The domain expert evaluations performed for the two exhaustive approaches are leveraged to evaluate the non-lattice versions. The results indicate that the non-lattice versions have increased precision than their exhaustive counterparts even though they do not capture some of the potential inconsistencies that the exhaustive approaches identify.

1 Introduction

Biomedical terminologies like Gene Ontology, SNOMED CT and NCI thesaurus have received an increase use in terms of knowledge management; data integration, exchange and semantic interoperability; and decision support and reasoning in biomedicine [1–4]. Biomedical terminologies are constantly curated to reflect the state-of-the-art knowledge of the particular domain that they represent. Though great care is taken to make sure the terminologies reflect the biomedical knowledge accurately, it is inevitable that errors will be introduced due to the manual effort involved in maintaining them combined with how complex terminologies have become over time. It can be laborious to manually audit a modern biomedical terminology due to their size and complexity and hence, automated approaches are preferred.

Non-lattice-based auditing methods have been effectively employed for quality assurance purposes on various biomedical terminologies[5–11]. These methods focus on subgraph fragments in terminologies that are error prone. In contrast, exhaustive methods focus on the entire terminology without restricting to such subgraphs. Leveraging the lexical features of concepts in non-lattice-subgraphs, various inconsistencies have been uncovered. The advantage of non-lattice-based methods over most other terminology auditing methods is that they are capable of not only identifying errors, but also suggesting remediation measures. Therefore, they require much less manual review effort from domain experts. According to the results of non-lattice-based methods, it is unquestionable that they are effective in uncovering errors in biomedical terminologies. However, a comprehensive comparison of non-lattice-based methods against exhaustive methods has not yet been performed to prove their effectiveness in uncovering inconsistencies.

In previous work, we have developed two exhaustive approaches to audit the hierarchical *is-a* relations in Gene Ontology (GO): one is a lexical-based inference approach[12,13], and the other is a subsumption-based sub-term inference framework[14]. In this paper, we implement these two approaches on non-lattice subgraphs to detect potential inconsistencies in GO. Leveraging the previous evaluations performed by domain experts, we perform a comparison of the effectiveness of the exhaustive version and non-lattice version for both approaches.

[⋆] Corresponding author. Email: licong.cui@uth.tmc.edu

2 Background

2.1 Methods to audit biomedical terminologies

Many approaches have been proposed to audit biomedical terminologies[15]. Abstraction networks which are summary graphs of terminologies have been extensively explored to perform quality assurance[16–18]. Zhe et al. have worked on identifying trapezoid structures in the hierarchies of a pair of terminologies to identify missing concepts[19] . Bodenreider has introduced a method to uncover missing *is-a* relations through inference of logical definitions constructed using lexical features of concept names[20] . Zheng et al. have introduced a transformation-based method that leverages Unified Medical Language System (UMLS) knowledge to identify missing hierarchical relations in terminologies in the UMLS[21]. Peng et al. have proposed a new algorithm to predict new GO terms and connect them to existing GO[22]. Mougin et al. have reasoned over relationships to identify redundant relations in GO and detected missing relations by using compositional structure of the concept names[23]. Xing et al.'s work combined dynamic programming with topological sort to detect redundant relation in biomedical terminologies including GO[24]. More recently, deep learning has been explored to audit biomedical terminologies. Zheng et al. have proposed a method that leverage deep learning to predict the concept names of new concepts that comply with the naming convention of the terminology[25]. Liu et al. have introduced a deep learning approach that can predict the placement of a new concept in the hierarchy of SNOMED CT[26].

2.2 Non-lattice subgraphs

Being a lattice is considered a desirable property for a well-formed terminology[5,27,28]. A terminology forms a lattice if any pair of concepts have a unique maximal shared descendant and a unique minimal shared ancestor. A pair of concepts is known as a non-lattice pair, if the they have more than one maximal shared descendant or minimal shared ancestor. For example, in Figure 1, concepts A and B form a non-lattice pair since they share two maximal common ancestors E and F.

The non-lattice pair (A, B) defines a non-lattice subgraph as follows. First, the maximal common descendants (*mcd*) of the non-lattice pair is obtained. For the non-lattice pair (A, B), the maximal common descendants $mcd(A, B) = \{E, F\}$. Then we reversely compute the minimal common ancestors (*mca*) of concepts E and F. This yields us $mca(mcd(A, B)) = \{A, B, C\}$. Then all the concepts and relations between $mcd(A, B)$ and $mca(mcd(A, B))$ is aggregated to form the non-lattice subgraph. This yields a non-lattice subgraph with six concepts $\{A, B, C, D, E, F\}$.

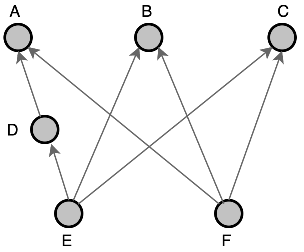

Figure 1: An example non-lattice subgraph.

2.3 Non-lattice-based auditing methods

Previously, we have investigated a number of non-lattice-based methods to audit biomedical terminologies. Four lexical patterns in non-lattice subgraphs were investigated to uncover missing *is-a* relations and missing concepts in SNOMED CT[5]. The same approach was applied to NCI thesaurus introducing two more lexical patterns[6]. We also introduced a method that leverages enriched lexical attributes of concepts in non-lattice subgraphs in SNOMED CT to uncover *is-a* relation inconsistencies[7]. We applied similar approaches on the NCI thesaurus while also investigating the inheritance of lexical attributes from all ancestors and leveraging role definitions of concepts.

Next we discuss the two exhaustive methods from previous work that we will compare later with their non-lattice-based counterparts.

2.4 Lexical-based inference framework

To audit *is-a* relations in GO, we introduced a lexical-based inference framework[12,13]. In this approach, we represented the name of each concept using two models: set-of-words and sequence-of-words. For each model, we generated hierarchically-linked and -unlinked Partial Matching Concept-Pairs (PMCPs), (A, B), such that both A and B have the same number of words, and contain at least one word in common and a fixed number of different words $(n = 1, 2, 3, 4, 5)$. The linked and unlinked concept-pairs further infer corresponding linked and unlinked term-pairs (ITPs) respectively. If the same ITP is inferred by a linked and unlinked PMCP, this is considered as an inconsistency. Applying this approach to the March 28, 2017 release of GO, a total of 5,359 potential inconsistencies were found by the set-of-words model and 4,959 were found by the sequence-of-words model. A random sample of 250 potential inconsistencies identified through this method was evaluated by domain experts to validate their correctness. The results showed that the set-of-words model achieved a precision of 53.78% while the sequence-of-words model achieved a precision of 57.55%.

2.5 Subsumption-based sub-term inference framework

In previous work[14], we developed a subsumption-based sub-term inference framework (SSIF) to audit GO. In SSIF, we represented each concept A with a sequence-based representation $E(A) = [e_1, e_2, e_3, ..., e_n]$, where each element is either a word or a sub-concept. We leveraged part-of-speech tagging, sub-concept matching and antonym tagging to construct the sequence-based representation for each concept. Then, we introduced three conditional rules: Monotonicity, Intersection, and Sub-concept that utilized the sequence-based representation to uncover problematic *is-a* relations. Here, we briefly discuss the three rules.

Monotonicity rule suggests A *is-a* B if both A and B have the same number of elements in their sequence-based representations $E(A)$ and $E(B)$ respectively, their corresponding elements A_i and B_i are either equal or if they are sub-concepts (discussed below) A_i *is-a* B_i, and $E(A)$ does not contain an element which is an antonym of any element of $E(B)$.

Intersection rule suggests a missing *is-a* relation between a concept A and an intersecting concept X as follows. Suppose that A has a pair of ancestors B and C. The intersecting concept X of B and C is defined as another concept that contains the lexical properties of both B and C and is lexically the most general concept that is a descendant of both B and C. Therefore, since A is also a descendant of both B and C, this rule suggests that A *is-a* X. Note that A should not have an element which is an antonym of an element of X.

Sub-concept rule suggests a missing *is-a* relation among a concept and its sub-concept. We say that a concept B is a sub-concept of a concept A if B is a proper substring of A. Suppose that if the sub-concept B is the last element of A, all other elements of A are either sub-concepts or belong to parts-of-speech noun or adjective, and A does not have an element which is an antonym of any element of B, then the Sub-concept rule suggests A *is-a* B.

Applying SSIF to the October 3, 2018 release of GO, 819; 691; and 669 potential inconsistencies were uncovered for Monotonicity, Intersection, and Sub-concept rules respectively. Domain experts evaluated a random sample of 210 potential inconsistencies uncovered by SSIF and the results showed that SSIF achieved a precision of 60.61%, 60.49%, and 46.03% for Monotonicity, Intersection, and Sub-concept rules respectively.

3 Methods

We first extract all the non-lattice subgraphs from both the March 28, 2017 and October 3, 2018 releases of Gene Ontology (the same versions used in the exhaustive approaches) leveraging an efficient large-scale non-lattice detection algorithm[28]. Then we develop the non-lattice version of the two approaches as follows.

3.1 Non-lattice lexical-based inference framework

For both set-of-words model and sequence-of-words model, the non-lattice version of the lexical-based inference framework is as follows. A pair of hierarchically-linked or -unlinked concepts (A, B) form a non-lattice PMCP if both A and B have the same number of words, contain at least one word in common and a fixed number of different words ($n = 1, 2, 3, 4, 5$), and are in the same non-lattice subgraph. From a non-lattice PMCP, a non-lattice ITP is derived by removing the common words across the two PMCPs. We say that an inconsistency exists in a non-lattice subgraph if it contains a linked non-lattice PMCP and an unlinked non-lattice PMCP, which infers the same non-lattice ITP. Figure 2 shows an example non-lattice subgraph that exhibits this scenario. Note that the concepts of linked PMCP are in green and the concepts of the unlinked PMCP are in red. This can be obtained by both the set-of-words and sequence-of-words models.

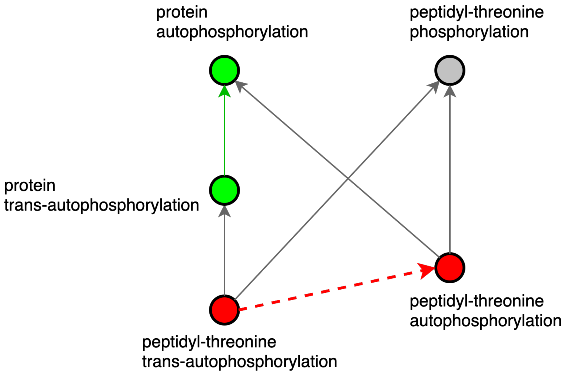

Figure 2: A missing *is-a* relation (dashed link in red) identified by the non-lattice lexical-based inference framework. Concept pair of linked PMCP is in green and concept pair of unlinked PMCP is in red. The suggestion here is that *peptidyl-threonine trans-autophosphorylation* <u>is-a</u> *peptidyl-threonine autophosphorylation*.

3.2 Non-lattice subsumption-based sub-term framework

Exhaustive SSIF proposed three conditional rules: Monotonicity, Intersection, and Sub-concept to uncover potential *is-a* inconsistencies. Here, we define their non-lattice-based counterparts as follows.

Non-lattice-based Monotonicity rule suggests A *is-a* B if concept A and concept B satisfy the following conditions:

- they are in the same non-lattice subgraph;
- they have the same number of elements in their sequence-based representations $E(A)$ and $E(B)$ respectively;
- their corresponding elements A_i and B_i are either equal or if they are sub-concepts, both of them are in the above-mentioned non-lattice subgraph where A_i *is-a* B_i; and
- $E(A)$ does not contain an element which is an antonym of any element of $E(B)$.

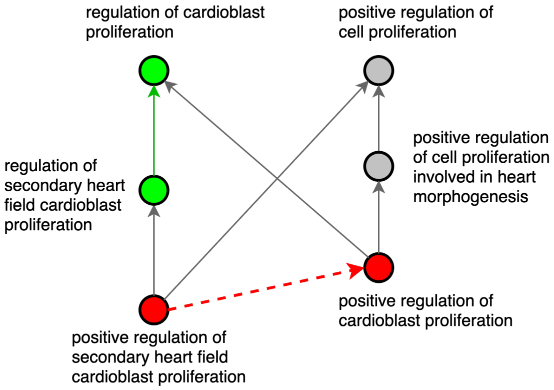

Figure 3: A missing *is-a* relation (dashed link in red) identified by the Monotonicity rule of the non-lattice SSIF. Concepts in green exists as sub-concepts of concepts in red. The suggestion here is *positive regulation of secondary heart field cardioblast proliferation* <u>is-a</u> *positive regulation of cardioblast proliferation*.

Figure 3 shows a non-lattice subgraph exhibiting Monotonicity rule. This example shows a suggested missing *is-a* relation between concepts A = *positive regulation of secondary heart field cardioblast proliferation* and B = *positive regulation of cardioblast proliferation*. Note that with the sequence-based representation, concept A here is represented as $E(A)$ = [*positive, (regulation of secondary heart field cardioblast proliferation)*], i.e., the first element is *positive* and the second element is *regulation of secondary heart field cardioblast proliferation* which is a sub-concept. Similarly, concept B is represented as $E(A)$ = [*positive, (regulation of cardioblast proliferation)*]. Since A and B are both in the same non-lattice subgraph, have two elements each, their first elements are equal and the second elements are sub-concepts having an *is-a* relation which also exists in the same non-lattice subgraph (denoted by green circles), we suggest a missing *is-a* relation among these two concepts (denoted by red dashed line).

Non-lattice-based Intersection rule suggests a missing *is-a* relation as follows. Suppose that A has a pair of ancestors B and C whose intersecting concept is X. If all A, B, C, and X are in the same non-lattice subgraph, this rule suggests a missing relation A *is-a* X if A does not have an element that is an antonym of an element of X. Figure 4 shows a non-lattice subgraph exhibiting Intersection rule. For instance, concepts B = *regulation of establishment of cell polarity* and C = *regulation of establishment or maintenance of cell polarity regulating cell shape* (denoted by green circles) which are in the same non-lattice subgraph has an intersecting concept X = *regulation of establishment of cell polarity regulating cell shape* which is also in the same non-lattice subgraph. Concept A = *regulation of establishment of bipolar cell polarity regulating cell shape* which also exists in the same non-lattice subgraph is a descendant of both B and C. Therefore, we suggest a missing *is-a* relation between the concepts X and A (denoted by the red dashed link).

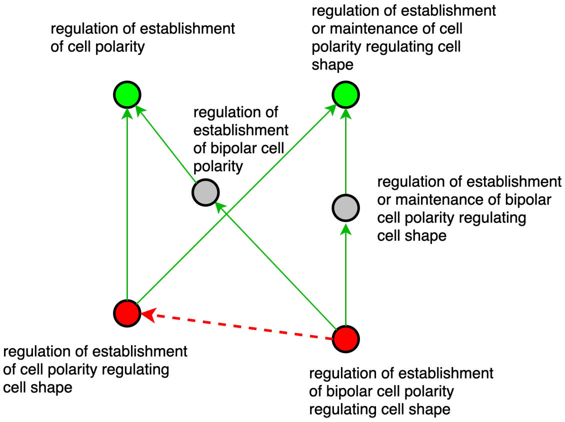

Figure 4: A missing *is-a* relation (dashed link in red) identified by the Intersection rule of the non-lattice SSIF. Concepts in green are the ancestor concepts. The suggestion here is *regulation of establishment of bipolar cell polarity regulating cell shape* <u>*is-a*</u> *regulation of establishment of cell polarity regulating cell shape*.

Non-lattice-based Sub-concept rule suggests a missing *is-a* relation between a concept A and its sub-concept B if

- both of A and B are in the same non-lattice subgraph;
- the sub-concept B is the last element of A, and all other elements of A are either a sub-concept or noun or adjective; and
- A does not have an element which is an antonym of any element of B.

For instance, Figure 5 contains a non-lattice subgraph exhibiting Sub-concept rule. Here, concept A = *positive regulation of phenotypic switching by regulation of transcription from RNA polymerase II promoter* is represented by the sequence-based representation as $E(A)$ = *positive, (regulation of phenotypic switching by regulation of transcription from RNA polymerase II promoter)*. i.e. the last element of A is a subconcept B = *regulation of phenotypic switching by regulation of transcription from RNA polymerase II promoter* which also exists in the same non-lattice subgraph. The remaining element *positive* is an adjective. Therefore, based on the non-lattice Sub-concept rule, we suggest there exists a missing *is-a* relation between A and B (denoted by the dashed red line).

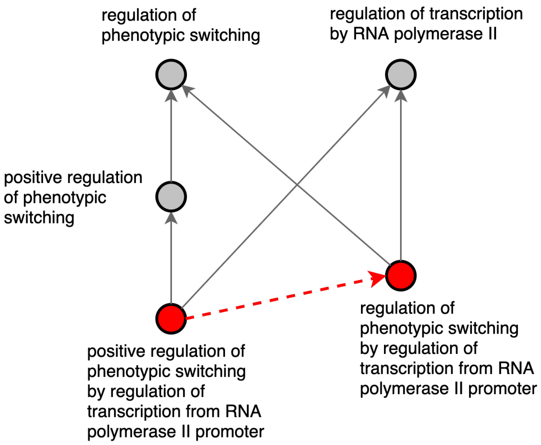

Figure 5: A missing *is-a* relation (dashed link in red) identified by the Sub-concept rule of the Non-lattice SSIF. The suggestion here is *positive regulation of phenotypic switching by regulation of transcription from RNA polymerase II promoter* *is-a* *regulation of phenotypic switching by regulation of transcription from RNA polymerase II promoter*.

3.3 Evaluating non-lattice-based methods

To evaluate the performance of the non-lattice versions of lexical-based inference framework and SSIF, we leverage the evaluations performed on their exhaustive counterparts. If a potential inconsistency identified through the non-lattice approach exists in the evaluation sample of the exhaustive approach, then, that particular sample is considered for evaluating the non-lattice approach. So, we simply take the intersection between the non-lattice results with the exhaustive evaluation set to identify the evaluation samples obtainable by the non-lattice approach. Then, based on the number of valid inconsistencies and false positives, we compute the precision.

4 Results

In this section, we compare the potential inconsistencies uncovered by the exhaustive and non-lattice versions of the lexical-based inference framework and SSIF.

Table 1 shows potential inconsistencies identified by exhaustive and non-lattice lexical-based inference frameworks. For example, the exhaustive approach has identified 5,359 potential inconsistencies with the set-of-words model while the non-lattice-based approach has identified 1,875 potential inconsistencies with the same model.

Table 1: A comparison of all potential inconsistencies uncovered by exhaustive and non-lattice lexical-based inference frameworks.

	Set-of-words	**Sequence-of-words**
Exhaustive	5,359	4,959
Non-lattice	1,875	1,691

Table 2 displays potential inconsistencies uncovered by the three rules of exhaustive and non-lattice versions of SSIF. For instance, the exhaustive version of Monotonicity rule identified 819 potential inconsistencies while the non-lattice version of the Monotonicity rule uncovered 354 potential inconsistencies.

Table 2: A comparison of all potential inconsistencies uncovered by exhaustive and non-lattice SSIF.

	Monotonicity	**Intersection**	**Sub-concept**
Exhaustive	819	691	669
Non-lattice	354	679	75

Table 3 shows a comparison of performance between the exhaustive and non-lattice versions of lexical-based inference framework. For example, the exhaustive version achieved a precision of 53.78% when the set-of-words model was used. The non-lattice-based approach achieved a precision of 58.97% for the set-of-words model.

Table 3: A comparison of the performance of exhaustive and non-lattice lexical-based inference framework.

	No. of potential inconsistencies		**No. of valid inconsistencies**		**Precision**	
	Exhaustive	**Non-lattice**	**Exhaustive**	**Non-lattice**	**Exhaustive**	**Non-lattice**
Set-of-words	238	39	128	23	53.78%	58.97%
Sequence-of-words	212	28	122	20	57.55%	71.43%

Table 4 shows a performance comparison of the exhaustive and non-lattice versions of SSIF. For instance, the exhaustive approach achieved a precision of 60.61% with the Monotonicity rule while the non-lattice-based approach achieved 61.54%.

Table 4: A comparison of the performance of exhaustive and non-lattice SSIF.

	No. of potential inconsistencies		**No. of valid inconsistencies**		**Precision**	
	Exhaustive	**Non-lattice**	**Exhaustive**	**Non-lattice**	**Exhaustive**	**Non-lattice**
Monotonicity	99	39	60	24	60.61%	61.54%
Intersection	81	79	49	49	60.49%	62.03%
Sub-concept	63	8	29	5	46.03%	62.50%

5 Discussion

In this paper, we implemented the non-lattice versions of our previous exhaustive lexical-based inference framework and subsumption-based sub-term inference framework for auditing *is-a* relations in GO, and performed a comparison between exhaustive and non-lattice-based approaches.

5.1 Performance comparison

From the performance comparison in Table 3 for the lexical-based inference approach, it can be seen that the non-lattice approach has outperformed the exhaustive approach considerably. The precision of the non-lattice-based approach is better by 5.19% for set-of-words model and 13.88% for the sequence-of-words model. However, the non-lattice-based approach only uncovers 35% and 34% of the potential inconsistencies that the set-of-words and sequence-of-words of the exhaustive approach identifies respectively. For instance, by both the set-of-words and sequence-of-words models, the exhaustive version identifies the missing *is-a* relation *diadenosine polyphosphate catabolic process is-a small molecule catabolic process* which the non-lattice-based approach does not.

From Table 4, it can be seen that the non-lattice-based approach exceeds the precisions by 0.93% for Monotonicity, 1.54% for Intersection and 16.47% for Sub-concept rules. However, it can also be seen that the non-lattice-based approach identifies 42%, 96%, and 11% of the potential inconsistencies that Monotonicity, Intersection, Sub-concept rules of the exhaustive approach finds respectively. For example, with the Sub-concept rule, the exhaustive version identifies the missing *is-a* relation *skeletal muscle cell differentiation is-a muscle cell differentiation* that the non-lattice version does not.

Therefore, from the results it is clear that the two non-lattice-based approaches perform better than their exhaustive counterparts in terms of precision. It seems that this performance increase also depends on the method itself since some methods (e.g. Sub-concept rule of SSIF) has gained more than the others.

It can be also seen that non-lattice-based approaches miss some of the potential inconsistencies that exhaustive approaches are able to uncover. We did not expect non-lattice subgraphs would capture all kinds of inconsistencies that exists in biomedical terminologies. However, we expected that the concentration of errors in non-lattice subgraphs would be higher than that in the general terminology and therefore, if the same method is applied exhaustively and inside non-lattice subgraphs, the non-lattice version would achieve a better precision.

It is easier to analyze a non-lattice subgraph than analyzing the entire terminology. If non-lattice subgraphs have a higher concentration of errors, such analysis may lead to the identification of new types of inconsistencies. Strategies to uncover and fix inconsistency types identified by analyzing non-lattice subgraphs may be applied exhaustively to identify more inconsistencies that non-lattice subgraphs itself do not capture. Therefore, non-lattice-based approaches may influence future exhaustive approaches in turn uncovering much more inconsistencies than NLS approaches alone can uncover.

5.2 Level differences of concepts and their superconcepts in the potential missing *is-a* identified

Since the non-lattice-based methods focus on substructures of a terminology and exhaustive methods works on the entire terminology, we also performed a level-based analysis of the potential missing *is-a* relations identified by exhaustive and non-lattice-based methods. For a descendant and an ancestor of a potential missing *is-a* relation, we first computed the level in the hierarchy for the descendant and the ancestor. The level of a particular concept is the number of concepts in the longest path from root to the particular concept. For a particular potential missing *is-a* relation, we subtract the level of the ancestor from the descendant to get the level difference.

Figure 6 shows a plot of the level differences and the number of *is-a* relations for the lexical-based inference approach. It can be seen that the lexical-based inference approach favors low level differences for both exhaustive and non-lattice-based methods. This means that the concepts in the potential missing *is-a* relations tend to be closer to each other in the hierarchy of the ontology. Other than the set-of-words model in the non-lattice-based approach, all other models have most number of potential missing *is-a* relations with a level difference of 0.

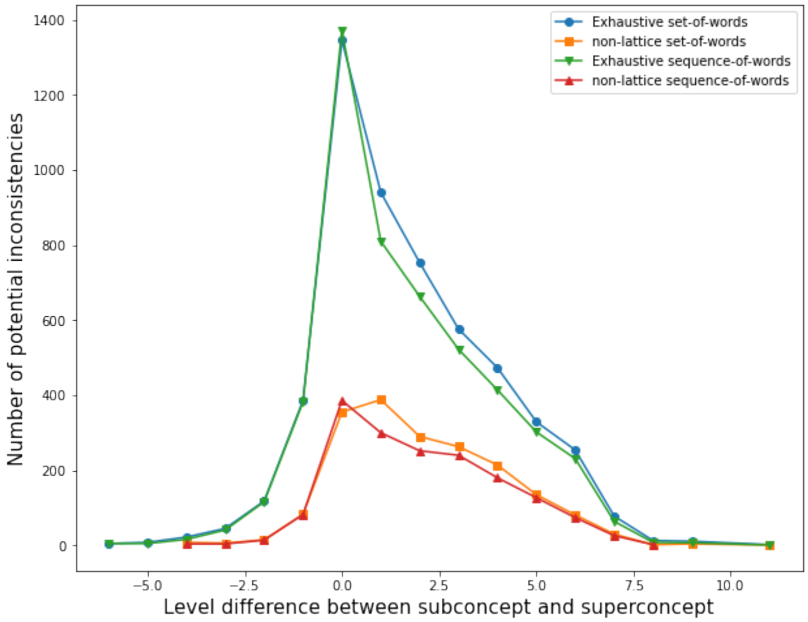

Figure 6: Distribution of level differences of descendant and ancestor in potential inconsistencies uncovered through exhaustive and non-lattice version of lexical-based inference framework.

Figure 7 displays the plot for level differences and the number of *is-a* relations for SSIF. It seems SSIF also tends to suggest hierarchically closer concepts as potentially missing *is-a*. Other than the Sub-concept rule in the exhaustive approach, all others have most number of potentially missing *is-a* with a level difference of 1.

5.3 Future work

In this work, we leveraged two previous exhaustive methods to compare non-lattice-based and exhaustive methods. We leveraged evaluations from the exhaustive methods to evaluate the non-lattice-based methods. Because of this, the

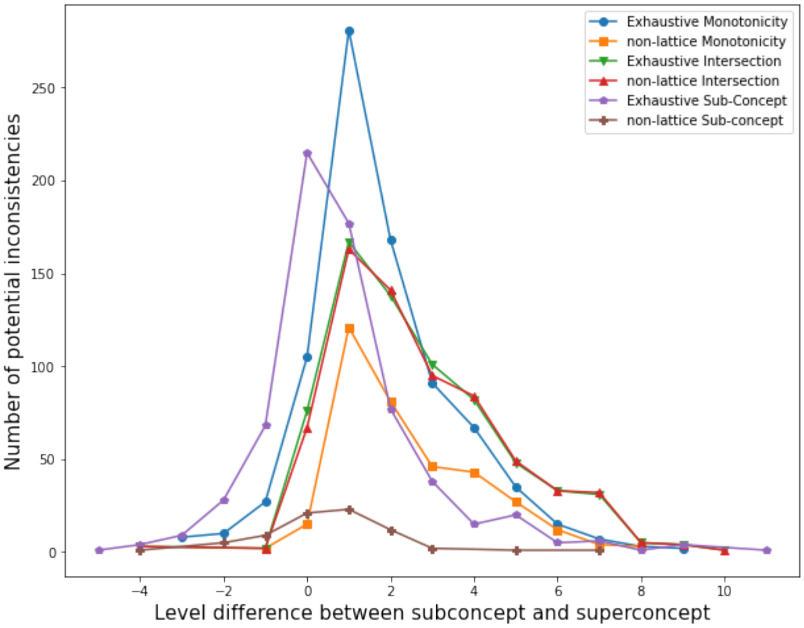

Figure 7: Distribution of level differences of descendant and ancestor in potential inconsistencies uncovered through exhaustive and non-lattice versions of SSIF.

number of evaluation samples for the non-lattice-based methods were on the smaller size. In the future we would like to perform a thorough domain expert evaluation exclusively for both the methods.

In this work, we also focused on one terminology: Gene Ontology. In the future we expect to see whether these results are terminology specific by performing a similar comparison on other major biomedical terminologies like SNOMED CT and NCI thesaurus.

6 Conclusion

In this paper, we performed a comparison of exhaustive methods vs non-lattice-based methods for auditing biomedical terminologies. We implemented non-lattice versions of two of our previous exhaustive works in auditing the Gene Ontology: (1) lexical-based inference framework and (2) subsumption-based sub-term inference framework. We leveraged the domain expert evaluations performed for the exhaustive methods to evaluate the non-lattice-based methods. The results indicate that non-lattice-based methods achieved better precisions than their exhaustive counterparts though, they do not capture some of the potential inconsistencies identified by exhaustive approaches.

Acknowledgment

This work was supported by the National Science Foundation (NSF) through grant IIS-1931134, and the 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.

References

1. Gene Ontology Consortium and others. The gene ontology (GO) project in 2006. Nucleic acids research. 2006;34(suppl 1):D322–D326.
2. Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's thesaurus and ontology. Journal of Web Semantics First Look 1_1_4. 2003;.
3. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics. 2006;121:279.
4. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support.

Yearbook of medical informatics. 2008;p. 67.

5. Cui L, Zhu W, Tao S, Case JT, Bodenreider O, Zhang GQ. Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT. Journal of the American Medical Informatics Association. 2017;24(4):788–798.
6. Abeysinghe R, Brooks MA, Talbert J, Cui L. Quality assurance of NCI Thesaurus by mining structural-lexical patterns. In: AMIA Annual Symposium Proceedings. vol. 2017. American Medical Informatics Association; 2017. p. 364.
7. 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.
8. 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.
9. Cui L, Abeysinghe R, Zheng F, Tao S, Zeng N, Hands I, et al. Enhancing the Quality of Hierarchic Relations in the National Cancer Institute Thesaurus to Enable Faceted Query of Cancer Registry Data. JCO Clinical Cancer Informatics. 2020;4:392–398.
10. Zheng F, Abeysinghe R, Cui L. A Hybrid Method to Detect Missing Hierarchical Relations in NCI Thesaurus. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2019. p. 1948–1953.
11. Zheng F, Abeysinghe R, Sioutos N, Whiteman L, Remennik L, Cui L. Detecting missing IS-A relations in the NCI Thesaurus using an enhanced hybrid approach. BMC Medical Informatics and Decision Making. 2020;20(10):1–11.
12. Abeysinghe R, Hinderer EW, Moseley HN, Cui L. Auditing subtype inconsistencies among gene ontology concepts. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2017. p. 1242–1245.
13. Abeysinghe R, Zheng F, Hinderer EW, Moseley HN, Cui L. A lexical approach to identifying subtype inconsistencies in biomedical terminologies. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2018. p. 1982–1989.
14. Abeysinghe R, Hinderer III EW, Moseley HN, Cui L. SSIF: Subsumption-based Sub-term Inference Framework to audit Gene Ontology. Bioinformatics. 2020;36(10):3207–3214.
15. Amith M, He Z, Bian J, Lossio-Ventura JA, Tao C. Assessing the practice of biomedical ontology evaluation: Gaps and opportunities. Journal of biomedical informatics. 2018;80:1–13.
16. Halper M, Gu H, Perl Y, Ochs C. Abstraction networks for terminologies: supporting management of "big knowledge". Artificial intelligence in medicine. 2015;64(1):1–16.
17. Ochs C, Geller J, Perl Y, Chen Y, Agrawal A, Case JT, et al. A tribal abstraction network for SNOMED CT target hierarchies without attribute relationships. Journal of the American Medical Informatics Association. 2014;22(3):628–639.
18. Ochs C, Geller J, Perl Y, Chen Y, Xu J, Min H, et al. Scalable quality assurance for large SNOMED CT hierarchies using subject-based subtaxonomies. Journal of the American Medical Informatics Association. 2015;22(3):507–518.
19. He Z, Geller J, Elhanan G. Categorizing the relationships between structurally congruent concepts from pairs of terminologies for semantic harmonization. AMIA Summits on Translational Science Proceedings. 2014;2014:48.
20. Bodenreider O. Identifying Missing Hierarchical Relations in SNOMED CT from Logical Definitions Based on the Lexical Features of Concept Names. In: ICBO/BioCreative. vol. 2016; 2016. .
21. Zheng F, Shi J, Yang Y, Zheng WJ, Cui L. A transformation-based method for auditing the IS-A hierarchy of biomedical terminologies in the Unified Medical Language System. Journal of the American Medical Informatics Association. 2020;27(10):1568–1575.
22. Peng J, Wang T, Wang J, Wang Y, Chen J. Extending gene ontology with gene association networks. Bioinformatics. 2016;32(8):1185–1194.
23. Mougin F. Identifying redundant and missing relations in the gene ontology. In: MIE; 2015. p. 195–199.
24. Xing G, Zhang GQ, Cui L. FEDRR: fast, exhaustive detection of redundant hierarchical relations for quality improvement of large biomedical ontologies. BioData mining. 2016;9(1):1–18.
25. Zheng F, Cui L. Exploring Deep Learning-based Approaches for Predicting Concept Names in SNOMED CT. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2018. p. 808–813.
26. Liu H, Geller J, Halper M, Perl Y. Using Convolutional Neural Networks to Support Insertion of New Concepts into SNOMED CT. In: AMIA Annual Symposium Proceedings. vol. 2018. American Medical Informatics Association; 2018. p. 750.
27. Zhang GQ, Bodenreider O. Large-scale, exhaustive lattice-based structural auditing of SNOMED CT. In: AMIA Annual Symposium Proceedings. vol. 2010. American Medical Informatics Association; 2010. p. 922.
28. Zhang GQ, Xing G, Cui L. An efficient, large-scale, non-lattice-detection algorithm for exhaustive structural auditing of biomedical ontologies. Journal of biomedical informatics. 2018;80:106–119.