NEO: Systematic Non-Lattice Embedding of Ontologies for Comparing the Subsumption Relationship in SNOMED CT and in FMA Using MapReduce

Wei Zhu, Guo-Qiang Zhang, PhD, Shiqiang Tao, MS, Mengmeng Sun, Licong Cui, PhD
Department of EECS, Case Western Reserve University, Cleveland, OH
Division of Medical Informatics, Case Western Reserve University, Cleveland, OH

Abstract. A structural disparity of the subsumption relationship between FMA and SNOMED CT’s Body Structure sub-hierarchy is that while the is-a relation in FMA has a tree structure, the corresponding relation in Body Structure is not even a lattice. This paper introduces a method called NEO, for non-lattice embedding of FMA fragments into the Body Structure sub-hierarchy to understand (1) this structural disparity, and (2) its potential utility in analyzing non-lattice fragments in SNOMED CT. NEO consists of four steps. First, transitive, upper- and down-closures are computed for FMA and SNOMED CT using MapReduce, a modern scalable distributed computing technique. Secondly, UMLS mappings between FMA and SNOMED CT concepts are used to identify equivalent concepts in non-lattice fragments from Body Structure. Then, non-lattice fragments in the Body Structure sub-hierarchy are extracted, and FMA concepts matching those in the non-lattice fragments are used as the seeds to generate the corresponding FMA fragments. Lastly, the corresponding FMA fragments are embedded to the non-lattice fragments for comparative visualization and analysis. After identifying 8,428 equivalent concepts between the collection of over 30,000 concepts in Body Structure and the collection of over 83,000 concepts in FMA using UMLS equivalent concept mappings, 2,117 shared is-a relations and 5,715 mismatched relations were found. Among Body Structure’s 90,465 non-lattice fragments, 65,968 (73%) contained one or more is-a relations that are in SNOMED CT but not in FMA, even though they have equivalent source and target concepts. This shows that SNOMED CT may be more liberal in classifying a relation as is-a, a potential explanation for the fragments not conforming to the lattice property.

Introduction

Ontologies serve as a knowledge source in many biomedical applications including information extraction [1], information retrieval [2], data integration [3], data management [4], and decision support [5]. The Unified Medical Language System (UMLS) [6] is the largest integrated repository of biomedical ontologies. It includes the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), the Foundational Model of Anatomy (FMA), and the Gene Ontology. Among these ontologies, SNOMED CT is the most comprehensive clinical terminology system used worldwide. FMA provides a unifying framework for the nature of the diverse entities for human anatomy.

However, SNOMED CT and FMA are not in complete agreement when it comes to the usage of is-a classification. For example (Fig. 1), both the concepts “Cardiac atrium” and “Interatrial septum” are in SNOMED CT and in FMA, as mapped by UMLS. While the relation “Interatrial septum” is-a “Cardiac atrium” is asserted in SNOMED CT, FMA does not include this relation. Instead, FMA asserts a different relation only: “Interatrial septum” is a constitutional part of “Cardiac atrium.”

The purpose of this paper is to provide a systematic study of such a structural disparity between FMA and SNOMED CT, motivated by the observation that the is-a relation in FMA has a tree structure, but the corresponding relation in SNOMED CT is not even a lattice (see Fig. 2 in Background). We introduce a method called NEO, Non-Lattice Embedding of Ontologies, for systematic structural embedding of FMA fragments into the Body Structure sub-hierarchy, to understand this structural disparity and its potential utility in analyzing non-lattice fragments (Fig. 2) in SNOMED CT.

NEO builds on existing work in two areas. One is UMLS mappings between FMA and SNOMED CT concepts, which provides equivalent concepts used in the structure embedding. The other is the cloud computing technology called MapReduce, for scalable, large-scale ontology quality assurance work [11, 12].

After identifying 8,428 equivalent concepts between Body Structure and FMA, we found 2,117 shared is-a relations and 5,715 relation mismatches. 73% of Body Structure’s non-lattice fragments contained one or more is-a relations that are not in FMA. Our results show that SNOMED CT may be more liberal in classifying a relation as subsumption (is-a), a potential explanation for the fragments not conforming to the lattice property.
1 Background

SNOMED CT. SNOMED CT is the most comprehensive, multilingual clinical healthcare terminology in the world, developed by the International Health Terminology Standard Development Organization (IHTSDO) [15]. It provides consistent representation of clinical content in electronic health records and has been used in more than fifty countries. It contains over 311,000 active concepts organized into 19 sub-hierarchies including Clinical Finding and Body Structure. The version used in this work is the US edition of SNOMED CT released in March 2014, which contains over 30,000 concepts in the Body Structure sub-hierarchy.

FMA. FMA is a domain ontology of anatomy with concepts and relationships representing the phenotypic structure of the human body, developed by the Structural Informatics Group at the University of Washington [16]. The version used in this work is FMA v3.1, which contains more than 83,000 concepts. Relationships between FMA concepts include “is a,” “part of,” “constitutional part of,” and “regional part of.”

Alignment of Anatomy in SNOMED CT and FMA. The representations of anatomy in FMA and SNOMED CT were compared using lexical alignment and structural validation in [7]. The shared concepts across the two systems were identified lexically through exact match and after normalization, and validated by the structural similarity using is-a and “part of” relations. Different from ontology alignment work like [7], in this paper, we directly use the associations provided by UMLS to establish the correspondence between concepts in SNOMED CT’s Body Structure and FMA, and then structurally embedding FMA into SNOMED CT’s Body Structure using non-lattice fragments.

UMLS. UMLS is the largest integrated repository of biomedical vocabularies, developed by the US National Library of Medicine [6]. The 2014AA release of UMLS, used for this study, covers over 2.9 million concepts from more than 140 source vocabularies including SNOMED CT and FMA. Each UMLS concept has a concept unique identifier (CUI), and maps to the corresponding concept names in the various source vocabularies. For example, the concept “Lymph node of trunk” has a CUI “C0729854.” It associates to three concept names in SNOMED CT: “Lymph node of trunk,” “Lymph node structure of trunk,” and “Lymph node structure of trunk (body structure)” (SNOMED CT code: 312502003), and one concept name in FMA: “Lymph node of trunk” (FMA code: 66131). Such information is provided in the distribution file MRCONSO, which contains each UMLS concept, its CUI, source vocabulary (SAB), concept name (STR), and other attributes.

Ontology Quality Assurance. A well-formed ontology often has a lattice structure in terms of the subsumption relationship (e.g., “is-a”) [9, 10, 11], that is, every pair of concepts in the ontology have one maximal common descendant (or maximal lower bound) and one minimal common ancestor (or minimal upper bound). For example, FMA forms a lattice in terms of the is-a relation. Although desirable, most ontologies do not satisfy the lattice property. For instance, 90,465 non-lattice pairs were identified in the Body Structure sub-hierarchy for the March 2014 release of SNOMED CT US edition [11]. In the lattice-based evaluation [9, 10, 11], missing concepts were identified as one of the main reasons for the non-lattice property (non-lattice pairs). Fig. 2 is a SNOMED CT non-lattice fragment. The double-circled concepts “Tissue specimen from breast” and “Tissue specimen from heart” legitimately share the two features of being a kind of tissue specimen and a kind of specimen from trunk. However, these concepts share two minimal common ancestors: “Tissue specimen” and “Specimen from trunk,” highlighted in pink. To make it lattice-conforming, one could add a missing concept “Tissue specimen from trunk” (dashed in Fig. 2). In this paper, to further explore potential causes of non-lattice fragments in SNOMED CT, we take advantage of both the non-lattice fragments identified in SNOMED CT’s Body Structure and FMA’s lattice structure, and embed FMA fragments into SNOMED CT’s non-lattice fragments. Here a non-lattice fragment consists of a non-lattice pair of concepts, their maximal common ancestors, as well as all other concepts between them.

2 Methods

Our method NEO for non-lattice embedding of ontologies consists of four steps. First, transitive, upper- and down-closures are computed for FMA v3.1 and SNOMED CT’s Body Structure (03/2014 version) using MapReduce. Secondly, equivalent concepts between FMA and SNOMED CT’s Body Structure are identified using UMLS mappings.
Thirdly, non-lattice fragments in the Body Structure sub-hierarchy are extracted, and FMA concepts matching the non-lattice fragments are used as the seeds to generate the corresponding FMA fragments. Lastly, the corresponding FMA fragments are embedded to the non-lattice fragments for comparative visualization and analysis.

Computing Closures. Given a concept \( c \) in an ontology, we use \( c^\uparrow \) and \( c^\downarrow \) to denote its upper-closure and down-closure respectively, with respect to the hierarchical order of the ontology. By “\( c \)’s upper-closure” we mean all the ancestors of \( c \) excluding itself, and by “\( c \)’s down-closure” we mean all the descendants of \( c \), excluding itself. For SNOMED CT’s Body Structure, the transitive, upper- and down-closures are calculated in terms of the is-a relation. For FMA, the relations including “part of,” “regional part of,” “constitutional part of,” “systemic part of,” “member of,” and “branch of” are also used to calculate the closures in addition to the is-a relation. These relations are included in order to provide a reference point for the relationships between matched concepts in non-lattice fragments and their corresponding FMA fragments. Such closures are also used directly for the rendering and analysis of embedded (or merged) fragments.

Sequential algorithms for computing transitive closures, such as Floyd-Warshall algorithm, has a cubic time complexity. Therefore, it is time-consuming to use it for large ontological structures such as SNOMED CT (>300k concepts, >450k is-a relations) and FMA (>83k of concepts, >2.5m of relations). We develop a parallel, distributed algorithm to compute transitive closure using MapReduce [12]. This algorithm consists of two main steps. First a hash map of concepts and their direct parents is stored in each computing nodes. Then in the map phase, each mapper reads in a concept and recursively collects its ancestors, level by level, until no direct parents can be found. In the reduce phase, each reducer emits all concept-ancestor pairs. Computing transitive closure for Body Structure and FMA each took less than 30 seconds. Also, upper- and down-closures are calculated using MapReduce to generate concept-closure pairs \((c, c^\uparrow)\) and \((c, c^\downarrow)\), which are used for generating a fragment from any given collection of seed concepts.

Identifying Equivalent Concepts using UMLS. We use the distribution file MRCONSO provided by UMLS (2014AA release) to extract equivalent concepts between SNOMED CT and FMA. A concept in SNOMED CT and a concept in FMA are considered equivalent if they share the same CUI. A total of 8,428 equivalent concepts are identified. These equivalent concepts are used to extract FMA fragments corresponding to the non-lattice fragments in SNOMED CT’s Body Structure, as described in the following subsection.

Extracting Non-lattice Fragments for SNOMED CT’s Body Structure Sub-hierarchy. We use non-lattice pairs and their maximal lower bounds identified for SNOMED CT’s Body Structure in [11] to generate non-lattice fragments. Given a non-lattice concept pair \( C = \{c_1, c_2\} \) and the set \( L \) of their maximal lower bounds, the corresponding non-lattice fragment is computed using the formula:

\[
C \cup L \cup \bigcup_{c \in C, l \in L} \{c^\downarrow \cap l^\uparrow\}
\]  

(1)

Since there are 90,465 non-lattice pairs, a sequential approach to computing non-lattice fragments may take several hours. To speed up, we generate non-lattice fragments in parallel using MapReduce:

- First, two hash maps are distributed to every computing node. One hash map stores concepts and their upper-closures, and the other hash map stores concepts and their down-closures.
- Then, in the map phase, each mapper reads in a non-lattice concept pair \( C \) and their maximal lower bounds \( L \), finds down-closures for the concepts in \( C \) and upper-closures for the concepts in \( L \) from the hash maps, and performs set operations to obtain the non-lattice fragment.
- Finally, in the reduce phase, each reducer emits the non-lattice pairs and their non-lattice fragments.

Extracting Corresponding FMA Fragments. After equivalent concepts in FMA and SNOMED CT’s Body Structure sub-hierarchy are identified, we find all matching concepts in FMA for each non-lattice fragment in Body Structure. Using these matching concepts as seeds, we construct the corresponding FMA fragment in the following way. Suppose \( S \) is the set of FMA concepts matched with those in a non-lattice fragment. First we identify the maximal and minimal of concepts in \( S \) by the formulas \( S - \bigcup_{s \in S} s^\uparrow \) and \( S - \bigcup_{s \in S} s^\downarrow \), respectively. Then the corresponding FMA fragment is obtained similarly as Eq. (1) for computing a non-lattice fragment.

Embedding FMA Fragments to Non-lattice Fragments. Given a non-lattice fragment in SNOMED CT’s Body Structure, the corresponding FMA fragment is embedded to the non-lattice fragment to visualize and compare the structures of the two fragments. For the embedding (or merging) of the two fragments, we not only need the matched concepts from both fragments, but also the matching relations. When mapping the relations, we distinguish matched
“is-a” and mismatched “is-a” relation. A matched “is-a” relation is one that is in both SNOMED CT and FMA. A mismatched “is-a” relation is one that is in SNOMED CT but not in FMA (which could be relations other than “is-a”).

Then, we merge the two fragments (both concepts and relations) together and render them using topological sort, a well-known rendering algorithm for directed acyclic graph. The rendering of the merged fragments is implemented using the svg (scalable vector graphics) drawing library D3 (http://www.d3js.org) (see Fig. 4 for an example of the merged fragments).

3 Results

Matching Concepts and Relations. A total of 8,428 equivalent concepts were identified among over 30,000 concepts in SNOMED CT’s Body Structure and over 83,000 concepts in FMA. To illustrate, in Fig. 3, concepts 1, 3, 5, 7 in SNOMED CT are equivalent to concepts 2, 4, 6, 8 in FMA, respectively. We also identified 7,832 relations having equivalent source and target concepts between SNOMED CT’s Body Structure and FMA. Among these 7,832 relations, 2,117 (27%) are is-a relations in both SNOMED CT and FMA (i.e., matched), and 5,715 are is-a relations in SNOMED CT but not is-a relations in FMA (i.e., mismatched). As illustrated in Fig. 3, the is-a relation between concepts 1 and 3 in SNOMED CT is matched with the is-a relation between 2 and 4 in FMA, but the is-a relation between concepts 5 and 7 in SNOMED CT is mismatched with non is-a relation between 6 and 8 in FMA.

Mismatched Relations in Non-lattice Fragments. Among 90,465 non-lattice fragments in SNOMED CT’s Body Structure [11], 65,968 (73%) contained one or more is-a relations that are in SNOMED CT but not in FMA, even though they have equivalent source and target concepts. This shows that SNOMED CT may be more liberal in classifying a relation as subsumption, which may cause the fragments not conforming the lattice property. Fig. 5 shows 10 frequently mismatched relations in SNOMED CT (is-a) and FMA (not is-a). For example, the relationship of Structure of pelvic viscus and Structure of abdominal viscus has a is-a relation in SNOMED CT, but “member of” relation in FMA.

Visualization of Merged Fragments. Fig. 4 displays the merged ontological fragments determined by the non-lattice pair “Cardiac chamber (91744000)” and “Cardiac septum (10746000)” in SNOMED CT’s Body Structure. The blue nodes and edges represent concepts and is-a relations in SNOMED CT, respectively. The red nodes and edges represent concepts and is-a relations in FMA, the green edges represents other relations in FMA. The gray nodes and edges represent matched concepts and is-a relations between SNOMED CT and FMA. The dotted gray edges represent non-exact matching relations between them, that is, is-a relations in SNOMED CT but other relations in FMA. There is no blue node in Fig. 4, since all the concepts in the non-lattice fragment determined by the given non-lattice pair have matched concepts in FMA.

It can be seen from Fig. 4 that the pair “Cardiac chamber” and “Cardiac septum” has three maximal common descendants: “Interventricular septum,” “Atrioventricular septum,” and “Interatrial septum.” It is worth noting that removing
1. Structure of pelvic viscus (8279000) member of Structure of abdominal viscus (19203006)
2. Small intestinal structure (30315005) regional part of Intestinal structure (113276009)
3. Uterine structure (35039007) member of Structure of pelvic viscus (8279000)
4. Penile structure (18911002) regional part of Male external genitalia structure (90418005)
5. Chest wall structure (78904004) constitutional part of Thoracic structure (51185008)
6. Male internal genitalia structure (38242008) regional part of Male genital structure (90264002)
7. Large intestinal structure (14742008) regional part of Intestinal structure (113276009)
8. Pelvic wall structure (3665003) constitutional part of Pelvic structure (12921003)
9. Abdominal wall structure (83908009) constitutional part of Abdominal structure (113345001)
10. Endometrial structure (2739003) constitutional part of Structure of uterine wall (245485002)

Figure 5: 10 frequently mismatched relations in FMA occurring in non-lattice fragments in SNOMED CT as is-a relation.

any of the dotted gray edges or blue edges will reduce the number of maximal lower bounds. For example, removing the dotted is-a relation from “Interatrial septum” to “Cardiac septum” will result in two (instead of three) maximal common descendants in SNOMED CT. If mismatched is-a relations are removed in Fig. 4, we obtain a lattice conforming fragment.

Discussion. Our approach of embedding of FMA fragments to SNOMED CT’s Body Structure sub-hierarchy can be applied to other domains to structurally compare two similar ontologies using order fragments. Our results indicate that one cause for the occurrence of non-lattice fragments in SNOMED CT might be due to its more liberal use of is-a relations (than in FMA).

Conclusion. This paper presented NEO, a systematic method to structurally embed FMA fragments into SNOMED CT’s Body Structure sub-hierarchy using non-lattice fragments. 73% of non-lattice fragments contain is-a relations that are not found in FMA. This shows that SNOMED CT is more liberal in classifying a relation as subsumption, a potential reason for inducing non-lattice fragments in Body Structure.

Acknowledgement. The authors acknowledge partial support by the Case Western Reserve University CTSA Grant UL1TR000439.

References
[1] Cui L. Ontology-guided health information extraction, organization, and exploration. PhD Dissertation, Department of EECS, Case Western Reserve University, August 2014.
[2] Cui L, Tao S, Zhang GQ. A Semantic-based Approach for Exploring Consumer Health Questions Using UMLS. AMIA Annual Symp Proc 2014, pages 432-441, 2014.
[3] Zhang GQ, Cui L, Lhatoo S, Schuele S, Sahoo S. MEDCIS: Multi-Modality Epilepsy Data Capture and Integration System. AMIA Annual Symp Proc 2014, pages 1248-1257, 2014.
[4] Jayapandian C, Chen CH, Dabir A, Lhatoo S, Zhang GQ, Sahoo S. Domain Ontology As Conceptual Model for Big Data Management: Application in Biomedical Informatics. International Conference on Conceptual Modeling 2014, 144-57, 2014.
[5] Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Geissbuhler A, Kulkowski C, editors. IMIA Yearbook of Medical Informatics 2008. Methods Inf Med 2008;47(Suppl 1):67-79.
[6] Bodenreider, O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl 1), D267-D270, 2004.
[7] Bodenreider O, Zhang S. Comparing the representation of anatomy in the FMA and SNOMED CT. AMIA Annual Symp Proc 2006, pages 46-50, 2006.
[8] Jiang G, Chute CG. Auditing the semantic completeness of SNOMED CT using formal concept analysis. J Am Med Inform Assoc 2009;16(1):89-102.
[9] Zhang GQ and Bodenreider O. Using SPARQL to Test for Latticess: application to quality assurance in biomedical ontologies. The Semantic Web-ISWC 2010, pages 273-288, 2010.
[10] Zhang GQ and Bodenreider O. Large-scale, exhaustive lattice-based structural auditing of SNOMED CT. American Medical Informatics Association (AMIA) Annual Symposium, pages 922-926, 2010.
[11] Zhang GQ, Zhu W, Sun M, Tao S, Bodenreider O, Cui L. MaPLE: A MapReduce Pipeline for Lattice-based Evaluation of SNOMED CT. IEEE International Conference on Big Data 2014, pages 754-759, 2014.
[12] Tao S, Cui L, Zhu W, Sun M, Bodenreider O, Zhang GQ. Mining Relation Reversals in the Evolution of SNOMED CT Using MapReduce. AMIA Joint Summits on Translational Science 2015.
[13] Luo L, Mejino JLV, Zhang GQ. An Analysis of FMA Using Structural Self-Bisimilarity. J. Biomedical Informatics 2013;46(3):497-505.
[14] Mortensen JM, Mark AM, Natalya FN. Ontology Quality Assurance with the Crowd. In First AAAI Conference on Human Computation and Crowdsourcing. 2013.
[15] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Stud Health Technol Inform Vol. 121, pages 279-90, 2006.
[16] Rosse C, Mejino JLV. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. Journal of biomedical informatics 2003; 36(6):478-500.