Proceedings of AMIA Annual Symposium 2006:804-808.

Enhancing Biomedical Ontologies through Alignment of Semantic Relationships: Exploratory Approaches

Lowell Vizenor[1], Ph.D., Olivier Bodenreider[1*], M.D., Ph.D., Lee Peters[1], MS, Alexa T. McCray[2], Ph.D.
[1]National Library of Medicine, Bethesda, Maryland
[2]Harvard Medical School, Boston, Massachusetts

Objective: *This paper investigates several methods for aligning Metathesaurus relationships with their counterparts in the UMLS Semantic Network. Unlike the categorization link defined between Metathesaurus concepts and Semantic Network types, no such correspondence exists between the relationships at these two levels of the UMLS.* ***Methods:*** *The first approach attempts to elicit the semantics of Metathesaurus relationships through an examination of their relata at different levels: concept, high-level ancestors and semantic types. The second approach examines the frequency of association between a given Semantic Network relationship and the actual relationships observed in the Metathesaurus between the concepts categorized by these semantic types.* ***Results****: A total of 139 relationships are present in the Metathesaurus. Using the methods described in this paper, 80 (58%) could be aligned with Semantic Network relationships. The remaining relationships are vocabulary internal, used, for example, for vocabulary management or to indicate strictly lexical relationships. The work reported here is a first step in the attempt to build a more comprehensive ontology of biomedical relationships.*

INTRODUCTION

Two complementary yet independent Unified Medical Language System® (UMLS®) knowledge sources are the Metathesaurus® and the Semantic Network. The Metathesaurus is a large repository of interrelated concepts coming from one hundred biomedical vocabularies. The Semantic Network, by contrast, is a small, manually curated high-level network of 135 semantic types and 54 semantic relationships. The UMLS editors assign categorization links, which thereby relate the two structures. More precisely, every Metathesaurus concept is assigned to at least one semantic type, independently of its hierarchical position in a source vocabulary. The rationale for this two-level structure is to provide a uniform semantics to the concepts "regardless of the particular structure of the source vocabulary" [1].

* corresponding author: olivier@nlm.nih.gov

In addition to the over one million Metathesaurus concepts there are also a number of Metathesaurus relationships—most of which come from the individual source vocabularies. In this case, however, the UMLS does not directly link Metathesaurus relationships to Semantic Network relationships. One consequence of this is that Semantic Network relationships cannot be used in a straightforward way to validate Metathesaurus relationships. For example, as shown in Figure 1, one auditing method for the UMLS is to simply check the compatibility between a relationship asserted between two concepts in the Metathesaurus (R_M) and the possible relationships defined in the Semantic Network (R_{SN}) between the semantic types of these two concepts. Intuitively, R_M is expected to be equivalent to or subsumed by R_{SN}. However, since no equivalence or subsumption relations are defined between relationships across the two levels of the UMLS, validation on a large scale is not easily accomplished [2, 3]. The objective of this study is to explore methods for establishing such relations (e.g., R_M equivalent to R_{SN}, R_M more specific than R_{SN}).

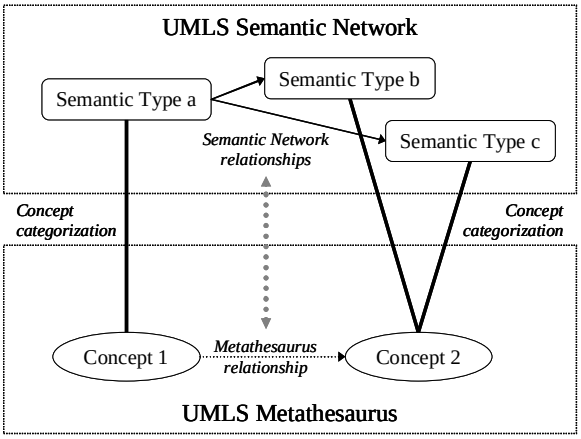

Figure 1 -- Metathesaurus and Semantic Network relationships

This work is part of a broader and on-going project of enriching the Semantic Network through the development of an ontology of biomedical relationships. We develop a number of related methods for aligning

Metathesaurus relationships with Semantic Network relationships—a preliminary step in the development of such an ontology of biomedical relationships.

BACKGROUND

The general framework of this study is that of an ontology alignment. However, in contrast to most existing ontology alignment methods [4], our primary focus is the alignment of relationship (not concepts) across ontologies. Because the two knowledge sources compared in this study represent knowledge at widely different levels of granularity, it is difficult to use existing tools.

This study also represents an attempt to align the two UMLS knowledge sources under investigation in [5]. There, however, the authors used sets of concepts (and not relationships) to align concepts and semantic types. To our knowledge, the only other comparison of relationships across biomedical ontologies is a study between relationships in the Foundational Model of Anatomy and GALEN [6]. In this work, the authors identified patterns of relationships between equivalent concepts across ontologies and derived correspondences between relationships from their frequency of association. This associative technique is one of several approaches we are using in this paper.

Some progress has been made recently in defining the relationships used in biomedical ontologies. Smith and colleagues [7] provide formal definitions for ten such relationships (*Is_a*, *Part_of*, *Located_in*, *Contained_in*, *Adjacent_to*, *Transformation_of*, *Derives_from, Preceded_by*, *Has_participant*, and *Has_agent*), which as a whole represents a small ontology of biomedical relationships that is currently part of the OBO ontology library [8]. There exist other ontologies of relationships as part of ontologies such as GALEN [6] and the UMLS Semantic Network [9].

MATERIALS

Most vocabularies integrated into the UMLS contribute thesaural relationships (e.g., parent/child, broader/narrower than) to the Metathesaurus. In addition, some vocabularies contribute specified relationships such as *isa* and *location_of*. Relations in the Metathesaurus are represented bidirectionally. In practice, each relation (C_1, *rel*, C_2) is mirrored by a relation (C_2, *rel'*, C_1), where *rel'* is the inverse of *rel*. Relationships present in several vocabularies include *isa*, *location_of*, *ingredient_of*, *manifestation_of* and *mapped_to*). The semantics of the Metathesaurus relationships are implicit; that is, no definitions are given for the relationships used. A total of 4,328,245 direct relations involving specified relationships are represented in the 2005AC version of the UMLS. Our first step was to establish the list of all relationships used in the Metathesaurus paired with their inverses, because no such list is provided as part of the UMLS distribution.

The semantics of the relationships in the UMLS Semantic Network are explicit. Each of the Semantic Network relationships has an inverse, a textual definition, and a list of semantic types that are linked by the relationship. The relationships are organized in a hierarchy which comprises five high-level categories: functionally related to (e.g., *treats*), physically related to (e.g., *contains*), spatially related to (e.g., *adjacent_to*), temporally related to (*precedes*) and conceptually related to (e.g., *analyzes*). There are about 7,000 relations defined in the Semantic Network.

METHODS

Our methods for aligning relationships can be summarized as follows. The first, Metathesaurus-centric approach consists of eliciting the semantics of Metathesaurus relationships by examining their relata at different levels: concept, high-level ancestors and semantic types. The second, Semantic Network-centric approach examines the frequency of association between a given Semantic Network relationship and the actual relationships observed in the Metathesaurus between the concepts categorized by these semantic types.

Metathesaurus-Centric Approach

1) Manual elicitation. We created two random samples of a maximum of 50 relations per Metathesaurus relationship to be evaluated by the authors. (50 represents a manageable number of relations to be examined manually and random selection ensures the representativity of the sample). The authors carefully studied the way these relationships are used within a given terminology to determine the meaning of the relationship within that terminology. In those cases where the usage was clear, it was possible to link the Metathesaurus relationship to a Semantic Network relationship and to identify the type of relation between them (e.g. semantically equivalent, more specific than or more general than). For example, the relata of the Metathesaurus relationship *causative_agent_of* appeared to be infectious agents and pharmacologic substances on the one hand and disorders on the other. This relationship so understood corresponds to the Semantic Network relationship *causes*, which is defined as follows: "Brings about a condition or an effect. Implied here is that an agent, such as for example, a pharmacologic substance or an organism, has brought about the effect" [10].

2) Abstraction at the level of high-level concepts. In order to characterize the relata, we compute the lowest common ancestor for the domain and range of each Metathesaurus relationship in a given source. To this end, we construct a graph of all the ancestors of the relata on the side the domain and on the side of the range, respectively. Second, for each ancestor, we record the frequency (number of concepts from the domain/range having this concept as their ancestor) and distance between this concept and the original domain/range concept. Finally, we identify the ancestor in the graph for which the frequency is maximal and the depth minimal.

Consider the relationship *access_instrument_of* in SNOMED CT. The lowest common ancestors are *Endoscope* (domain) and *Procedure by method* (range). In practice, this automatic approach allows us to understand the prototypical relationship *access_instrument_of* as being a relationship between endoscopes and medical procedures. In some cases, the prototypical relationship is uninformative, because the lowest common ancestor is the root of the terminology. In other cases, there exists so much dispersion that the semantics of the relationship cannot be elicited by this method.

3) Abstraction at the level of semantic types. The relata can also be characterized at a higher level by their semantic types. For each Metathesaurus relationship, we compute the distribution of the semantic types of the concepts on the side of the domain and on the side of the range, respectively. For example, all 1,600 Metathesaurus relations involving the SNOMED CT relationship *access_instrument_of* have *Medical Device* as the semantic type of the range concept. Domain concepts, in contrast, are essentially divided between *Therapeutic or Preventive Procedure* (898 cases) and *Diagnostic Procedure* (699 cases). In the remaining 3 cases, the semantic type of the domain concept is *Health Care Activity*. In order to find the equivalent of *access_instrument_of* in the Semantic Network, we need to examine the possible Semantic Network relationships defined between *Therapeutic or Preventive Procedure* and *Diagnostic Procedure* on the one hand and *Medical Device* on the other. In this case, we find *uses*, defined as "Employs in the carrying out of some activity. This includes applies, utilizes, employs, and avails." [10]

Semantic Network-Centric Approach

Semantic types are high-level categories to which all Metathesaurus concepts are linked. It is therefore possible, starting from a given semantic type, to obtain its extension, i.e., the set of concepts that have been assigned this semantic type. From a given Semantic Network relation (T_1, rel_{SN}, T_2), we extract E_1 and E_2, the extensions of T_1 and T_2, respectively. We then examine what relations (C_1, rel_M, C_2) are represented in the Metathesaurus, pairwise, between one concept C_1 from E_1 and a concept C_2 from E_2. We obtain a set of Metathesaurus relationships {rel_{M1}, rel_{M2}, …} along with frequency information for each relationship in this set. Of these, the Metathesaurus relationships rel_M associated with the original Semantic Network relationship rel_{SN} with a high frequency constitute the candidates for identifying counterparts of rel_{SN} in the Metathesaurus.

For example, the Semantic Network relationship *ingredient_of* is defined between the domain semantic type *Substance* and the range semantic type *Clinical Drug*. Additionally, *ingredient_of* is inherited downwards along the *isa* hierarchy of semantic types. The union of the extensions of the semantic types on the side of the domain and range of *ingredient_of* contains 260,505 and 160,994 Metathesaurus concepts, respectively. The Semantic Network relationship *ingredient_of* is frequently associated with several Metathesaurus relationships, including *active_ingredient_of*, *dose_form_of* and *ingredient_of*. Of note, other Metathesaurus relationships linking substances to drugs (e.g., *metabolizes* and *has_contraindication*) are also associated with the Semantic Network relationship *ingredient_of*, although with a lesser frequency.

EXTENDED EXAMPLE

In order to illustrate this general methodology, we take an extended example of a single Metathesaurus relationship and show how each method is used to help determine its semantics and identify a correspondence with Semantic Network relationships. The SNOMED CT relationship *finding_site_of* is used throughout this example.

The first step in determining the semantics of *finding_site_of* is to look at the way this relationship is used in SNOMED CT. In this case, we study the list of fifty randomly selected relations. Here is a sample of the domain and range pairs for *finding _site_of*:

- *Endocrine structure* / *External endometriosis*
- *Gallbladder structure* / *Malignant tumor of gallbladder*
- *Skin structure* / *Epithelioma basal cell*
- *Stomach wall structure* / *Gastromalacia*

In this case, we determined that *finding _site_of* is a specification of the Semantic Network relationship *location_of*, defined between anatomical structures and disorders. In other words, *finding _site_of* is more specific than *location_of*.

Ideally, we would like to show that our conclusion is consistent with the source terminology and the UMLS as a whole. To this end, we abstract the relationship to the level *of high-level concepts* by computing the lowest common ancestor for the domain and range of each Metathesaurus relationship in the source under investigation. As expected, all fifty domain concepts of *finding_site_of* in SNOMED CT are descendants of *Anatomical structure*. The lowest common ancestor to the fifty range concepts is the root of the SNOMED CT vocabulary, indicating that the range concepts belong to several distinct hierarchies in SNOMED CT. In fact, while *Gastromalacia* and *Malignant tumor of gallbladder* are under the top-level concept *Clinical finding*, *Epithelioma basal cell* is subsumed by *Morphological abnormality*, itself under the top-level concept *Body structure*. This finding is not inconsistent with our previous claim that *finding_site_of* is a specification of *location_of*, but it also provides no positive evidence to support it directly.

We now abstract to the level of Semantic Types. On the side of the domain of *finding_site_of*, the most frequent semantic types are *Body Part, Organ, or Organ Component*, *Body System*, *Body Location or Region*, *Body Space or Junction* and *Tissue*. Analogously, *Injury or Poisoning*, *Disease or Syndrome*, *Finding*, *Congenital Abnormality* and *Neoplastic Process* are the most frequent semantic types on the range side. The only Semantic Network relationship defined between the domain and range semantic types above is *location_of*. Moreover, out of the 63,655 pairs of Metathesaurus concepts related by *finding_site_of*, 99.5% have their semantic types related by *location_of* (when any relationship is defined between their semantic types at all). At the level of abstraction of the semantic types, we find strong evidence to support the correspondence between the Metathesaurus relationship *finding_site_of* and the Semantic Network relationship *location_of*.

Additional evidence can be found when examining this correspondence using the Semantic Network-centric approach. We create the set of Metathesaurus concepts categorized by the semantic types in the domain and range of the Semantic Network relationship *location_of*, respectively. The relationships existing in the Metathesaurus between these sets of concepts include *finding_site_of*, predominantly, but also *procedure_site_of*, and the Metathesaurus relationships *location_of* and *isa*. This relative dispersion does not contradict our prior finding, but rather indicates that other Metathesaurus relationships (e.g., *procedure_site_of*) represent a specialization of the Semantic Network relationship *location_of*. The presence of *isa* in association with *location_of* is somewhat unexpected as there is an important ontological distinction between diseases (processes) and anatomical structures (entities). However, this is explained when we note that abnormal anatomical structures are often also considered diseases (e.g., *Bladder fistula isa Bladder disease*).

RESULTS

A total of 139 relationships are present in the Metathesaurus. Most of the 45 English vocabularies that include relationships have just a small number. The largest number of relationships are contributed by SNOMED CT (62), LOINC (15) the National Drug File – Reference Terminology (15), the University of Washington Digital Anatomist (8), and RxNorm (7). 116 are unique to a specific vocabulary and 23 are found in at least two vocabularies, e.g., *component_of* is found in SNOMED CT, PDQ, and LOINC.

Using the methods described above, we were able to align 80 (58%) of the Metathesaurus relationships with Semantic Network relationships. In some cases this alignment is at a coarse level of granularity, e.g. *metabolic_site_of* is *more specific than functionally_related_to*, in other cases, the relationships are *roughly equivalent to* each other, e.g., *focus_of* and *issue_in*, and in 27 cases there exists an *identical* relationship in the Semantic Network, e.g., *affects*, *process_of*, *ingredient_of*.

The remaining 59 Metathesaurus relationships fall into a number of additional categories. Some are lexical relationships (e.g., *british_form_of*, *permutated_term_of*, *xml_form_of* and *suffix_of*), some are mapping relationships (e.g., *see_from* and uniquely_mapped_from), and others are used strictly for vocabulary management purposes (e.g., classifies, *moved_from, replaces*). These relationships are not associated with any particular pair of semantic types and do not converge towards any particular Semantic Network relationships. They are also not useful candidates for inclusion in an ontology of biomedical relationships.

DISCUSSION

The methods employed in this study combine manual and automated techniques. In some cases, even using multiple methods, it is difficult to discern the underlying semantics of the Metathesaurus relationships. The methods described here provide a good characterization of the meaning of a relationship, but they are not a substitute for an explicit definition. The developers of individual vocabularies have invested a good deal of effort in linking pairs of terms with these specific relationships. Their usefulness for vo-

cabulary-specific applications would be greatly enhanced if their semantics were made explicit. Clearly defined relationships would also make it possible for these originally vocabulary-specific relationships to be fully and accurately integrated with other vocabularies, thereby further broadening their usefulness.

We have explored a number of methods for aligning two UMLS knowledge sources. We have conducted this work, first, to improve the usefulness of the vocabulary-specific relationships in the context of the UMLS. Second, we intend for the methods described here to be a first step in identifying and classifying biomedical relationships beyond the existing 54 Semantic Network relationships. The goal is to use the augmented set of relationships as the basic building blocks for a broader and more comprehensive ontology of biomedical relationships. This work is important since it will go some way toward ensuring that the ontology of biomedical relationships will remain relatively stable as new terminologies are added to the UMLS and as changes are made to existing terminologies. There are already some efforts toward developing an ontology of relationships that are specific to biomedicine [6, 11, 12] and there is ongoing research on the underlying ontological principles [7, 13-15]. The work reported here is an effort to contribute to these investigations. An ontology of biomedical relationships promises to add logical and ontological rigor to biomedical ontologies, to bring existing terminologies and ontologies more closely in-line with one another, and to serve as a resource for the construction of future vocabularies.

CONCLUSIONS

In this study we have employed a number of methods to determine the semantics of Metathesaurus relationships and to align relationships between the Semantic Network and the Metathesaurus. These methods should be seen as complementary. In the ideal case, all the methods would point to a single semantic interpretation of a given relationship. In practice, some methods work better than others for some cases. In some cases, the various methods might even produce inconsistent results, due to inconsistencies in the vocabularies under investigation. The hope is that, in combination, these methods will provide, first, a comprehensive strategy for guiding the alignment of Metathesaurus relationships and Semantic Network relationships, and, second, will serve as a good starting point for the development of a comprehensive ontology of biomedical relationships.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM). This work was done while Lowell Vizenor was a visiting fellow at the Lister Hill National Center for Biomedical Communications, NLM, NIH.

References

1. McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med 1995;34(1-2):193-201
2. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41-51
3. McCray AT, Bodenreider O. A conceptual framework for the biomedical domain. In: Green R, Bean CA, Myaeng SH, editors. The semantics of relationships: an interdisciplinary perspective. Boston: Kluwer Academic Publishers; 2002. p. 181-198
4. Noy NF. Tools for mapping and merging ontologies. In: Staab S, Studer R, editors. Handbook on Ontologies: Springer-Verlag; 2004. p. 365-384
5. Bodenreider O, Burgun A. Aligning knowledge sources in the UMLS: methods, quantitative results, and applications. Medinfo 2004;11(Pt 1):327-31
6. OpenGalen: http://www.opengalen.org/
7. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biol 2005;6(5):R46
8. Open Biomedical Ontologies: http://obo.sourceforge.net/
9. UMLS Semantic Network: http://semanticnetwork.nlm.nih.gov/
10. UMLS: http://umlsks.nlm.nih.gov/
11. Mungall CJ. Obol: integrating language and meaning in bio-ontologies. Comparative and Functional Genomics 2004;5(6-7):509-520
12. Rosse C, Mejino JL, Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform 2003;36(6):478-500
13. DOLCE: http://www.loa-cnr.it/DOLCE.html
14. Schulz S, Hahn U. Part-whole representation and reasoning in formal biomedical ontologies. Artif Intell Med 2005;34(3):179-200
15. Smith B, Grenon P. The cornucopia of formal-ontological relations. Dialectica 2004;58(3):279-296