Medinfo 2004:327-331
© 2004, IMIA

Aligning Knowledge Sources in the UMLS: Methods, Quantitative Results, and Applications

Olivier Bodenreider[a] and Anita Burgun[b]

[a] *U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA*
[b] *Laboratoire d'Informatique Médicale, Université de Rennes I, Rennes, France*

Abstract

The UMLS Semantic Network and Metathesaurus are two complementary knowledge sources. While many studies compare relationships across the two structures, their alignment has never been attempted. We applied two methods based on lexical and conceptual similarity to aligning the Semantic Network with the UMLS Metathesaurus. Approximately two thirds of the semantic types could be aligned by lexical similarity. Conceptual similarity suggested mappings in all but ten cases. Potential applications enabled by the alignment are discussed, namely auditing the consistency between the Semantic Network and the Metathesaurus and extending the Semantic Network downwards. The relative contribution and limitations of the two methods used for the alignment are also discussed.

Keywords

UMLS, Alignment, Hierarchical relationships.

Introduction

One of the distinctive features of the Unified Medical Language System® (UMLS®) compared to other terminology systems is the existence of two complementary yet independent knowledge sources: the Metathesaurus®, a large repository of inter-related concepts coming from some 60 families of biomedical vocabularies, and the Semantic Network, a small, manually curated set of high-level categories – called semantic types – and relations expressing definitional knowledge. Most noticeably, the two structures are related together by categorization links assigned by the UMLS editors. Each concept integrated in the Metathesaurus is assigned (at least) one semantic type from the semantic network, 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].

Because it consists of broad categories, the Semantic Network is often presented as the overarching knowledge structure, while the Metathesaurus, containing essentially finer-grained concepts, is represented underneath it. Accordingly, the categorization relation of a concept to a semantic type is generally interpreted as *is a kind of*, semantic types subsuming concepts [2]. Not all Metathesaurus concepts, however, are fine-grained. Concepts close to the root in vocabulary hierarchies are indeed at the same level as most semantic types. Thus, in addition to the categorization links between the Semantic Network and the Metathesaurus, one could imagine a more direct junction, realized by those high-level concepts which have an equivalent in the Semantic Network. For example, the concept *Vitamins* (C0042890) is equivalent to the semantic type *Vitamin* (T127).

Our objective is to investigate the equivalence between semantic types and concepts. In other words, we want to align two UMLS knowledge sources: the Semantic Network and the Metathesaurus. The underlying hypothesis is that there exists some resemblance, both lexical and conceptual, between the two structures. Lexical similarity exists when a concept and a semantic type have similar names (e.g., *Vitamins* and *Vitamin*). Conceptual similarity comes from the equivalence between the categorization relationship (between a semantic type and a concept) and hierarchical relationships in the Metathesaurus (between a high-level concept and finer-grained concepts). In practice, conceptual similarity is measured by the overlap between the set of concepts having a given semantic type on one hand, and the set of descendants of a given concept in the Metathesaurus on the other.

In this paper, we applied two methods based on lexical and conceptual similarity to aligning the Semantic Network with the UMLS Metathesaurus. After a brief review of related work, we present these two methods and some qualitative results. We then give an extended example before discussing the applications of the alignment of these two structures.

Background

The general framework of this study is that of the alignment of knowledge structures. The following clues, mentioned by several authors (e.g., [3, 4]) can be used to assess the equivalence between concepts across knowledge structures:

- their names are similar;
- their definitions are similar;
- their relations (both hierarchical and associative) to other concepts are similar.

Similarity based on concept names can be assessed by requiring an exact match or using edit distance. Sophisticated techniques such as normalization, based on knowledge about term variation, can be used. Word-sense ambiguity is a limitation of this method and may result in false positives. Similarity based on definitions is difficult to assess automatically in the absence of a formal representation (e.g., in description logics [5]). Conceptual similarity is based on the relations of a concept to other concepts. For hierarchical relations, methods have been developed for comparing taxonomies. For associative relations, similarity is based on shared slots, i.e., similar associative relations to other concepts.

In aligning the Semantic Network with the Metathesaurus, our task differs from the general case. Although semantic types all have a textual definition, this is not the case of the Metathesaurus concepts, ruling out the use of definition-based similarity. In many cases, the associative relationships among concepts in the Metathesaurus are not precisely labeled, making it difficult to compare them reliably to associative relationships among semantic types. Therefore, the two methods available for aligning the Semantic Network with the Metathesaurus are lexical similarity and conceptual similarity based on hierarchical relations.

Our contribution is not to develop a novel technique for aligning knowledge structure, but rather to adapt existing techniques to the specificity of two biomedical knowledge structures, the Semantic Network and the Metathesaurus.

Materials and Methods

Preparing the UMLS knowledge sources

The UMLS[1] has been developed and maintained by the U.S. National Library of Medicine since 1991. The version of the UMLS used in this study is 2003AA, released in January 2003. In this version, the Semantic Network comprises 135 semantic types and 558 relations among them. In contrast, the Metathesaurus comprises 875,255 concepts and more than 12 million relations among concepts.

As noted in other studies [e.g., 6], cycles can be found among the hierarchical relations in the Metathesaurus. Since a directed acyclic graph is required for computing reliable lists of descendants for Metathesaurus concepts, we use a slightly modified version of the Metathesaurus from which the links responsible for the cycles have been removed.

Mapping Semantic Type names to the Metathesaurus

Each semantic type name was mapped to the Metathesaurus by first attempting an exact match and, if necessary, a normalized match. Normalization makes semantic type names compatible with concept names by abstracting away from such inessential differences as punctuation, word order, and case and hyphen variation.

However, some semantic type names exhibiting coordination (e.g., *Disease or Syndrome*) are not expected to map to a single Metathesaurus concept. To address this issue, we decomposed the semantic type names. For example, the names *Disease* and *Syndrome* were extracted from the original semantic type *Disease or Syndrome*. Additionally, when present in a semantic type name to be decomposed, modifiers were distributed as required. For example, *Body Space or Junction* was transformed into *Body Space* and *Body Junction*. All mappings were reviewed manually by the authors for accuracy and disambiguation in case of mapping to multiple concepts.

Establishing sets of concepts

The extension of a semantic type is the set of concepts that have been assigned this semantic type. It is easily obtained by a simple query on the UMLS table MRSTY. In contrast, establishing the list of all descendants for a given concept requires computing the transitive closure on hierarchical relationships. Parent-child and broader-narrower relationships are used interchangeably in this process.

Comparing sets of concepts

In order to compare the extension of a semantic type (E_{st}) with the set of all descendants of a Metathesaurus concept (D_c), we computed similarity coefficients measuring the degree of overlap between the two sets. Various coefficients have been developed, each having slightly different mathematical properties. All coefficients, however, compare the cardinality of the intersection to that of each set. In this study, we use three similarity coefficients: cosine, Jaccard, and Dice [7]. In all three cases, their value varies from 0 (indicating disjoint sets) to 1 (indicating total overlap). The three similarity coefficients are defined as follows:

$$\mathrm{Sim}_{\mathrm{cosine}}=\frac{\mathrm{AB}}{\sqrt{\mathrm{A.B}}} \qquad \mathrm{Sim}_{\mathrm{Jaccard}}=\frac{\mathrm{AB}}{\mathrm{A+B-AB}} \qquad \mathrm{Sim}_{\mathrm{Dice}}=\frac{2*\mathrm{AB}}{\mathrm{A+B}}$$

where A and B represent the cardinality of the two sets and AB that of their intersection.

It is beyond the scope of this paper to present a detailed, qualitative analysis of the comparison between E_{st} and D_c. Only quantitative results are reported.

Results

Lexical similarity

Out of the 135 semantic type names, 32 contain the conjunction *or*. None of these complex names mapped to a Metathesaurus concept. Transforming these complex names as described earlier yielded 69 simple names. A total of 172 names was mapped to the Metathesaurus. After manual review, 106 mappings were deemed relevant (e.g., semantic type *Organism*

[1] http://umlsinfo.nlm.nih.gov

to concept *Living Organisms*) while no valid mapping could be found for 66 names (e.g., *Biologic Function*, *Temporal Concept*). Ten ambiguous mappings were disambiguated manually.

Conceptual similarity

For the similarity between the extension of a semantic type (E_{st}) and the set of all descendants of a Metathesaurus concept (D_c), the three coefficients were systematically computed for each (E_{st}, D_c) pair and gave generally consistent results. We only report cosine values for brevity. The top cosine values observed for each semantic type ranged from .0094 to .9943. The frequency distribution of the top cosine values for the 135 semantic types is shown in Figure 1. Example of similarity values for (E_{st}, D_c) pairs are given in Table 1.

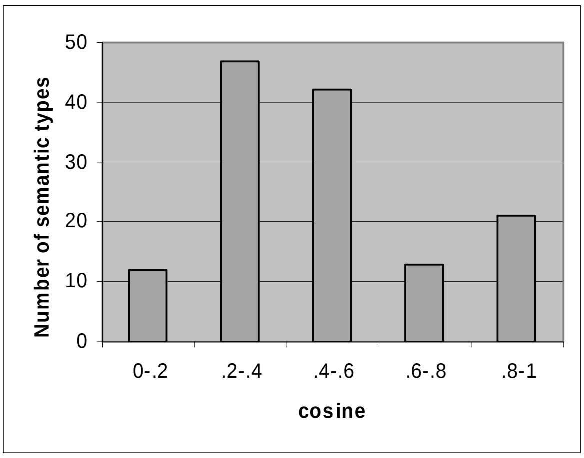

Figure 1 – Frequency distribution of the top cosine values for the 135 semantic types

Table 1 – Similarity between semantic types and concepts

Sem. type	Concept		Cosine
Amphibian	*Amphibia*	C0002668	.9943
Reptile	*Leptosauria*	C0999079	.9729
Gene or Genome	*Cancer Genes*	C0919431	.6781
Anatomic Structure	*Organ*	C0178784	.5447
Immunologic Factor	*Immunology*	C0152036	.3242

Although reflected in the value of the similarity coefficients, the number of concepts specific to each set (i.e., not in their intersection) must be considered for selecting (E_{st}, D_c) pairs. For example, the highest cosine value for *Gene or Genome* is with *Cancer Genes* (.6781) and the intersection of the sets contains 358 concepts. The cosine value of the same semantic type with the concept *Genes* is indeed slightly lower (.6466), but the intersection of their two sets contains more concepts (472). The difference in the cosine values comes form a larger number of concepts specific to each set in the latter case.

Lexical vs. conceptual similarity

In 106 cases, a semantic type name (or part thereof) was successfully mapped to a Metathesaurus concept through lexical similarity. In 60 of these cases, the concept mapped to was present in the top 25 concepts identified as candidates for this semantic type through conceptual similarity. In ten cases, the concept mapped to had no descendants and therefore could not participate in conceptual similarity. In the remaining 36 cases, semantic type name and concept names exhibited lexical similarity and limited conceptual similarity.

Extended example

The highest cosine value (.9943) was observed between the semantic type *Amphibian* and the concept *Amphibia*. More precisely, out of the 1135 concepts assigned the semantic type *Amphibian*, 1124 were common to the 1126 descendants of the concept *Amphibia*. This almost perfect case of overlap between the extension of a semantic type and the descendants of a concept is ideal for examining in detail the few outliers.

The two descendants of *Amphibia* not assigned the semantic type (ST) *Amphibian* are *Tadpoles* (ST: *Invertebrate*) and *Toad licking* (ST: *Pharmacologic Substance*). *Tadpoles* was assigned the parent semantic type of *Amphibian* instead of the most precise semantic type, but it should be part of the extension of *Amphibian*. *Toad licking*[2] is correctly not categorized as *Amphibian*. Its presence among the descendants of *Amphibia*, though surprising, can be explained by an inaccurate child relationship to *Bufo* (a toad). This relationship, although used in a vocabulary to create a hierarchy, is neither taxonomic not partonomic. It simply is an associative relationship, possibly useful for information retrieval tasks, but detrimental in this context.

In the extension of the semantic type *Amphibian*, ten concepts are not descendants of the concept *Amphibia*. Of these, seven are descendants of the concept *unclassified rana*. Although categorized as *Amphibian*, this concept has no parents, thus no hierarchical relations to the concept *Amphibia*. This is a case of missing hierarchical relation in the Metathesaurus. Two parents of the concept *Amphibia*, *Amphibians and Reptiles* and *Tetrapoda* are incorrectly categorized as *Amphibian* (in both cases, their semantics is broader). Even more surprisingly, the concept *Class reptilia* is incorrectly categorized as *Amphibian*, not *Reptile*.

Discussion

Although aligning knowledge structures may constitute an interesting endeavor in and of itself, even more interesting are the applications enabled once the structures are aligned,

[2] The skin of Bufo alvarius allegedly contains a hallucinogenic tryptamine

namely auditing the consistency between the Semantic Network and the Metathesaurus and extending the Semantic Network downwards. The relative contribution and limitations of the two methods used for the alignment will also be briefly discussed.

Auditing consistency

Several authors have reported semantic inconsistencies between Semantic Network and Metathesaurus relations. Cimino, for example, detected inconsistent hierarchical relations in the Metathesaurus based hierarchical relations between corresponding semantic types and suggested additional Semantic Network relations based on the existence of corresponding relations in the Metathesaurus [6]. In a previous study, we found that 13% of the Metathesaurus relations (both hierarchical and associative) were in violation of Semantic Network relations [8].

However, these studies analyzed Metathesaurus relations one at the time rather than by sets. Comparing sets may provide a more global auditing method for the consistency between the two structures. Ideally, there would be a large overlap between the extension of a semantic type (E_{st}) with the set of all descendants of a Metathesaurus concept (D_c). The pair (*Amphibian*, *Amphibia*) is an illustration of this situation. Most often, however, D_c fails to completely cover E_{st} and contains specific concepts. In this case, a more detailed analysis is needed to determine the exact causes. A limited review led us to identifying four major causes:

- A wrong semantic type assignment results in concepts present in E_{st} and not in D_c.
- A missing semantic type assignment results in concepts present in D_c and not in E_{st}.
- A wrong hierarchical relation in the Metathesaurus results in concepts present in D_c and not in E_{st}.
- A missing hierarchical relation in the Metathesaurus results in concepts present in E_{st} and not in D_c.

Most of these situations were encountered in the extended example presented earlier.

Extending the Semantic Network downwards

Many studies conducted recently have proposed various enhancements of the Semantic Network. While most studies suggest adding relations [e.g., 9], some also suggested adding new semantic types, often for improving the coverage of a given subdomain [e.g., 10]. The alignment with the Metathesaurus provides a more general, domain-independent method for extending the Semantic Networks downwards. After selecting the concept corresponding to a given semantic type through lexical and conceptual similarity, it is likely that the first-generation descendants of this concept be reasonable candidates for becoming subtypes of the semantic type. For example, from the concept *Chromosomal and cytologic alterations* (identified as a possible mapping for the semantic type *Cell or Molecular Dysfunction*), the following first-generation descendants may arguably provide subtypes for *Cell or Molecular Dysfunction*: *Extracellular alteration*, *Membrane alteration*, *Cytoplasmic alteration*, and *Genetic alteration*. The remaining descendant is of lesser interest here (*Abnormal cell*).

In practice, however, the process of extending the Semantic Network downwards may be difficult to automate for several reasons. There will often be a large number (several dozen) of first-generation descendants. In a well-formed hierarchy, only a fraction of them should become subtypes. While selecting these concepts requires expertise of the domain, such an approach may facilitate the work of experts, compared to creating a hierarchy from the top down. Additionally, because hierarchical relations in the Metathesaurus sometimes reflect the particular view of a given source vocabulary rather than taxonomic or meronomic relations, subtype candidates selected by this method must be reviewed manually for accuracy and completeness.

Relative contribution and limitations of each method

The alignment based solely on lexical similarity is subject to two major types of errors. False Positives come from polysemy (e.g., the semantic type *Idea* [thought] and the concept *Idea* [organism]). False negatives come from missing synonymous terms in the Metathesaurus (e.g., the semantic type name *Anatomical Junction* is not a name for the concept *Body Junction*), missing Metathesaurus concepts (e.g., for metaclasses such as *Organism Attribute* in the Semantic Network), or the inability of matching algorithms to handle differences in concept names.

To address this issue, it is possible to complement the lexical matching with other techniques. For example, traversing a well-organized terminology such as the Medical Subject Headings (MeSH), starting from a close subsumer of the semantic type of interest and navigating downwards may constitute a useful strategy, albeit requiring manual selection by a domain expert. For example, the concept *Pharmaceutical Preparations* whose MeSH definition is ‘drug intended for human or veterinary use, presented in their finished dosage form’ corresponds to the semantic type for *Clinical Drug*. It can be reached from the top-level concept *Chemicals & Drugs* in MeSH.

In this study, we proposed a more automatic solution based on conceptual similarity. The limitations of this method lie in the difficulty of determining a threshold on similarity coefficients for selecting the mappings automatically. However, we showed that, in presence of lexical similarity, the conceptual similarity method was able to identify the lexical mapping as part of the top 25 concepts in 57% of the cases. Again, this method may require some degree of manual intervention, but considerably less than for an entirely manual mapping. Moreover, conceptual similarity was able to detect mappings in the absence of lexical similarity. For example, although no lexical mapping was found for the semantic type *Injury or Poisoning*, the concept *Injury, poisoning, and procedural complications* was identified as a possible match by conceptual similarity with a cosine value of .6866.

Future work

This preliminary, quantitative analysis of the alignment of the Semantic Network and the Metathesaurus must be followed by an in-depth, qualitative analysis. We also plan to pursue the potential applications presented above: auditing semantic type assignment and hierarchical relations in the Metathesaurus and extending the Semantic Network downwards.

References

[1] McCray AT, Nelson SJ. The representation of meaning in the UMLS. *Methods Inf Med* 1995;34(1-2):193-201.

[2] Burgun A, Bodenreider O. Aspects of the taxonomic relation in the biomedical domain. In: Welty C, Smith B, editors. *Collected papers from the Second International Conference "Formal Ontology in Information Systems"*: ACM Press; 2001. p. 222-233.

[3] Maedche A, Staab S. Measuring Similarity between Ontologies. In: *Proc. of the European Conference on Knowledge Acquisition and Management - EKAW-2002. Madrid, Spain, October 1-4, 2002*: Springer; 2002. p. 251-263.

[4] Noy NF, Musen MA. PROMPT: algorithm and tool for automated ontology merging and alignment. *Proc of AAAI* 2000:450-455.

[5] Cornet R, Abu-Hanna A. Usability of expressive description logics--a case study in UMLS. *Proc AMIA Symp* 2002:180-4.

[6] Cimino JJ. Auditing the Unified Medical Language System with semantic methods. *J Am Med Inform Assoc* 1998;5(1):41-51.

[7] Rasmussen E. Clustering algorithms. In: Frakes WB, Baeza-Yates R, editors. *Information retrieval : data structures & algorithms*. Englewood Cliffs, N.J.: Prentice Hall; 1992. p. 419-442.

[8] 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.

[9] Zhang L, Perl Y, Halper MH, Geller J, Cimino JJ. Enriching the structure of the UMLS semantic network. *Proc AMIA Symp* 2002:939-43.

[10] Yu H, Friedman C, Rhzetsky A, Kra P. Representing genomic knowledge in the UMLS semantic network. *Proc AMIA Symp* 1999:181-5.

Address for correspondence

Olivier Bodenreider, National Library of Medicine,
8600 Rockville Pike, MS 43, Bethesda, MD 20894, USA.
Email: olivier@nlm.nih.gov. Phone: (301) 435-3246.