Finding Semantic Inconsistencies in UMLS using Answer Set Programming

Halit Erdoğan
Faculty of Engineering and
Natural Sciences, Sabanci University,
Istanbul, TURKEY

Olivier Bodenreider
US National Library of Medicine,
NIH, Bethesda, USA

Esra Erdem
Faculty of Engineering and
Natural Sciences, Sabanci University,
Istanbul, TURKEY

Abstract

We introduce a new method to find semantic inconsistencies (i.e., concepts with erroneous synonymity) in the Unified Medical Language System (UMLS). The idea is to identify the inconsistencies by comparing the semantic groups of hierarchically-related concepts using Answer Set Programming. With this method, we identified several inconsistent concepts in UMLS and discovered an interesting semantic pattern along hierarchies, which seems associated with wrong synonymy.

Introduction

The Unified Medical Language System (UMLS) is a terminology integration system, which includes two sources of semantic information: the Metathesaurus and the Semantic Network. The UMLS Metathesaurus was assembled by integrating some 150 source vocabularies; it contains more than 2 million concepts (i.e., clusters of synonymous terms coming from multiple source vocabularies identified by a Concept Unique Identifier). The UMLS Metathesaurus contains also more than 36 million relations between these concepts, such as *parent*, *child_of*, *broader*, *narrower*. More than 7.5 million hierarchical relations are represented in the Metathesaurus. The Semantic Network is a network of 135 Semantic Types (STs) organized in a tree structure. Each Metathesaurus concept is assigned at least one ST. Groupings of STs, called semantic groups (SGs), represent subdomains of biomedicine such as *Anatomy*, *Chemicals and Drugs*. Each ST belongs to exactly one SG. For example, the concept *Eye* belongs to the ST *Body part, organ, or organ component* which belongs to the semantic group *Anatomy*. The concept *Retina*, a child of *Eye*, belongs to the same ST with *Eye*.

Since different sources of information contribute to the UMLS, there exists inconsistencies, such as erroneous synonymy, wrong categorization, ambiguity and so forth. For example, *Capsule of adrenal gland* is an anatomical concept found in the Foundational Model of Anatomy (FMA) and the NCI Thesaurus. In the FMA, it is defined as a subclass of *Capsule*. Once integrated in the UMLS, *Capsule of adrenal gland* appears as a child of the concept *capsule (pharmacologic)*, for which "Capsule" is also a name. Of course, *Capsule* is an ambiguous name used by both anatomy and pharmacology specialists. In fact, a search for "capsule" in the UMLS yields 4 concepts. Surprisingly, none of these concepts pertains to macroscopic anatomical structures. The issue here is both the absence in the UMLS of a concept for the membranous layer surrounding an organ and the wrong association in the UMLS Metathesaurus of this meaning with the pharmacologic concept *capsule (pharmacologic)*.

Motivated by the *capsule* example, our objective is to find semantic inconsistencies in UMLS from the perspective of their hierarchical relations and to show through examples how semantically inconsistent concepts can help reveal erroneous synonymy relations. The specific contribution of this paper is to leverage the semantic groups for identifying inconsistencies and to consider not only the semantics directly ascribed to a concept, but also the semantics it inherits from its ancestors. We introduced an inconsistency definition for Metathesaurus concepts based on their hierarchical relations and compute all such inconsistent concepts. After that we manually review some of the inconsistent concepts to determine the ones that have erroneous synonymy relations such as wrong synonymy.

Various research groups have investigated quality in the UMLS, addressing issues including terminological cycles (Bodenreider 2001), ambiguity of concepts (Cimino 2001), concept categorization (Gu et al. 2004). Consistency across hierarchies has been addressed by (Cimino, Min, and Perl 2003), while (McCray and Bodenreider 2002) has studied the consistency of Metathesaurus relations against Semantic Network relations. More recently, the semantic groups have been used for analyzing the consistency of Metathesaurus relations (Vizenor, Bodenreider, and McCray 2009). Most closely related to our work is a study of the validity of concepts associated with multiple semantic groups (Mougin, Bodenreider, and Burgun 2009). Our work uses the same approach to assess the validity, not of single concepts, but of pairs of hierarchically-related concepts.

Computational Methods

We say that a UMLS concept is inconsistent if the following two conditions hold for the concept: (1) it belongs to two different semantic groups either directly or via its ancestors; (2) it does not have any inconsistent ancestor (i.e., the inconsistency of the concept is not due to inheritance, it

```
% define the concepts C with some inconsistent ancestor C1
descendantOfInconsistent(C) :- descendant(C,C1), inconsistent(C1).

% identify the groups G (except conc) that a concept C belongs to,
% such that C is not a descendant of an inconsistent ancestor
groupOfConcept(C,G) :- hasCategory(C,T), hasGroup(T,G), set(n,C), G!= conc.
groupOfConcept(C,G) :- not descendantOfInconsistent(C),
    descendant(C,C1), hasCategory(C1,T), hasGroup(T,G), G != conc.

% a concept is originally inconsistent if it belongs to two different groups G and G1
orgn_inconsistent(C) :- groupOfConcept(C,G), groupOfConcept(C,G1), G<G1.
```

Figure 1: Defining original inconsistencies in ASP

is original).

To compute all inconsistent concepts, we need some method to find all the ancestors of a concept and their semantic groups, and check the inconsistency of each ancestor. However, there are over 2 million concepts in the Metathesaurus and some of them have too many ancestors (over 800); and thus it may not be practical to generate all the ancestors of each concept.

We have introduced a new method for computing inconsistent concepts: first, it divides the set of all UMLS concepts into smaller sets (e.g., of size 20,000); then, for each set, it computes in parallel all ancestors of its elements in the whole UMLS graph that are not inconsistent. Therefore, with this method, we compute the inconsistent concepts as we compute their ancestors and check their inconsistency.

We have realized this method using Answer Set Programming (ASP) (Lifschitz 2008). The idea is to define the ancestors of concepts, and the inconsistent concepts by means of (possibly recursive) rules, and then call an existing ASP system (e.g., the ASP solver CLASP) to find inconsistencies based on these definitions. Figure 1 shows a sample ASP definition of original inconsistency for a set of concepts. Since recursion is allowed in ASP, we can define hierarchical relations (e.g., `descendant`). Furthermore, since ASP has default negation (`not`), we can define original inconsistencies without generating all paths to their ancestors.

Results

We have identified 334,396 inconsistent concepts. Out of these concepts, 81,512 concepts are inconsistent due to the following reason: the semantic group of the parent differs from that of the source concept, and no ancestor of the concept is inconsistent. For example, the concept *Anti-purkinje cell antibody* is one of these 81,512 concepts: its semantic group is *Chemicals and Drugs*, whereas its parent *Purkinje Cells* belongs to the semantic group *Anatomy*; furthermore, no ancestor of *Anti-purkinje cell antibody* is inconsistent.

To identify the erroneous synonymy relations, first we have obtained the distribution of inconsistencies by semantic group of the source concept. After that, we have refined this map of inconsistencies by looking at the semantic group of the parent of inconsistent concepts in reference to that of the source concept. By manual review of pairs of concepts with inconsistent *child_of* relations, we have identified four errors (called wrong synonymy) of the same type, including the "capsule" error mentioned in Introduction.

Further discussion on results can be found in (Erdogan, Erdem, and Bodenreider 2010).

Discussion

We have defined inconsistency of concepts in UMLS and developed an ASP program that identifies them. Interestingly, we have discovered that the four instances of wrong synonymy we have identified exhibit a pattern of "semantic rupture" along the hierarchical structure of the terminology. By semantic rupture, we mean that, along one hierarchy, the source concept belongs to a given semantic group, its parent concept does not, but one of the parents of the parent belongs to the same group as the source concept. We hypothesize that such pattern of semantic rupture might be a good marker for wrong synonymy, and plan to test it.

Acknowledgments: This research was supported by the Intramural Research Program of NIH/NLM, and TUBITAK Grant 108E229.

References

Bodenreider, O. 2001. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. In *Proc. of AMIA Symp.*, 57–61.

Cimino, J. J.; Min, H.; and Perl, Y. 2003. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. *Biomed. Inform.* 36:450–461.

Cimino, J. J. 2001. Battling Scylla and Charybdis: the search for redundancy and ambiguity in the 2001 UMLS metathesaurus. In *Proc. of AMIA Symp.*, 120–124.

Erdogan, H.; Erdem, E.; and Bodenreider, O. 2010. Exploiting UMLS semantics for checking semantic consistency among UMLS concepts. In *Proc. of MedInfo*.

Gu, H.; Perl, Y.; Elhanan, G.; Min, H.; Zhang, L.; and Peng, Y. 2004. Auditing concept categorizations in the UMLS. *Artif Intell Med* 31:29–44.

Lifschitz, V. 2008. What is Answer Set Programming? In *Proc. of AAAI*, 1594–1597.

McCray, A. T., and Bodenreider, O. 2002. *A conceptual framework for the biomedical domain*. Kluwer. 181–198.

Mougin, F.; Bodenreider, O.; and Burgun, A. 2009. Analyzing polysemous concepts from a clinical perspective: application to auditing concept categorization in the umls. *Biomed. Inform.* 42:440–451.

Vizenor, L. T.; Bodenreider, O.; and McCray, A. T. 2009. Auditing associative relations across two knowledge sources. *Biomed. Inform.* 42:426–439.