Comparing Inconsistent Relationship Configurations Indicating UMLS Errors

James Geller, PhD[1], C. Paul Morrey, MS[1], Junchuan Xu, MD[1], Michael Halper, PhD[2], Gai Elhanan, MD[1], Yehoshua Perl, PhD[1], George Hripcsak, MD[3]

[1]NJIT, Newark, NJ; [2]Kean University, Union, NJ; [3]Columbia University, New York, NY

Abstract

The goal of this paper is to audit null-annotated parent-child pairs in the UMLS Metathesaurus. We have developed techniques for identifying suspicious pairs with high likelihood of errors by using inconsistencies between the hierarchical relationships of the Metathesaurus and the Semantic Network. Two formal conditions, called semantic inversion and lack of ancestry are investigated. Analyzing two corresponding samples shows that semantic inversion is significantly more likely to indicate an error than lack of ancestry, which in turn is more likely to indicate errors than a consistent configuration. We also discuss cases of parent-child pairs with semantic inversion that may be corrected by disambiguating the child.

Introduction

The Unified Medical Language System® (UMLS) [1] is a large terminological database containing medical terms from many sources, e.g., SNOMED CT, LOINC, and NCI among many others. Currently, the UMLS Metathesaurus (META) contains about 143 source terminologies with more than 1.8 million concepts and over 7.5 million strings. The UMLS Semantic Network (SN) [2] consists of 135 semantic types, connected by IS-A relationships to form two trees. The two resources are related by the assignment of one or more semantic types from the SN to each concept in the META. Management of the UMLS content is of the utmost importance to its users, who depend on its quality for the correct performance of their systems. Because the UMLS consists of so many terminologies, inconsistencies are likely to occur. The UMLS editors may attempt to prevent them from being entered during the time of integration or try to find them after the fact by auditing.

The interplay between the META and the SN can be leveraged to support auditing of the META through automated verification of internal consistency. In some cases, the auditing can pinpoint inconsistencies that can be readily addressed. In other cases, the methods used can merely suggest potential problems. Automated methods can help to focus the limited resources of human review on the cases most likely to need attention [3].

In [3] an approach was described that compares the parent-child relationships, annotated explicitly as "isa" relationships, between concepts in the META with the IS-A relationships between their assigned semantic types in the SN. (We will use "**isa**" when we are referring to the database annotation of META. When we are referring to the idea of the relationship or its use in the SN we will use "**IS-A**.") In this paper, we are focusing on parent-child relationships which are annotated with "null," meaning that no relationship attribute (RELA) is available.

We distinguish between two structural inconsistency conditions, which we call *semantic type inversion (short: semantic inversion)* and *lack of ancestry*. In the former case, the semantic type of the parent in META is less general than the semantic type of the child in META, as expressed by the IS-A relationship between the two semantic types in the SN. In the latter case there is no hierarchical relationship between any of the semantic types of the parent and any of the semantic types of the child.

The question addressed here is whether semantic inversion and lack of ancestry can be used as predictors of inconsistencies of parent-child relationships. Secondly, we are raising the question whether semantic inversion is a better predictor of inconsistency than lack of ancestry. In this paper, one sample of each kind of inconsistency is audited, as well as a control sample.

Background

Authoring medical terminologies, ontologies and meta-terminologies such as the UMLS Metathesaurus is a difficult, error-prone, human resource-intensive task. Most such *useful* repositories are by far too large and financial resources too limited to allow a team of auditors to work through a complete terminology and verify the conceptual and relational correctness of all its elements. Nevertheless, auditing is an important task to ensure the quality of a terminological resource. Thus, methods are needed to determine "areas" of a terminology which are most likely to contain errors. Concentrating scarce auditor resources on those areas will result in the most cost-effective auditing process. A general approach to recognizing such "areas" is to discover structural anomalies in a terminology by using computer algorithms. If it can be verified that concepts with structural anomalies are indeed more likely to exhibit errors, then auditors can concentrate on such concepts. The purpose of this paper is to demonstrate such a structural method, applied to the UMLS.

In the META, all hierarchical relationships are represented in the file MRREL. According to Bodenreider [4] hierarchical relationships in the MRREL file are recorded according to their origin. Those relationships which are presented as hierarchical in the source terminology are recorded in MRREL as parent/child relationships. Relationships of source terminologies which are deemed as hierarchical by UMLS editors are recorded as broader/narrower.

In this study, we will concentrate on the parent/child hierarchical relationships, i.e. on a single child and a single parent at one time. In other words, we are using a *relationship-centric approach.* We note that a *concept-centric approach* is also possible (see Discussion). In fact, we will concentrate on the parent direction of the inverse pairs of relationships, pointing from a child concept to a parent concept (with a value of PAR in the MRREL file). Many of these relationships have some additional annotations. Among possible annotations there are *inverse_isa, has_part, has_branch,* etc. Those annotations are coming from the source terminologies. For example, some source terminologies, e.g. SNOMED, NCI Thesaurus, UWDA, NDDF and LOINC use IS-A as their hierarchical relationships. The isa annotations for parent relationships in the MRREL file are typically coming from those source terminologies.

Table 1: Distribution of PAR relationship annotation

REL	RELA	#rows	%
PAR	codesystem_of	1911	0.058
PAR	has_branch	5327	0.16
PAR	has_member	3631	0.11
PAR	has_part	19436	0.59
PAR	has_subtype	7023	0.21
PAR	has_tributary	1659	0.051
PAR	inverse_isa	1467247	44.9
PAR	(null)	1760486	53.89

Table 1 shows a breakdown of the parent-child relationships by rows in the UMLS table MRREL for version 2008AB. Notably, 1,760,486 pairs have relationships which are annotated with "null," meaning no RELA is available. In our previous work [3] only PAR relationships with isa annotation were used. However, the big (absolute) majority of relationships has the null annotation. If the unknown distribution of annotations for unlabeled PAR relationships follows the distribution of the labeled PAR relationships then there is a high probability that almost all represent IS-A links.

Thus, a null-annotated PAR relationship is not guaranteed to represent an IS-A relationship. However, if the IS-A relationship between the semantic types of the parent and of the child is inconsistent with the relationship between the parent concept and the child concept, this could indicate a possible error and should be reviewed by a human auditor. Of course, the auditor should keep in mind that there are other possible labels than IS-A.

In [3], we distinguished between six kinds of errors. Here, we define four error types: 1) Incorrect child semantic type, 2) Incorrect parent semantic type, 3) Incorrect parent-child relationship and 4) Child requires disambiguation. The first three are the same as in [3]. According to [8], if an ambiguity appears within one source, then a second ST is added. If the ambiguity comes from a contradiction between two sources, then two concepts should be created to disambiguate the source concept. While in the analysis of [3] a semantic type assignment is added whenever a parent or child is missing a semantic type, we disambiguate the ambiguous concept.

Methods

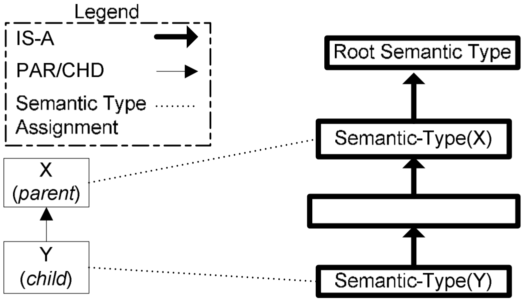

Figure 1: Possible Consistent Configuration

The primary purpose of this paper is to study whether the presence of structural inconsistencies of different kinds predicts different likelihoods of errors in three samples of parent-child relationships. Ideally, we would assume that whenever X is a parent concept of Y, then the semantic type of X is either identical to, or an ancestor (parent, grandparent, etc.) of the semantic type of Y (Figure 1). This is called a consistent configuration. We will distinguish between two kinds of inconsistencies: 1) Semantic inversion and 2) Lack of ancestry.

Semantic Inversion: T he semantic type of the child is an ancestor of the semantic type of the parent. Figure 2 shows a case of semantic inversion, which is visually recognizable by the crossing of the two "Semantic Type Assignment" (dotted) lines.

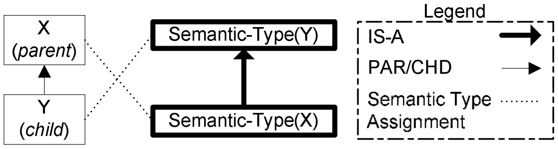

Figure 2: Semantic type inversion

Lack of Ancestry: There is no ancestor relationship between the semantic type(s) of the parent and the semantic type(s) of the child **at all**. Figure 3 shows one possible case of lack of ancestry. The semantic types in this case may be any two semantic types which are not hierarchically related by an ancestor-descendant relationship. There are different possible cases, e.g. cousin, uncle, etc. In this paper we do not distinguish between them. Lack of ancestry implies that this is neither a case of semantic inversion nor a consistent configuration for this parent-child pair.

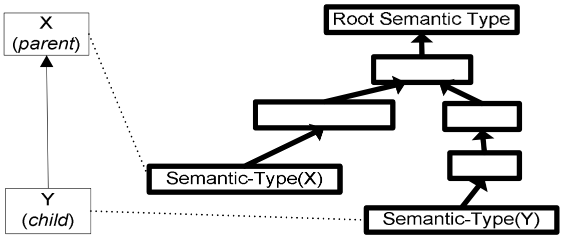

Figure 3: Lack of Ancestry

It would appear that semantic type inversion is a more severe problem than lack of ancestry. Lack of ancestry expresses a lack of knowledge about how two concepts should hierarchically/semantically relate to each other. Semantic inversion expresses the claim that a concept that was asserted to be more general than another concept is semantically more specialized than that concept. This is a direct contradiction between the asserted knowledge and the semantic knowledge about these two concepts. We stress, however, that a case of semantic type inversion is not automatically wrong. Due to the economy principle of the Semantic Network, cases of semantic type inversion may express valid medical knowledge [9]. Similarly, child concepts marked as NOS or NEC may cause valid cases of semantic type inversions.

Thus, a sample of parent-child relationships exhibiting semantic inversion is expected to contain more errors than a comparable sample exhibiting lack of ancestry, and both are expected to contain more errors than a sample with consistent configurations.

Hypothesis 1: A random sample of parent-child pairs with null annotation and semantic inversion will contain more errors than a comparable sample of parent-child pairs with lack of ancestry.

Hypothesis 2: A random sample of parent-child pairs with null annotation and lack of ancestry will contain more errors than a comparable sample of parent-child pairs with consistent configurations.

We analyzed three random samples of 100 parent-child pairs each, all with null annotations. The samples are for (1) semantic inversion, (2) lack of ancestry, and (3) consistent configurations. We found the problematic configurations (1 & 2) with a technique based on the UMLS numbering of semantic types [3]. We used the UMLS 2008AB release for our sample selection. As to the statistical method used to establish the significance of our results, we used a chi-square test with a continuity correction to compare the error rates in the three samples. [5]

Results

We found 38,323 cases of semantic inversion and 544,441 cases of lack of ancestry. There are 1,359,991 cases of parent-child pairs with consistent configurations. The sum of the numbers (1,942,755) exceeds the number of PAR (null) relationships (1,760,486 in Table1), because each concept may have multiple semantic types. Thus, a relationship may be qualified by more than one configuration (semantic inversion, lack of ancestry, or consistent).

Out of the 100 randomly selected parent-child pairs with semantic inversion that were analyzed, a domain expert determined that 84 contain errors. An analysis of these errors (Table 2) shows that they may be corrected in one of four ways: (1) The semantic type of the child is too general and needs to be more specific. (2) The semantic type of the parent is too specific and needs to be more general. (3) The parent-child relationship may be genuinely wrong, which means an error in the source terminology. (4) The child requires disambiguation. In our sample we did not find cases where ambiguity [8] could be repaired by assigning a new semantic type to the parent or the child concept, or cases of missing SN relationships, as in [3]. Neither was there a need for disambiguating the parent in any case.

Table 2: Wrong instances in sample with inversion

Error	Description	Count	Percentage
1	Wrong CHD ST	58	69%
2	Wrong PAR ST	16	19%
3	Wrong p/c rel	7	8%
4	Ambiguous CHD	3	4%
Total		84	100%

Table 3 provides examples of all four cases of semantic inversion errors and their corresponding corrections. The remaining 16 (16%) error cases do not fall into any of our four types of potential errors. In all of these cases we judged the PAR-CHD relationship to be of a non-IS-A type. Interestingly enough, only very few of the 16 cases fall into any of the acceptable categories in Table 1. For example, *Sleep Disorders* **{Disease or Syndrome}** is a child of *Mental disorders* **{Mental or Behavioral Dysfunction}**, and while sleep disorders may have an association with mental disorders and vice versa, a PAR-CHD IS-A relationship would be incorrect. Since an association between the concepts seems valid, we do not recommend marking *Mental Disorders* as incorrect parent. However, the relationship does not fit any of the typical META hierarchical relationship attributes.

Table 3: Examples of error types (1-4) detected by analyzing cases of semantic inversion

Parent **{ST of Parent}**	Child **{ST of Child}**	Change to correct error
Phytophthora {**Alga**}	Phytophthora megakarya {**Plant**}	Change CHD ST to **Alga**
Paraproteins {**Immunologic Factor**}	Myeloma Proteins {**Biologically Active Substance**}	Change PAR ST to **Biologically Active Substance**
Soft tissue neoplasms benign NEC {**Neoplastic Process**}	Lipogranuloma {**Disease or Syndrome**}	Mark PAR/CHD relationship as erroneous
Obesity monitoring NOS **{Therapeutic or Preventive Procedure}**	Target weight discussed **{Health Care Activity}**	Requires disambiguation; CHD synonyms are finding / procedure / regimen

Out of the 100 randomly selected parent-child relationships exhibiting lack of ancestry, 60% of the relationships were judged by a human domain expert to be correct, while 40% were wrong. Table 4 shows the breakdown of the sample according to the four kinds of errors.

Table 4: Wrong instances in sample with lack of ancestry

Error	Description	Count	Percentage
1	Change CHD ST	12	30%
2	Change PAR ST	18	45%
3	Wrong p/c rel	10	25%
4	Ambiguous CHD	0	0%
Total		40	100%

The random sample of 100 parent child pairs with consistent configurations contained only one case which was considered by a human domain expert to be a possible error. The semantic inversion sample had statistically significantly more errors than the lack of ancestry sample (p<0.001), which in turn had statistically more errors than the sample with consistent configuration (p<0.001). This confirms Hypotheses 1 & 2, that semantic inversion is a stronger indicator of potential problems than lack of ancestry, which in turn is a stronger indicator of potential problems than a consistent configuration.

Discussion

Sample Choice

In two of our papers [3,7] two different conditions were used for determining the condition of lack of ancestry. In [3] the approach is relationship-centric. The comparison is done between the semantic types of the child and the semantic types of the parent in each parent-child relationship. In [7] the focus is on checking the correctness of the semantic type assignments of a child concept. There we compare the types of the child concept to the set of types of all the parent concepts of this child.

Since the paper deals with inconsistencies between relationships on the META level and the Semantic Network level, our choice in this paper is to use the relationship-centric approach. As a result, we followed the model of [3] with regards to the choice of samples. However, in our auditing, we of course consider the possibility of having additional parent concepts of the child concept. It is possible that the semantic type of the child is identical to or a sub-type (IS-A) of a semantic type of one of the other parents. In such a case, the auditor will rule that there is no error. That is, in spite of the lack of ancestry in the original relationship the modeling is correct.

Interpretation

Structural features of the SN which appear inconsistent with structural features of the META lend themselves to automatic techniques for the identification of cases of semantic inversion or lack of ancestry. This automation helps focus auditing resources on concept sets where errors are expected. However, not all inconsistencies are equally valuable in predicting errors.

Analyzing a sample of parent-child pairs with semantic inversion led to the detection of relatively more errors than for sample pairs exhibiting lack of ancestry. Neither semantic inversion nor lack of ancestry is an accurate indicator that errors will occur. However, 84% of cases of the severe indicator of semantic inversion were found to be erroneous. The practical implication is to concentrate scarce auditing resources on cases of semantic inversion first and then move on to cases of lack of ancestry, if resources are still available. According to *this* study, auditing of parent-child pairs may be omitted for cases with consistent configuration, as the likelihood of finding errors appears too low. More research is needed to potentially find a more refined condition of consistent configuration which may have a higher probability of errors, than the general case used here.

Many of the errors detected can be corrected by changing the semantic types assigned to concepts. However, a number of detected errors can only be fixed by removing a parent-child relationship, or by removing its labeling as parent-child relationship, since it is not correct, even with the most general possible interpretation of the parent-child relationship. By the UMLS policy, this requires changes in the source terminologies, by communicating with the organization in charge.

Previous research concentrated on the isa annotated parent-child relationships [3]. Our study shows that even pairs with null annotation can provide a rich source for auditing when they are inconsistent with SN hierarchical relationships. In the future we plan to investigate the distribution of PAR-CHD relationship types among null-annotated pairs and compare it to the distribution among the annotated ones (Table 1), as this might affect the percentage of relationships for which semantic type inversion may indicate an error.

In this study we also defined a new type of potential error: Child requires disambiguation (Ambiguous CHD). This error came to light due to the analysis of semantic type inversions. One such example (Table 3) is the concept *Target weight discussed* which has the semantic type of **Health Care Activity**. However, a close look at the synonyms of the concept reveals that they are: *Target weight discussed (finding) / … (procedure)* and *… (regimen/therapy),* and while the parent would be appropriate for a procedure, **Health Care Activity** or **Therapeutic or Preventive Procedure** would clearly be inappropriate semantic types for a finding.

In a pure IS-A hierarchy it is hard to imagine justified cases of semantic inversion. However, by definition, the META does not rely on a single type of relationship for the hierarchical structure. Table 1 lists the available types of PAR-CHD relationships. The fact that most of the 16 remaining error cases of semantic type inversion have parent-child relationships not fitting with the choices in Table 1 suggests that further research is required into the relationship attributes of hierarchical relationships. *Sleep Disorders* has numerous additional parents, many of which do not fit under an IS-A relationship, nor under the other types. While our study utilized a relationship-centric approach, the example highlights that additional problems may exist within the immediate neighborhood, that merit investigation.

Conclusions

The structural interplay of the SN and META is valuable in discovering inconsistencies in the META. In two samples of 100 parent-child pairs, 84% and 40% were found to be wrong. This is not a coincidence; they were automatically retrieved by satisfying two conditions of inconsistency with the SN, one more severe than the other. In future work we will distinguish between semantic type inversion where one semantic type is a parent versus where it is an ancestor of the other semantic type.

Acknowledgments

This work was partially supported by the United States National Library of Medicine under grant R 01 LM008445-01A2.

References

1. Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The Unified Medical Language System: An Informatics Research Collaboration. JAMIA. 1998;5(1):1-11.
2. McCray AT. An upper level ontology for the biomedical domain. Comp Funct Genom. 2003. 4:80-4.
3. Cimino JJ, Min H, Perl Y. Consistency across the Hierarchies of the UMLS Semantic Network and Metathesaurus. JBI. 2003 December;36(6):450-61.
4. Bodenreider O. Circular Hierarchical Relationships in the UMLS: Etiology, Diagnosis, Treatment, Complications and Prevention. In: Bakken S, editor. Proc 2001 AMIA Annual Symposium; 2001 November; Washington, DC; 2001. 57-61.
5. Snedecor GW, Cochran WG. Statistical Methods. 8th ed. Iowa State University Press; 1989.
6. 2007AC Section 3 UMLS Semantic Network. [cited March 11, 2008]; Available from: **http://www.nlm.nih.gov/research/umls/meta3.html**
7. Chen Y, Gu H, Perl Y, Geller J, Halper M, Structural group auditing of a UMLS semantic type's extent. JBI. 2009;42(1):41-52.
8. McCray AT, Nelson SJ, The representation of meaning in the UMLS. Meth Inform Med. 1995;34(1-2):193-201.
9. 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. 222-33.