MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

Methods for Exploring the Semantics of the Relationships between Co-occurring UMLS Concepts

Anita Burgun, Olivier Bodenreider

U.S. National Library of Medicine, Bethesda, MD, USA
{burgun, olivier}@nlm.nih.gov

Abstract

Objective: To characterize the relationships among UMLS concepts that co-occur as MeSH descriptors in MEDLINE citations (1990-1999). Design: 18,485 UMLS concepts involved in 7,928,608 directed pairs of co-occurring concepts were studied. For each directed pair of concepts C1-C2: (i) the "family" of C1 was built, using the UMLS Metathesaurus, and we tested whether or not C2 belonged to C1's family; (ii) we used the semantic categorization of Metathesaurus concepts through the UMLS Semantic Network and Semantic Groups to represent the semantics of the relationships between C1 and C2. Results: In 6.5% of the directed pairs, the co-occurring concept C2 was found within the "family" of C1. Detailed results are given. The most frequent co-occurrences involved "Chemicals & Drugs" and "Chemicals & Drugs", as well as "Disorders" and "Chemicals & Drugs". Discussion: This work takes advantage of both symbolic and statistical information represented in the UMLS, and analyzes their overlap. Further research is suggested.

Keywords:

UMLS; Semantics; MeSH; co-occurrences.

Introduction

Knowledge associated with a concept, i.e. its definition and its relationships with other concepts refers to both symbolic representation and statistical information, from which semantic spaces can be constructed [1, 2]. The Unified Medical Language System® (UMLS) can be seen as an attempt to combining symbolic knowledge and statistical information about the biomedical domain. Symbolic knowledge is provided by the Semantic Network (SN), and by the symbolic interconcept relationships in the Metathesaurus. Statistical information is represented by the co-occurrences, mainly co-occurrences between MeSH descriptors in MEDLINE®. For each pair of MeSH descriptors, the frequency of co-occurrence in MEDLINE citations is recorded in the UMLS and can be used as a surrogate for the strength of the relationships. Therefore, co-occurrences are an important source of knowledge that has the potential to complement the limited set of symbolic relationships, and should benefit from characterization of their semantics to be fully usable.

The objective of this study is to propose a methodology for characterizing the relationships among UMLS concepts that co-occur in MEDLINE (referred to as COC relationships), from two perspectives: 1) To compare the COC relationships to existing symbolic relationships in the Metathesaurus. Part of the concepts that co-occur with a given concept are expected to belong to its semantic space, also named its family, which refers in the UMLS to symbolic relationships in the Metathesaurus [3]. We use a broad definition of family, so that family includes not only the set of strict relatives, but also concepts that are related by associative relationships. As a metaphor, this notion of family applied to a person P would encompass ascendants, descendants, siblings, uncles, cousins, and also friends, parents' friends and children's friends. When someone is seen (co-occurs) with P, he should have high probability to belong to P's family. 2) To classify semantically the COC relationships according to the UMLS Semantic Network. In addition, broader categories than UMLS Semantic Types (STs) are used to provide higher-level categorization.

Materials

Data about co-occurring concepts were selected from the UMLS MRCOC file, with the following criteria: 1) the unique source that was taken into account was MEDLINE, 2) the period was 1990-1999, 3) the concepts had to be starred MeSH descriptors. Using those criteria, 18,485 UMLS concepts were selected. Each MeSH descriptor corresponds to a concept in the Metathesaurus, which makes it possible to process the UMLS concepts within their UMLS semantic environment (links to other concepts, semantic types). The 18,485 selected concepts participated in 7,928,608 directed pairs of co-occurring concepts (directed meaning that both C1-C2 and C2-C1 are represented), i.e. 3,964,304 non-directed pairs of co-occurring concepts. In this study, we used the 2000 release of the UMLS Knowledge Sources [4].

Methods

We analyzed co-occurrence relationships using three progressively decreasing levels of semantic granularity. The **family of a concept** provides the most specific information but its scope is limited to the symbolic relationships represented in the Metathesaurus. The **Semantic Network** systematically provides categorization for the concepts, and potentially instantiates the semantics of co-occurrences. Complexity of the resulting representation may be reduced by aggregating the information into broader **Semantic Groups**.

Categorization based on the family of a concept

This approach aims to compare, for each concept C among the 18,485 relevant UMLS concepts, the concepts that co-occur with C to the concepts that belong to its "family". The family of a given concept C is built not only from existing relationships in MRREL, but also from additional, more complex relationships, in order to increase the chances of overlap between C's family and its co-occurring concepts.

Existing relationships in MRREL are either hierarchical relationships: PAR (parent), CHD (child), RB (broader), RN (narrower), hierarchically-related: SIB (sibling), or non-hierarchical, essentially associative relationships: RO (other). All these relationships are direct, i.e. are 1-level relationships. We defined 3 types of more complex relationships, listed in table 1:

- **Redefined hierarchical** (or hierarchically-related) relationships. By their meaning and their use, PAR and RB are very close, and CHD and RN are very close. Therefore, three redefined hierarchical or hierarchically-related relationships have been implemented: Ancestor1, Descendant1, and Extended siblings. Ancestors1 (ANC_1) result from the union of Parents and Broader concepts. The mention 1 means that only one level in the hierarchy is explored. Descendants1 (DES_1) result from union of Children and Narrower concepts. Extended siblings (SIB_X) are descendants1 of the ancestors1.
- **Multiple-level** relationships derive from a recursive definition of Ancestors or Descendants, all the way to the top or to the bottom of hierarchies.
- **Combined** relationships lead to "relatives" that are more distant from C than directly related concepts, but still belong to its "family". They are *uncles* (the extended siblings of the ancestors1) and *cousins* (the descendants of uncles), *other related of ancestors1* (AOT), *other related of descendants1* (DOT).

The family of a given concept C is therefore the set of all the concepts that are related to C by any of the above relationships. They may be clustered according to 3 axes. The first axis, H, represents all the ancestors and descendants. The second axis, A_H, represents lateral relatives with siblings, uncles and cousins. The last one, A_O, represents the other related concepts (RO, AOT, DOT).

Whereas some relationships, such as hierarchical relationships, are symmetric, others are not symmetric (e.g. uncles, since we do not define a nephew relationship), which justifies processing directed pairs of co-occurring concepts. For each concept C2 that co-occured with C1, its belonging to C1's family was tested for each type of family relation. The Perl 5 program is based on the object-oriented model of UMLS developed at NLM [5].

Additionally, for some of the relationships that are both co-occurrences and family relationships, relationship attributes recorded in the Metathesaurus may provide information about the nature of the relationship. For example, *Addison's Disease* and *Hyponatremia* are co-occurring concepts, and *Hyponatremia* belongs to *Addison's Disease*'s family, the relation being RO, with the attribute "clinically associated with".

Table 1 Additional family relationships

Relationship	Definition
Ancestor 1 (first level)	$ANC_1 = PAR + RB$
Ancestor	$ANC_n = ANC_1 \circ ANC_{n-1}$
Descendant 1	$DES_1 = CHD + RN$
Descendant	$DES_n = DES_1 \circ DES_{n-1}$
Extended Sibling	$SIB_X = DES_1 \circ ANC_1$
Uncle	$UNC = SIB_X \circ ANC_1$
Cousin	$COU = DES_1 \circ UNC$
Other Related of Anc1	$AOT = RO \circ ANC_1$
Other Related of Des1	$DOT = RO \circ DES_1$

Categorization based on the Semantic Network

The UMLS Semantic Network (SN) provides a high-level semantic structure for representing the biomedical domain. In order to provide all potential semantic relationships for all pairs of STs, we used the fully developed set of relationships between STs resulting from the transitive closure of the SN graph, according to the methodology described in [6]. For example, the six potential semantic relationships between the STs "Disease or Syndrome" and "Pharmacologic Substance" are shown in figure 1. Semantics provided by the SN graph may be used to instantiate semantics of the co-occurrence relationship. For example, *Addison's Disease* and *Adrenal Cortex* are co-occurring concepts that have no relationships in the Metathesaurus. According to their STs, the semantics of the co-occurrence relationship is "has location". A limitation of

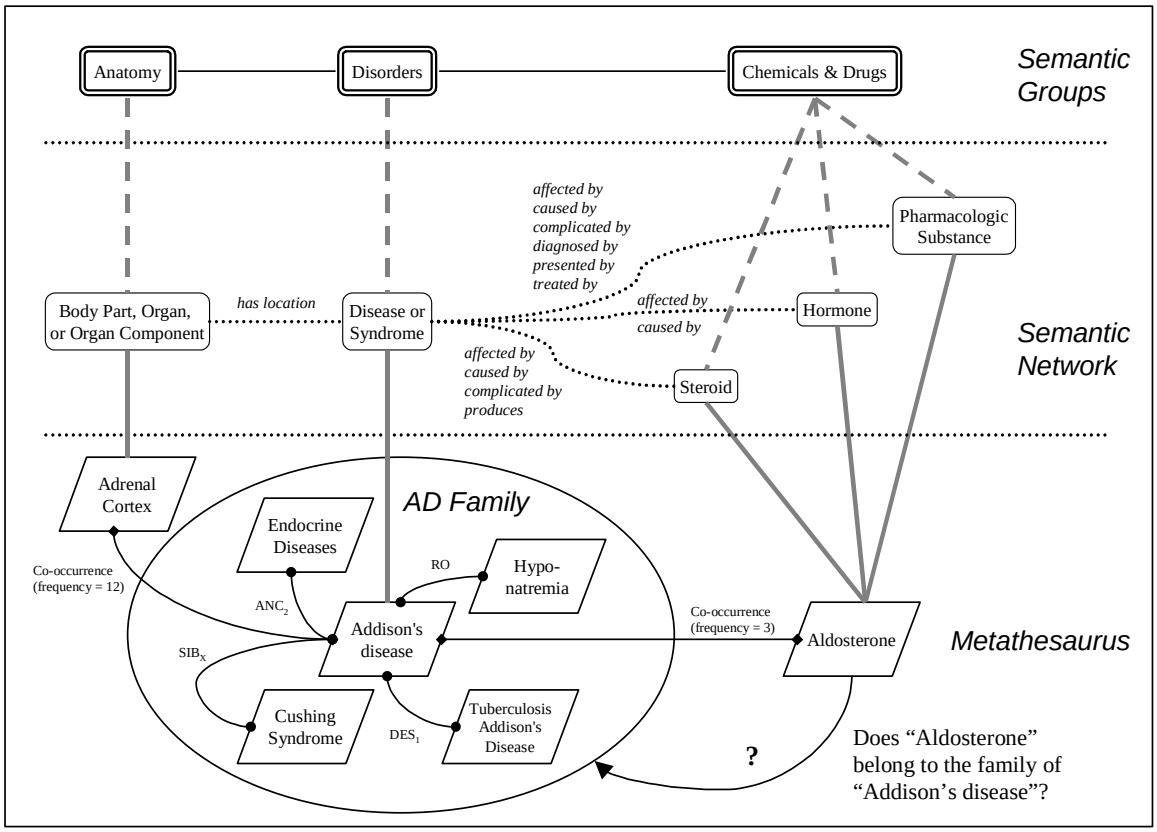

Figure 1- Overview of the methodology applied to the relationships of "Aldosterone" to "Addison's disease"

this method is the fact that a concept may be assigned several STs, which prevents exact counting.

Categorization based on Semantic Groups

To categorize semantically the relationships between co-occurring concepts, we reused the Semantic Groups initially established for displaying MEDLINE co-occurrence information [7], and used for organizing concepts in the UMLS Semantic Navigator [8].

The 134 STs are clustered into 15 Semantic Groups (SG), which can be defined as clusters of STs, and not necessarily as supertypes of STs. For example, in the SN:

- *Body System* is a subtype of *Functional Concept*, which is a subtype of *Idea or Concept*, and thus a subtype of *Conceptual Entity*;
- *Body Part, Organ, or Organ Component* is a subtype of *Fully Formed Anatomical Structure*, which is a subtype of *Anatomical Structure*, and thus a subtype of *Physical Object*,

However, *Body System* and *Body Part, Organ or Organ Component* are gathered in the same SG, which is "Anatomy".

In addition to simply organizing the UMLS, Semantic Grouping aims to provide a partition of Metathesaurus concepts, i.e. each concept essentially belongs to one and only one group [9]. Only 4913 concepts are assigned more than one SG in the whole UMLS, thus SGs provide a limited combination of categories for the 3,964,304 distinct, non-directed pairs of co-occurring concepts.

Combining the frequency of co-occurrence between two concepts with the belonging of concepts to a SG, we were able to compute the frequency of co-occurrences between pairs of SGs.

Figure 1 provides an overview of the whole methodology.

Results

Categorization based on the family of a concept

The proportion of directed pairs of co-occurrences where the concept C2 belongs to the family of C1 is 6.5% (511,673/ 7,928,608). Among those 511,673 pairs,

- most of them belong to A_H axis: extended siblings (14%), uncles (16%), or cousins (42.5%).
- 124,296 co-occurring concepts are close relatives, corresponding to 1-level relationships in the Metathesaurus (PAR, RB, CHD, RN, RO). They represent the inner circle in figure 2.
- 94,937 are "related concepts" (A_O axis): 26,633 directly (OTH), the others are related to direct ancestors (47,978) or are related to direct descendants (31,300).

Additionally, 94.6% of the concepts have at least one co-occurring concept which is in their family, i.e. only 427 concepts among 18,485 have none of their co-occurring concepts that belong to their families. One concept may be linked to another one by several types of family relationships.

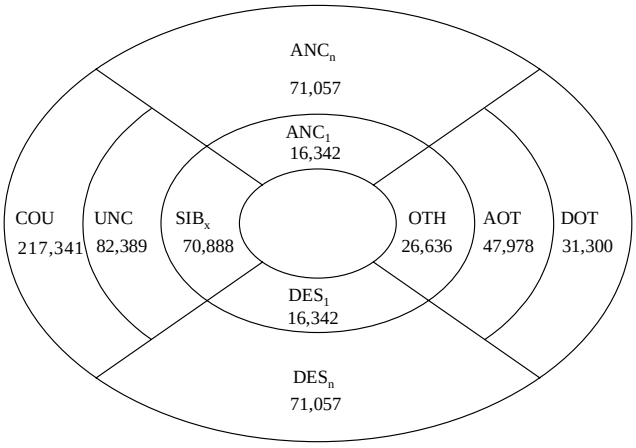

Figure 2. Co-occurring concepts belonging to relatives

More specifically, relationship attributes represented in the Metathesaurus are used to provide semantics for some of the relationships that are both co-occurrences and family relationships. Examples of attributes include "isa", "part of", or "has location". Nevertheless, the large majority of relationships have no relationship attribute. For example, among the 26,636 relationships that are qualified as OTH relationships, 78% are not defined, whereas "clinically associated with" represents 12% of the OTH relationships, "has location" represents 5.5%, "sign symptom complex" represents 1.2% and the other attributes represent less than 1%.

Categorization based on the Semantic Network

Since a UMLS concept may be assigned more than one Semantic Type, the 3,964,304 distinct, non-directed pairs of co-occurring concepts instantiate 7,293,481 pairs of STs, 80% of which are instantiation of allowable links among pairs of STs, "allowable" referring to links that do fit the SN structure. Roughly half of the distinct pairs of STs generated by co-occurrences can be represented by allowable relationships according to the SN. In other terms their semantics can be inferred, more or less precisely, from the SN. 92% of the distinct allowable SN links are represented among co-occurring concepts. Frequent relations are "interact with", "affects", "causes", and "complicates".

Categorization based on Semantic Groups

The 3,964,304 distinct, non-directed pairs of co-occurring concepts are categorized into 119 non-directed pairs of SGs among the 120 possible pairs. The association "Geographic Areas"- "Genes & Molecular Sequences" is not represented. A pair of SGs subsumes 5 ("Geographic Areas"- "Devices") to 585,155 ("Chemicals & Drugs" – "Chemicals & Drugs") distinct pairs of concepts. The pairs of SGs that subsume more than 100,000 rows are displayed in figure 3; they represent 9 pairs, they involve 6 SGs ("Procedures", "Living Beings", "Physiology", "Disorders", "Anatomy", and "Chemicals & Drugs"), and they represent 60% of all the co-occurrences.

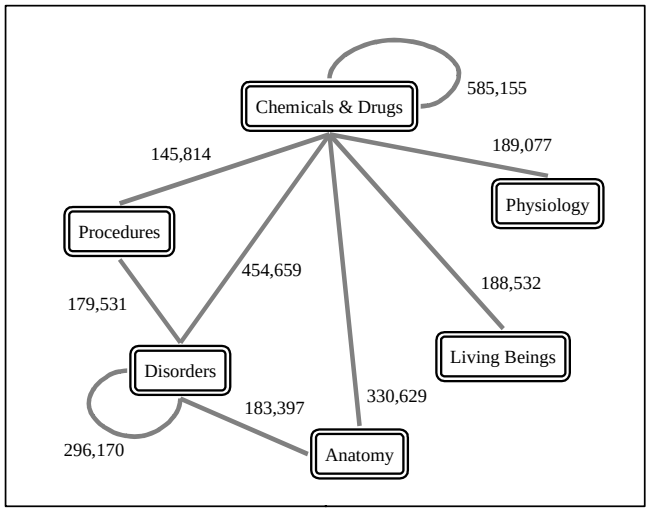

Figure 3- Pairs of Semantic Groups in MEDLINE co-occurrences (without taking into account frequency)

The same analysis was performed, taking into account the frequency of each pair of co-occurring concepts in MEDLINE. The ten pairs of SGs that gather roughly 2/3 of MEDLINE co-occurrences involve the same 6 SGs. The range of the 25 first pairs of SGs is globally unchanged. Nevertheless, some SG pairs are more represented when taking into account frequency, such as "Chemicals & Drugs" – "Genes & Molecular Sequences".

The semantics of the relationships can be partially inferred from the pairs of SGs. For example, the relationship between "Disorders" and "Anatomy" should be mainly "has location". Nevertheless the relationship between "Disorders" and "Chemicals & Drugs" remains ambiguous, since it could be either "treated by" or "caused by".

Discussion

The method presented in this paper was applied to MEDLINE citations, but it is applicable to other sources as well. For example, it could be applied to any kind of documents (patient records, Web documents, etc), assuming that they are indexed or indexable with UMLS concepts, or with any terminology that is one of the UMLS sources. This method is language independent since it is based on interconcept relationships and not on terms. It is also vocabulary independent, since it does not use any vocabulary specific features but relies on properties of UMLS concepts. That was our rationale for using SGs rather than top MeSH categories.

This approach takes advantage of both symbolic and statistic information. The SN Relationships provide semantics for co-occurrences. Approaches are based on semantic links resulting from the transitive closure of the

SN graph, even if the relevance of the semantic relations that can be inferred by this method has to be evaluated, according to the low rate of redundancy between Metathesaurus relationships and co-occurrences. This approach is somewhat limited by the fact that roughly one fourth of Metathesaurus concepts are assigned several STs. As a result, pairs of co-occurring concepts may generate many combinations of STs, which makes it difficult to get a precise view of the semantics of co-occurrences. Similar issues are addressed by Mendonça and Cimino while extracting medical knowledge from MEDLINE co-occurrences [10]. In our study, we used clusters of STs that partition the UMLS.

Our results show a low rate of redundancy between direct relationships in Metathesaurus (MRREL) and co-occurrences. Overlap is limited since (1) selection of hierarchically-related descriptors is limited by indexing rules, (2) some associative relationships are not systematically represented in the UMLS, e.g. manifestation, adverse effect, etc, and (3) co-occurrence may be accidental. Moreover, few co-occurring concepts were found in the set of "other related" concepts (4% of the co-occurring concepts that belong to families). This proportion remains low even when the notion of "other related" is extended to the "other related" of parents (14 %) or to the "other related" of children (7%). These results show that Metathesaurus symbolic knowledge cannot systematically help select relevant links between co-occurrences or provide more refined semantics. This could also suggest that additional non-hierarchical links could be instantiated among the Metathesaurus concepts. A more detailed analysis of associative relationships is to be done.

Work on co-occurrences would take advantage of Semantic Interpretation tools [11]. In Semantic Interpretation, semantic rules establish a correspondence between a linguistic item and a SN relation. Assuming that MeSH descriptors can be identified in the abstract, and that a syntactic relationship can be found between them, a semantic relationship could be inferred.

Acknowledgments

This research was supported in part by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine.

References

[1] Landauer TK, Dumais ST. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 1997, 104, 2:211-240

[2] Chute CG, Yang Y. An evaluation of concept based latent semantic indexing for clinical information retrieval. *Proc. 17th Annual Symposium on Computer Applications in Medical Care*, 1993; pp 639-643

[3] Nelson SJ, Tuttle MS, Cole WG, Sherertz DD, Sperzel WD, Erlbaum MS, Fuller LL, Olson NE. From meaning to term: Semantic locality in the UMLS Metathesaurus. *Proc. 16th Annual Symposium on Computer Applications in Medical Care*, 1992, pp 209-213

[4] UMLS Knowledge Sources, 11th edition, 2000, National Library of Medicine, Bethesda MD

[5] Bodenreider O. An object-oriented model for representing semantic locality in the UMLS; submitted Medinfo 2001

[6] McCray AT, Bodenreider O. A conceptual framework for the biomedical domain. In *Semantics of Relationships*, Myeng SH and Green R Eds Kluwer; to appear

[7] UMLS Knowledge Sources, 6th edition, 1995, National Library of Medicine, Bethesda MD

[8] Bodenreider O. A semantic navigation tool for the UMLS. *Proc. AMIA Fall Symposium*, 2000, pp 971. (umlsks.nlm.nih.gov → Resources → Semantic Navigator)

[9] McCray AT , Burgun A, Bodenreider O. Aggregating UMLS Semantic Types for reducing conceptual complexity; submitted Medinfo 2001

[10] Mendonça EA, Cimino JJ. Automated knowledge extraction from MEDLINE citations. *Proc. AMIA Fall Symposium*, 2000, pp 575-579.

[11] Rindflesch TC, Aronson AR. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. *Proc. 18th Annual Symposium on Computer Applications in Medical Care*, 1994, 240-244

Address for correspondence

Anita Burgun, National Library of Medicine, 8600 Rockville Pike (MS 43), Bethesda, MD 20894 - USA.

e-mail: burgun@nlm.nih.gov