Formalizing UMLS Relations using Semantic Partitions in the context of task-based Clinical Guidelines Model

Anand Kumar, Matteo Piazza, Silvana Quaglini, Mario Stefanelli

Laboratory of Medical Informatics, Department of Computer Science, University of Pavia, Pavia, Italy.

Abstract

Objectives To formalize the relations between the different Semantic Types in the Semantic Network of the Unified Medical Language System (UMLS) in the context of computer interpretable task-based clinical guidelines model.

Design We used the Semantic Type Collections as our basis in the formalization. We defined some operators based on the relations which we considered were applicable to all the Semantic Types in the Collection.

Measurement We separated the relations dealing with the Semantic Types Diagnostic Procedure, Laboratory Procedure and Therapeutic or Preventive Procedure. We calculated the ratio of the total relations in UMLS Semantic Network to the adjacent relations of these Semantic Types and also the percentage of Semantic Types who have an adjacent relations with these Semantic Types.

Result Without the consideration of Semantic Type Collection, the total adjacent relations for these three Semantic Types was 6.87% of the total and these covered almost half of the total Semantic Types on an average. With the consideration of Semantic Type Collection, we were able to represent these relations and also cover the whole UMLS Semantic Network.

Conclusion With the formalization of the mapping operators for the adjacent Semantic Types with respect to the three Semantic Types, the next step would be to map the non-adjacent ones to those three Semantic Types following a path with least distance.

Index Terms Clinical Practice Guidelines, UMLS, Semantic Types, Semantic Network, Graph Theory, Minimal Spanning Network

I. INTRODUCTION

Creation of task-based CIGMs [1] requires effective medical ontologies and terminologies [2]. Integration of the standard terminology systems into a unified knowledge representation system is achieved in Unified Medical Language System (UMLS), designed by the National Library of Medicine. UMLS Semantic Network (SN) (January 2003) consists of 132 Semantic Types (STs) and 54 possible links. These form a graph with 2-tree structure. The vertices consist of the STs and the edges consist of the links between them. The corresponding complete graph would contain more than 6000 edges. The SN serves as a high-level abstraction for the Metathesaurus (META), which is the UMLS concept repository.

In section II, we mention some fundamentals of graph theory which would be needed as a theoretical basis while we develop our arguments further. In section III, we deal with the representations of UMLS SN as proposed by other groups and how we adapt it for our use. In section IV, we deal with the issue of UMLS SN restriction from the context of task-based CIGMs to three STs – *Diagnostic Procedure*, *Laboratory Procedure* and *Therapeutic or Preventive Procedure* and then in section V we deal with the formalizations which are needed for the same, keeping in mind the already existing relations between the UMLS STs. Section VI deals how this work is applicable to the ongoing work regarding the ontological aspects of CIGMs, represented in DAML+OIL. We conclude in Section VII mentioning the ongoing effort to map the STs not adjacent to the three STs mentioned above, in order to cover the complete UMLS SN.

II. FUNDAMENTALS OF GRAPH THEORY

A graph is a pair $G = (V,E)$ of sets satisfying $E \subseteq [V]^2$; thus the elements of E are two-element subsets of V.The elements of V are the Vertices

or Nodes of the graph G, and the elements of E are it's edges.

Let G(V,E) and G' = (V',E') be two graphs. If V' ⊆ V and E' ⊆ E, the G' is a subgraph of G, represented as G' ⊆ G.

If G'⊆ G and G' contains all edges xy ∈ E and x,y ∈ V', then G' is the induced subgraph of G.

A graph G' is a spanning subgraph to graph G if G'⊆ G and vertices of G' spans all of G (the vertices V of G), that is V'=V. A minimal spanning graph is a spanning subgraph such the sum of all edge costs is minimal.

In a graph, each edge may be associated with a cost, that depends on the "meaning" of the edge, for example, a quantitative measure of the distance between two vertices. In the current context, since the edges do not have a quantitative weight assigned to them, we consider all the edges to have a weight of one.

III. REPRESENTATION OF UMLS SN

The work of Chen, Perl et al and Geller, Perl et al are significant in reducing the complexity of the SN. In order to describe a scenario with the simplified representation, they gave a few definitions. [Table 1] This work helps to simplify the UMLS SN visual representation and that it appeals to the human interpretation, as pointed out by their evaluation study.

The work is based on the concept of subgraphs. The idea of Induced subnetwork is based on the Induced Subgraph, which essentially means that in the subgraph, all the edges are considered which are present between the nodes in that subgraph. But in any one of those subgraphs, all the UMLS STs cannot be considered. Since the UMLS META is connected to the UMLS STs, the whole META in the UMLS SN is not covered in one Induced Subnetwork or Collection Subnetwork. The STs in the UMLS SN are the vertices and the concepts in META are connected to those vertices, and therefore, if we reduce the number of vertices in the Induced Subnetwork or Collection Subnetwork, we will not be able to cover the whole META.

The work done towards the partition and development of Induced Subnetwork, Collection Subnetwork and Collection Environment, quite obviously, wasn't meant to cover the whole META at the same time, but to simplify the UMLS SN representation.

When we deal with the task-based CIGMs, the text source of the GLs includes the terms which can belong to any of the UMLS STs. This meant that the there is a need to implicitly or explicitly map the knowledge which belonged to the different STs into executable tasks for the task-based CIGMs. Therefore, to interpret the text and to develop the CIGM based on the Induced Subnetwork, we need all the vertices of the UMLS SN.

Considering this requirement and the work done with cohesive partition, we developed a Minimal Spanning Subnetwork of the UMLS SN which could be used specifically in the context of CIGMs.

We will first try to justify the issue and then give a formal description of the technique used to create the Minimal Spanning Subnetwork.

IV. RESTRICTION OF UMLS SN FOR TASK-BASED GUIDELINES

From the GL point of view, most of the GLs suggested actions can be mapped into the terminology associated with the STs – *Laboratory Procedure*, *Diagnostic Procedure* and *Therapeutic or Preventive Procedure*. These STs are subclassifications of the ST *Health Care Activity*.

Thus in terms of the GL-based task determination, the tasks suggested in the GLs can be linked to one of these STs. Other STs which are used relatively less frequently in GLs are related to *Occupational Activity*, which contains *Educational Activity*, *Governmental or Regulatory Activity* and *Research Activity*. And even these are related to *Health Care Activity*. For example, *Research Activity* actually helps to determine the *Health Care Activity* by giving the "Strength of Evidence".

This fact led us to consider these three STs as the basis for the creation of the Minimal Spanning Subnetwork. Thus the goal was refined to connect every other ST in the UMLS SN to these three STs.

However, for this task, we didn't consider every ST separately. We took the STCs already defined and studied the relations mentioned in the UMLS SN between these STCs to the three STs.

We found out that a lot of the relations which existed between the STs belonging to the STCs and these three STs could be combined together. Based on this idea, we defined some operators which would help the mapping of the STs present in the STCs to the three STs. These

operators were then formalized based on the existing relationships in the UMLS SN.

V. FORMALIZED RELATIONS

In the January 2003 edition of UMLS SN, we found 6718 adjacent relations between the different STs. When we consider the adjacent relations just to the three STs, we found there are 179, 179 and 104 relations for *Diagnostic Procedure*, *Laboratory Procedure* and *Therapeutic or Preventive Procedure* respectively. This means that these relations comprised 2.66%, 2.66% and 1.55% of the total relations mentioned in the UMLS SN, putting the total to 6.87%. Even though the number of relations for *Diagnostic Procedure* and *Laboratory Procedure* are identical, not all the relations of those two STs are identical.

In the same edition, the number of STs in the UMLS SN are 132. Among those STs, the number of STs which have a direct relation to the three STs mentioned above are 79, 67 and 51 for *Diagnostic Procedure*, *Laboratory Procedure* and *Therapeutic or Preventive Procedure* respectively. Thus, the percentage of nodes having adjacent relation to these three STs are 60.8%, 50.8% and 38.6%, putting the mean at 50.1%.

These numbers signify that even without considering STCs or related subnetworks, with mere 6.87% of the total 6718 relations, we are able to define the adjacent relations of almost half of the STs in the UMLS SN with *Diagnostic Procedure*, *Laboratory Procedure* and *Therapeutic or Preventive Procedure*.

However, this meant taking care of three main concerns. Firstly, the 6.87% of relations which we considered had a lot of similarities as regards their meaning and this could be further simplified. Secondly, this still meant more than 100 relations of each of the three STs and therefore we couldn't define few operators without further simplication. Thirdly, the adjacent STs still consisted of only half of the total STs present in the UMLS SN and therefore with this framework we couldn't cover all the terms in the UMLS META.

These were the reasons to adopt the STC and related specifications. This leads to consider all the STs in a STC to have common relations, as mentioned in Rule 1.

RULE 1: *If a relation holds for the most generalized class in a STC (i.e. the class gives name to STC), we assume the relation holds for its children, and then for the entire STC.*

For example, the STC – *Biologically Active Substance* consists of *Biologically Active Substance*, *Receptor*, *Vitamin, Enzyme, Neuroreactive Substance* or *Biogenic Amine*, *Hormone and Immunologic Factor*. All the STs belonging to this STC have three relations to the *ST Diagnostic Procedure*, namely – analyzes, assesses_effect_of, measures. Therefore, we can assign these relations as the relations for the whole STC.

However, the work is not finished here due to two main concerns. Firstly, these three relations have similar meaning and therefore we could define a common operator for mapping the STs belonging to this STC into *Diagnostic Procedure*. Secondly, this example is relatively simpler. The reason for this is explained by what is called "Structurally uniform" by Chen, Perl et al, that is when the STC and STG are the same. Therefore, the relations between the STs of the STC and *Diagnostic Procedure* are identical. Therefore, it is easy to understand the relations and the use of Rule 1 in such situations.

On the lines of the above arguments, we created some operators. For the example mentioned above, the operator was "Determination OF" or DOF. All the three relations – analyzes, assesses_effect_of, measures can be put together as DOF, and therefore by the use of DOF, the STs present in the STC – *Biologically Active Substance* can be mapped into *Diagnostic Procedure*. The DOF operator has been formalized in table4.

In this way, we formalised other different operators which were based on the relations mentioned in the UMLS SN.

COF (Cause OF) ≡ {inv(result_of), affects, complicates}

UOF (Use OF) ≡ {uses}

MOF (Management OF) ≡ {treats, prevents}

Therefore for STC-Biologic Funtion, the formalization would be as follows.

STC-Biologic Funtion ≡ {Biologic Function}

Diagnostic Procedure = measures (Biologically Active Substance)

⇒ Diagnostic Procedure = DOF (STC- Biologically Active Substance)

Extending the theory from the point of view of representation, we formulated Rule 2.

RULE 2: *If the same relation holds for a certain class and it is repeated in all its sub-classes it is possible to simplify the graph and the formalization making the relation explicit just once for the most general class.*

Based on this, we could show the relations based on the three STs mentioned above in Figure. Thus, with the help of Semantic Type Collections and Induced Subnetworks, with the defintion of Semantic Relation Collections and using them in the context of three STs, we are able to give a simplified representation of the UMLS SN, which is specific for the task-based CIGMs.

VI. GUIDELINE – SPECIFC ONTOLOGY

Apart from the general ontology among the UMLS STs, we created the GL-specific ontology. We used the 1999 WHO-International Society of Hypertension Guidelines for the Management of Hypertension.

On the one hand, we were able to define the context in which the tasks were to be performed in the GL text by using the is-a structure. While on the other, the base classes being *Laboratory Procedure*, *Diagnostic Procedure* and *Therapeutic or Preventive Procedure*, the connection to the UMLS STs was maintained even in the GL-specific ontology.

The GL text of the WHO-International Society of Hypertension deals with the classification of hypertensive patients based on their Blood Pressure; risk factor assessments based on accompanying pathologies; and therapeutic measures for the management of hypertension.

The above-mentioned operators have been used in order to create the task-based ontology, a part of which is represented in the Figure.

VII. CONCLUSION

We are able to draw different quantitative and qualitative conclusions based on the results. To be extended….

REFERENCES

[1] Peleg M, Tu S, Bury J et al.Comparing Computer-interpretable Guideline Models: A Case-study Approach. J Am Med Inform Assoc. 2003 Jan-Feb;10(1):52-68.

[2] Nigel S., Michel C, and Jean PB. Which coding system for therapeutic information in evidence-based medicine. Computer Methods and Programs in Biomedicine 2002; 68(1):73-85.

[3] Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc. 1998;5(1):1-11.

[4] Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993;32:281-91.

[5] Zhang L, Perl Y, Halper MH, Geller J, Cimino JJ. Enriching the Structure of the UMLS Semantic Network. Proc AMIA Symp. 2002;:939-43.

[6] Geller J, Perl Y, Halper M, Chen Z, Gu H. Evaluation and application of a semantic network partition. IEEE Trans Inf Technol Biomed. 2002 Jun;6(2):109-15.

[7] Chen Z, Perl Y, Halper M, Geller J, Gu H. Partitioning the UMLS semantic network. IEEE Trans Inf Technol Biomed. 2002 Jun;6(2):102-8.

[8] Gu H, Perl Y, Halper M, Geller J, Kuo F, Cimino JJ. Partitioning an object-oriented terminology schema. Methods Inf Med. 2001 Jul;40(3):204-12.

[9] Bechhofer S, Horrocks I, Goble C, Robert S. OilEd: a Reason-able Ontology Editor for the Semantic Web. Proceedings of KI2001, Joint German/Austrian conference on Artificial Intelligence, September 19-21, Vienna. Springer-Verlag LNAI Vol. 2174, pp 396--408. 2001.

[10] Kumar A, Ciccarese P, Quaglini S, Stefanelli M, Caffi E, Boiocchi L. Relating UMLS semantic types and task-based ontology to computer-interpretable clinical practice guidelines. Proc MIE 2003. Accepted.

Table 1 Concepts proposed by Chen, Perl et al and Geller, Perl et al

Concepts	**Definition**
Cohesive partition	For a group of STs to be cohesive, it should have a unique root, i.e., one ST which all other STs in the group are descendants of.
Semantic Type	An abstract conceptual entity comprising the set of all semantic types with the exact

Group	same set of relationships.
Semantic Type Collection	Each semantic-type collection is an abstract conceptual entity representing a set of semantic types in the SN. Each will have a unique root and will thus be cohesive. Rule 1: Each semantic-type group with a nonleaf unique root becomes a semantic-type collection. Rule 2: If a leaf semantic type is in a Singleton in the structural partitioning, and its parent semantic-type group does not have a DNI root, then it is added to the semantic-type collection which contains its parent.

Table 2 UMLS Semantic Types and Definition most used in the Clinical Practice Guidelines

UMLS ST	**Definition**
Diagnostic Procedure	A method, procedure or technique used to determine the nature or identity of a disease or disorder. This excludes procedures which are primarily carried out on specimens in a laboratory.
Laboratory Procedure	A procedure method or technique used to determine the composition, quality, or concentration of a specimen, and which is carried out in a clinical laboratory. Included here are procedures which measure the times and rates of reactions.
Therapeutic or Preventive Procedure	A procedure method or technique designed to prevent a disease or a disorder, or to improve physical function, or used in the process of treating a disease or injury.

Table 3 Some of the UMLS Relations used as adjacent relations for Diagnostic Procedure and their Definition

UMLS Relations	**Definition**
affects	Produces a direct effect on. Implied here is the altering or influencing of an existing condition, state, situation, or entity. This includes has a role in, alters, influences, predisposes, catalyzes, stimulates, regulates, depresses, impedes, enhances, contributes to, leads to, and modifies.
analyzes	Studies or examines using established quantitative or qualitative methods.
assesses_effect_of	Analyzes the influence or consequences of the function or action of.
complicates	Causes to become more severe or complex or results in adverse effects.
diagnoses	Distinguishes or identifies the nature or characteristics of.
Evaluation_of	Judgment of the value or degree of some attribute or process.
measures	Ascertains or marks the dimensions, quantity, degree, or capacity of.
prevents	Stops, hinders or eliminates an action or condition.
result_of	The condition, product, or state occurring as a consequence, effect, or conclusion of an activity or process. This includes product of, effect of, sequel of, outcome of, culmination of, and completion of.
treats	Applies a remedy with the object of effecting a cure or managing a condition.
uses	Employs in the carrying out of some activity. This includes applies, utilizes, employs, and avails.

Table 4 Formalization of DOF (Determination OF)

DOF (Determination OF) is semantically equivalent to the set of the following UMLS semantic relations: {diagnoses, measures, analyzes, assesses_effect_of, evaluation_of}.
That is, { [DOF implies [(diagnoses) or (analyzes) or (measures) or (assesses_effect_of) or (evaluation_of)]} and {[(diagnoses) or (analyzes) or (measures) or (assesses_effect_of) or (evaluation_of)] implies DOF}
That is, [DOF → (diagnoses ∨ analyzes ∨ measures ∨ assesses_effect_of ∨ evaluation_of)] ∧ (diagnoses ∨ analyzes ∨ measures ∨ assesses_effect_of ∨ evaluation_of) → DOF]
That is,

[DOF ↔ (diagnoses ∨ analyzes ∨ measures ∨ assesses_effect_of ∨ evaluation_of)]
That is, DOF (Determination OF) ≡ {diagnoses, measures, analyzes, assesses_effect_of, evaluation_of}
That is, DOF ≡ SRC(STC), where SRC means "Semantic Relations Collection" associated to the STC, that is the set {diagnoses, measures, analyzes, assesses_effect_of, evaluation_of}.
That is, DP = DOF (STC) meaning: "diagnosis" is a "determination of" where "determines" = "is the Determination OF" = DOF $\forall x \in X \quad \exists y \in Y \quad y = f(x)$ $\forall aa \in AA \quad \exists dp \in DP \quad dp = DOF(aa)$ Meaning that: for each element aa ∈ STC-Anatomical Abnormality exists at least one Diagnostic Procedure which could determine it, where: DP = DOF(STC-Anatomical Abnormality) = DOF(anatomical abnormality) ∨ DOF(acquired abnormality) ∨ DOF(congenital abnormality)

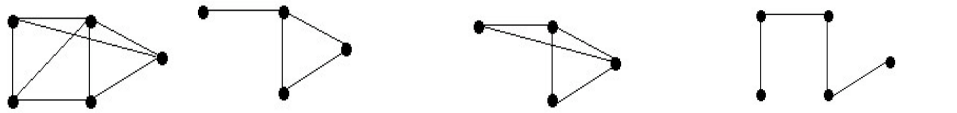
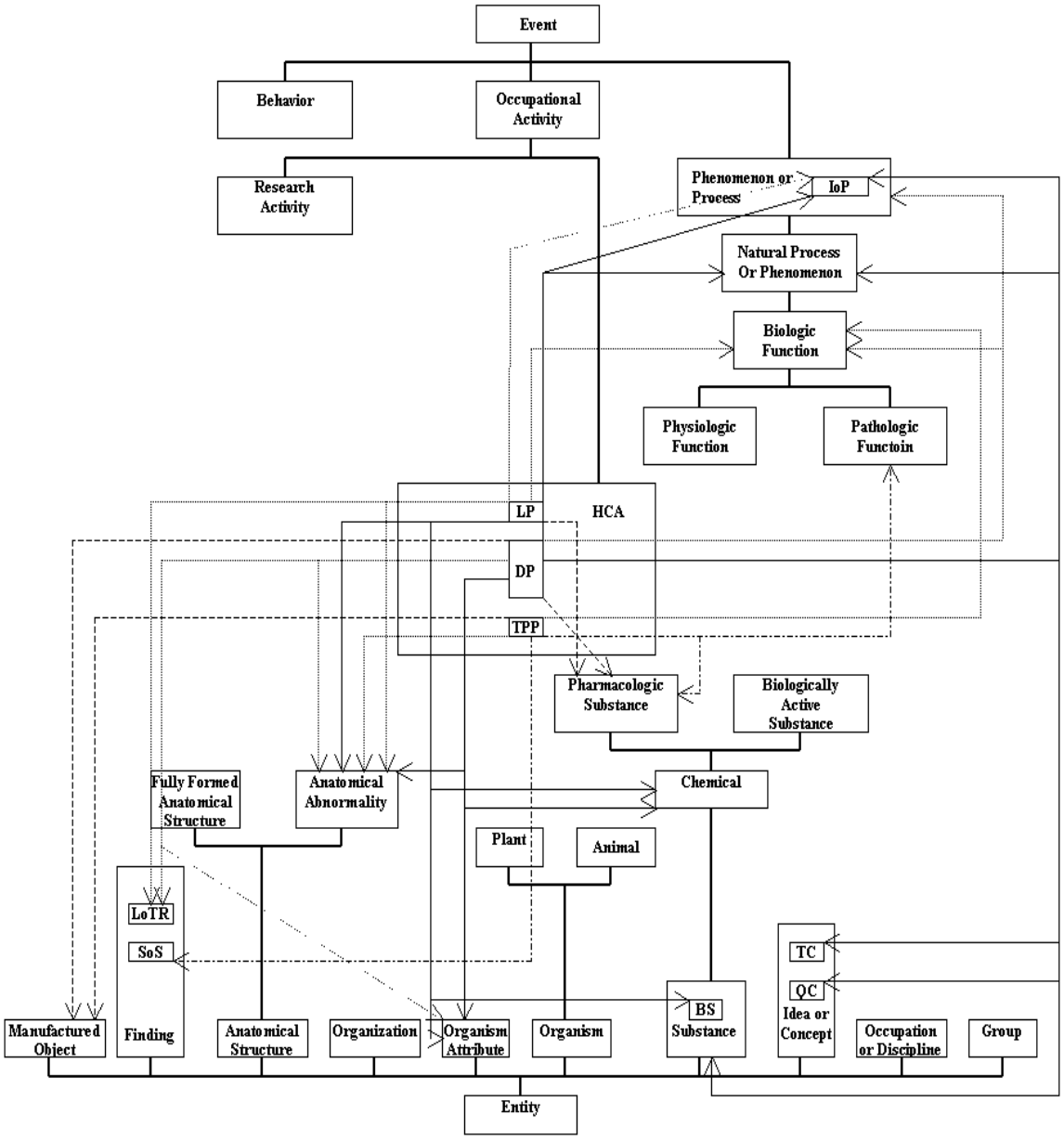


Fig 1. Representation of Graph, Subgraph, Induced Subgraph and Spanning Subgraph


Event
Behavior
Occupational Activity
Research Activity
Phenomenon or Process
IoP
Natural Process Or Phenomenon
Biologic Function
Physiologic Function
Pathologic Functoin
LP
HCA
DP
TPP
Pharmacologic Substance
Biologically Active Substance
Fully Formed Anatomical Structure
Anatomical Abnormality
Chemical
Plant
Animal
LoTR
SoS
TC
QC
BS
Manufactured Object
Finding
Anatomical Structure
Organization
Organism Attribute
Organism
Substance
Idea or Concept
Occupation or Discipline
Group
Entity