

Using the UMLS Semantic Network as a Basis for Constructing a Terminological Knowledge Base: A Preliminary Report

Giuseppe Carenini and Johanna D. Moore*
University of Pittsburgh, Department of Computer Science
Pittsburgh, PA 15260

Sharing and reuse of knowledge bases is recognized in Artificial Intelligence and Medical Informatics as beneficial, but difficult. Reusing an existing knowledge base can save considerable time and effort during the knowledge engineering phase, and facilitates integration of systems. However, the degree to which knowledge can be shared among different applications is still mainly an empirical question [1]. In this paper, we describe the preliminary results of our attempt to reuse the UMLS Semantic Network [2, 3] as an ontology for the knowledge base of a patient education system.

INTRODUCTION

We are involved in a research effort whose goal is to improve patient compliance by better educating patients about their disease, possible therapies, and medications [4]. Initially, we are focusing on migraine patients who require periodic interaction with their physicians for effective management of their condition. To determine the types of information patients desire and the types of explanations that are most effective, we are basing the design of the system on extensive use of empirical data, including ethnographic studies of doctor-patient interactions [5], interviews with patients, and interviews with physicians. The patient education system provides a tailored, interactive patient handout describing important aspects of the individual patient's care, including information about the diagnosis, the therapy plan, and the potential side effects of medication. The handout and all answers to subsequent patient queries about it are automatically generated using an explanation planner that tailors all of its utterances to the individual patient and the context created by prior explanations [6]. Information about individual patients is collected by a history-taking program that gathers extensive knowledge of their medical history in electronic form.

To support the generation of explanations, the system must have extensive knowledge pertaining to

the particular disease, symptoms, therapies, and drugs it explains, but must also have a wide array of general medical knowledge. This knowledge must be represented in a declarative form where the meanings of terms and relations between them are well defined. These representational capabilities are provided by terminological subsumption languages, which have proven useful in many systems intended to provide explanations [7]. Since constructing such a knowledge base is a huge task, we wished to make use of previous and contemporary research efforts aimed at representing medical knowledge.

Musen [1] identified several aspects of knowledge that are sharable, including: lexicons, ontologies, inference syntax, tasks and problem-solving methods. For our application, we required a lexicon, i.e., the terms used for referring to entities in the domain, and an ontology, i.e., the structure of relationships among the entities to which the terms refer. We decided to make use of the UMLS Semantic Network because it provides extensive coverage of medical knowledge. While UMLS Net was not originally designed as a reusable ontology and lexicon, we felt that it could be adapted for our purposes because it provides a set of general medical concepts and the relationships between them.

Our goal was to exploit the extensive coverage of the UMLS Net as a basis for acquiring much of the general medical knowledge needed by our system. To do this, we must re-represent UMLS concepts and relations in the knowledge representation language used by our system. Once we have done this, we can subordinate the knowledge specific to migraine and its treatment to the generic concepts acquired from UMLS. In this paper we focus on the acquisition process, and the modifications that were required in order to use knowledge expressed in the UMLS Net as a basis for building a knowledge base in a sophisticated terminological language, LOOM [8]. Because the process of translating a knowledge source from one format to another is tedious and error prone, and we expect new versions of the UMLS Net to be released pe-

* The research described in this paper was supported by Grant No. R01 LM05299-01 from the National Library of Medicine of the National Institutes of Health

riodically, we wished to automate as much of the acquisition process as possible.

We believe that the results of this effort are useful for two reasons. First, the extensive medical knowledge provided in the UMLS is represented in a form more readily usable by others building medical knowledge based systems. This is because knowledge representation systems such as Loom provide an array of inferencing capabilities, access functions, and development tools that facilitate the construction, use and evolution of large knowledge bases. Second, the modifications that are required in order to express medical knowledge using the representational constructs of a language with a model-theoretic semantics may be of use in further development of UMLS.

TRS's IN MEDICINE

A terminological representation system (TRS) is a type of frame system in which the meanings of concepts (frames) and relations between them are unambiguously determined by explicit notational devices that have a well-defined semantics (see [9] for a model-theoretic semantics of a generic TRS.) Based on these semantics, a TRS can compute several useful relationships between structured concepts automatically based solely on the definitions of those concepts. For example, concept c_1 is said to *subsume* concept c_2 iff the set denoted by c_1 includes the set denoted by c_2 . Subsumption of structured concepts is the basis for the fundamental terminological inference of *classification*, which allows a TRS to determine the proper position of a new concept in a preexisting taxonomy. Similarly, a TRS can position a new instance under all of the concepts to which it belongs (the *recognition* inference).*

Previous efforts to represent medical knowledge in a TRS met with difficulties, mainly due to the fact that TRS's developed prior to Loom were dominated by the "dogma" [10] that the expressiveness of the language should be restricted in order to assure the tractability of terminological inferences. For this reason, the expressive power of TRS's was severely limited and turned out to be insufficient for representing medical knowledge [11, 12]. In particular, there were several representational features that were considered necessary but were not provided, including: distinction between definitional and non-definitional properties, number intervals as role restrictions, use of disjoint covers by

* For a detailed description of these inference procedures and their usefulness, see [9].

the classifier and the representation of sequences. We do not expect to encounter the same sort of difficulties because Loom's implementors have made a concentrated effort to overcome such limitations and Loom includes several of the representational features that were missing from earlier TRS's. The philosophy of Loom is to provide a maximally expressive language, leaving to the users the burden of dealing with the intractability of some terminological inferences.

EXPLOITING THE UMLS NET

Initially, we desired to implement a fully automatic procedure for using UMLS Net to construct a Loom knowledge base, in order to guarantee correctness and to minimize the effort involved in updating our knowledge base when new releases of the UMLS Net became available. Unfortunately the current format of the UMLS Net does not allow such an automatic process. Here we examine the reasons why this process cannot be fully automated, and present a semi-automatic method for acquiring knowledge from UMLS Net, indicating points where intervention by the knowledge engineer (KE) is required.

Problems with Automated Acquisition

In the UMLS Net, relations are specified by two kinds of information: a super-relation and the pairs of concepts that the relation can relate. For instance, the specification for the relation "causes" is:

Super-relation: functionally-related-to
Concept pairs related:
[FUNGUS—PATHOLOGIC FUNCTION]
[VIRUS—PATHOLOGIC FUNCTION]
[RICKETTSIA OR CHLAMYDIA—PATHOLOGIC FUNCTION]
[BACTERIUM—PATHOLOGIC FUNCTION]
[INVERTEBRATE—PATHOLOGIC FUNCTION]
[SUBSTANCE—PATHOLOGIC FUNCTION];
[SUBSTANCE—CONGENITAL ABNORMALITY]
[SUBSTANCE—ACQUIRED ABNORMALITY]
[SUBSTANCE—INJURY OR POISONING]
[MANUFACTURED OBJECT—PATHOLOGIC FUNCTION]
[MANUFACTURED OBJECT—CONGENITAL ABNORMALITY]
[MANUFACTURED OBJECT—ACQUIRED ABNORMALITY]
[MANUFACTURED OBJECT—INJURY OR POISONING]

This representation is incompatible with terminological formalisms where a relation is logically interpreted as a binary predicate that can relate only instances of two concepts (called the domain and the range of the relation) or pairs of concepts that are subsumed by them. Thus, there is no syntactic mapping between UMLS relations and those required by a TRS, and consequently no fully automatic acquisition process is feasible.

The Acquisition Method

In the method presented in Figure 1, the following terminology is used: *REL* is the set of all the relations in the UMLS Net. As noted above, a UMLS

Automatic Phase of Acquisition Algorithm: *Acquisition of Loom terms from relations expressed in the UMLS unit record format [2].*

For all $rel \in REL$

(A) For all $D_j \in \overline{D_{rel}}$ with $\overline{R_j} = \bigcup \{R_i \mid (D_j, R_i) \in \overline{P_{rel}}\}$; (in Loom ($: or R_1 \dots R_N$))

(B) Build new pairs $P_j = (D_j, \overline{R_j})$ and collect them as a new $\overline{P_{rel}}$. ;(see the new $\overline{P_{rel}}$ for the relation “causes”)

Non-Automatic (Aided) Phase of Acquisition Algorithm:

(C) Provide a name for each of the $\overline{R_j}$ that are medically meaningful.

(D) Express the relationship between the pairs $(D_j, \overline{R_j})$ in a form compatible with the semantics of Loom:

(1) Whenever possible find or define a meaningful concept, SD , that subsumes the maximum number of D_j and a meaningful concept, SR , that subsumes the corresponding $\overline{R_j}$.

Define $rel \mid \text{domain}(rel) = SD$ and $\text{range}(rel) = SR$

(2) Whenever possible restrict the value of rel in the definition of subconcepts of SD ($subSDs$) by means of subconcepts of SR ($subSRs$) according to the relationships expressed by the P_j .

(3) For any D_j not subsumed by SD show how the corresponding P_j is represented in the knowledge base in an implicit way.

Figure 1: The automated acquisition method

relation, rel , is specified by its super-relation and by $\overline{P_{rel}}$: a set of pairs $P_i = (D_i, R_i)$ identifying the concepts that it can relate. $\overline{D_{rel}}$ is the set of distinct D_i .

The automatic phase of the algorithm should be self-explanatory. We call the second phase of the method “aided” because we have implemented an interface that facilitates the KE in steps C and D. This system allows KEs to easily access information about any rel that has been collected during the automatic phase. For instance, they can ask questions such as: “What is the $\overline{D_{rel}}$ of a relation?” “What is the $\overline{R_j}$ corresponding to a particular $D_j \in \overline{D_{rel}}$?” “What is the superconcept of any subset of $\overline{D_{rel}}$?” etc. The system (see Figure 2) has been integrated with a graphical interface in which all entities that appear on the screen are mouse sensitive. By clicking on entities, the user can obtain their definition (textual or graphical), their documentation and the graphical display of related ISA hierarchies, etc. This increases the KE’s ability to examine and understand the system’s response in the context of the UMLS ISA hierarchy.

Note that in the “aided” phase, the role of the user is fundamental. In step C, the KE must pick meaningful names for the $\overline{R_j}$, i.e, the union of the ranges of the relation rel for a D_j . Moreover, any execution of step D on a new relation may present new representational problems requiring a solution that depends on the semantics of the particular

relation.

TWO DETAILED EXAMPLES

To make the acquisition process clearer, here we work through its application on two UMLS relations.

The relation “causes”: The UMLS definition of the “causes” relation was shown above. The result of the application of the automatic phase of the method to this definition is:

Pairs of relates $(D_j, \overline{R_j})$:

- [FUNGUS—PATHOLOGIC FUNCTION]
- [VIRUS—PATHOLOGIC FUNCTION]
- [RICKETTSIA OR CHLAMYDIA—
PATHOLOGIC FUNCTION]
- [BACTERIUM—PATHOLOGIC FUNCTION]
- [INVERTEBRATE—PATHOLOGIC FUNCTION]
- [SUBSTANCE—PATHOLOGIC FUNCTION,
CONGENITAL ABNORMALITY,
ACQUIRED ABNORMALITY,
INJURY OR POISONING]
- [MANUFACTURED OBJECT—
PATHOLOGIC FUNCTION,
CONGENITAL ABNORMALITY,
ACQUIRED ABNORMALITY,
INJURY OR POISONING]

In step C, the KE recognizes the concept [Pathologic Function, Congenital Abnormality, Acquired Abnormality, Injury or Poisoning] (one of the $\overline{R_j}$) as a medically meaningful concept and provides a name for it: PROCESS-OR-PHENOMENON-REQUIRING-MEDICAL-ATTENTION (PPRMA). Then in step D, the superconcepts SD and SR must be selected. Using the interface to query the UMLS knowledge base, the KE determines that PHYSICAL-OBJECT and PROCESS-OR-PHENOMENON subsume all the D_j and all the $\overline{R_j}$, respectively. Moreover they seem to express adequately the notion of

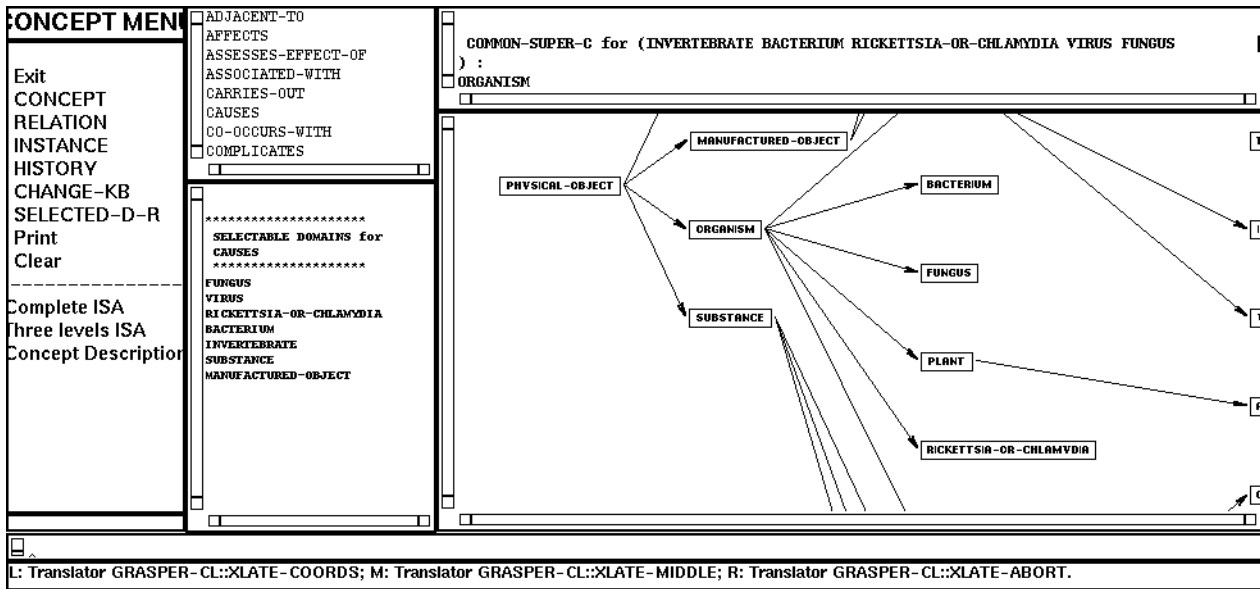


Figure 2: Interface for the knowledge engineer

“cause” expressed by the P_j (i.e., they are neither too general nor too specific). So, PHYSICAL-OBJECT is selected as SD and PROCESS-OR-PHENOMENON is selected as SR . Note that although PPRMA subsumes all of the \bar{R}_j , it is not selected as SR for two reasons: first, it is not in the UMLs Net, and second it seems to be too restrictive for a relation as general as “causes”.

Now the KE looks for meaningful abstractions of the D_j and the \bar{R}_j (substep D2). In this case, all the \bar{R}_j are subsumed by PPRMA. So it is reasonable to define a subconcept of SD that has the relation “causes” restricted to PPRMA. Such a concept can be named PATHOLOGIC-AGENT. Second, considering the D_j , there is a distinction between D_j that are organisms and D_j that are not. The former have “causes” restricted to PATHOLOGIC-FUNCTION, whereas the latter have “causes” restricted to PPRMA. Thus the KE can define a new concept, LIVING-PATHOLOGIC-AGENT, that has two parent concepts, ORGANISM and PATHOLOGIC-AGENT, and where “causes” is restricted to PATHOLOGIC-FUNCTION.

We claim that the definition of these two new concepts (i.e., PATHOLOGIC-AGENT and LIVING-PATHOLOGIC-AGENT) closely follows what is represented in the UMLs Net and is in some ways more precise. For example, consider the concept BACTERIUM. In the original UMLs representation for “causes”, BACTERIUM was paired with PATHOLOGIC-FUNCTION, indicating, in straight terminological interpretation, that if a BACTERIUM participates in a “causes” relation, it causes a PATHOLOGIC-FUNCTION. But

note that it is somewhat imprecise to say that all bacteria “cause” a PATHOLOGIC-FUNCTION. There are in fact bacteria that aid BIOLOGICAL-FUNCTION. Therefore, in our terminological representation, we should not restrict “causes” for BACTERIUM to PATHOLOGIC-FUNCTION. However, whenever we have a particular BACTERIUM that is known to cause a PATHOLOGIC-FUNCTION, we would like the system to recognize it as a LIVING-PATHOLOGIC-AGENT (i.e., an ORGANISM that is also a PATHOLOGIC-AGENT).

The relation “diagnoses”: The result of the application of the automatic part of the method to “diagnoses” is the following:

- Pairs of relates (D_j, \bar{R}_j):
- [DIAGNOSTIC PROCEDURE—
PATHOLOGIC FUNCTION,
CONGENITAL ABNORMALITY,
ACQUIRED ABNORMALITY,
INJURY OR POISONING]
 - [LABORATORY PROCEDURE—
PATHOLOGIC FUNCTION,
CONGENITAL ABNORMALITY,
ACQUIRED ABNORMALITY,
INJURY OR POISONING]
 - [SIGN—
PATHOLOGIC FUNCTION,
CONGENITAL ABNORMALITY,
ACQUIRED ABNORMALITY,
INJURY OR POISONING]
 - [SYMPTOM—
PATHOLOGIC FUNCTION,
CONGENITAL ABNORMALITY,
ACQUIRED ABNORMALITY,
INJURY OR POISONING]
 - [PHARMACOLOGIC SUBSTANCE—
PATHOLOGIC FUNCTION]
 - [PROFESSIONAL OR OCCUPATIONAL GROUP—
PATHOLOGIC FUNCTION]

Note that some compound \bar{R}_j correspond to PPRMA, a concept defined in the previous example. Thus we are done with step C. In step D, the only SD

that seems to appropriately cover all the D_j is the concept ANYTHING. But ANYTHING is too general to be a domain for the “diagnoses” relation. Therefore, closer examination of the semantics implied by the $\overline{P_j}$ is necessary. In this case, we propose DIAGNOSTIC-PROCEDURE as SD for “diagnoses” and PPRMA as its SR . Now it is necessary to show how the $\overline{P_j}$ whose D_j are not subsumed by DIAGNOSTIC-PROCEDURE are implicitly represented in the knowledge base (substep D3). The concept DIAGNOSTIC-PROCEDURE might have, among others, two relations “has-agent” and “has-result” (the inverse of the UMLS relations “carries-out” and “result-of” respectively). It is consistent with the UMLS Net to restrict the “has-agent” relation to the concept PROFESSIONAL-OR-OCCUPATIONAL-GROUP, and the “has-result” relation to the concept FINDING. Such a definition of DIAGNOSTIC-PROCEDURE would implicitly represent the $\overline{P_j}$ whose D_j are: PROFESSIONAL-OR-OCCUPATIONAL-GROUP, SIGN and SYMPTOM. For example, a PROFESSIONAL-OR-OCCUPATIONAL-GROUP implicitly “diagnoses” a PPRMA because it is the agent who performs the DIAGNOSTIC-PROCEDURE. SIGN and SYMPTOM implicitly “diagnose” a PPRMA because they are the results (i.e., subconcepts of FINDING) of the DIAGNOSTIC-PROCEDURE. Similar solutions can be found for LABORATORY PROCEDURE and PHARMACOLOGIC SUBSTANCE, but space limitations preclude a detailed discussion here.

RESULTS AND FUTURE WORK

We have successfully applied the acquisition method to 10 UMLS relations. We have found that representing a relation in LOOM often introduces new concepts that may be useful in the definition of other relations. Thus we expect the acquisition process to be iterative. At the outset, we believed we could automate much more of the process of exploiting the UMLS Net, and that a more direct mapping between the two representations existed. However, we still believe that the use of the UMLS Net is shortening the time taken to build a medical knowledge base with extensive coverage and to ensure its correctness.

There are still several important issues requiring further work. First, in the UMLS Net, the inheritance of a restriction of a relation can be blocked for any subconcept [2]. We have not yet considered how the acquisition algorithm can be modified to assist the KE in distinguishing between necessary, sufficient and default conditions in the definition of a concept. Second, some UMLS relations express

knowledge related to actions (e.g., diagnose). An alternative approach we have begun to explore is to represent this knowledge explicitly as concepts that correspond to actions. Finally, once we finish the acquisition process we will have to place our migraine-specific knowledge under the general knowledge acquired from the UMLS Net. As in [13], we expect that this process will require us to add new concepts and new relations to the Net.

Reference

- [1] M. A. Musen. Dimensions of knowledge sharing and reuse. *Computers and Biomedical Research*, 25:435–467, 1992.
- [2] National Library of Medicine. *UMLS Knowledge Sources - 3rd Experimental Edition*, 1992.
- [3] A. T. McCray and William T. Hole. The scope and structure of the first version of the UMLS semantic network. In *Proceedings of the 14th SCAMC*, pages 126–130, Washington, DC, 1990.
- [4] J. D. Moore and S. Ohlsson. Educating patients through on-line generation of medical explanations. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, Bloomington, Indiana, August 1992.
- [5] D. E. Forsythe. Using ethnography to build a working system: Re-thinking our basic design assumptions. In *Proceedings of the 16th SCAMC*, Baltimore, MD, 1992.
- [6] G. Carenini and J. D. Moore. Generating explanations in context. In *Proceedings of the International Workshop on Intelligent User Interfaces*, pages 175–182, Orlando, Florida, January 4-7 1993. ACM Press.
- [7] K. R. McKeown and W. R. Swartout. Language generation and explanation. *Annual Review of Computer Science*, 2:401–449, 1987.
- [8] ISX Corporation. *Loom User's Guide*, 1991.
- [9] W. A. Woods and J. G. Schmolze. The KL-ONE family. *Computers and Mathematics with Applications*, 23(2-5):133–177, 1992.
- [10] J. Doyle and R. S. Patil. Two theses of knowledge representation: language restrictions, taxonomic classification, and the utility of representation services. *Artificial Intelligence*, 48(3):261–298, 1991.
- [11] I. J. Haimovitz, R. S. Patil, and P. Szolovits. Representing medical knowledge in a terminological language is difficult. In *Proceedings of the 12th SCAMC*, pages 71–75, 1988.
- [12] Y. Jang and R. S. Patil. KOLA: A knowledge organization language. In *Proceedings of the 13th SCAMC*, pages 71–75, 1989.
- [13] C. Friedman. The UMLS coverage of clinical radiology. In *Proceedings of the 16th SCAMC*, pages 309–313, Baltimore, MD, 1992.