A UMLS-based Knowledge Acquisition Tool for Rule-based Clinical Decision Support System Development

SOUMEYA L. ACHOUR, MS, DVM, MICHEL DOJAT, ENG, PHD, CLAIRE RIEUX, MD, PHILIPPE BIERLING, MD, PhD, ERIC LEPAGE, MD, PhD

Abstract Decision support systems in the medical field have to be easily modified by medical experts themselves. The authors have designed a knowledge acquisition tool to facilitate the creation and maintenance of a knowledge base by the domain expert and its sharing and reuse by other institutions. The Unified Medical Language System (UMLS) contains the domain entities and constitutes the relations repository from which the expert builds, through a specific browser, the explicit domain ontology. The expert is then guided in creating the knowledge base according to the pre-established domain ontology and condition-action rule templates that are well adapted to several clinical decision-making processes. Corresponding medical logic modules are eventually generated. The application of this knowledge acquisition tool to the construction of a decision support system in blood transfusion demonstrates the value of such a pragmatic methodology for the design of rule-based clinical systems that rely on the highly progressive knowledge embedded in hospital information systems.


Clinical decision support systems (CDSSs) have been shown to be very helpful to medical practitioners. Knowledge acquisition and modeling play leading roles in the development of such knowledge-based systems. Unfortunately, two key factors limit the development of CDSSs and their integration into hospital information systems—the time and effort required by medical experts to create and maintain knowledge bases, and the extreme difficulty of sharing and reusing the validated knowledge bases because of their idiosyncrasies and lack of clarity for users outside the originating institution.

The latter factor is considered the main obstacle to daily use of CDSSs. The goal of our work is the design of a knowledge acquisition tool to facilitate the creation and maintenance of a knowledge base by the medical expert, and its subsequent sharing and reuse by other medical institutions. Reuse may be approached from several directions; in particular:

- Through generic abstractions. Lexicons, ontologies, and problem-solving methods are developed and then adapted to fit the specific needs of a typical medical domain application.

- Through the standardization of the knowledge representation. Several standards have been proposed, such as KIF, Ontolingua, and Arden Syntax.

- Through specific models, such as GLIF or Proforma, to represent certain kinds of clinical guidelines and support their dissemination.

Our project relies on the first two approaches.
medical reasoning process and medical expertise. The research project GAMES\textsuperscript{11} views the knowledge acquisition process as the construction of two models—the epistemological model, entailing the knowledge required to perform a particular task, and the computational model, containing data structures and algorithms designed to allow the computerized execution of that task. The PROTÉGÉ\textsuperscript{12} environment allows developers to configure available problem-solving methods that are mapped to domain ontologies and generates task-specific knowledge acquisition tools. In parallel to research into architectures for knowledge acquisition, considerable work is under way to create medical ontologies, a key component in building and reusing knowledge-based systems. For instance, the representation and integration language GRAIL, from the GALEN project, is specially designed to support models of medical terminology and achieve reuse of taxonomies.\textsuperscript{4,13}

The Unified Medical Language System (UMLS) constitutes a knowledge corpus that is useful for building knowledge bases. Our approach, less general than that of PROTÉGÉ, is designed for medical experts without specific training in medical informatics. It allows them to configure rule-based templates that are mapped to the UMLS-based domain-specific lexicon that they have previously constructed. Medical logic modules (MLMs) written in Arden Syntax are eventually generated to facilitate reuse and sharing of this knowledge base.

Medical logic modules have been used to generate clinical alerts, interpretations, and diagnoses.\textsuperscript{14–18} For several years, numerous attempts have been made to facilitate the creation of MLMs with respect to the Arden Syntax. Interesting clinical experiences with MLM editors have been reported.\textsuperscript{19–21}

Our project differs from previous work in that it integrates the generation of MLMs into the general process of knowledge acquisition (ontology and knowledge base). In this article we report on our knowledge acquisition methodology; the use of this methodology to build a CDSS applied to blood transfusion; and a preliminary test of our system.

In blood transfusion, the evolution of both medical knowledge and government regulations imposes a continuous adaptation of the CDSS that wholly justifies the use of our approach. Several CDSS have been proposed in this field.\textsuperscript{22,23} Like all classical rule-based systems, these systems were difficult to maintain and extend. No specific knowledge acquisition technique was used for their development. One application embedded in the HELP system concerns a computerized monitor blood ordering system based on a critiquing approach.\textsuperscript{24–26} A set of knowledge tools is proposed to the expert for the definition and the maintenance of the knowledge base, and a local data dictionary can be used. All these CDSSs in blood transfusion turn essentially on the question of whether to transfuse. In addition to this issue, we studied the choice of the qualifier (e.g., phenotyped or irradiated) of the product to be transfused. Our CDSS is currently being tested at the Henri Mondor Hospital in Créteil, France.

**Methodology**

Our methodology for knowledge acquisition and representation relies on two steps. First, a domain ontology is developed—the expert selects the entities and relations present in UMLS that are relevant to constructing the domain ontology, and introduces new terms and relations if they are needed. Second, the domain knowledge is constructed—according to the pre-established domain ontology, the expert is guided in creating the knowledge base via condition–action rule templates. Corresponding MLMs are eventually generated automatically. Figure 1 represents our methodology.

**Construction of the Domain Ontology Using UMLS**

To facilitate sharing and reuse, we have chosen UMLS, elaborated to facilitate the integration of information from multiple biomedical sources, to constitute the repository of entities and relations from which the expert builds the domain ontology. Three components constitute the UMLS. The Metathesaurus contains a collection of biomedical concepts and their inter-relations. The Semantic Network, through its high-level semantic types, provides a consistent categorization of all concepts represented in the Metathesaurus. The Specialist Lexicon is an English-language lexicon with biomedical terms. The lexicon entries for each word or term record syntactic, morphologic, and orthographic information. The UMLS provides a shared source of terminology widely used in clinical applications.\textsuperscript{27–30}

To navigate between concepts through the concept hierarchy and through the semantic network, we have designed a specific navigator.\textsuperscript{31}

The UMLS Navigator

The conceptual model of our navigator (Figure 2) is based on two components of the UMLS—the Metathesaurus and the Semantic Network. We have adopt-
ed an object-oriented representation. Concept is the main class of our Metathesaurus model. It includes several attributes (for example, ConceptDefinition, ConceptLinks). It reflects properties present in UMLS; each concept has a unique identifier. Different terms with the same meaning are linked to the same concept identifier. We have reified the link (LinkConceptToVariants) between a concept and these terms (preferred term, synonyms, etc). The relation (ConceptLink) between two concepts has three parameters represented as classes: a ConceptRelation (e.g., parent, child, general, specific), a ConceptRelationAttribut (e.g., isa, inverse_isa), and an InformationSource (e.g., Mesh98) that provides it. Our representation of the UMLS Semantic Network is then a reified graph whose edges are the semantic relations (SemanticLinks) such as affected_by and occurs_in, and whose vertices are the semantic types (SemanticType) or semantic relations (SemanticRelation).

Domain Ontology Construction
The construction of the domain ontology requires four steps. First, the expert identifies the medical terms used in the specific domain. Using our specific browser (Figure 3), the expert looks in the UMLS Metathesaurus for the corresponding concepts, selects relevant concepts, and adds them to create a list. To facilitate comprehension and enrich the list, the expert then identifies for each concept its ascending hierarchy (i.e., its parents). Depending on the source vocabulary in the Metathesaurus, a concept can be found at several levels in a specific ascending hierarchy. Figure 4 shows the different possible ascending hierarchies for the term hemolytic autoimmune anemia. The expert selects the hierarchy that corresponds to the semantic meaning that he or she associates with a specific concept, or constructs a hierarchy by choosing the concepts from the different sources. Because strong standardization is not

---

**Figure 1** Knowledge acquisition process and decision support system design. Top left, The domain ontology is built via the UMLS browser. Top right, The knowledge acquisition interface allows the instantiation of a general problem-solving method using the domain ontology. Medical logic modules (MLMs) are automatically created and organized by the knowledge manager. Bottom, Example of a clinical decision support system (CDSS) architecture in which the knowledge base is generated using our methodology. The arrival of new data (1) triggers the event procedure (2). The relevant medical logic modules (MLMs) (3), which represent the knowledge base, are then executed by the inference engine (4). A results table containing a summary of the diagnosis and the order for blood products is provided (5). Initial data are sent automatically by the hospital information network or are entered manually by a physician.
always possible in medicine, a medical expert may
insert new terms or extend the initial hierarchy cor-
pus to take into account particular considerations.32
For term insertion, two constraints must be respect-
ed—the new concept must be linked to a semantic
type from the UMLS Semantic Network and must
have a parent. If a specific classification is construct-
ed, it is saved in the UMLS format (MRCXT table)
with a specific tag (PERSO) in the raw source (SAB).
All the concepts extracted from the UMLS or created
by the expert have a local identifier and are stored in
a local table. This local table allows the link with the
clinical database of the hospital. When introducing
new terms and classification, the expert has the
responsibility for checking their adequacy and their
consistency with the set of information contained in
PERSO. The expert also assigns the semantic types
from the Semantic Network for any new concept
added. As semantic types are general, an intermedi-
ate layer, named categories, has been added to intro-
duce more specific types. The categories correspond
to the concepts located at the top of the hierarchy of
the contextual list selected from those present in the
Metathesaurus or created by the expert. For instance,
Anemia Sickle Cell is initially linked to the Diseases or Syndrome semantic type. We have introduced the category Hematologic Diseases as a subclass of Disease or Syndrome semantic type, to which Anemia Sickle Cell is linked (Figure 5).

In the fourth and final step, following the classification of the terms according to an ascending hierarchy, all possible relations among these terms are automatically determined. For this purpose, a query is performed on the UMLS Semantic Network, to extract relationships between semantic types and concepts useful for representing the expert’s knowledge. Then, the expert selects the relevant relationships between concepts present in the Metathesaurus. Once again, the expert may add new relationships to this contextual list (identified by the label PERSO).

During the ontology design, a local identifier is automatically created for each acquired concept. This identifier is stored in the slot data of the MLM during the knowledge acquisition process. A local table is then designed to link and translate these local identifiers to the clinical database identifiers.

From the development of the domain ontology, three different structures are obtained—the hierarchy of terms and concepts used in the domain, which have an ascending hierarchic context, connected by the relation “inverse_isa”; the concepts located at the top of the hierarchy that have been selected, which will compose the list of categories that subsume the medical terms used; and the UMLS Semantic Network excerpt covering the concepts selected for the domain, such that each term has a semantic type and a category.

For the collection of terms present ($n = 120$), in the blood transfusion domain, the medical expert found corresponding UMLS Metathesaurus concepts for 88 (73 percent). The blood transfusion domain used high-
ly specialized terms, such as “transfusion of red blood cell plasma depleted,” “transfusion of red blood cells phenotyped,” and “patient erythrocyte alloimmunized,” which were not present in the version of the Metathesaurus used. The coverage of disease terms qualified by such adjectives as minor, severe, or chronic was uneven.

Domain Knowledge Construction

Starting from the predefined domain ontology, the expert creates the domain knowledge needed by the application. We have modeled the knowledge with production rule templates. Our knowledge acquisition tool guides the user through three steps of rule
description: documentation following the Arden Syntax definitions, specification of the condition part of each rule, and specification of the action part of each rule. The entities that compose the conditions and action parts of the rules come from the domain ontology and are linked to the semantic types.

To illustrate the process of rule specification, let us start with the rule template (Figure 6)

\[
(A_1) \rightarrow (A_3)
\]

After selecting \(A_1\) for the condition part, the expert can choose between \(b_α\) and \(b_β\). The selection of \(b_α\) leads to three possible choices—\(b_1\), \(b_2\), or \(b_3\). The expert chooses the concepts \(b_1\) and \(b_3\) for the condition part and then chooses \(A_3\), \(b_δ\), and \(b_7\) for the action part, leading to Rule 1:

\[
\text{IF condition (}b_1\text{)} \text{ present AND condition (}b_3\text{)} \text{ present THEN action (}b_7\text{)} \text{ activated (Rule 1)}
\]

Generation of Medical Logic Modules

An MLM, equivalent to a single rule in a rule-based expert system, contains enough medical knowledge and data to make a single clinical decision. Medical logic modules can be independent of one another but also can be linked by CALL statements. An MLM is an ASCII file composed of slots grouped into three categories—Maintenance, Library, and Knowledge. The category parts of the MLM are acquired via the knowledge acquisition tool. Figure 7 shows the interface for the construction of Rule 2.

To facilitate maintenance and avoid redundancy or inconsistency, MLMs are organized depending on the concepts they manipulate. The name of the MLM is stored in the slot filename of the category Maintenance, and the concepts used in this MLM (e.g., \(b_1\), \(b_3\), and \(b_7\) for Rule 1) are stored in the slot keywords of the category Library. The concepts appearing in the condition part are linked to the corresponding MLMs by the relation “has as rule” to facilitate the retrieval of specific MLMs. The goal of each MLM, represented by its action, is stored in the slot purpose of the category Library. The concepts appearing in the action part are linked to the corresponding MLMs by the relation “the action of.” This allows the expert to retrieve easily all MLMs that trigger a specific action. Finally, this knowledge is formalized according to the Arden Syntax with MLMs linked to the domain ontology.

One hundred ten rules (translated in MLMs) make up the knowledge base. The indexing and the referencing of the terms that compose each MLM are used to identify possible redundancies and inconsistencies during the creation of a new MLM. In the current prototype, the checking for redundancy and inconsistency is performed manually. Inheritance property is not used during MLM construction.

Experience with a Clinical Decision Support System for Blood Transfusion

The transfusion of blood products is an inescapable therapy for certain pathologies but has the potential to induce undesirable immunologic and infectious effects, like HIV or hepatitis. The best guarantee of a patient’s safety is strict adherence to transfusion guidelines. In addition, the transfusion of blood products is an expensive therapy that, in terms of public health, must be prescribed advisedly. A CDSS that assists in the prescription of blood products
seems to be an appropriate tool for enhancing patient safety and reducing health care costs. However, because of continual expansion of medical knowledge about blood transfusion and changes in government regulations, decision rules for blood transfusion must be frequently updated.

At the Henri Mondor Hospital in Créteil, France, approximately 30,000 blood products are distributed to 5,000 patients each year. This volume justifies the development of a CDSS integrated into the hospital information network to assist clinicians in blood product prescriptions. We constructed a CDSS for blood transfusion using the methodology previously described.

Architecture of the Clinical Decision Support System
Starting from blood data (essentially hemoglobin and platelet concentrations) and the patient’s state (disease, current therapy, etc.), the system indicates whether transfusion is required and, if so, which type and quantity of product to transfuse. The system works according to the data-driven and application-driven principles of the MLM controller that schedules MLM execution. The CDSS is composed of the three modules shown in Figure 1: the user interface to retrieve information from the patient database, the knowledge base that contains expertise for blood transfusion in MLMs, and the inference engine that executes the relevant MLMs. Two modes are available—alert (data driven) and consultation (application driven).

Alert
As shown in Figure 1, when biological data are automatically sent into the patient database (1), the event procedure is triggered (2). This activates several MLMs (3) executed by the inference engine (4). Actions are performed, and a table of results is generated (5). Depending on the seriousness of the diagnosis, an alarm signal could be set.

Consultation
The physician keys in data about the patient and the prescription, specifying the type and quantity of products to be transfused. Then the system determines whether the transfusion is appropriate and the type, qualifier, and quantity of the product to be transfused. Details of system decisions are displayed to the physician. If the physician disagrees, he or she can order other products and justify the choices in a brief report. A complete report that includes the recommendations of both the system and the physician is sent to the blood bank.

Preliminary Results
A first evaluation was performed to validate the coherence and the correctness of the CDSS. For that, 30 orders were studied, and the reliability of transfusion order was determined (98 percent for platelet and 97 percent for red cell transfusion). A discrepancy about the qualifiers of the transfusion product—such as phenotyped or compatibilized product—was observed between the system and the physicians in 18 of 30 cases (60 percent).

Analysis of these discrepancies revealed a need to add more information about the patient’s status and to modify rules in the knowledge base. Twenty new orders were then evaluated using the modified CDSS. Some disagreements persisted between physician’s orders and CDSS conclusions (40 percent), but agreement between the CDSS and the experts who reviewed the orders was 100 percent. This result was explained by the fact that physicians did not follow the current transfusion guidelines. A working committee was appointed to modify current practice to adhere to new transfusion guidelines. The prototype is currently being evaluated at the Henri Mondor Hospital.

Discussion
We argue that our approach is a practical way to build CDSSs. It avoids constructing the domain specific ontology from scratch, limits the introduction of idiosyncratic terms and relations, and combines reuse and sharing with domain-specific lexicons familiar to clinicians. In our experiment, the UMLS provides a useful corpus of medical knowledge for designing a domain-specific lexicon. In parallel with efforts to define a specific knowledge representation language for clinical concepts, the UMLS provides a pragmatic way to facilitate knowledge sharing and reuse. Many attempts to quantify the content coverage of the UMLS have been made, in several fields—clinical radiology, laboratory terminology, surgical procedures, and in hypertension notes. These projects showed limitations of Metathesaurus terms in specific areas, demonstrating the need to add new terms and relations for particular uses of the UMLS. Nevertheless, like other investigators, we consider the UMLS a useful formal framework that is readily available to medical institutions and continuously updated by the National Library of Medicine. The introduction of new terms, besides those found in UMLS, reflects the specific needs of experts and respects the philosophy of the UMLS, which concatenates several databases to gather the particularities of each of them.
We have designed a domain-specific interface to allow the domain expert to input directly or modify existing knowledge in a CDSS using domain ontology. Our knowledge acquisition tool uses rule templates, which appear well adapted to the decision-making process in the clinical applications we envisage. Our tool is designed for medical experts without specific training in medical informatics and allows their active participation in the process of CDSS construction. This can facilitate future integration of the CDSS into the hospital system. Domain knowledge is automatically transformed into MLMs in Arden Syntax so that it can be reused and shared by other institutions.

Our knowledge acquisition tool is not only an MLM editor; it allows the acquisition of knowledge through the predefined domain ontology. It is in the same vein of the work of Thurin et al.,35 who proposed a tool for building MLMs from a GALEN-based terminology. Arden Syntax is mature enough to be used successfully in large working projects in clinical institutions. However, it does have drawbacks, such as a style of rule representation that is not necessarily adapted to all clinical applications, low expressiveness of temporal primitives, and possible interactions between MLMs that may pose long-term maintenance problems.

In our system, the introduction of new terms, during domain ontology acquisition, can generate redundancies and conflicts with terms that are already present in the database. During domain knowledge acquisition, some conflicts may appear between rules. In the current implementation, the domain expert receives relatively little help from the system in solving redundancies and conflicts. Temporal aspects of the domain are currently not taken into account. Clearly, our tool should be extended to incorporate specific mechanisms to deal with these drawbacks.

To date, we have used the prototype only in collaboration with two medical experts involved in this project in our hospital. Their involvement since the start of the project clearly biases their fully positive opinion. However, we believe that our tool will be accepted by end users in those medical fields in which decision support systems have to be easily modified by medical experts themselves.

Creating domain ontologies that are reusable in new medical applications is a challenging task.36 Our experiment provides a practical example of how large resources, such as the UMLS, can be used to design a domain ontology and how that ontology can be put to use. Results in the area of blood transfusion demonstrate that the UMLS provides a useful and manageable corpus of medical knowledge for domain ontology construction.

References