Modeling Principles fòr QMR Medical Findings

Anne-Marie Rassinoux[1], Ph.D., Randolph A. Miller[1], M.D.
Robert H. Baud[2], Ph.D., Jean-Raoul Scherrer[2], M.D.
[1]Division of Biomedical Informatics, Vanderbilt University, Nashville, TN
[2]Medical Informatics Division, University Hospital of Geneva, Switzerland

Structured representation of medical information is essential for ensuring the accuracy and reliability of computerized decision support applications. Such systems require input that is error-free and clinically pertinent. This paper reviews existing medical models, particularly those exploited for natural language understanding, and highlights modeling features important to future indexing of medical texts with controlled vocabularies. A hybrid representation derived from existing frame-based and conceptual-graph-based systems is proposed to represent relevant medical terms as used by experts.

INTRODUCTION

Mediating between the language of users and the language used to express medical information in computer-based medical systems is a challenging endeavor[1]. This problem also occurs at the level of exchanging information between two systems which use different vocabularies to express relevant medical information. One solution consists of building a language-independent model describing the meaningful concepts of the concerned domain together with the relationships among the concepts. Such a conceptual model would act as an "interlingua", facilitating the mapping among electronic medical vocabularies. Further, it would support natural language processing (NLP) techniques, which allow the content of medical texts to be automatically structured into this formal representation scheme.

The model proposed in this paper derives from two independent efforts undertaken by the authors in the past. It combines the frame-based system developed by Miller, Masarie, et al.[2,3] with the conceptual-graph-based approach taken by Rassinoux et al. in developing the RECIT analyzer[4], which has been furthermore adjusted to the GALEN model[5]. The merged expertise has led to the specification of a computationally tractable medical model, capable of recognizing distinctions among complex medical terms generated by experts as well as ensuring the integrity of retrievals from free medical texts. This paper presents observations and desiderata for a final, combined system that is currently under development, and discusses important issues rather than presenting final conclusions or a systematic assessment of the proposed model.

BACKGROUND

Different nomenclatures and thesauri are currently used to help standardize aspects of clinical practice and organize the literature[1,6,7,8,9]. These traditional controlled vocabularies are generally specified through surface-form expressions, which are in most cases noun phrases. Such expressions do not reflect clearly the underlying concepts that they designate. They do not embody a complete and computationally tractable definition of what a term is, nor define, in computationally useful ways, how one term can differ from another one. Generally, the only implicit link occurs when a parent term is related to its children through their relative positions in a classification hierarchy. A deeper representation, more meaningful than the traditional tree structure, must be developed in order to model the intricate concepts of medicine. This task requires the understanding and careful specification of the structure of medical concepts. The specification must detail the attributes, values and relationships that are allowed to occur within components of concepts. It must also limit allowed instantiations of the relationships in order to prevent "nonsense" terminology from being acceptable. Attempts to add a semantic structure over existing controlled vocabularies in order to fully exploit them have already been described[10,11]. A more challenging approach consists of exploring how formal systems could be used for representing the concepts underlying medical terminology as evidenced by the following works:

- Miller, Masarie, et al., in early work supported by the UMLS project, have described a frame-based interlingua[2,3] to map equivalent concepts between controlled clinical vocabularies. This system is based on the assumption that clinically relevant statements about patients contain at least one identifiable central concept and central concepts can serve as focus for mapping between medical vocabularies. Generic finding frames were used to specify how a central concept may be expressed and also be qualified by linguistic

0195—4210/96/$5.00 © 1996 AMIA, Inc.

modifiers. Over 750 generic frames were created for describing the medical meaning of a test set of 1,500 medical terms for general internal medicine identified from the Quick Medical Reference (QMR)[7] lexicon, as well as portions of the HELP PTXT lexicon[8], and parts of the DXplain lexicon[9].

- Cimino et al. have constructed the Medical Entities Dictionary (MED)[12], a hybrid of terminology and knowledge, using a semantic network based on the Unified Medical Language System (UMLS)[1], with a directed acyclic graph to represent multiple hierarchies. Each concept node in the MED graph can be viewed as a frame, and may have links to nodes other than parent-child nodes through the semantic relationships. This system, which is beginning to reach critical mass (it currently contains 32,765 conceptual components), is in active clinical use at the Columbia Presbyterian Medical Center (CPMC).
- Since 1992, Rector et al. have been developing, through the GALEN project, a fully compositional and generative system of medical concepts[5] which is expressed in a language-independent manner through the GALEN Representation and Integration Language (GRAIL) Kernel. One important feature of this medical model is that it attempts to restrict entries to valid combinations of concepts that form medically sensible expressions. Moreover, the notation of the GRAIL Kernel can directly be converted to that of conceptual graphs[13] and the set of criteria associated to a concept can be seen as a frame-like structure. The current version, which contains nearly 6,000 concepts, must nevertheless be extended in order to be useful in general clinical applications.
- Another important effort is the Canon Group's work toward a merged medical model[14] (i.e. a common model that represents a consensus among Canon Group members) for use in exchanging data and applications. The conceptual graphs formalism[13] was chosen as the representational notation for the initial effort. The merged model specifies canonical medical concepts through, first, the semantic classification and hierarchical organization of the concepts, and second, canonical graphs consisting of terminological knowledge about the structure of the concepts and their semantic relationships with each other. A core model for radiological findings has then been developed, as chest radiography reports were used as the initial experiment.

A few comments can be made with respect to the afore-mentioned models. First, the common underlying goal of these models is to provide a repository for both patient data and medical knowledge through the specification of concepts and relationships between them. For this, several formalisms such as semantic nets, frames, or Sowa's conceptual graphs are currently used. Several attractive features (such as readability and straightforward semantics) have established the conceptual graph formalism[13] as possibly the most commonly used linguistic representational notation in the medical-informatics community. Second, as these conceptual models define medical knowledge, they create excellent opportunities for exploiting NLP techniques to understand the medical information embedded in natural language free texts. In particular, the frame-based interlingua system developed by Miller, Masarie, et al.[2,3] was successfully used to map among the "pseudo" natural language embedded in QMR[7], HELP[8], and DXplain[9] terms. The text processor developed by Friedman et al.[15] allows the impression section of chest x-rays to be mapped into unique medical concepts defined in the MED[12]. Likewise, the RECIT analyzer[16] developed by author AMR at the Geneva University Hospital, has also been adjusted to the GALEN model[5] in order to reinforce its ability to ground its semantic components on a solid medical model.

MODEL REQUIREMENTS

Cimino et al.[17] argue that a medical vocabulary must have synonymy, domain completeness and multiple classification, providing consistent views and explicit relationships other than hierarchical relations, while remaining unambiguous and non-redundant. The MED[12] and the GALEN[5] models significantly meet these requirements through their formal structures. Nevertheless, while considering these requirements as important, the review of the informational content and structure of the frame-based interlingua system[2,3] has led to the delineation of additional essential criteria which are discussed below.

A Bottom-Up Versus a Top-Down Approach

Conceptual modeling can be considered from both a top-down and a bottom-up approach. An important design criterion is to find a balance between a complete but complex semantic representation of full medical texts and a partial representation which is nevertheless useful for decision support. Although many groups have been working on medical language processing, very few useful and practical systems exist at the present time. Indeed, the strong medical constraints to be error-free and accurate have slowed the overall development. Moreover, in developing systems for processing terminology, the final end-point cannot simply be the translation of terms from source "A" to the format or structure of source "B". Practical feedback is required, and such feedback can only be provided by running clinical systems. These considerations point to the need to develop medical applications based on a limited and well-defined domain, answering to a precise goal, in order to yield concrete outcomes. That is why the authors have focused their efforts towards recognizing sensible information related to the QMR findings[7], with the aim of being able to index medical texts with these findings for eventual use in diagnostic decision support. Therefore, as opposed to a top-down approach aiming at building a medical model based on a priori conceptual organization of medicine (which is time and labor-intensive, and without a mechanism for feedback regarding success or failure), the strategy used here is based on a bottom-up approach consisting of collecting all the relevant axes and terms that clinicians might use to describe any and all medical concepts embedded in QMR terms. This approach allows the information specified in the QMR terms to be directly extracted and represented, thus ensuring the robustness of the representation as the generic frames directly fit instances of concepts defined through QMR terms. This enumerative, ad hoc, method has led to a flat series of frame descriptions, which are in return not easy to maintain. In order to keep a consistent view of the overall frames database, a conceptual layer acting as a multilevel hierarchy is being added to the frame system.

Terminological, Conceptual, and Knowledge Levels of Modeling

Describing medical concepts entails covering all the lexical items that can be used to express these concepts, while providing sufficient semantic structure (expressed via the specification of concepts and their relationships) to disambiguate similar variants of separate concepts. Moreover, the ability to distinguish between a normal and abnormal finding, as well as the specification of the methods used to elicit concepts in a medically meaningful fashion, are features rarely taken into account in existing models, while being of paramount interest in reasoning about medical concepts. The specification of all these distinctive features is incorporated through the consideration of different levels of modeling.

The terminological level, consisting of providing the model with the linguistic characteristics of the medical concepts, supports the specification of the concept names as well as complementary information (given under the form of definitions or fundamental criteria) useful to unambiguously recognize the medical concept encoded in a generic frame. The conceptual level, specifying the meaningful relationships a generic concept can have with the others, constitutes an important part of the description of the semantics of the domain under consideration. This set of relationships is supported in our system by the specification of allowed qualifiers, which can be optionally added to an instantiated concept to make its meaning more precise clinically. Finally, the knowledge level specifies intrinsic information which is attached to a given generic frame. In order to assist decision support applications that abstract structured findings from natural language texts describing patient cases, it is extremely important for a system to incorporate an explicit representation of constraints, such as standard or default values. Such information is contained in the specification of the status, as well as the methods together with their degree of reliability which are used to elicit the generic finding frame.

The distinctions among these levels are directly taken into account in the new proposed structure for the generic frames (as shown in the subsequent section through the frame slots) in order to address practical goals of producing structured data which can be directly used by decision support applications.

THE PROPOSED APPROACH

The structure of the proposed model addresses the above requirements through the specification of a hybrid representation consisting of frame structures (concept names, slots, and fillers) and a series of conceptual hierarchies integrating these various components.

```
genericFrame (#ConceptName,
        [Status: #StatusDescriptor,
        methods: #MethodsList,
        fundamentalInfo: #RequiredFields,
        basicDefinition: #Definition,
        qualifiers: #QualifiersList ] ).
```

Figure 1 - Basic structure of generic frames

The slots here are tags that unambiguously specify what information can be embedded in the fillers. The slots displayed in bold characters in Figure 1 are always present in a concept description, whereas the slots in italics are optional. Depending on the complexity of the clinical information to be represented, the fillers can have the format of simple values or conceptual graphs. A description of these different slots follows below.

The Core of the Description

The frame structure intends to unambiguously formalize the medical concepts which are clinically relevant (i.e. whose formulation provides enough information to start suggesting hypotheses). For example, *AbdominalPain* is a relevant concept, whereas *Pain* is too general. Indeed, a chart containing "There was severe substernal pain and intermittent right upper quadrant pain. The pain was colicky and radiated to the scapular area." might present difficulties to a system that represented characteristics of pain at a general level, since it would not be able to disambiguate which pain was colicky (the abdominal pain, since chest pain is not described as colicky) or which pain radiated to the scapula (the right upper quadrant pain, since it is the colicky pain referred to in the sentence).

Review of previous frames led to the recognition of two general categories of concepts differing from their status: existential or quantitative. Existential concepts either occur or do not occur in a given patient, as for example, *AbdominalPain* or *DrugAbuse*, whereas quantitative concepts specify measurement of clinical parameters, such as *AbdominalLymphNodeSize* or *Appetite*.

Reliability of the Representation

For each generic frame, the **status** reflects explicitly the "default normal value" of the considered medical concept. The allowable status for an existential concept is typically *PresenceOrAbsence*, which can be further specialized into *TrueOrFalse*, *NormalOr-Abnormal* or *PositiveOrNegative*, as respectively specified for the concepts *DrugAbuse*, *Affect* or *FungusCulture*. For a quantitative concept, the status is specified through qualitative possible states as well as ranges of allowed numeric values when relevant. For example, the following set: *[absent, decreased, normal, increased]* is used to describe the quantitative status of the concept *Appetite*. Representing the "default normal finding" together with its corresponding "abnormal findings" adds reliance to the overall description of the generic frames.

Moreover, our interlingua system, like the MED[12], designates the **methods** used to elicit concepts in a medically meaningful fashion, as well as the degree of reliability associated to each of them. For example, *an abdominal mass on palpation* is not the same concept as *an intraluminal mass in the colon*, even though they are both in a sense "intra-abdominal masses". However, *an intraluminal mass in the colon* is a unique concept, whether discovered by barium contrast studies, CT Scan, or colonoscopy.

Finally, some features are more essential to a generic concept than are others. In such circumstances, a **fundamental information** slot is added to the generic frame structure. This slot can specify two kinds of information: a site and a subcategory. For example, *AbdominalTopographicalSite* is a required site that further specifies the concept *AbdominalPain*, whereas the specification of a subcategory for *NAMEDDrug* particularizes the concept *DrugAbuse*.

Additional Details

As natural language (NL) is highly compositional (an underlying concept can be expressed in NL by strings of characters not necessarily contiguous), it is crucial to exhibit a **basic definition**, which fully specifies the meaning embedded through the concept name. Such a lexical definition, expressed in the conceptual graph formalism, provides a normalization which is specifically useful to detect similar terms in order to avoid redundancy. For example, the concept *AbdominalAorticAneurysmByImaging* can lexically be decomposed as an aneurysm which is located in the abdominal aorta and which is specifically determined by some imaging procedures.

The **qualifiers** slot provides additional contextual information which can be divided into two sorts: general and local qualifiers. General qualifiers can be seen as well-defined features that are useful across a number of generic concepts. For example, *severe acute abdominal pain* denotes the concept *AbdominalPain* qualified with the general modifiers *Severity* and *Chronicity*. A limited set of values for such qualifiers is defined external to each generic finding concept, and a measure of the distance between the different possible values is given. For example, the possible values for *Severity* are: *[mild, moderate, severe]*. These values are characterized as "progressive deviations", which implies that a "gray zone" may exist between clinical expressions such as *mild* and *moderate*, but it is unlikely that confusion would occur between *mild* and *severe*. Similarly, the values: *[decreased, normal, increased]* are defined as "opposite extremes around normal", so that *not increased* can be interpreted to mean *normal or decreased*, and a *decreased* value can be unambiguously interpreted as ruling out *increased* (at a given point in time). The local qualifiers denote some specific relationships (such as *isInfluencedBy* or *radiatesTo*), whose the set of corresponding values is strongly dependent on one specific generic concept. The list of factors that influence *ChestPain* is quite different than the list of factors that influence *AbdominalPain* (even though there is a small overlap), that is why these factors are defined as local qualifiers in each associated generic frame.

CONCLUSION AND FUTURE WORK

Since a large portion of medical information remains embedded in natural language texts (either under the form of controlled vocabularies or free medical texts), it is paramount to correctly interpret and then represent these medical information sources through a uniform and formal structure. The current proposal explores how to recast the generic interlingua frame system, initially developed by Miller, Masarie, et al.[2,3], into a more computationally tractable model which introduces conceptual graphs to represent and standardize important aspects of medical information. This approach fully exploits the representation of "normal" findings as well as the compositional aspects of medical information and constraints on such information.

Only the representation issues have been discussed in this paper. The use of such a model by language processing techniques will be explored through the use of the RECIT system[4], for which the generic frames can be seen as valid conceptual schemata useful to accurately build the sound representation of medical sentences. This language-independent representation will then open the way toward knowledge-oriented applications which add new functionality by moving from data to concepts.

Acknowledgments

This work is supported by a grant from the "Fonds National Suisse de la Recherche Scientifique". Work on the generic frame schema was originally supported through NLM Contract N01-LM-6-3522.

References

1. McCray AT, Nelson SJ. The Representation of Meaning in the UMLS. Meth Inform Med. 1995; 34: 193-201.
2. Miller RA. A Computer-based Patient Case Simulator. Clin Research. 1984; 32: 651A.
3. Masarie FE, Miller RA, Bouhaddou O, Giuse NB, Warner HR. An Interlingua for Electronic Interchange of Medical Information: Using Frames to Map between Clinical Vocabularies. Comput Biomed Res. 1991; 24(4): 379-400.
4. Baud RH, Rassinoux A-M, Wagner JC et al. Representing Clinical Narratives Using Conceptual Graphs. Meth Inform Med. 1995; 34: 176-186.
5. Rector AL, Nowlan WA, Glowinski A. Goals for Concept Representation in the GALEN project. In Safran C, ed. Proc of SCAMC 93. New York: McGraw-Hill, 1993: 414-418.
6. Rothwell DJ. SNOMED-Based Knowledge Representation. Meth Inform Med. 1995; 34: 209-213.
7. Miller RA, Massarie FE, Jr. Use of the Quick Medical Reference (QMR) Program as a Tool for Medical Education. Meth Inform Med. 1989; 28(4): 340-345.
8. Pryor TA, Gardner RM, Clayton PD, Warner HR. The HELP system. J Med Syst. 1983; 7:87.
9. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain: An Evolving Diagnostic Decision-Support System. J Am Med Informatics Assoc. 1987; 258:67-74.
10. Campbell KE, Musen MA. Representation of Clinical Data Using SNOMED III and Conceptual Graphs. In Frisse ME, ed. Proc of SCAMC 92. New York: McGraw-Hill, 1992: 354-358.
11. Joubert M, Miton F, Fieschi M, Robert J-J. A Conceptual Graphs Modeling of UMLS Components. In Greenes RA et al., ed. Proc of MEDINFO 95. Alberta: HC&CC, 1995: 90-94.
12. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based Approaches to the Maintenance of a Large Controlled Medical Terminology. J Am Med Informatics Assoc. 1994; 1: 35-50.
13. Sowa JF. Conceptual Structures: Information Processing in Mind and Machine. Reading, MA: Addison-Wesley Publishing Company, 1984.
14. Friedman C, Huff SM, Hersh WR, Pattison-Gordon E, Cimino JJ. The Canon Group's Effort: Working Toward a Merged Model. J Am Med Informatics Assoc. 1995; 2: 4-18.
15. Friedman C, Cimino JJ, Johnson SB. A Conceptual Model for Clinical Radiology Reports. In Safran C, ed. Proc of SCAMC 93. New York: McGraw-Hill, 1993: 829-833.
16. Rassinoux A-M, Wagner JC, Lovis C, et al. Analysis of Medical Texts Based on a Sound Medical Model. In Gardner RM, ed. Proc of SCAMC 95. Philadelphia: Hanley&Belfus, Inc., 1995: 27-31.
17. Cimino JJ, Hripcsak G, Johnson SB, Clayton PD. Designing an Introspective, Multipurpose, Controlled Medical Vocabulary. In Kingsland III LC, ed. Proc of SCAMC 89. Washington: IEE Computer Society Press, 1989: 513-518.