Compositional and Enumerative Designs for Medical Language Representation

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

Medical language is in essence highly compositional, allowing complex information to be expressed from more elementary pieces. Embedding the expressive power of medical language into formal systems of representation is recognized in the medical informatics community as a key step towards sharing such information among medical record, decision support, and information retrieval systems. Accordingly, such representation requires managing both the expressiveness of the formalism and its computational tractability, while coping with the level of detail expected by clinical applications. These desiderata can be supported by enumerative as well as compositional approaches, as argued in this paper.

These principles have been applied in recasting a frame-based system for general medical findings developed during the 1980s. The new system captures the precise meaning of a subset of over 1500 medical terms for general internal medicine identified from the Quick Medical Reference (QMR) lexicon. In order to evaluate the adequacy of this formal structure in reflecting the deep meaning of the QMR findings, a validation process was implemented. It consists of automatically rebuilding the semantic representation of the QMR findings by analyzing them through the RECIT natural language analyzer, whose semantic components have been adjusted to this frame-based model for the understanding task.

INTRODUCTION

Medicine is a domain involving a huge amount of information, most of which is still expressed through textual forms. Understanding and extracting the meaning embedded in these texts is a continuous challenge to researchers in medical informatics[1]. Standardization efforts towards reducing the expressiveness and peculiarities inherent in medical language have led to the emergence of two major methods of organizing medical information. On the one hand, different thesauri or controlled medical vocabularies (CMVs) - such as the UMLS Metathesaurus or the ICD classification - are now available, affording an extensive set of relevant terms to express patient-specific observations. On the other hand, more formal semantic models for medical concept representation - such as the Medical Entities Dictionary (MED) or the GALEN model - have come to light, fostered by the need to transcend words and phrases and to capture their "meaning". These conceptualization efforts result in language-independent and compositional systems for modeling the intricate concepts of medicine.

The counterbalancing features underlying the two approaches for medical concept representation relate mainly to breath of coverage and depth of representation, which respectively involve enumerative and compositional strategies. Actually, CMVs allow rapid and easy incorporation of new terms without disturbing the general representational architecture. But, enumerative description performed through language-surface form entails redundancy and inconsistency which can impede the overall maintenance of such vocabularies. Besides, representing medical concepts in a more computationally meaningful manner implies decomposing and structuring information in a formal way, which is suitable for manipulation by computer programs. This constitutes a more labor-intensive and time-consuming task. Therefore, it is necessary to limit the medical subject domain for fine modeling to yield concrete outcomes in a reasonable period of time.

This paper presents a challenging effort undertaken by the authors to recast the frame-based system initially developed by Miller, Masarie, et al.[2,3]. The objective is to obtain a more computationally tractable model which introduces conceptual graphs[4] to represent and standardize the various compositional aspects of medical information. The checking and adjustment have been manually performed for 750 generic medical finding frames that capture the meaning of 1500 selected QMR "surface-level" findings. One way to validate the accuracy and tractability of the new frame-based system is to use Natural Language Processing (NLP) techniques to check the meaning of QMR findings against this system.

1091—8280/97/$5.00

BACKGROUND TO THE FRAME-BASED SYSTEM

Characteristics of the QMR Vocabulary

The QMR vocabulary (which is a superset of the original INTERNIST-I vocabulary)[5] was created to describe possible (reported) patient findings in diseases in general internal medicine. It contains over 4500 clinical manifestations, including patient symptoms, physical findings, and laboratory test results. This vocabulary was derived from extensive manual literature review and serves the purpose of providing input for the QMR diagnostic program[5]. Such a vocabulary fits the characteristics of enumerative systems. Terms are mainly described through noun phrases consisting principally of medical phrases with generally accepted definition and usage, as shown in Figure 1.

Finding: DYSPNEA PAROXYSMAL NOCTURNAL
This phrase corresponds to the medical expression "paroxysmal nocturnal dyspnea", which describes an *acute onset of inappropriate shortness of breath or similar difficulty in breathing occurring at night.*

Finding: ORTHOPNEA
This term describes a *discomfort in breathing which is brought on or aggravated by lying flat.*

Figure 1 - QMR findings and their clinical definition

Moreover, it is worth noting that the language used to express these findings is strongly stereotyped and has not strictly applied the syntactic formative rules of English. In particular, conventional orders of certain words are not followed (in order to maintain a new form of internal consistency for word order), and prepositions are less frequently used. These surface observations already suggest that semantic categories appear to be more appropriate to determine the details of interpretation of these noun phrases as syntax is used in a "fancy" way.

Evolution of the Frame-Based System

In order to capture the clinical complexity of the QMR findings, Miller, Masarie et al. developed a frame-based interlingua[2, 3], which has been further used to facilitate the translation between CMVs. This system limits itself to collecting - through a bottom-up approach reviewing each existing QMR finding - a core set of central concepts considered as relevant to recognize any and all sensible information embedded in the QMR findings. For this, it is assumed that any clinically relevant statement about patients contains at least one identifiable central concept. Figure 2 shows an example of a generic frame, followed by the list of QMR terms which are candidate to map this structure.

DYSPNEA
Generic Frame: last edited on */*/* by *****
Allowable Status: Presence Or Absence
Normal Status: Absent
Method(s)
Name: Cardiopulmonary Observation
Reliability: 4
Qualifier(s)
Pattern Of Occurrence; Time Duration Qualitative; Time Duration Quantitative; Influence On Dyspnea; Time Of Day; Time Onset Qualitative

DYSPNEA ABRUPT ONSET
DYSPNEA ACUTE RECURRENT ATTACK <S> HX
DYSPNEA AT REST
DYSPNEA AT REST RELIEVED BY RECUMBENCY
DYSPNEA EXERTIONAL
DYSPNEA IMPROVEMENT AFTER HEMOPTYSIS HX
DYSPNEA PAROXYSMAL NOCTURNAL
ORTHOPNEA
DYSPNEA RELIEVED BY SQUATTING HX

Figure 2 - Initial generic frame structure

The generic frame structure provides the backbone for describing the fundamental characteristics associated with the central concepts. This structure integrates both the status description of the considered medical concept (i.e. its "default normal value", usually describing clinical findings as normal or abnormal conditions affecting anatomical sites) and the methods used to elicit such a concept in a medically meaningful fashion, as well as the potential qualifiers which can be applied to this central concept. The qualifiers lists (also called item lists[3]) are useful to encapsulate fine details. Such qualifiers are maintained apart from the generic frames as they specify well-defined features often applicable across a number of generic frames. The qualifiers description incorporates both a limited set of values as well as a header stating the logical relationship among the components. For example, the qualifier *'Time Duration Qualitative'* is represented through the following values: *Acute, Subacute, Chronic* linked by the header *ExactlyOneOf.*

The thorough and enumerative method used to build the frame-based system insures the richness and accuracy of the resulting model. Indeed, the builder of the knowledge base system (usually referred to as the expert) was concerned primarily with the extraction of relevant concepts from the test set of QMR terms (and some terms from DXplain and

HELP as part of the UMLS project) without being compelled to apply some protocol instructions. However, this approach limited development of a fully language-independent and computationally tractable system of medical concept representation. On the one hand, it appears that the concept system itself is not clearly separated from the precise language used for specifying its components. The extensive use of complex linguistic names to label central medical concepts (such as *'Left Ventricular End Diastolic Internal Diameter'*), as well as qualifiers (such as *'Timing Within Systole Or Diastole'*) blurs the separation between the concepts to be represented, and the linguistic terms and mechanisms used to refer to those concepts. Moreover, the separation between concepts and relationships is masked by the use of equivocal labels (such as *'Influence On Dyspnea'*). On the other hand, the flat enumeration of generic frames, making use of a large amount of conceptual entities which are not structured in a hierarchical framework, causes trouble for maintenance and navigation through the system itself.

Facing these drawbacks, a new structure[6] has been developed by the authors. The result, based on the example shown in Figure 2, is displayed below in Figure 3.

```
genericFrame('Dyspnea',
  [existentialStatus:
        [allowableStatus:'PresenceOrAbsence',
        normalStatus:absent],
  definition: ['Difficulty', [actsOn('Breathing')]],
  methods:['CardiopulmonaryObservation'('4')],
  qualifiers:[hasProcessPattern('ProcessPattern'),
        hasChronicity('Chronicity'),
        hasDuration('Duration'),
        isInfluencedBy(['Hemoptysis','Exercise',
                'BodyPosition','Rest']),
        occursDuring('TimeOfDay'),
        hasOnset('TypeOfOnset')] ] ).
```

Figure 3 - Revised generic frame structure

For the sake of clarity, the nature of the manipulated information is highlighted by considering two kinds of generic frames, differing by the type of their status. On the one hand, the *existential frames* describe findings which may or may not occur for a given patient. On the other hand, the *quantitative frames* describe clinical parameters which can be measured. Moreover, except for the slot *qualifier*, the other slots contain mandatory information that helps in recognizing, in a non-ambiguous way, the current generic frame.

DEALING WITH COMPOSITIONALITY

As emphasized in Figure 3, the recasting of the frame-based system dealt mainly with transforming a rather enumerative description into a more structured system, which fits most of the desiderata highlighted by Cimino[7]. The main innovations and their issues are discussed below.

Hierarchy of Concepts

Even if the frame structure used to represent the central medical concepts is convenient to express a first level of description (through slots and fillers), allowing then the initial structure to be inverted according to some criteria, this representation is nevertheless not easy to maintain. Therefore, a hierarchically-structured view of, at least, all the primitive concepts which are useful to describe more complex medical information has been implemented. The high level of this multiple hierarchy (i.e. lattice) first delimits conceptual entities from relationships, thus determining straight-away the atomic objects handled by any compositional process. Second, it separates medical concepts from the modifiers which serve to precisely describe these concepts. Such a subclassification reflects the two main parts of the frame-based system (i.e. the generic frame structure and the qualifiers description) and allows for specifying the weight given to the information, in particular for its potential use by NLP tools. In addition, the part of the hierarchy listing the methods is especially detailed, as such methods play an important clinical role in eliciting the central concepts.

Formal Definitions

In order to be able to exploit (with a computer) the meaning of complex medical expressions, formal definitions are introduced. At this level, it is important to delineate definitional knowledge from assertional knowledge[7]. The literal definition, added in the frame-based system, only reflects the terminological (also called lexical or literal) meaning embedded in the central concept name. For example, the concept *Dyspnea* refers to *a difficulty* (Greek prefix "dys") *in breathing* (Greek root "pnea"). Such a definition, acting as definitional knowledge, is often not complete enough to describe the full clinical meaning of the treated concept. This meaning, referring to the assertional (also called encyclopedic or contextual) knowledge is explicitly expressed in the model itself (through the slots methods, qualifiers...), which establishes the context and circumstances in which the central concept should occur in the clinical reality.

This literal definition presents some interesting features. First, it is expressed through the Conceptual Graph (CG) formalism[4], which allows a convenient graphical representation of concepts linked through relationships. This formalism offers a rich representation as conceptual graphs can be arbitrarily large. It also supports various kinds of operations, in particular, contraction and expansion, which are especially important in handling definitions. Second, this definition is of paramount importance in retrieving the different linguistic expressions of the central concept from textual documents, especially when this concept is expressed with a multi-word phrases (that is to say, consisting of more than one word). For example, the definition related to the central concept *Dyspnea* (see Figure 3) is heavily relevant to extract this concept from the sentence "*The patient presents some difficulties in breathing at night.*", using a semantic-oriented medical language processor such as the RECIT system[8].

Finally, having a compositional model allows equivalent definitions to be expressed and maintained at the conceptual level, thereby eliminating the need to provide the system with numerous lexical variants, as discussed in the next section.

Hierarchy Annotation

The hierarchy annotation is particularly important for ensuring that tools with access to textual sources, such as retrieval engines or natural language processors, function correctly. Indeed, it consists in an extensive enumeration of synonyms and related terms (expressed through single words or multi-word phrases) which are used to refer to concepts, and are stored in the so-called dictionaries. For example, the concept *Dyspnea* can be annotated in English by the following linguistic expressions: *dyspnea*, *breathlessness*, *shortness of breath*, etc. However, noun phrases such as *discomfort in breathing* are not mandatory, as long as the literal definition of *Dyspnea*, as well as the annotations of the primitive concepts composing this definition, are provided by the system.

Scope of Relationships

An important aspect of compositionality is handled through the notion of relationships and is emphasized in the frame-based system by replacing qualifiers such as '*Influence on Dyspnea*' into the relationship *isInfluencedBy* which points to the set of relevant concepts (see Figure 3). At this level, it is important to clarify and enforce the scope of relationships in order to avoid misinterpretation of the meaning specified in the generic frames. For instance, the generic frame *AbdominalPain* (whose literal definition is *[Pain,[hasLocation(Abdomen)]]*) embeds the qualifier *Periodicity*, whose *Colicky* is a possible value. Saying that an *AbdominalPain* can be *Colicky* does not infer any information on the concept *Pain* in particular. Therefore, the relationship *hasPeriodicity* must strictly link the full concept *AbdominalPain* (and not part of its definition) to the possible value *Colicky* in order to avoid wrong interpretation.

EVALUATION OF THE FRAME SYSTEM

The evaluation of the frames' content is essential for the consistency, extension, and sharing of the overall system. The global process is reported in Figure 4, and is discussed according to the expressiveness and computational tractability of the revised frame system.

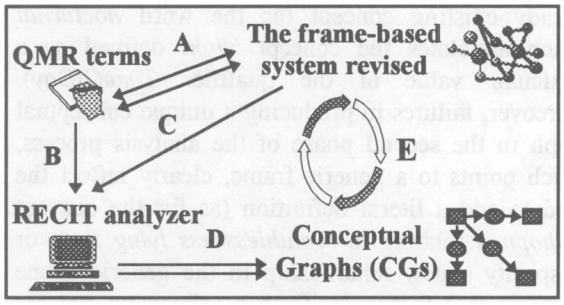

Figure 4 - The global evaluation process

The way the frames were created[3], and then reviewed[6] - by checking each frame's content to the set of QMR terms, candidates to be instantiated through this frame (link A in Figure 4) - constitutes a first validation of the expressiveness of the system in grasping the QMR terms meaning. This manual validation is then reinforced through the use of the RECIT multilingual analyzer[8], which automatically analyzes and stores the semantic content of the QMR terms under the form of CGs (link D in Figure 4). This last process, useful to validate both the granularity and tractability of the frame system is presented below.

The RECIT analyzer first applies "Proximity Processing" rules to group neighboring words together, and second links these semantic fragments into a sound structure expressed through conceptual graphs. For the task of analyzing the QMR terms (link B in Figure 4), the semantic components of the RECIT analyzer have been grounded directly from the revised frame-based system (link C in Figure 4). This latter connection emphasizes the computational

tractability of the frame system, and has been facilitated by our previous experience in relying on the GALEN model[8]. For the present experiment, the generic frames are integrally used as valid conceptual schemata, useful to accurately build the sound representation of medical sentences. The compatibility rules used in the first analysis phase are also extracted from this structure.

RESULTS

As a preliminary test, 200 QMR findings, instantiating nearly 50 generic frames, were given as input to the RECIT system. The results were reviewed for the two analysis phases. In particular, failures during the proximity processing phase generally occur because of lack of specifying a particular concept (as the concept *Orthopnea*, further classified in the hierarchy as a child of the concept *Dyspnea*), or lack of a specific annotation for an already existing concept (as the word *nocturnal* which annotates the concept *Night* defined as a particular value of the qualifier *TimeOfDay*). Moreover, failures in producing a unique conceptual graph in the second phase of the analysis process, which points to a generic frame, clearly reflect the need to add a literal definition (as for the concept *Orthopnea* relating to *breathlessness lying flat*), or to specify a new relationship in the generic frame structure (see Figure 3). Such a refinement process of the hierarchy, dictionaries, literal definitions, and generic frame structure, can be considered as a feedback loop from the NLP system to the model as illustrated by link E in Figure 4.

Finally, as the model evaluation lies on the result of the RECIT analyzer, the performance of this analyzer toward dealing with the medical jargon has also been readjusted, facilitated by the fact that such analyzer was specifically designed for this task[8].

CONCLUSION

This paper reassigns the importance of compositional and enumerative designs for medical language representation, respectively between the modeling process and the linguistic annotation process (which underlies any concept model intended to be used by some NLP tools[8]). It clearly emphasizes the benefits of managing a fully compositional and tractable model of medical concept representation, in parallel with an enumerative dictionary of synonyms and related terms, in order to handle the intricacy of the medical language.

The automatic validation process of the frame-based system, using the RECIT medical language analyzer, allows both the expressiveness and tractability of the model to be checked. This experiment promotes NLP tools, whose generation has also been successfully applied for this task[9], as quality assessment processes of concept models.

Acknowledgments

This work is supported by grant number 8220-046502 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. Dr. Miller's current work is supported in part by NLM contract 1 R01 LM06226-01A1.

References

1. Spyns P. Natural Language Processing in Medicine: An Overview. Meth Inform Med, 1996; 35(4/5): 285-301.
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. Sowa JF. Conceptual Structures: Information Processing in Mind and Machine. Reading, MA: Addison-Wesley Publishing Company, 1984.
5. Miller RA, Masarie FE, Jr. Use of the Quick Medical Reference (QMR) Program as a Tool for Medical Education. Meth Inform Med, 1989; 28(4): 340-345.
6. Rassinoux A-M, Miller RA, Baud RH, Scherrer J-R. Modeling Principles for QMR Medical Findings. In: Cimino JJ (ed). Proceedings of the 1996 AMIA Annual Fall Symposium (formerly SCAMC). Philadelphia: Hanley & Belfus, Inc. 1996: 264-268.
7. Cimino JJ. Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. Proceedings of the Fourth International Conference on Natural Language and Medical Concept Representation, Jacksonville, Florida, January 19-22, 1997: 257-267.
8. Rassinoux A-M, Wagner JC, Lovis C, et al. Analysis of Medical Texts Based on a Sound Medical Model. In: Gardner RM (ed). Proceedings of SCAMC 95. Philadelphia: Hanley&Belfus, Inc., 1995: 27-31.
9. Baud RH, Rodrigues J-M, Wagner J, et al. Validation of Concept Representation Using Natural Language Generation. Proceedings of the 1997 AMIA Annual Fall Symposium (formerly SCAMC). Nashville, TN, October 25-29, 1997.