Analysis of Medical Texts Based on a Sound Medical Model

Anne-Marie Rassinoux[1], Ph.D., Judith C. Wagner[1], M.S., Christian Lovis[1], M.D.,
Robert H. Baud[1], Ph.D., Alan Rector[2], M.D., Ph.D., Jean-Raoul Scherrer[1], M.D.,
[1]Medical Informatics Division, University Hospital of Geneva, Switzerland
[2] Medical Informatics Group, Department of Computer Sciences, University of Manchester, UK

Automatic understanding of natural language is a complex task due to the presence of ambiguities. In particular, semantic ambiguities which are often immediately and unconsciously solved by human beings, are raised when analyzing natural language sentences by computer. The latter has to know the implicit and contextual information in order to resolve these difficulties.

Nowadays in medicine, a considerable effort is deployed to model semantic contents of the medical domain. Such a task is usually performed separately from linguistic considerations.

The goal of this paper is to highlight the key issues of basing a medical language processing system on a sound semantic model. To illustrate the requirements and advantages of such a conceptual approach to the analysis process, the experiment conducted to adjust the RECIT analyzer to the GALEN model is shown.

INTRODUCTION

Natural language texts are largely used in the medical domain to communicate relevant information on patients. Automatically understanding their content represents a substantial aid for decision making (in particular for the retrieval of similar cases from the database) and for clinical documentation. Some important solutions for natural language processing (NLP) in medicine have already emerged[1,2,3] in the last decade. However, medical text analysis requires the consideration of both the characteristics of the medical domain and the particularities of medical language. The first point is concerned with the description of the semantics of the medical domain. Such a modelling process necessitates the definition of the kind of information that must be represented, as well as the level of detail required to describe such information. The second point implies the consideration of the particularities of medical sublanguage[4]. The latter is characterized by a compact and concise style (close to telegraphic style) which reflects the physician's need to communicate pertinent information using a limited amount of space and time. As a result, medical texts may be expressed ignoring scholastic rules, especially from the syntactic point of view. They consist essentially of juxtaposed noun phrases and prepositional phrases, using adverbs and abbreviations. Due to the rather imprecise syntax of medical language, semantic approaches offer new perspectives for medical text analysis. This is why the need for verifying semantic constraints against a solid medical model for analysis purposes arises.

The intent of this paper is to emphasize the benefits of using such modelling information for NLP purposes. The importance of modelling for natural language understanding has already been recognized by the authors[5] and it is practically applied to strengthen the inference of the RECIT analyzer by using the medical content of the GALEN model. First, both systems are briefly described. Then, the peculiarities of using modelling information for medical language analysis are emphasized as well as the gap between linguistic and conceptual views.

ANALYSIS OF MEDICAL JARGON: THE RECIT SYSTEM

The RECIT system (a French acronym for *REprésentation du Contenu Informationnel des Textes médicaux*) aims at transforming a sentence from a sequence of words to a conceptual representation. The analysis process is split into two phases, each dealing with specific features of medical language[6].

The Proximity Processing Phase

The first analysis phase, called *Proximity Processing* (PP), is a deterministic phase which combines the application of non-conventional syntactic procedures with the checking of semantic compatibilities in order to group neighbouring words together. Its main aim is to highlight syntagmatic expressions, considered as meaningful from the medical content viewpoint. The analyzer begins with a lexical preprocessing step, the aim of which is to locate the unambiguous syntactic category of words or groups of words. Consequently, the grouping of words constituting the core of proximity processing can be applied. As soon as two syntactic constituents are detected in a noun phrase (such as noun + noun, noun + adjective or noun + noun complement), a check is triggered in order to retrieve the correct semantic relationships linking these two underlying concepts. Such implicit relationships are directly defined through the so-called semantic compatibilities. These rules (see Figure 1) consist of two parts. The first clarifies the syntactic structures which can support the expression of the concepts and the relationship in a specific language (*en* for English, *fr* for French...). The second part gives the sensible combination of a pair of concepts linked by a relationship (for example an *acquired lesion* can have as *location* a *body structure*).

These compatibility rules present linguistic and conceptual key issues. On the one hand, they take advantage of the way a syntactic category can co-occur with others and how the semantic information can be combined through the specification of semantic relationships. On the other hand, they constitute an important part of the description of the semantics of the domain under consideration. The result of proximity processing, given already a partial interpretation of the sentence in the form of meaningful fragments, is the starting point for the next phase.

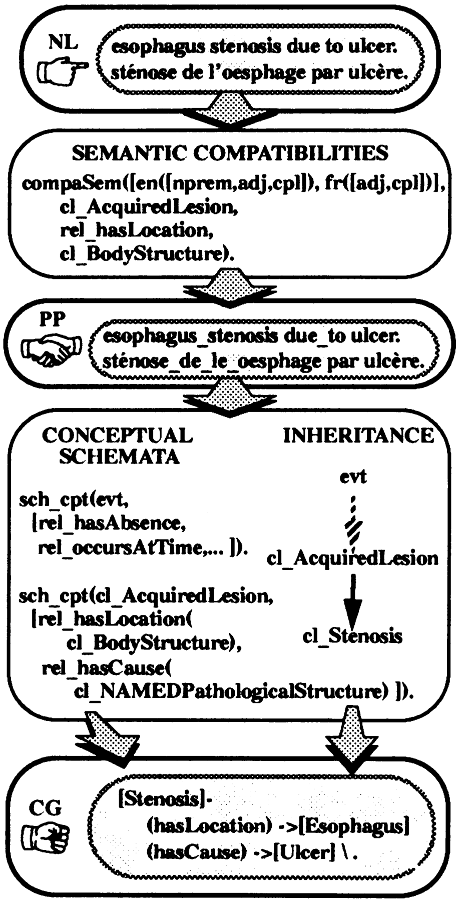

Figure 1. The two-phase process of the analysis

The Conceptual Graph Building Phase

The second analysis phase deals with the building of a sound representation of the sentence meaning into the formalism of *Conceptual Graphs* (CGs), as defined by J. F. Sowa[7]. For this, the system begins by highlighting the different concepts embedded in the syntagmatic structures built during the proximity processing phase. Then it has to select the heading concept and to establish the links this latter has with the others, in order to build the corresponding CG.

In order to perform these tasks, canonical conceptual schemata have been defined for each useful concept representing the meaning of the text (see Figure 1). They are similar to the frame-structures as they allow the description of the semantic relationships that a concept can have with the others in a specific context. These schemata are enhanced by an inheritance mechanism in the typology of concepts.

Connecting the different constituents of the sentence to the heading concept is then performed by mapping these constituents with the roles described in the conceptual schema of the heading concept. At this stage, the relationships expressed by a lexical word (such as a preposition or a conjunction, also called explicit relationships) are properly handled, even in the case of semantic ambiguity (for instance, the semantic ambiguity of the French preposition *par*, which can express a place, a manner, or a cause, etc., is solved in the example shown in Figure 1, through the only sensible relationship *hasCause* described in the schema of *acquired lesion*).

This phase of building CGs is mainly made possible by taking full advantage of the known semantics in a domain of knowledge like medicine. Therefore, it allows ungrammatical structures to be taken into account insofar as they are semantically consistent.

MODELLING THE MEDICAL DOMAIN: THE GALEN MODEL

Medicine constitutes a field where several attempts have been made to formalize the underlying knowledge. The most widespread technique consists of defining a semantic network, the backbone of which is a typology of concepts. This kind of modelling process requires a number of conceptualisation efforts from specialists of the domain. The final goal is to obtain a solid medical knowledge base, able to be used as a common structure for the exchange of knowledge. There are no unique methodological principles to constrain the acquisition of such a universal medical model. This fact is emphasized by numerous existing systems, such as UMLS[8] or SNOMED[9] and more recently the *General Architecture for Language, Encyclopaedias and Nomenclatures in medicine* (GALEN[10]) project. These medical models, more or less independent of any natural language, are intended to be used as a knowledge server for different applications. In particular, the GALEN model was used at first by the authors for the development of a multilingual medical natural language generation system[11]. This successful experiment was then extended to the analysis of medical language sentences.

The GALEN Model

GALEN is a project of the Advanced Informatics in Medicine (AIM) program of the European Union (EU). 'It is developing a 'Terminology Server' to manage language-independent shared systems of concepts for clinical applications'[10]. The medical concepts are represented in the COncept REference (CORE) model, which is both compositional and generative. Indeed, it allows composite concepts to be defined through the combination of basic concepts linked by relationships. Moreover, its generative capacities allow sensible composite concepts to be sanctioned at the highest possible level in the model, without having to store them explicitly. The GALEN Representation and Integration Language (GRAIL) Kernel is the formalism for representing concepts in GALEN. The GALEN representation is independent of any specific natural language. Moreover, every combination of concepts specified in the model must be 'sanctioned' in order to be 'coherent' with respect to the model. Such sanctions are expressed by statements in the model at three nested levels: grammatical, sensible and necessary. The first one sanctions at a high level what is grammatically reasonable to be said (such as, *a disease can be located in a part of the body* but not *in color*). The next one provides the semantics of what is medically sensible (such as, *a temperature may have a quantitative value* (see Figure 2)). Finally, the last one is the most constraining level which provides additional information about the real medical world (such as, *the tibia is a structural component of the leg*).

Temperature
hasQuantity TemperatureValue : **sensible**

Figure 2. Sensible statement in the GALEN model

MODELLING FOR ANALYSIS PURPOSES

The first step towards the adjustment of the RECIT analyzer to the GALEN model, consists in specifying which kinds of knowledge described in such a medical model can be directly used by the analysis process, and how such information is managed to correspond to the analyzer needs.

The Typologies

Typologies of concepts and relationships constitute the basic conceptual skeleton around which the semantics of the domain can be expressed. Their reusability for different applications is the foundation of numerous projects[12, 13]. However, using an existing typology for applications other than the one for which it has been built, is known as a difficult task[14]. Indeed, each conceptualisation effort is guided by a peculiar point of view on the semantics of subsumption. In particular, in the RECIT system, the conceptual hierarchy of origin was structured in a tree forest describing the *actors*, the *medical events*, the *attributes*, the *values* and the *modalities*. This kind of structure was appropriate for the analysis purpose, especially to identify the importance of the information during the analysis process (for example, actors and medical events constitute the basic meaningful entities of the medical domain which can possibly be described by qualifiers, temporal information and so on). In order to retain the analysis strategies set up by the RECIT system, a necessary alignment of the GALEN typology onto the five fundamental divisions of the previous RECIT typology is advocated. This has been realized through the specification of pointers at the highest possible level of the GALEN typology.

A similar solution has been established to categorize the GALEN relationships in the four categories treated by the RECIT analyzer (modality relationships, thematic and attribute relationships, temporal relationships and finally inter-sentence relationships).

The Sanctioning by Statements

The GALEN sensible statements present strong similarities with the conceptual part of the compatibility rules and with the content of conceptual schemata (see Figures 1 and 2). They allow a relationship to be set up between each pair of sensible concepts. This relationship can be any of those described in the typology of relationships.

Analyzing medical texts aims at extracting information which is pertinent from the medical viewpoint. In order to allow the analyzer to handle medical jargon and to correctly predict the semantics of additional phrases, it is necessary to check the natural language input against specific description of statements. This is why only sensible statements will be used during the proximity processing phase to group together sensible medical facts. For example, the semantics of the phrases *arm amputation* or *amputation of the arm* is completely checked against the following sensible statement:

Amputating
actsOn ExtremityBodyPart: sensible

At the level of conceptual schemata, the sanctioning may be more extensible, given that their use is to build the most complete CG expressing all the semantic dependencies established between the meaningful terms of the analyzed sentence. In this respect, the criteria representing essential characteristics of a concept in the GALEN model can be seen as the similar structure described in the conceptual schemata. This means that a conceptual schema for a concept is now built by joining the statements, whatever their level, defined for this concept.

The Definition of Composite Concepts

A composite concept in the GALEN model is built from any combination of concepts and relationships which have already been sanctioned as sensible. If such a composite concept (such as *Nephrectomy*) has been annotated by words in the treated language (such as *néphrectomie* for French or *nephrectomy* for English) it will be recognized as a single entity during the analysis process. As the major goal of NLP is to allow the resulting base to be queried for specific information retrieval, it is useful to replace the composite concept by its definition. For example, by using the following definition given by the model:

Nephrectomy is:
Excising actsOn Kidney

the query about the possible characteristics of the kidney will be able to retrieve information related to nephrectomy. This operation is performed on the CG generated by the analysis through conceptual expansion.

CONCEPTUAL VERSUS LINGUISTIC KNOWLEDGE

The advantage of using a medical model for analysis purposes is important for the definition of the Medical Linguistic Knowledge Base (MLKB[15]). Indeed, the latter is intended as a recipient for all the necessary declarative knowledge used during the analysis of medical texts. The clear separation between conceptual and linguistic knowledge parts has been recognized by the authors as being of paramount interest for a potential extension of the MLKB to other domains as well as to other European languages. Adopting the GALEN model as the conceptual part of the MLKB implies that the conceptual knowledge is not scattered everywhere in the analysis process but centralized in the domain model. Nevertheless, as its use is oriented towards NLP, linguistic features are imperative.

Annotation of the Model

In order to bridge the gap between how things are described in the model and how things are expressed in natural language, it is necessary to attach linguistic knowledge to the CORE model. Such linguistic annotations must take place both at the level of the typology and of the statements defined in the GALEN model. Nevertheless, it is obvious that this information must be clearly distinguished from the conceptual model.

On the one hand, the typology annotation allows concepts to be annotated by words or expressions available in the different languages treated together with their syntactic properties. These annotations constitute the basis for the automatic building of the multilingual dictionaries which are used during the analysis process. On the other hand, the statements annotation allows syntactic structures to be defined in order to specify how the model statements (i.e. concept-relationship-concept combinations) can be expressed in the different languages. Such annotations present similarities with the syntactic part of the compatibility rules conceived in the former version of the RECIT system. Nevertheless, the statement annotations are more particularly an annotation of the relationship itself at a high level which can always be refined by defining a more restrictive conceptual context (refutation of inheritance).

Detailed Degree of the Conceptual Representation

In order to grasp all the information precisely described with natural language, it is important to have a refined typology. However, it is often difficult to specify at its creation how detailed a typology would be for its future use. That is why a fairly high level model is required in order to obtain an overview of the field concerned. Furthermore, the possibility to refine it for specific purposes is a necessary property of the model, which is true for the GALEN model.

Moreover, the way medical information is modeled in the GALEN project would be considered as the degree of required detail to express the medical facts. This consensus on the representation allows the inference mechanism to be delimited during the analysis process.

Uniformity of Representation Towards the Model

Natural language is rather permissive. It allows things to be expressed without explicitly specifying the underlying context. Such implicit information is easily completed by any human being through context or by default. However a computer program only knows what it has been fed with. This is why the use of a complete medical model is particularly important to reproduce medical reasoning and to prepare further queries. For example, the complete meaning of the utterance *acute pain* can be defined in relation with the model as being a pain which has a value for chronicity which is acute. In order to analyze this expression during the proximity processing phase, a specific linguistic relationship *hasChronicityState* is created. Such a relationship is then specified in the model through the setting up of the corresponding definition (see Figure 3). Therefore, the operation of relational expansion allows the CG representation built for this utterance to be compatible with the way things are expressed in the GALEN model.

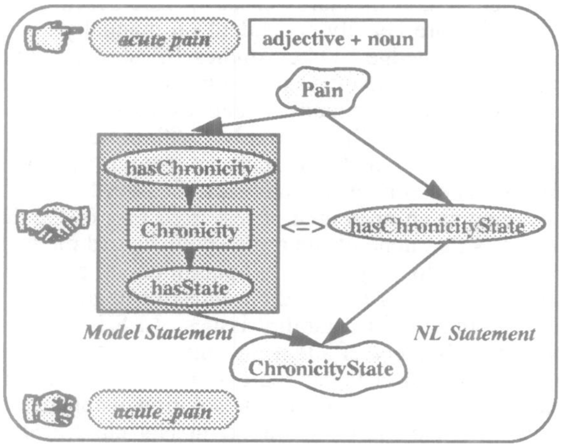

Figure 3. Handling implicit information

This converging expressiveness between the CGs resulting from the analysis and the information expressed in the GRAIL formalism can be exploited to extend the model through natural language input. Indeed, when the checking of a sentence against the model fails due to lack of knowledge in the model, such a sentence can be used to feed the model. The mastery of such a knowledge acquisition process is, however, known to be difficult, and further experiments in this direction are necessary.

CONCLUSIONS

In this paper, we have presented a model-based approach to the processing of medical language texts. This approach presents several impacts on the resulting quality of the conceptual analysis of medical texts. On the one hand, this quality becomes strongly dependent on the quality of the model, both for the sharpness and accuracy of the resulting knowledge representation, and for the extension of the multilingual dictionaries. On the other hand, this quality is also connected to the ability of the analyzer to handle ill-formed utterances as well as medical jargon. Nevertheless, this type of model-based solution allows us to foresee the rapidly growing ability of the medical language analyzer to process more and more elaborate medical reports as the model knowledge regularly increases. Finally, adapting the multilingual RECIT analyzer to the GALEN model, on which a multilingual generator[11] of medical language has already been developed, is a promising step towards automatic translation in different languages.

References

1. Schröder M. Knowledge-based Processing of Medical Language: A Language Engineering Approach. In: Ohlbach H-J (ed). Sixteenth German Workshop on Artificial Intelligence (GWAI 92). Proceedings. Berlin: Springer-Verlag. 1192; 221-34.
2. Zweigenbaum P, Consortium Menelas. MENELAS: an access system for medical records using natural language. Comput Meth Prog Biomed. 1994; 45: 117-20.
3. Sager N, Lyman M, Nhàn NT, Tick J. Medical Language Processing: Applications to Patient Data Representation and Automatic Encoding. Meth Inform Med. 1995; 34:140-6.
4. Grishman R, Kittredge R. Analyzing Language in Restricted Domains: Sublanguage Description and Processing. Hillsdale, NJ: Lawrence Erlbaum Associates, 1986.
5. Baud RH, Lovis C, Alpay L et al. Modelling for Natural Language Understanding. In: Safran C (ed). SCAMC 93. Proceedings. New York: McGraw-Hill. 1993; 289-93.
6. Rassinoux A-M, Juge C, Michel P-A et al. Analysis of Medical Jargon: The RECIT System. In: Barahona P, Stefanelli M, Wyatt J (eds). AIME 95. Proceedings. Berlin: Springer. 1995; 42-52.
7. Sowa JF. Conceptual Structures: Information Processing in Mind and Machine. Reading, MA: Addison-Wesley Publishing Company, 1984.
8. McCray AT, Nelson SJ. The Representation of Meaning in the UMLS. Meth Inform Med. 1995; 34:193-201.
9. Rothwell DJ. SNOMED-Based Knowledge Representation. Meth Inform Med. 1995; 34:209-13.
10. Rector AL, Solomon WD, Nowlan WA et al. A Terminology Server for Medical Language and Medical Information Systems. Meth Inform Med. 1995; 34:147-57.
11. Wagner JC, Solomon WD, Michel P-A et al. Multilingual Natural Language Generation as Part of a Medical Terminology Server. In: Greenes RA, Peterson H, Protti D (eds). MEDINFO 95. Proceedings. Alberta: HC&CC, Inc. 1995; 100-4.
12. Lenat DB, Guha RV. Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project. Reading, MA: Addison-Wesley Publishing Company, 1990.
13. Rector AL, Nowlan WA, Glowinski A. Goals for concept representation in the GALEN project. In: Safran C (ed). SCAMC 93. Proceedings. New York: McGraw-Hill. 1993; 414-18.
14. Zweigenbaum P, Bachimont B, Bouaud J, Charlet J, Boisvieux J-F. Issues in the Structuring and Acquisition of an Ontology for Medical Language Understanding. Meth Inform Med, 1995; 34:15-24.
15. Baud R, Lovis C, Rassinoux A-M et al. Toward a Medical Linguistic Knowledge Base. In: Greenes RA, Peterson H, Protti D (eds). MEDINFO 95. Proceedings. Alberta: HC&CC, Inc. 1995; 13-17.