Medical Informatics Europe '97
C. Pappas et al. (Eds.)
IOS Press, 1997

ADM-INDEX: An automated system for indexing and retrieval of medical texts

SEKA L.-P.[a], COURTIN C.[a], LE BEUX P.[a]

[a] *Laboratoire d'Informatique Médicale, faculté de Médecine, Université de Rennes I, Avenue du Professeur Léon Bernard, 35043 Rennes cedex, France*

Abstract : ADM-INDEX is a system for indexing and retrieval of Patients Discharge Summaries (PDSs) by using linguistic methods (morphologic, syntaxic and semantic processing). The ADM-INDEX knowledge base is a restructuring of a diagnostic aid knowledge base (ADM) in order to allow the linguistic analysis of medical texts. The ADM system is a comprehensive medical knowledge base which has been developped since 1972 at the University Hospital of Rennes and which has been the first professional videotex medical diagnostic aid in France. After linguistic analysis, ADM-INDEX build the index table with thesaurus wording, medical words, concepts and phrases, unknown words contained in each PDS. The benefit of using those different elements is to improve information retrieval. Although our system is constructed with the ADM dictionary, it can be easily applied to other medical nomenclature or thesaurus. In this paper, we present on the one hand the ADM-INDEX knowledge base which is constituted by rules, a dictionary and a thesaurus, and on the other hand, the process of indexing and retrieval information.

1. Introduction

The semantic field of medicine is very wide and complex. A large number of its activities consists in producing medical reports written in natural language (Patients Discharge Summaries or PDS). PDSs describe the patients state of health. They are important documents in so far as they are firstly used for the patient's medical follow -up, and secondly as synthesis and self-teaching tool and also documents for communication, clinical research, epidemiological studies, evaluation of medical care.

Considering the abondance of information contained in each PDS, it is necessary to store and access them in a selective and judicious way through a quick and efficient system. This system will provide a significant help to physicians in accomplishing their task.

Elaborating such a system implies not only to solve problems concerning medical language (paraphrases, ambiguities..) [1] for this system must select the judicious concepts which would enable to represent the document's content, but also the use of medical nomenclature. The knowledge bases of medical decision systems such as INTERNIST [2], DXPLAIN [3], QMR [4], ADM [5] may be used as a starting point to build a corpus of entity and concepts.

We chose A.D.M. base (Diagnostic Medical Aid) as the core of our system because on one side it was developped in the Medical Informatics Laboratory, University of Rennes and on the other side it is both qualitatively and quantitavely rich (12.000 diseases, syndroms, undesirable effects and clinical forms by a 130.000 entities nomenclature, a 60.000 words dictionary). In spite of this abondance in terms, the ADM base does not lend itself to the analysis of medical texts (lack of syntactic and semantic information) which are written in natural language. ADM-INDEX is a restructuration of ADM base entity dictionary in order to adapt it to linguistic analysis of documents. We will show how we can detect concept and/or medical phrases, how we build index register and how we use it for information retrieval. We will end with the implementation and one of the applications of such a system which is necessary to have an automatic indexing and cross reference between our information and knowledge bases on a web server.

2. Material and methods

The knowledge base is composed of a dictionary, rules and a thesaurus. In a system as ours, the dictionary plays an essential role because it contains morphological, syntactic and semantic information. These information are useful to the different stages of analysis for the exact recognition of phrases. Besides, it must be a springboard for the inference and the deduction in so far as it will provide the necessary elements for starting the deduction and/or inference process. Inference and deduction are two important process for an indexing system using linguistic technics [6].

The dictionary : ADM-INDEX dictionary (around 60.000 words) essentially aims at detecting medical concepts and phrases whatever form it has in the text. Our dictionary inludes the ADM one in which words are classified by families (around 24.000 families). The links between the different words of a family are based on synonymy and inflexion. There are two categories of words : simple words and complex words which are subdivided into compound words and associated words.

- Simple words (around 45.000) : Among them, there are meaningless words (determiners,....) and non essential words (mostly adverbs).
- Complex words : they are composed by a group of simple words. There are two types of this kind of word.

. The Compound words (around 900) were created to avoid the dissociation of words. It enables to express a very strong link between its components on one side and on a other side to deal with synonymies between, for instance "Fievre jaune" (compound word) and "Amaril" (simple word). A compound word is very rigid. It neither accepts partial synonymy (synonymous of its components), nor allows a permutation in the order of its components.

. The Associated words (around 750) are like compound words but they are less rigid. Indeed, associated words enable to take into account synonymies between components.

We add syntactic and semantic information to the different constituents of ADM dictionary in order to make them more qualitative.

Constituents definition : There are three types of constituents which are concept[1], expressions and simple words. Each constituent will be defined according to the group of following elements.

. CODE_SIG	: code associated to the constituent
. CATEG	: indicates whether the constituent is a word, a concept or a expression,
. TYPE	: indicates whether the constituent is a void word or not
. [MEDIC]*[2]	: indicates whether the constituent is a medical term or not
. [CATEGRAM]*	: grammatical category of the constituent
. [CAT_SEM]*	: semantic category of the constituent
. [PERHIERAG]*	: hierarchical fathers according to the generic link
. [PERHIERAP]*	: hierarchical fathers according to the partitive link
. [MOT]	: constituent's word
. [LOC]*	: localization of the constituent in comparison with the others
. [PREC]*	: specified concepts categories
. [DEF]*	: definition of the constituent if necessary
. [OPP]*	: concepts or words to which it is opposed
. [IMP]*	: implied concepts
. [CAUSE]*	: causes that generates it

Figure 1 : Representation of ADM-INDEX dictionary constituents

We also use the operators /// and // to represent the different meanings and possible cases. The symbol /// represents the different cases which are mutually exclusive and the symbol // represent the different possibilities within each case. They are useful to take into account the different contexts in which a concept, an expression or simply a word is used. This makes the concept belong to several families at the same time although it is registered in only one.

---

1. A concept is a scientific or linguistic term which definition is well specified and which represents a group of objects or ideas.
2. * means that the rubric is optional

Their use prevent from having several inputs for a same word in the dictionary if there are several meanings for it. For instance, in French, the word "Secretaire" (Secretary) can have 3 meanings : *boss assistant*, *writing desk* and *of secretary bird*. This word will only be registered once in the dictionary. It will be represented as follows :

. CODE_SIG	: 27139000
. CATEG	: Concept
. TYPE	: non avoid
. CATEGRAM	: Substantive
. CAT_SEM	: human_being 1/// Animal 2/// Object 3
. PERHIERA	: human_being 1/// bird 2 /// furniture 3
. MOT	: Secretary
. LOC	: Administration 1
. DEF	: boss assistant 1/// see: secretary bird 2/// writing desk 3

What follows is the translation of this representation : when SECRETARY is used in the administration field, then it means the boss assistant; when it is used in an animal context, then it is the synonymous of "Secretary bird"; when it is used in a furniture context, then it means "writing desk". The numbers link the different characteristics with the concept according to the semantic category considered.

The dictionary's elements attribute to a given term in the dictionary, Morphological, syntactic and semantic information. *Morphological information* is considered within the families (a family contains each form of a word). The CODE_SIG links a word with a family. *Syntactic information* is indicated in the field CATEGRAM (substantive, verb, adjective, prefix, ...). *Semantic information* is taken into account for the use of compound and associated word, the links existing between the different concepts (CAT-SEM, CAUSE...), the explicit definition of some concepts, the use of /// and // operators.

This way to represent words and concepts will make it possible to better identify them in the texts. Moreover, using syntagms (compound and associated words) makes it possible to well specify the idea or notion which is expressed and to reduce the cases of polysemies. Besides, structuring the dictionary with the family system gives all the inflexion forms of a word. We do not need to make a special lexical treatment to identify the words. This has the advantage of accelerating the process.

We will detect the different terms through rules and transformations. Fives rules and three transformations (Permutation, Reduction and Substitution) will accomplish the appropriate treatments to detect the good terms. The definition and use of these three transformations can be justified by the fact that constituents of the dictionary are mainly in their minimal form.

The thesaurus : ADM-INDEX thesaurus wording are hierarchically organised. This organisation is mainly based on generic "IS_A " and partitive "PART_OF" relations. This term hierarchy is based on the definition of the different concepts. Wording hierarchy is very important in an indexing process in so far as it not only makes it possible for sons to inherit their fathers's properties but also to prefer exact concept to broad concept. Besides, within the thesaurus, similarity between terms is a necessary element because it reduces the silence[LI][1] risks when searching for information.

3. Implementation and results

Indexing : It is based on a certain number of process. Presentation of these process will be made according to the way they follow each other.

- The process of cutting and recognizing words. It enables to divide the text into sentences. Each sentence is then divided into words in order to carry out the recognition of each word. At this stage, the spelling of unknown words may be corrected.
- Syntactic and semantic analysis : this unit is composed of two secondary units
  - The syntactic "segmentor" which divides the sentences into parts (nominal groups) comparable to the dictionary headword. As regards medical texts, trying to use complete

[1] . Silence : the fact that nothing is proposed or not enough relevant answers are proposed when the base is consulted

syntactic analyser is illusive because texts often do not comply with natural language's grammar. Hence, the necessity to use a syntactic "segmentor". This segmentor is based on strong markers which are Verb, conjunction, preposition, predicative expression and punctuation sign. The benefit of using a syntactic "segmentor" is to reduce the unnecessary trials and avoid false concept recognition.

- The semantic conceptual analyser, from the sentence segmentation, detects concepts or phrases. Compound words are firstly identified, then associated words. Terms identification comes in this order because the links between the constituents of a compound word are stronger than those of associated words.

. compound word recognition : this stage will begin with the choice, within each nominal group, of a particular word called "*Principal*". The *Principal* belongs to the category of Substantive, Prefix or Adjective because these three categories are most likely to be in the first place in an expression. If there are several *Principals* in one group, priority depends on the rank of the *Principals* . The *Principals* accelerate the recognition process of compound words. They are used to consult the dictionary. Consulting aims at providing all headwords of compound words beginning with a given *Principal* or one of its inflexions. We will apply the compound word rule to this phrases list in order to only select the good terms.

. Associated word recognition : this recognition consists in consulting the dictionary with all words and compounds words (already detected) of the sentence. The result of the consulting will be composed of all associated words containing the word considered or one word of its family. We apply the different associated word rules to this phrases list in order to only select the good terms. When there are still isolated concepts, these latter will be replaced by their fathers in the sentence through generic and partititve link of the thesaurus in order to search for other possible associated words.

However, when two compounds or associated words are detected and if one of them is lexically contained in the other, the longest is selected because it is in general the most precise and it reflects better what has been expressed.

Once compound and associated words are detected, we will search for A.D.M. terminology[1] wording. Their detection will happen the same way as in the associated words detection because we consider wording as associated words.

• Index generation : The creation of index table will occur according to a method which results from an association of methods that already exist [7]. It consists in only keeping, as elements which could belong to the index table, concepts and/or medical terms, medical words, unknown words and very precise wording of the thesaurus ("son wording" are more precise than "father wording"). This not only compresses the index table but it also make far more significant matching. Including unknown words index table will produce "noise"[2] but we would rather have noise than silence. Each index is linked to a list. The list includes the text reference, the number of sentences in which the index appears with its nature.

Retrieval process : This process fistly capture the user's request and extracts all concepts, medical terms and thesaurus wording. Secondly, we create for each concepts and thesaurus wording a complete semantic consultation set through the similarity links in the thesaurus. We will consult the index table with these sets which allow to select all documents semantically close to each other through the consultation wording. Once documents are selected, they are given a rating according to the number and the nature of each index contained in each of them.

4. Applications and results

ADM-INDEX knowledge base representation is based on RDBMS ORACLE relational model. The system has been developped in PRO C. The actual system is used in ICONOWEB project [8] which contains around 3.500 clinical cases. ADM-INDEX is considered as the heart of the indexing-retrieval engine of the multimedia ICONOWEB project. ICONOWEB's main objective is to build multimedia clinical case documents and to put them at the disposal of medical students so that they can exploit them at best.

---

[1] . Terminolgy wording is in the thesaurus

[2] . Noise means proposing too many non judicious answers as solution when the base is consulted.

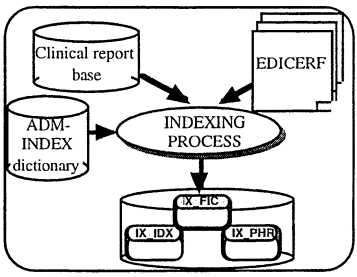

Fig 2: Indexing process in Iconoweb

ADM-INDEX is used for indexing two kinds of texts : the book of medical imaging references (EDICERF) and clinical cases with images and descriptions (ICONOCERF); each case is structured in chapters (History, Author, Context, Diagnostic....) and stored in a relational database. ADM-INDEX makes links between clinical reports and A.D.M. database, EDICERF books and A.D.M. database, clinical reports and EDICERF books. These links are very useful. Indeed, when you search information, the query process analyses conceptualy the request and shows (after a matching process) all clinical cases (by chapter : Diagnostic, Context ...), EDICERF books (by chapter : Pediatry ....) and ADM deseases which contain the query concepts. Thus, you can navigate easily on the Web server.

If a concept is present in the Context chapter, the system may present the report without diagnosis and commentaries and allow the user to propose a diagnosis. In this case, ADM-INDEX plays an important role. In fact, it makes a parallel between the user's answer and the concepts in the diagnostic and then gives an answer and a correction.

5. Discussion

Starting with an existing ADM entities dictionary has the advantage of covering pratically all medical fields and being complete. It takes into account the particularities of medical language and common sense language. Besides, the system is not only linked with A.D.M. terminology. It is possible to replace this terminology by another thesaurus or medical terminology. We have developped a tool which translate a terminology in ADM-INDEX formalism. Thus, we can adapt easily the system to the ICD9 and ICD10 (International Classification of Diseases), to SNOMED (Systematized NOmenclature of MEDecine) and to the MESH (MEdical Subject Headings), which all have french translation. Although, the system is based on french language, the translation of concepts given in the most important terminologies may give a powerful indexing and retrieval system on english texts.

In the extraction unit, the selection method reduces silence as much as possible. That is very important because the lack of document can be a great disadvantage if these documents contain judicious information. On the other side, it may increase the noise for general queries, which dœs not matter because the user may refine his queries with more precise concepts.

References

[1]. Ghazi Joseph: Vocabulaire du discours médical, structure, fonctionnement, apprentissage. Edition Didier Eudition, 1985

[2]. Miller R. A., M.D., Pople H. E., Jr., Ph.D., Myers J. D., M.D. : *INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine.* The New England Journal of Medicine. 1982; 307: 468-477.

[3]. Miller R. A., Masarie F. E., Myers J. D. : Quick Medical Reference (Q.M.R.) for diagnostic assistance MD Comput. 1986; 3: 34-48

[4]. Barnett G. O., M.D.; Cimino J. J., M.D.; Hupp T A., M.D.; Hoffer E. P., M.D. : *DXPLAIN An evolving diagnostic decision-support system* JAMA, July 3, 1987 (vol 258, N° 1)

[5]. Lenoir P., Riou C., Fresnel A.: *L'aide au diagnostic médical (ADM). Modalités et perspectives.* Médecine de l'homme N° 135.

[6]. Jayez Jacques: *L'inférence en langage naturel.* Ed. Hermès, Paris, 1988

[7]. Hersh W. R., Hickam D. H., Leone T. J.: *Words, concepts or both: optimal indexing units for automated information retrieval.* In Proceedings SCAMC 93, pp. 644-648, 1993

[8]. Duvauferrier R., Rambeau M., André M., Denier P., Le Beux P., Coussement A., Caillé J. M., Robache P., Morcet N. : *Iconothèques et ouvrages multimédia sur serveur et cd-rom en imagerie médicale (l'expérience française).* J. Radiol 1995; 76 (12) : 1079-85.