Spyns P, Ceusters W. Document Management in Healthcare: Presentation of the DOME Project. Proceedings of AMICE (Amsterdam Medical Informatics Congress Europe) Amsterdam, The Netherlands, Nov 27-29, 1995;:359-368.

Document Management in Healthcare: Presentation of the DOME project

Peter Spyns and Werner Ceusters
RAMIT vzw.
c/o Division of Medical Informatics
University Hospital Gent
De Pintelaan 185 (5K3), B-9000 Gent (Belgium)
Peter.Spyns@rug.ac.be

0. Abstract:

The ever evolving technology makes it possible nowadays to store large amounts of information electronically. However, the health sector, where enormous quantities of information are passed around as free text documents, does not (or at a limited scale) make use of the modern document management or information retrieval systems. Can the use of language technology [4,13,19,23] bridge the gap between the hospital staff in need of a user friendly and intelligent way to retrieve information and the already well developed area of large document handling systems [1] ? This paper presents the DOME project (MLAP #63-221), a preparatory study on the introduction of document handling systems enhanced with natural language technology in the field of Medicine. Special attention will be paid to a list of objectives retained for a future implementation of such a new generation medical record handling system.

1. Introduction

Paper records, as they exist today, are a result of more than a hundred years of evolution. The paper record systems have some strong points: they are highly portable and can contain almost any kind of information; text, numbers, codes, graphs and images. Almost anything a clinician may requires, can be put into it. The content can be structured in different ways according to local needs. The concept of a record is well known by all healthcare professionals in Europe who are likely to understand all parts of any record system with no or little education, as long as the language is understood. They know the "interface", how it is structured, where to expect information and what information to expect. But the paper record has also several weak points. It is only available to one person at a time, if available at all. Several studies have indicated that more than 20 % of the paper records in hospitals may be unavailable when needed. As the information content increases, the possibility to find a specific piece of information within a record also decreases, and large paper records are almost impossible to handle. In addition, the information is difficult to use for secondary purposes, including quality control, decision support, research, management, etc. Also, in some countries, there is a language obstacle. Record information may be transferred between professionals working with, and writing in, different languages.

There are several reasons for making a computerised healthcare record system: to increase the availability and usage of the information in the record; to facilitate healthcare professionals' operation; to improve the quality of the information as well as of health care; and to be able to use the vast amount of information collected in the healthcare records for the benefit of the individual person as well as of the whole medical community [14,28].

2. Presentation of the DOME project and consortium

DOME is an exploratory study for RTD work in the area of document handling. The focus of the work is a document management system for health care applications in the hospital context. The goal of the project is:

- to carry out a definition study of the target system.
- to provide a detailed plan for the development of this system (= future DOME 2 project)

The target system will exploit state-of-the-art language technology, and will strike a balance between practical and economical feasibility from the hospital's point of view, and completeness and robustness from the end-user's point of view. Although the target system will be a language engineering RTD system with a market potential, the main goal is to contribute to an improvement in health care, and the needs of the health care industry will be taken as a starting point. The project will concentrate on automating the processing of reports filled out by physicians who performed a certain procedure.

The automation will tackle two different aspects of this point.

- computational aids for processing of the reports. In the area of text processing the intention is to automate the extraction of the reports´ information to store it in a database or knowledge base where it can be accessed by consultants, researchers, hospital administrators, etc. This requires substantial language processing and knowledge engineering capabilities [22]. The major issue is to integrate such technologies in cost effective systems that are of use to health care establishments and that can enhance the quality of the health care.
- computational aids for the production of the reports. Tools that assist physicians and their staff in the fast, reliable production of possibly more standardised reports could increase the accuracy and usefulness of these reports. Specialised language checkers with sufficient background knowledge both about the domain and about the nature of an ´ideal´ report could check reports for completeness and consistency. Ideally, one would want to combine free data entry with quality checking. The study aims at identifying which combination of these techniques is relevant for satisfying the user needs.

The partners in the DOME project are:

- HCRC (Language Technology Group) University of Edinburgh
- Service d'Informatique Médicale, Assistance Publique - Hôpitaux de Paris (project coordinator)
- Irish Medical Systems, Dublin
- Centre d'Informatique Hospitalière, Hôpital Cantonal Universitaire de Genève
- RAMIT vzw, c/o Department of Medical Informatics, University Hospital Gent

The DOME consortium is the union of the main competence in Europe about NLP in the medical domain [8,16,26,29,31]. This is a unique occasion, by mutual potentiations of the partners, to develop a demonstrator of what is feasible today, to achieve a decisive proof of concepts in this domain and to really open a niche in the general market for medical document management systems. To achieve this purpose, already four deliverables have been written:

- D0.1 is an intermediate instanciation of the final report summarising more systematically and from different points of view the work reported on in the other deliverables. It also includes a detailed description of the service area and an estimation of the expected impact of the future implementation project. In annex, a list of document handling systems and vendors or suppliers is provided.
- D1.1 offers the general survey of the field:
    - a presentation of the various assessment methodologies used for the survey of the field
    - a description of the current situation in the concerned hospitals
    - an outline of the ideal system from an end-user's point of view, a synthesis of the market study
    - the synthesis of the wishes concerning the ideal document handling system expressed by the 4 user groups (Medical Advisory Boards) taking into account an assessment of current hospital practice.
    - some annexes providing more illustrative material and information.
- D1.2 contains the statements of objectives as they will be presented below (cf. section 4)
- D2 concerns the language engineering requirements:
    - description: a general overview of the proposed system
    - general requirements: a general overview of the proposed system
    - linguistic engineering requirements: a sketch of technology outwit the language engineering technology needed for this application
    - possible extensions: ways of extending the proposed application, and the technology needed for that.

3. Perspective

Although each observation is fixed and non-changeable, the patient record itself is a fast changing, complex document, especially during a hospital visit. Dealing with this type of document requires a different document handling technology.

Moreover, our study of representative sets of clinicians from five European countries revealed that the possibility of building, storing and retrieving integrated multimedia patient dossiers should be the long-term goal of any hospital management system. This adds further requirements on the authoring tools provided to clinicians, as well as on the storage, search and display facilities provided to users.

The best way to visualise this type of electronic health care record and a user's interactions with it is by visualising a WWW page, with information about a patient, and links to other reports, to raw lab results, to video clips of brains scans or recordings of irregular heart beats, etc. The project will closely follow developments in these areas by co-opting representatives of key vendors into its advisory board.

DOME 2 will instead be concerned with functions of multimedia document handling systems that were identified as crucial by our medical advisors and which can be improved through natural language engineering. These tasks will be presented in the following section.

4. Objectives of the DOME 2 project

4.1 Introduction

The DOME 2 project will design and implement a system which will offer a palette of user-level services in a variety of hospital environments, in a consistent way and in a proper interface to a variety of existing systems (Hospital Information Systems, Document Management Systems, Picture Archiving and Communication Systems, Ward Information Systems, Decision Support Systems, etc.). While the system will integrate available systems, the added-value will come from the inclusion of modular Natural Language Processing (NLP) based tools and services which will be realised gradually during the project.

The core of the system will be an advanced database system which supports the creation, storage and retrieval of information in different modalities [11]. Several possible candidate systems have already been identified and their suitability will be further evaluated in the course of DOME 1. Starting from a proper study of available systems will ensure that the development work on DOME 2 is compatible with established and emerging standards in the document handling industry --- especially document interchange standards like SGML, Hytime and DSSSL, and architectural standards like OLE and OpenDoc --- without tying ourselves too closely to one particular system.

The core system is best described from a user point of view as a multimedia, hypertextual patient record, including dynamic content-based retrieval facilities [6,7]. This system offers access to patient information in a structured and flexible way which best satisfies spontaneously expressed user demands.

4.2 Report-based coding assistance

Most reports entered into the patient's dossier have (international or national) medical codes associated with them. Assignment of these codes is mostly done by hand, by the clinicians in charge of writing the reports. This is time consuming; but, more importantly, it is often done inconsistently, thus defeating the purpose of the code assignment. Automatic code assignment would not necessarily increase the efficiency of the code assignment process, but would increase its consistency. Semi-automatic code assignment would be less time consuming, and would probably result in slightly higher consistency than in manual encoding .

For this medical coding, some text analysis is needed --- a shallow analysis for high-speed coding; a deep analysis for slower, more accurate coding [21]. At specified events (or on request), a coding program will examine the contents of the patient record to propose codes according to the enforced coding system (e.g., the International Classification of Diseases). This is based in particular on NL analysis of the contents of the reports.

4.3 NLP-based indexing and retrieval: content-based document search

Current document management systems allow for document search on the basis of (Boolean combinations of) diagnostic items in particular text fields. These items can be words or phrases and their thesaurus equivalents, with limited control for

morphological alternations (singular vs. plural) and soundex facilities (find a file on a patient called ``Meijer'' --- or Mayer, Meyer, Maya).

But this content-based search does require appropriately structured and annotated thesauri and other concept organisations. Many of these are available from other Medical Informatics projects and will be adapted for incorporation in the retrieval component. The use of more sophisticated NL analysis tools will enable better performance, in particular by relying on semantic and conceptual knowledge of the domain [12,18].

4.4 Expert-text component

It often happens that pertinent medical information items are not explicitly mentioned in reports because they are self evident by the context. It is the case, for instance, of implicit knowledge which can be inferred from the context through a pragmatic analysis. Some common sense reasoning is required, but full reasoning capabilities are too far from current state-of-the-art techniques. Improving hypertext in “expert-text" by adding inference capabilities to the future system will be studied. DOME 2 could provide a rule-based mechanism for basic inferences based on text contents, in particular time dependent resolutions.

4.5 Activity Management Board: data extraction

There are ever increasing requirements for hospitals and clinicians to provide performance indicators (how many beds occupied in a particular period; how many diabetics seen by which doctor; how many X-rays performed on behalf of the orthodental department) [27]. These data are cumbersome to collect. But with patient reports mapped into templates, the search and concomitant statistical analysis of such data becomes easier.

The NLP task involved here is similar to the message understanding tasks commonly used for evaluation purposes. The experience gained there for the construction of such templates will be used; anonymised corpora and templates prepared as part of this task will be made available to the wider research community. The language technology developed as part of projects like MENELAS [24] will be fine-tuned to this particular message understanding task.

4.6 Multilinguality

The DOME project is definitely oriented towards European use, meaning presently 11 languages. Multilinguality will be present from the design phase on. Any prototype will be made available in some of the following European languages: English, French, German and Dutch. This multilinguality will be achieved by combination of natural language analysis with natural language generation services.

There are several aspects to the users' request for polylingual or multilingual systems. A simple one is that doctors sometimes want to pass on reports and other textual material to colleagues elsewhere. A similar functionality is required when patients are treated in a country other than their own. Since one of the goals of the European EHCR is to guarantee mobility of patients throughout Europe, the ability to translate retrieved documents is an important one, but this functionality can be achieved by

means of a translation component which provides a rough-and-ready translation of the retrieved document.

A more complex aspect to multilinguality is that some hospital departments (e.g. in Belgium) are truly bilingual, with reports being written in French or Dutch, depending on the language of the patient rather than on the native language of the doctor. But even for a French speaking patient the lab may send a Dutch report to the doctor about that patient, and the results of that report will have to be incorporated in the final report in the patient's own language. This calls for the integration of documents of more than one language in a single indexing, storage and retrieval mechanism.

4.7 Authoring Tools: Quality Assurance of input reports

Much of the NL technology that needs to be developed for the indexing and retrieval tasks can also be used to provide quality assistance in the authoring of various reports - from spell checkers that know medical terms and names of relevant hospitals, to content checkers that know what type of information should be present in a document of a given type and can alert authors to missing items or can provide guidance in the form of automatic template provision for certain documents. Also, much of the information in hospital documents comes from other documents. Given flexible access to this other material when writing new documents was also identified by our user group as a high priority.

4.8 Authoring Tools

Automated Formatting and Link Creation for integrating new reports into the hypertextual patient record. The bottleneck in the application of a hypertext structure for the patient record is likely to be the creation of the hyperlinks at document entry. The authoring tools will suppress this bottleneck by relying on an analysis of text structure and contents to insert a new report into the current hypertext network.

4.9 Report entry by speech recognition (using commercially available systems)

Integration of speech input systems will be performed in languages and domains where commercial systems are available or can be extended by their manufacturers.

5. Conclusion

At evidence, despite the general market is expected to largely grow in the years to come, the health care sector for electronic document management systems is not yet as well developed. In no case the reason is that the needs are not so strong in this domain. On the contrary, user expectations for an electronic document management solution are really there. But why hospitals are not yet starting regular use of such systems ?

The main reason is certainly to be found in the fact that indexing and retrieval techniques in current use are not sufficient for the compulsion of medical records. Indeed, medical texts are strongly structured about their contents. The information they convey is expressed in a scientific language which encompass more than half of human words and concepts. They are too complicated to index by just a few indexes with a limited range of variation, because the retrieval in this situation is not precise enough to be considered as useful.

Full-text indexing techniques certainly improve over the manual indexing techniques. However, they are purely lexical approaches and their sensitivity is therefore limited. Synonym expressions, paraphrase and even metaphorical expressions are not rare in the medical language. Usage of modal expressions with suspicions, doubt, negations, etc. is frequent.

Definitively, the need is for conceptual indexing, dealing with multiple forms for any expression of a given fact or finding, transformed in a unique knowledge representation. Clearly stated, the following tools are necessary for such an approach. First, analysis of medical utterances expressed in plain natural language by physicians [15,20,25,32]. Second, transformation of the results of the analysis phase into a knowledge representation of the initial text [3,17,30]. For that purpose a model of the domain should be defined in terms of medical concepts and relationships. Such systems issued of advanced NLP techniques and inference systems in the field of knowledge representation are still in a R & D phase. They are not yet mature enough in order to receive the large industrial investment they need to enter definitively on the market.

This is the main target of the DOME project to bridge the gap from the present R & D situation to some industrial commitment and investment in this domain. There is little doubt about the potential market when mature solutions are ready. But there is a real challenge on deciding how and when the new software technologies are to be launched.

6. Bibliography

[1] Alpay L, Baud R, Lovis Ch., (1994), *Let's meet the users with Natural Language Understanding*, in Barahona P. and Christensen J.P., (1994), Knowledge and Decisions in Health Telematics, IOS Press, 103 - 108 .

[2] Ball M. and Collen M., (1992), Aspects of the Computer-based Patient Record, Springer - Verlag

[3] Baud R. Rassinoux, A.-M. and Scherrer, J.R., (1992), *Natural Language Processing and Semantical Representation of Medical Texts*, Meth. of Information in Medicine, 31: 117 - 125.

[4] Baud R. and Rassinoux A.-M. and Scherrer J.-R, (1992), *Natural Language Processing and Medical Records*, in [13], 1362 - 1367 .

[5] Beckers W. and ten Hoopen A. (eds.), (1994), Ontwikkelingen in de Medische Informatica (XIVde Medisch Informatica Congres), VMBI/TMI

[6] Ceusters W, De Moor G, Thienpont G, Lapeer L, Bonneu R, Schilders L, (1992), *Electronic Healthcare Records and Multimedia: modes of implementation*. in [10], 359-362.

[7] Ceusters W, Bonneu R., De Moor G., Lapeer R., Thienpont G. *The challenge of the nineties: bringing Multimedia Healthcare Records to life*. In: Reichert et al. (eds) Proceedings of MIE 1993, Freund Publishing House, 594-599, 1993.

[8] Ceusters W, (1994), The Generation of multi-lingual specialised lexicons by using augmented lemmatizer-taggers, Deliverable Report MultiTale #1.

[9] Ceusters W., Deville G., Streiter O., Herbigniaux E. and Devlies J., (1994), *A Computational Linguistic Approach to Semantic Modelling in Medicine*, in [5], 311 - 319 .

[10] Hoopen ten AJ, Hofdijk WJ, Beckers WPA (eds), Proceedings of MIC '92, Publicon Publishing Rotterdam, 1992 .

[11] Levy A. and Lawrance D., (1992), *Information Retrieval*, in [2], 146 - 152 .

[12] Lindberg D. and Humphreys B., (1992), *The Unified Medical Language System (UMLS) and Computer-based Patient Records*, in [2], 165 - 175

[13] Lun K., Degoulet P., Pierre T. and Rienhoff O., (1992), Seventh World Congress on Medical Informatics, MEDINFO 92, Geneva, North Holland

[14] Mc Donald C., (1992*), Physicians' Needs for Computer-based Patient Records*, in [2], 3 - 11 .

[15] Morel-Guillemaz (Rassinoux) A-M, Baud R., Scherrer J.R., (1990), *Proximity Processing of Medical Texts*, Proceedings of MIE 90, 625 - 630 .

[16] Rassinoux A.M., (1994), *Natural Language Processing of Medical Texts within the HELIOS Environment*, Computer Methods and Programs in Biomedicine, 45 Suppl.: 79 - 96 .

[17] Rector A.L., Solomon W.D., Nowlan W.A. and Rush T.W., (1994), *A terminology server for medical language and medical information systems*, in [19]

[18] Rossi Mori A., (1994), *Co-operative Development of a shared Ontology for Medicine in CEN/TC251/WG2*, in [19]

[19] Safran C., Chute C. and Scherrer J.R. (eds.), (1994), Natural Language and Medical Concept Representation, Vevey, (Preprints of the IMIA WG6 Conference)

[20] Sager N. Friedman C. and Lyman M., (1987), Medical Language Processing: Computer Management of Narrative Data, Addison Wesley, Reading, Massachussets

[21] Sager N, Lyman M, Nhan N.T. and Tick L.J., (1994), *Medical Language Processing: Applications to Patient Data Representation and Automatic Encoding*, in [19]

[22] Friedman C., Johnson S.B., (1992), *Medical Text Processing: Past achievements, future directions*, in [2], 212-228

[23] Scherrer J.R., Coté R.A. and Mandil S.H. (eds), (1989), Computerized Natural Medical Language Processing for Knowledge Representation, North Holland

[24] Spyns P. Zweigenbaum P. and Willems J.L., (1992), *Representation and Extraction of Information from Patient Discharge Summaries by means of Natural Langage Processing*, in [10], 309 - 316, [in Dutch]

[25] Spyns P, and Willems J.L., (1994), *Natural Language Analysis of Dutch Medical Discharge Summaries: a prototype*, in [5], 291 - 300, [in Dutch]

[26] Spyns P. and Willems J.L., (1995*), Dutch Medical Language Processing: discussion of a prototype*, Proceedings of MEDINFO 95, (to appear)

[27] Ullian E., (1992), *Hospital Administrators' Needs for Computer-based Patient Records*, in [2], 30 - 35 .

[28] Van Ginneken A.M., (1994), *Computer Based Patient Records (*Synopsis), in Yearbook of Medical Informatics 94, 173-175

[29] Whittemore G., (1994), *The MENELAS English Natural Language Understander: Natural Language Understanding in the medical domain*, in The first World Congress on Computational Medicine, Public Health and Biotechnology, Austin, Texas

[30] Zweigenbaum P., Bachimont B., Bouaud J., Charlet J., Boisvieux J.F., (1994), Issues in the Structuration and Acquisition of an Ontology for Medical Language Understanding, in [19]

[31] Zweigenbaum P. et al., (1994), *MENELAS, an access system for medical records using natural language* , Computer Methods and Programs in Biomedicine.

[32] Zweigenbaum P. et al., (1995), *The MENELAS project (#A2023),* in The AIM Yearbook (forthcoming)