ELSEVIER

Computer Methods and Programs in Biomedicine 45 (1994) 117–120

computer methods and programs in biomedicine

MENELAS: an access system for medical records using natural language

Pierre Zweigenbaum[1]

DIAM — INSERM U.194 and SIM, Service d'Informatique Médicale, Assistance Publique — Hôpitaux de Paris, 91, boulevard de l'Hôpital, F-75634 Paris Cedex 13, France

Abstract

The overall goal of MENELAS is to provide better access to the information contained in natural language patient discharge summaries, through the design and implementation of a pilot system able to access medical reports through natural languages. A first, experimental version of the MENELAS indexing prototype for French has been assembled. Its function is to encode free text PDSs into both an internal representation and ICD-9-CM nomenclature codes. A preliminary evaluation shows the potential for reasonable coverage and precision. The MENELAS prototype will be enhanced and extended into a pilot system which will be tested in two hospital sites.

Key words: Natural language processing; Patient record; Knowledge-based systems; Information retrieval; ICD-9-CM; Conceptual graphs

1. Introduction

The dramatic expansion of health services over the past decades has created an information explosion for the health profession and a growing demand for detailed record-keeping. In response to this, paper charts are being slowly replaced by computerized medical records developed within Hospital Information Systems. The latter were developed primarily to handle numerically based information. However, a number of applications such as ongoing patient care, evaluation of health care and clinical research require information which is predominantly available in narrative patient discharge summaries (PDSs). This narrative is a widely available, low cost source of reliable information. However, with the current technology, there is no way to directly access the information contained in a free text format: a new sort of system is thus required to analyse PDS texts using natural language processing techniques.

2. Background

The automatic analysis of medical reports has been widely studied [1–3]. Whereas most early methods were essentially syntax-based, recent approaches focus on the semantic representation of medical texts [4–6]. They enable a deeper level of understanding, including the representation of in-

[1] This paper is written by P. Zweigenbaum on behalf of the consortium MENELAS.

0169-2607/94/$07.00
SSDI 0169-2607(94)01557-V

formation that was implicit in the texts but is evident for the target readers [6].

This project builds on previous work performed by the different partners either in the AIM Exploratory Action [7] or independently on the analysis of French, English and Dutch [8–10].

3. Design considerations

The overall goal of MENELAS is to provide better access to the information contained in PDSs through the following services:

- The management staff of the medical unit may consult its own electronic activity board.
- The clinical staff may also access the stored PDSs to retrieve a set of patients having specific characteristics, which allows some form of case based reasoning.
- The same facility allows a limited test of research hypotheses, on the basis of the available patient sample.
- Nomenclature codes are produced automatically for each PDS. This contributes to one of the AIM Tasks, 'data standardisation, classification and encoding.'

These services will be implemented as modular subsystems, which can be realised gradually. The test medical domain is coronary diseases.

The basis of the project is the design and implementation of a pilot system able to analyse the contents of medical reports and to store them in a database as a set of conceptual structures which represent their 'meaning'. These structures, which may be seen as the core representation of the narrative, may then be consulted to retrieve specific information contained in a PDS. An index is also produced according to the ICD-9-CM nomenclature. The twelve partners of MENELAS represent three linguistic groups: French, English and Dutch. Language-dependent components are connected to the rest of the system through well-specified interfaces in order to facilitate extension to other natural languages.

MENELAS adopts a knowledge-based approach to natural language understanding, and relies on a large body of linguistic and medical knowledge to perform its task. All semantic and domain knowledge is expressed in a common, general representation called conceptual graphs [11], which has already been used in medical systems [4,5,12].

4. System description

MENELAS is organised around three main systems:

- *The document indexing system* analyses a PDS and stores it in a database as a set of conceptual structures. It also produces the nomenclature indexes. This process can be run in batch mode.
- *The consultation system* allows physicians and administration staff to access PDSs by their contents through a user-friendly interface.
- *The administrator system* enables the customization of the knowledge base of the system in order to set up a new application or to maintain an existing application.

These systems are implemented in PROLOG, LISP, C++ and C, and run under Unix (currently, on SUN Sparc); the consultation system user interface runs on a Personal Computer with Windows.

5. Status report

The first phase of the project is now terminating, and a prototype version of most MENELAS components is available. The indexing prototype, which constitutes the core of the MENELAS prototype, has been integrated for French. The indexing system follows the standard division of natural language processing systems: morpho-syntactic, semantic and pragmatic analysers; the latter is language-independent.

The performance of the prototype was explored in two dimensions. Along one dimension, a subset of the functions of the indexing system for English were tested on a whole 475 English PDS corpus from the Royal Victoria Hospitals in Belfast. This evaluated the breadth of coverage that can be reached: >73% of the sentences in the corpus could receive a full syntactic parse. A similar

experiment on 56 French PDSs (Hôpital de la Timone, Marseilles, and Pitié-Salpêtrière, Paris) showed a comparable result of 70% parsed sentences.

Along the other dimension, we set up a method for evaluating globally the near-whole set of functions of the indexing system for French. The protocol is inspired from the third Message Understanding Conference [13,14]. The principle of this evaluation consists in comparing the behaviour of the system with the one of a human reader, disposing of the same texts, and performing the basic tasks of MENELAS: information retrieval and nomenclature code generation.

The first phase of the project mainly involved developing and tuning this evaluation procedure, with a restricted three PDS sample from the French PDS corpus. The procedure allowed us to monitor the progress of system performance on these PDSs while the lexicons and knowledge bases were developed. This gave an idea of the depth and precision of treatment that can be performed by the system, and of the time spent per PDS on knowledge development. Given appropriate knowledge, the system could correctly identify a set of 11 specific information items in a test PDS, as well as the six ICD-9-CM codes that a human coder had assigned to this PDS. During the second phase of the project, the procedure will be applied to a larger number of texts.

6. Lessons learned

Both sides of the evaluation have brought insights into the methodologies to apply in order to obtain better coverage and precision. On the one hand, even if we estimate that the percentage of fully parsed sentences could at best be raised to a ceiling of 80%, recovery methods that can process additional sentences will be useful. In the current design, partial parses from the French morpho-syntactic analyser can already be passed on to the semantic and pragmatic components, which can make sense of them independently.

On the other hand, what needs to be enhanced now is mainly the quality and volume of the lexicons and knowledge bases. The knowledge development process has proven to be time-consuming and error-prone, and a more precise methodology is being set up. Tools will also be needed to facilitate this process. The evaluation procedure is an important asset; two complementary directions are being explored: the development of consistency-checking tools, and the semi-automated reuse of available on-line knowledge resources [15-17].

7. Future plans

The rest of the project is divided in two sections. Phase 2 will complete and enhance components, knowledge bases and integration (first half of 1994). Emphasis will be put on the above-mentioned points. Phase 3 will evaluate the resulting pilot system with users in two hospital sites (second half of 1994).

8. Information about deliverables

Deliverable #9 on the first version of the MENELAS 'Linguistic and Medical Knowledge Bases' is available from the author.

Acknowledgements

We would like to thank the members of the AIM team of the European Commission, in particular Luciano Beolchi and Jens P. Christensen, for their help in starting and driving this project.

References

[1] Naomi Sager, Carol Friedman and Margaret S. Lyman, editors. Medical Information Processing — Computer Management of Narrative Data. Addison Wesley, Reading, MA, 1987.

[2] J.R. Scherrer, R.A. Côté and S.H. Mandil, eds. Computerised Natural Medical Language Processing for Knowledge Engineering, Amsterdam, 1989. North-Holland.

[3] F. Borst, N. Sager, N.-T. Nhàn, Y. Su, M. Lyman, L.-J. Tick, C. Revillard, E. Chi and J.-R. Scherrer. Analyse automatique de comptes rendus d'hospitalisation. eds, P. Degoulet, J.-C. Stéphan, A. Venot and P.-J. Yvon, Informatique et Gestion des Unités de Soins — Comptes Rendus du Colloque AIM-IF, Informatique et Santé, chapter 5, pp. 246-256. Springer-Verlag, Paris, 1989.

[4] R.H. Baud, A.M. Rassinoux and J.R. Scherrer. Natural language processing and semantical representation of medical texts, Methods Inform. Med. 31 (1992) 117-125.

[5] Martin Schröder, Knowledge-based processing of medical language: A language engineering approach. In Proceedings of GWAI'92. ed, D. Bonn, September 1992.

[6] Marc Cavazza, Laurent Doré and Pierre Zweigenbaum, Model-based natural language understanding in medicine. eds, K.C. Lun, P. Degoulet, T. Piemme and O. Rienhoff, Proc MEDINFO '92, pp. 1356–1361, Amsterdam, 1992. North Holland.

[7] Pierre Zweigenbaum, Marc Cavazza, Laurent Doré, Jacques Bouaud and David Sedlock, Natural language processing of patient discharge summaries (NLPAD) — extraction prototype. eds, J. Noothoven van Goor and J. Pihlkjaer Christensen, AIM Reference Book, pp. 277–286. IOS Press, Amsterdam, 1992.

[8] A. Bérard-Dugourd, J. Fargues, M.-C. Landau and J.-P. Rogala. Un système d'analyse de texte et de question/réponse basé sur les graphes conceptuels. eds, P. Degoulet, J.-C. Stéphan, A. Venot and P.-J. Yvon, Informatique et Gestion des Unités de Soins — Comptes Rendus du Colloque AIM-IF, Informatique et Santé, chapter 5, pp. 223–233. Springer-Verlag, Paris, 1989.

[9] C. Grover, J. Carroll and T. Briscoe, The Alvey Natural Language Tools Grammar. University of Cambridge, Computer Laboratory, Cambridge, 1992. 4th Release.

[10] Peter Spyns and Geert Adriaens, Applying and improving the restriction grammar approach for Dutch patient discharge summaries. ed, Antonio Zampolli, Proceedings of the 14th COLING, pp. 1264–1268, Nantes, France, July 23–28 1992.

[11] John F. Sowa, Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, London, 1984.

[12] K.E. Campbell and M.A. Musen, Representation of clinical data using SNOMED III and conceptual graphs. In SCAMC [18], pp. 354–358.

[13] Nigel Strang, An evaluation methodology for Menelas. Rapport de DEA, Informatique Médicale, Université Paris VI, 1993.

[14] Wendy Lehnert and Beth Sundheim, A performance evaluation of text-analysis technologies. Artificial Intelligence Magazine, (Fall), pp. 81–94, 1991.

[15] B.L. Humphrey and D.A.B. Lindberg, Building the Unified Medical Language System. In Proc of SCAMC'89, pp. 475–480. IEEE, 1989.

[16] Françoise Volot, Pierre Zweigenbaum, Bruno Bachimont, Mohamed Ben Saïd, Jacques Bouaud, Marius Fieschi and Jean-François Boisvieux, Structuration and acquisition of medical knowledge: Using UMLS in the Conceptual Graph formalism. In Proc of SCAMC'93. Mc Graw Hill, 1993.

[17] George A. Miller, Wordnet: An on-line lexical database. Int. J. Lexicography, 3 (4), 1990.

[18] Proc of SCAMC'92. McGraw Hill, 1992.