Menelas: Coding and Information Retrieval from Natural Language Patient Discharge Summaries*

Pierre Zweigenbaum (Ed.), Consortium MENELAS *
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
pz@biomath.jussieu.fr

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 analyse medical texts. This paper summarises the developments performed for this purpose during the project. It stresses the fundamental issue of the normalisation of medical information into a common, language-independent representation.

The project has resulted in a collection of components, knowledge bases and systems for processing natural language patient discharge summaries. It has also produced methodological results on the design of a normalised representation of medical information and its construction from textual input. Further developments and applications have already started in the context of two other European Union projects.

1. Introduction

Modern medicine uses an increasing quantity of information. This is due to the emergence of new, diverse and complementary investigation techniques (imaging, etc.), and mostly to growing numbers of patients treated for chronic diseases in the developed countries. The double goal of enhancing the quality of health care and of controlling costs implies a need for rapid access to all information concerning a patient’s history. This creates an increasing demand for better management of medical information resources, and of patient records. In the case of pathologies that are sufficiently limited and well standardised in their treatment, computerised medical records, developed in the framework of Hospital Information Systems, are being used. However, they require users to adhere strictly to a fixed, closed set of items to describe patient information.

Nevertheless, a number of applications such as ongoing patient care, evaluation of health care and clinical research use information which is predominantly available in narrative patient discharge summaries (henceforth PDSs), in the form of “free text.” These reports synthesise the relevant information concerning each patient, and contain

*Published in “Advances in Health Telematics, Maria Fernanda Laires, Maria Júlia Ladeira and Jens Pihlkjær Christensen (eds.), IOS Press, Amsterdam, pp. 82–89, 1995.”
a wealth of data regarding patient history, current diagnoses, and current and suggested subsequent treatments. They constitute a widely available, low cost source of reliable information, which is flexible and can adapt to medical progress. However, available industrial technology does not allow access to this textual information. A new sort of system is thus required to extract medical information from PDS texts and make it available for further processing. The MENELAS project aims to push forward the European research and development effort towards the achievement of such a system.

2. Objectives

The overall objectives of MENELAS fall into two categories: target, user-level services which will be created or enhanced by extracting medical information from free text, and advances in the enabling medical language processing technology. The larger issues and effort needed lie in the development of the enabling technology. Nevertheless, the project has taken care to include user-level service demonstration modules in its design. This helped to show the end-user benefits of the work and to focus the R&D work on the strategic points to be addressed to achieve such user-level services.

2.1. User-level services

Target, user-level services are organised around two main axes (Figure 1):

Coding: the automatic assignment of codes to each PDS. In most European countries, each patient stay must be indexed with codes from national or international nomenclatures and classifications. This coding task is imposed on health care professionals for external reasons, and constitutes extra work with little direct benefit. In this context, such a service can save much time and effort to them. In MENELAS, the production of ICD-9-CM codes from PDS texts has been addressed.

Information retrieval. Patient cases can be retrieved on the basis of the contents of the discharge summary. This may be used for clinical and research purposes: (i) the
clinical staff may access the stored PDSs to retrieve a set of patients having specific characteristics, which allows some form of human case based reasoning; (ii) the same facility allows a limited test of research hypotheses, on the basis of the available patient sample. The analysed contents of the texts may also be used to feed activity management tools.

These services can be implemented as modular subsystems, and be realised gradually. The project has focussed on the production of ICD-9-CM codes and a query-based retrieval functionality. The test medical domain is coronary diseases.

2.2. Enabling technology

At the technical level, the goals of the project were to adapt and develop advanced text analysis techniques to build a canonical representation of the medical information contained in PDSs. The key points of the chosen approach were the following.

Normalised core representation. The relevant medical information contained in the texts is extracted and structured according to a normalised, conceptual representation. This representation is abstracted away from its original linguistic form: it is robust with respect to the variability and flexibility of language and across different languages. It provides a normalised, systematic informational basis to support smoothly a palette of user-level services: classification or nomenclature codes (here, ICD-9-CM) can be extracted from it; queries can be matched against it to retrieve specific information contained in a PDS. A direct benefit is that the components which implement these services do not need to have any knowledge of the initial natural language used in the input text, nor do they have to cope with synonymy or paraphrase. As a consequence, this representation and its construction constitute the focal point of the project.

Full text processing. PDS texts may contain both full prose paragraphs, with grammatically complex sentences, and shorter noun phrase statements. Non-trivial syntactic phenomena, such as embedded controlled clauses or negation, can have a substantial impact on text meaning. It is therefore difficult to dispense from taking them into account. Parsers with advanced syntactic knowledge, which attempt full sentence parsing, are therefore useful here — and are more and more available in various languages.

In-depth, knowledge-based approach. 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 formalism called conceptual graphs [1]. The medical knowledge needed depends on the application domain.

Common architecture for multiple languages: maximise language-independent, shared components and knowledge bases. The twelve partners of MENELAS represent three linguistic groups: French, English and Dutch, addressed during the project. The core representation, and as a result, the code generation component and the retrieval component, are language-independent, as well as the medical knowledge description they share. Language-dependent components are connected to the rest of the system through well-specified interfaces in order to facilitate extension to other natural languages.

Reuse and adapt existing linguistic resources as much as possible. The project builds on previous work performed by the different partners either on medical language processing, both in the AIM Exploratory Action on French and independently on Dutch [2, 3], or on the analysis of general French and English [4, 5]. As a result, adaptation and integration of these independently developed components was an important task in the project. New developments have been performed where necessary.
3. Project Design and Development

3.1. Background

The automatic analysis of medical reports has been widely studied [6, 7]. Whereas most early methods were essentially syntax-based, recent approaches focus on the semantic representation of medical texts [8, 9, 10]. They enable a deeper level of understanding, including the representation of information that was implicit in the texts but is evident for the target readers [10]. Conceptual graphs are gradually emerging as a knowledge representation standard in medical systems [8, 9].

3.2. System description

MENELAS has a two-part organisation (Figure 2): analysis system and user-level services. In addition, administrator tools help to customise system knowledge bases.

Analysis System. The heart of MENELAS is the document analysis system, which analyses a PDS and stores it in a database as a set of normalised, conceptual representations. These representations hold the informational basis for the target services.

Target Services. While a large range of services could exploit the information extracted by the analysis system, the project has concentrated on two services: (i) the coding service produces nomenclature codes; (ii) the consultation service allows physicians and management staff to retrieve PDSs that satisfy content-based criteria.

3.3. Issues Encountered and Revised Design

The technical issues which arose during the project were of several kinds. On the one hand, the importance of some tasks was found to be even greater than originally estimated. This was the case of the specification of the information to be extracted from the PDSs (target information) and of the development of a representation of the medical concepts of the domain (ontology, knowledge base). This led to the design of new methodological principles to support such development (see section 4.2.). Interprocess communication issues in the integration task were also underestimated.

On the other hand, some techniques initially included in the system design proved unapplicable, and were replaced by better ones. As the first experiments were conducted, the methods used to build a canonical representation of the medical information contained in PDSs were found to impose too complex constraints on the development...
of semantic lexica. The question of the normalisation of the representation of medical information then became a central concern in the project (see section 4.2.).

Finally, the original architecture needed to be completed on some aspects. The addition of a PDS preprocessor now cleans up the texts and performs some fast and useful pre-syntactic operations before actual linguistic processing takes place.

Too strong inter-partner and inter-workpackage dependencies were also uncovered. The new semantic normalisation method and the reuse of a semantic analyser available for English both helped to reduce them.

Figure 3 shows the final architecture of the analysis system and coding service.

4. Results

The project has produced results of two types: software results, and methodological results, shortly described below.

Preliminary evaluation shows that the existing prototype displays promising results for automatically encoding PDSs into an existing classification such as ICD-9-CM and for information retrieval from natural language PDSs. Partial tests (up to syntactico-semantic analysis) could be performed on a large PDS sample (475 English texts and 100 French texts). In-depth tests of the whole analysis chain were performed on a smaller, 6 text sample, representing 75 sentences of up to 23 words. They consisted in comparing the user-level service results (coding and query answering) obtained by the system with the performance of health care professionals performing the same tasks: given a PDS text, assign codes to it and answer a fixed questionnaire. Results were satisfactory, but
must be confirmed by a larger-scale evaluation, which is currently being conducted.

4.1. Software results

From a software development point of view, a set of components, tools, and knowledge bases have been built. Given space constraints, we can only list them here.


The components have been implemented in PROLOG, LISP, C++ and C, and run on Unix; the consultation service user interface runs on a PC with Windows. Medical knowledge bases are written in the Conceptual Graph formalism.

4.2. Methodological results

A methodology for developing ontologies. The basis for the representation, and thus the processing, of medical information, is a structured description of medical concepts: an ontology. Such a fundamental description needs to be sound and principled: one needs to follow clear, guiding principles in its design. This issue is all the more important as it is both necessary and applicable to the full spectrum of medical information management, from patient databases to knowledge-based systems. A set of principles for ontology design have been produced out of the MENELAS experience [17]. They were followed to structure a more than 1500 concept ontology, on top of which over 450 knowledge models could be described.

A methodology for semantic normalisation. We have stressed the opposition between the variability of natural language and the normalised, systematic character of the representation built. The project has designed and implemented a powerful, heuristic method which ensures that all representations are built conform to the norm. The method consists in projecting linguistic relations to the normalised model of the domain, selecting the conceptual path which best links the corresponding notions. As a result, it can assign identical projections to linguistic paraphrases. At the same time, it can discriminate between linguistically similar expressions which have different meanings. Since much of its performance relies on the content and structure of the knowledge base, the methodological work on ontology design is again crucial here.

5. Conclusion

5.1. Lessons learnt.

The experience of the project sheds some light on the technical objectives set at the beginning of the work (see section 2.2.). Whereas the initial aim was to build a normalised core representation that would be as general as possible, we found out that the specification of a such a representation necessarily depends to a fair extent on the client services and the target information they require; as a result, the specification of this target information helps to design the normalised representation and system knowledge
bases. The large coverage of the full text processing parsers used was no doubt very useful; their robustness to agrammatical sentences could be improved by extending their recovery procedures beyond the ones already available in these components. The in-depth, knowledge-based approach is really linked to the task: representing correctly and univocally the target information, starting from the variability of natural language; the complexity of this approach cannot be lowered without changing this task. The common architecture for multiple languages successfully hosted the French and Dutch components, as specified in the original design, as well as the English components, connected at a different level of the analysis chain. One should not underestimate the cost of adapting and tailoring reused existing resources; however, project partners could successfully reinject parsers with large coverage lexica and grammars with a benefit/cost ratio much better than would have resulted from redescribing linguistic knowledge anew for each language addressed.

5.2. Perspectives and Future Work

Facilitating the production of a normalised encoding of patient medical information from national languages is a key enabling factor for harmonisation across Europe.

In the short term (1994–5), a first reuse of the MENELAS prototype is taking place at Lille, in Northern France, within the framework of the AIM integration project ISAR.

Apart from this, a core of project partners, in collaboration with partners from other consortia of both the health care (AIM) and the language engineering (LRE, MLAP) Telematics Community programmes, have jointly launched a feasibility study, including a market analysis, which will identify how MENELAS-like systems can be integrated into the electronic document processing flow of hospitals (MLAP project DOME).

Acknowledgements

We would like to thank all the hospital departments which provided us corpora, as well as 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


**Main Project Participants**

Jean-François Boisvieux, Pierre Zweigenbaum, Bruno Bachimont, Jacques Bouaud, Jean Charlet, Mohamed Ben-Said (INSERM U.194 and SIM, Service d’Informatique Médicale AP-HP, Paris, F), Pierre Dujols (GLIM, Groupe Linguistique, Informatique et Médecine, Montpellier, F), Marius Fieschi, Françoise Volot, Nigel Strang (CERTIM, Service de l’Information Médicale, Marseille, F), Pierre Le Beux, Denis Delamarre, Anita Burgun (LIM, Laboratoire d’Informatique Médicale, Rennes, F), Mark Keane, Brenda Nangle, Stephen Flinter (TCD, Trinity College Dublin, Dublin, IRL), Brendan McAllister, Stephen Mc Namara, Kevin O’Sullivan, Roy Johnston, Margaret McDermott (IMS, Irish Medical Systems, Dublin, IRL), Bridie Kelly (RVH, Royal Victoria Hospital, Belfast, UK), Jos L. Willems†, Peter Spyns (KUL, Katholieke Universiteit Leuven, Leuven, B), Antoine Ogonowski (ERLI, GSI-Erli, Charenton,