The SINAMED and ISIS Projects: Applying Text Mining Techniques to Improve Access to a Medical Digital Library

Manuel de Buenaga[1], Manuel Maña[2], Diego Gachet[1], and Jacinto Mata[2]

[1] Universidad Europea de Madrid – Escuela Superior Politécnica
28670 Villaviciosa de Odón, Madrid, España
{buenaga, gachet}@uem.es
[2] Universidad de Huelva – Dpto. Ing. Electrón., Sistemas Informáticos y Aut.
Escuela Politécnica Superior
21819 Palos de la Frontera, Huelva, España
manuel.mana@diesia.uhu.es, mata@uhu.es

Abstract. Intelligent information access systems integrate text mining and content analysis capabilities as a relevant element in an increasing way. In this paper we present our work focused on the integration of text categorization and summarization to improve information access on a specific medical domain, patient clinical records and related scientific documentation, in the framework of two different research projects: SINAMED and ISIS, developed by a consortium of two research groups from two universities, one hospital and one software development firm. SINAMED has a basic research orientation and its goal is to design new text categorization and summarization algorithms based on the utilization of lexical resources in the biomedical domain. ISIS is a R&D project with a more applied and technology-transfer orientation, focused on more direct practical aspects of the utilization in a concrete public health institution.

1 Project Goals

The SINAMED and ISIS projects are focused on information access on a specific biomedical domain: patient clinical records and related scientific documentation. These two projects have a strong interrelation and also different and complementary orientation. The SINAMED[1] project has a main orientation of basic research, focused on the design and the integration of automatic text summarization and categorization algorithms to improve access to bilingual information in the biomedical domain. The ISIS[2] project has a more applied and technology-transfer orientation, and its aim is the improvement in the intelligent access to the medical information, having in mind

[1] The research described in this paper has been partially supported by the Spanish Ministry of Education and Science and the European Union from the European Regional Development Fund (ERDF) - (TIN2005-08988-C02-01 and TIN2005-08988-C02-02).

[2] The research described in this paper has been partially supported by the Spanish Ministry of Industry and the European Union from the European Regional Development Fund (ERDF) - (FIT-350200-2005-16).

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 548 – 551, 2006.

doctors and patients as end users. It is focused on providing advanced and more effective tools than the current ones for the search, localization, use, and understanding of different sources of medical information.

2 The Medical Domain

The medical information is voluminous, heterogeneous and of extreme complexity. One of the factors with a major repercussion in the heterogeneity of the medical content is the source diversity. Each source (scientific papers, databases of summaries, structured or semi-structured databases, Web services or clinical records of patients) has several features. For example, the existence or not existence of an external structure for the document, the occurrence of free text together with structured data (tables with clinical results) or the length of the documents. These differences in domain, structure and scale hinder the development of robust and independent systems that facilitate the access to this kind of content.

Medical Documentation: Considering, for instance, the scientific medical articles, there are thousands of scientific journals in English language, and the problem grows if we consider other languages and other sources. Medline, the most important and consulted bibliographical database in the biomedical domain, constitutes a main example. Medline contains more than 13 million references, with an increment between 1.500 and 3.500 references per day. This huge volume of articles makes the experts difficult to take advantage of the whole published and interested information.

The Patient Clinical Record: The patient clinical record is defined as the set of documents (data, assessments and other type of information) that are generated throughout the assistance process of a patient. The system of clinical record sheets presents many drawbacks (unreadable information, chaos, absence of consistency, questionable availability, uncertain confidentiality guarantee, damage in the documents,…) that could largely be corrected with the usage of electronic clinical records. Some of the advantages of the electronic clinical record are: a better accessibility to the information and an improvement in the confidentiality; data homogenization; prescription filled in an automatic way; overall view of the patient; coordination of medical treatments; gathering of the whole information of a patient.

The combination of a scientific information system with the electronic clinical record would help doctors to make decisions, to decrease the mistakes and the clinical variability and to increase the patient's safety.

3 Text Mining Techniques

In the projects that we present in this paper, we propose to integrate text categorization and summarization techniques into the searching and browsing processes. We expect that a better organization could help users to feel less overwhelmed by the amount of information and to get a better understanding of the information available in the retrieved documents [1, 2].

Text Categorization: Automated text categorization can be applied, for example, to catalogue medical reports using standards descriptors, as the Medical Subject Headings (MeSH). However, the language variability and the lack of the needed data for an effective learning limits the effectiveness of these systems. Also, text categorization has rarely been applied in biomedical environment [3, 4] and the use of this technique on medical information writing in Spanish is virtually nonexistent.

The mentioned problems can be dealt with the use of lexical semantic resources. The techniques presented in these works are specially applicable to the medical information, since there are available specific resources as the Unified Medical Language System (UMLS).

Text Summarization: In information access environments, summaries (single-document or multi-document) have proved its utility, improving the effectiveness of several tasks, as ad hoc and interactive retrieval.

The application to the medical domain is fraught with a variety of challenges which do not had been dealt sufficiently in previous works [5, 6]. Among them, we stand out the following problems. The great part of the summarization systems handles documents wrote in a single language (English, fundamentally), although there are innumerable text collections and resources in other languages (Spanish, specially). Also, most of the systems has been conceived to deal with a restricted subdomain. Therefore, it is necessary to develop techniques that could be applied to broader domains or, at least, that could be easily adaptable from a subdomain to another. As in automatic categorization, we think that the integration of knowledge from resources as UMLS, which has some bilingual components, can play a key role in both problems.

4 The Projects

The SINAMED Project. propose the introduction of original and relevant improvings in the techniques and algorithms, and the specialization and adaptation needed for the specific application environment and the processing of bilingual information (English/Spanish). We are developing an environment for application and experiment of adequate dimension, working with documents of the biomedical domain: Medline, MedlinePlus/HealthDay (English/Spanish) and TREC/Genomics track. This environment integrates text analysis techniques developed with search tools facilitating the information access to the specific user needs. An evaluation of the application environment of each one of the different elements integrated according to general and specific standards of information retrieval, just like the ones used in TREC, and of the concrete operations of text categorization and summarization will be carried out.

The ISIS Project. aims to improve the intelligent access to the medical information, having in mind doctors and patients as end users. It is focused on providing advanced and more effective tools than the current ones for the search, localization, use, and understanding of different sources of medical information. Some interesting aspects are the integrated access to patient's clinical record and related health information.

We intend that, both doctor and patient, exploit the methods and techniques of text mining and intelligent analysis of document's content.

The main scientific and technological objectives of the project are organized around topics as, for example, integration of heterogeneous sources. In this case, the system under development will provide access, in an integrated way, to information coming from the clinical records, scientific articles and others publications concerning health. These sources have very different features as: free text, text endowed with certain external structure (for example, in scientific articles), blended free text with structured data (for example, in clinical results), etc.

The ISIS project has as partners the Universidad Europea de Madrid, Universidad de Huelva, Hospital de Fuenlabrada, a public health care institution with a high technological infrastructure, and Bitex (The bit and text company), a firm specialized in text processing. We are working together in order to decrease the overload of information using text summarization and categorization, improving the organization of answers, presenting groups of related documents and also integrating our algorithms with the SELENE Information System at Hospital de Fuenlabrada.

References

1. Aronson, A.R., Mork, J.G., Gay, C.W., Humphrey, S.M., Rogers, W.J.: The NLM Indexing Initiative's Medical Text Indexer. In: Proceedings of Medinfo, San Francisco (2004)
2. Maña, M.J., de Buenaga, M., Gómez, J.M.: Multidocument summarization: An added value to clustering in interactive retrieval. ACM TOIS, 22 (2), pp. 215-241 (2004)
3. Mostafa J., Lam, W.: Automatic classification using supervised learning in a medical document filtering application. Information Processing and Management 36, 3 (2000) 415-444
4. Ribeiro-Neto B., Laender, A.H.F., De Lima, L.R.: An Experimental Study in Automatically Categorizing Medical Documents. Journal of the American Society for Information Science and Technology 52, 5 (2001) 391-401
5. Elhadad, N., McKeown, K.R..: Towards generating patient specific summaries of medical articles. In: Proceedings of Automatic Summarization Workshop (NAACL), Pittsburgh, USA (2001)
6. Johnson, D.B, Zou, Q., Dionisio, J.D., Liu, V.Z., Chu, W.W.: Modeling medical content for automated summarization. Annals of the New York Academy of Sciences 980 (2002) 47-58