SMART DOCUMENT TECHNOLOGIES FOR EXTRACTING AND STRUCTURING DATA FROM PATIENT RECORDS

Opportunities for New Knowledge based Services

Denys Proux, Eric Cheminot, Caroline Hagege and Frederique Segond
Xerox Research Centre Europe, 6 Chemin de Maupertuis, 38240 Meylan, France

Keywords: Smart document technologies, Textual content processing, Structured information, Document conversion, Risk assessment, Healthcare, Hospital acquired infections.

Abstract: Costs related to healthcare are exploding. In Europe studies show that they will reach in 2050 up to 10% to 15% of the Gross Domestic Product. Several causes drive up these costs such as ageing population or the rise of chronic diseases. Adverse events jeopardizing patient safety in complex hospital workflows such as Hospital Acquired Infections are also a major contributor to these costs. But recent studies suggest that it might be possible to reduce this impact by 20 to 30% with a combination of appropriate processes and smart monitoring tools. Information Technologies are therefore seen as a key enabler to help improving information workflow and process efficiency inside hospitals.

1 THE EUROPEAN HEALTHCARE MARKET

The healthcare system in Europe is currently submitted to an extreme pressure. One explanation is the fact the population is getting older (the amount of people over 50 will increase by 35% between 2005 and 2050 and the amount over 85 will triple in the same period) and require more attention and cares. As a consequence, latest estimates foresee a rising of the global cost up to 16% of the European Gross National Product (GNP) until 2020 (currently it represents only 9%). Information and Communication Technology (ICT) which has already allowed significant process and cost improvements in other domains could be and will be a key element to help keeping the healthcare budget under control.

However to achieve this goal there is a clear need for new, more efficient and Smarter tools for proactively predicting diseases, for personalizing healthcare and more generally to empower the patient in his own safety and health.

Therefore challenges for ICT are many. Solutions should address various needs on a stepped way to electronic and structured information which will eventually open the door for analytics, simulation forecasting, prevention, etc. They range from:

i. moving from paper or isolated data to electronic information systems,

ii. moving from unstructured to structured information,

iii. moving from isolated database to interconnected information systems,

iv. finally building advanced services, such as forecasting and risk assessment systems, on top of this.

In this paper we focus on two aspects which are: moving from paper (or unstructured information) into structured electronic records, and leveraging recorded information through smart and value added services such as for example risk assessment tools. In that perspective two projects will be detailed where we have customized existing tools or created entirely new technologies and functionalities to address both parts of the workflow and value chain.

Proux D., Cheminot E., Hagege C. and Segond F..
SMART DOCUMENT TECHNOLOGIES FOR EXTRACTING AND STRUCTURING DATA FROM PATIENT RECORDS - Opportunities for New Knowledge based Services.
DOI: 10.5220/0003184906330636
In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence* (ICAART-2011), pages 633-636
ISBN: 978-989-8425-40-9

2 SERVICES FOR MEDICAL RECORDS MANAGEMENT

2.1 The Legacy Documents Conversion Project: Context

Our team in our Research Group started a collaboration with a business company developing a document management solution for group of hospitals in the middle-west of the UK. The deal is to outsource the management of legacy patient records. These documents are paper based, so the project in its original form aims to convert this unstructured information into an electronic format. The considered corpus represents almost 1,000,000 patient records (that are generally composed by a several documents coming from several departments such as radiology, surgery, intensive care, etc). The conversion is foreseen to be done on a 7 year period. Several scenarios are considered ranging from simple paper to digital conversion to more advanced content analysis and indexing features built on top of it.

2.2 Workflow and Implementation

The model that has been put in place is the "scan when requested" scenario. Each time a new request to access a given patient record is submitted all patient related documents are retrieved and scanned. Our contribution was to customize for that purpose a set of Smart Document Technologies that automate part of the document content indexing process in order to extracted meta-data that could be attached to these scanned documents in order to facilitate future queries and document retrievals. This set of technologies and their context of application is detailed below.

Glyphstamps for Document separation: Glyphstamps is a technology that consist in special tags printed on documents that encode rich information. It is used to identify information pieces in the set of document related to a patient. They are used to detect document boundaries but also to route documents for appropriate post processing according to their type.

FuzzyMatch: sometimes, a document category can be defined (at least partly) by a set of keywords. Unfortunately, when dealing with paper documents, we first have to perform an OCR to extract the content in a useful format. This generally leads to OCR errors so the idea of the FuzzyMatch technology is to be able to efficiently match searched keywords in OCRed documents while taking into account possible OCR errors (or orthographic errors in the original document). Distance with queried keywords is computed and frequent errors are learned and taken into account.

Categorization: this technology is based on a model previously trained on collections of already annotated documents. The aim is to detect automatically from document contents those that match with predefined and pertinent categories. For this project we use our own made categorizer that learns from human annotations to build probabilistic models It relies on a probabilistic generative model (Gaussier 2002) which can be seen as a hierarchical extension to PLSA (Probabilistic Latent Semantic Analysis (Hofmann 1999)) where both documents and words may belong to different categories.

Named Entity Detection: The goal of this step is to identify in the text some specific information that can be used to further enrich the indexing step. This information can be for example people names, social security numbers, locations, treatments, drug names, bacteria,, etc. This step can be done by pattern matching, but we found it more efficient to use a deeper analysis with the Xerox Incremental Parser (XIP) (Aït-Mokhtar 2002) which allows to design more complex and efficient detection and disambiguation rules as well a co-reference management.

All these technologies work on electronic documents which means that paper documents should be scanned and OCRed first. However some documents may also contain some handwritten parts which is problem as up-to-now, automating document content processing in this case is very difficult. Nevertheless we developed and tested some new technologies and algorithm such as **Handwritten Keyword Recognition** to try to address this issue. This technology works as a machine-learning image categorization process: sample documents are pre-processed to extract the bounding boxes of these words (graphical boxes around words) and searched – positive - words are annotated by an operator (Perronnin 2009). The system is then trained to recognize those words. We have reached very promising results, but due to the extreme variability of hand-writing (especially in the medical field), it will be reserved for a limited number of words, crucial for the process.

3 LEVERAGING INFORMATION AND KNOWLEDGE MANAGEMENT

3.1 Improvement Opportunities

Optimizing the document workflow in hospitals by converting legacy documents into a digital version, or unstructured into structured information to feed databases is the very first step to unleash the power of advanced services building on this information. It opens the door to a new set of applications that can provide the assistance of knowledge services thanks to automated data mining, and information analysis tools. Risk assessment is one example.

A project has been started in 2009 to develop an Hospital Acquired Infection (HAI) detector monitoring patient related reports produced inside hospitals (ALADIN-DTH). This project is funded in part by the French Research Agency (ANR) for 3 years. It is a unique collaboration between key partners with unique competencies in all aspect required to build such a system: an University Laboratory specialized into building multi-terminologies resources for the medical domain, a content French provider for Pharmaceutical information, a research Center specialized into designing Natural Language Processing and Semantic Management systems and 4 university hospitals providing real data and HAI expertise.

3.2 Key Technology behind this Project

The heart of this project is the Xerox Incremental Parser (XIP), which performs text mining. This parser is robust that is to say it has already been used in various projects to process large collections of unrestricted documents. It has been designed to follow strict incremental strategies when applying parsing rules. The system never backtracks on rules to avoid falling into combinational explosion traps which makes it very appropriate to parse real long sentences from scientific texts for example (Aït-Mokhtar 1997).

We have decided to use such a tool as HAI is a complex issue that implies for instant pieces of evidences appearing according to a strict chronology (e.g. a patient has a surgery, then 2 days latter some symptoms occur such as fever, then some specific antibiotics are prescribed, etc.). To establish these connections we need to have a certain level of understanding of the content of the document, simple keyword detection is not enough.

3.3 Addressing the Terminology Issue

To be able to detect pertinent information from text it is crucial to be able to address all terminologies in use in targeted hospitals. This is required to build appropriate lexicons that encompass pertinent terms characterizing an HAI or serving as pieces of evidence to conclude to an HAI suspicious case. One of the partners (CISMeF) provides this information: the Multi Terminologies Indexer is a generic automatic indexing tool able to tag an entire document with all terminologies necessary for the project. This server offer term identification covering the following terminologies: SNOMED 3.5 (International Systemized Nomenclature of human and veterinary MEDicine), MeSH (Medical Subject Heading), ICD10 (Classification of Diseases) and CCAM (French CPT), TUV (Unified Thesaurus of Vidal), ATC (Anatomical Therapeutic and Chemical Classification), drug names with international non-proprietary names (INN) and brand names, Orphanet terms (rare diseases), CIF (International Functional Terminology), CISP2 (International Classification for Primary care), DRC (Consultation results), MedlinePlus.

3.4 Temporal Sequence Detections

In this project the chronology of events is crucial to characterize a problem. Relying on an existing temporal processing system (Hagege 2008) we perform an adaptation of this system for French in order to be able to detect and normalize temporal expressions appearing in text. Three kinds of temporal expressions are considered:

1) Absolute dates (for example 10/03/2010)

2) Referential dates with reference to the utterance time. (for example “yesterday”)

3) Referential dates whose reference is another textual expression (for example, “two days after admission”).

Discovering and normalizing temporal expressions enable us to associate a time stamp to the event described in text. As a result, we can associate to the description of a potential risk factor for HAI a time stamp and check if it occurs after or before another specific event.

3.5 Temporal Sequence Detections

Risk indicators consist in the description of infection events such as mentioning a specific bacteria, antibiotic, symptoms such as fever etc. All the

elements must also occur according to a specific and strict time frame. For instance, the simple presence of a temperature rising is not enough to decide that we have an HAI case. But if this temperature rising is associated with the presence of a certain bacteria, and with the presence of a catheter, and if this temperature raising occurs at least 2 days after the admission of the patient in the hospital, then we have clues enough to suspect an HAI occurrence. The relationship between risk indicators is carried by the syntactic dependencies calculated by XIP.

The following figure illustrates this temporal aspect of information extraction.

Output of temporal sequence processing (addition of a time stamp):
<TEMPS id="12" val="T+7J">
La patiente a présenté , une semaine après son arrivée dans le Service , une infection urinaire à Klebsielle , accompagnée d' une pneumopathie nosocomiale (sans germe retrouvé au combicath) , pour lequel elle a bénéficié d' un traitement pendant une semaine par Rocéphine et Oflocet.
</TEMPS>

Output of potential candidates for HAI risk detection
ANTIBIOTIQUE(Rocéphine)
ANTIBIOTIQUE(Oflocet)
DISPOSITIF(combicath)
BACTERIE(Klebsielle)
INFECTION(infection urinaire)
INFECTION(pneumopathie)
INFECTION(pneumopathie)
GERME_INFEC(Klebsielle)
BACTERIA(Klebsiella)

Figure 1: Analysis results for a given sentence.

3.6 Next Steps

The current version of the system, or to be precise, current detections rules work at the sentence level. This means that link between risk factors are search inside the boundary of a sentence. But it can also happen that elements characterizing an HAI are distributed in different and distant section of the medical report. It is therefore necessary to build long distance dependency detection and a reasoning algorithm able to work on a full patient report. These risk scenarios are now refined with the help of medical experts.

4 CONCLUSIONS AND BEYOND

In this paper we described the needs for ICT solutions to help improving the management of medical information to the benefits of both patient safety and healthcare economy. In this context the improvement of internal hospital processes and the way information is managed can be a key contributor to drive down these costs.

We have described in this paper two on-going projects illustrating the benefit brought by Smart Document Technologies to two major challenges which are: the move from unstructured information (e.g. paper) to structured EMR, and the development of smart monitoring systems for risk assessment. These are just examples.

They are many others fields of applications for ICT such as remote monitoring of patients or improvement of an hospital workflow through assisted activity tracking and resources management. We are just at the beginning of the e-revolution of the traditional healthcare landscape.

ACKNOWLEDGEMENTS

ALADIN is a 3 year project funded by the French *Agence Nationale de la Recherche* (National Research Agency - ANR) in the context of the TecSan (*Technologies pour la Santé et l'Autonomie*) program.

REFERENCES

Aït-Mokhtar S., Chanod J. P., Roux C., 2002. Robustness beyond shallowness: incremental deep parsing. Cambridge University Press. Volume 8, Issue 3, Pages: 121 - 144. June 2002.

Gaussier E., Goutte C., Popat K., Chen F., 2002. A hierarchical model for clustering and categorising documents. *Proc. ECIR-02*. 229-247. Springer.

Haas J. P., Mendonca E. A., Ross B., Friedman C., Larson E., 2005. Use of Computerized Surveillance to Detect Nosocomial Pneumonia in Neonatal Intensive Care Unit Patients. Am J Infect Control, 2005; 33(8):439-43.

Hagège C., Tannier X. XTM a robust Temporal Text Processor. *Proceedings of CICLing 2008*, Haifa, Israel.

Hofmann T., 1999. Probabilistic Latent Semantic Analysis**.** *Proc 15th Conf. on Uncertainty in Artificial Intelligence*. 289-296. Morgan Kaufmann.

Perronnin F., Rodriguez J. A., 2009. Handwritten word-image retrieval with synthesized typed queries. ICDAR 2009 (International Conference on Document Analysis and Recognition). Barcelona, Sapin, July 26-29, 20.