
Addressing Risk Assessment for Patient Safety in Hospitals through Information Extraction in Medical Reports

Denys Proux, Frédérique Segond
Xerox Research Centre Europe
6, Chemin de Maupertuis, Meylan 38240, France
Denys.proux@xrce.xerox.com, Frederique.Segond@xrce.xerox.com

Solweig Gerbier, Marie Hélène Metzger
Service d'hygiène, épidémiologie et prévention des Hospices Civils de Lyon
Hôpital Henry Gabrielle - Villa Alice,
20 Route de Vourles BP 57, 69 230 Saint-Genis Laval cedex, France
solweig.gerbier@chu-lyon.fr, marie-helene.metzger@chu-lyon.fr

Abstract: Hospital Acquired Infections (HAI) is a real burden for doctors and risk surveillance experts. The impact on patients' health and related healthcare cost is very significant and a major concern even for rich countries. Furthermore required data to evaluate the threat is generally not available to experts and that prevents from fast reaction. However, recent advances in Computational Intelligence Techniques such as Information Extraction, Risk Patterns Detection in documents and Decision Support Systems allow now to address this problem.

Key words: Hospital Acquired Infections, Natural language Processing, Information Extraction, Risk Pattern.

1. INTRODUCTION

Patient's security is a key issue in hospitals and monitoring adverse events is a preliminary step of a corrective or preventive action. Only a qualitative and quantitative estimate of observed adverse events in hospital can help in deciding which measures to implement. For example, in France, the incidence of adverse events was estimated [1] to 6.6 per 1000 hospital days in 2004, from which 24.1% were Hospital Acquired Infections (HAI).

Hospital acquired infections represent an important part of adverse events in hospitals and monitoring procedures are in place in most of European countries. These procedures are mostly based on methods developed in the United States by the Centers for Disease and Control and Prevention (CDC) National Nosocomial Infection Surveillance System [2]. However, the important workload linked to these monitoring methods forced the hospitals to consider alternatives to these methods which are based on active report of HAI by the medical personnel or infection control experts. There is need for automation of part of the surveillance to backup Risk Management teams that often have not enough resources to efficiently perform this monitoring.

The use of Natural Language Processing techniques is one of the promising alternatives for monitoring adverse events in hospitals. Text Mining Techniques applied on medical reports specifically for risk assessment are still relatively new [3] because it assumes to have access to a very accurate and disambiguated terminology, to a list of factors characterizing a potential infection and finally it requires most of the time robust parsing capabilities to handle real life medical literature. Most of these systems are keywords based or based on simple pattern matching [4]. The identification and disambiguation of complex information such as HAI require not only having access to named entity recognition but also and mainly to the detection of specific semantic links appearing in text between these entities.

Therefore two key elements will be needed;

- a rich and standardized terminology to allow detecting inside text some meaningful pieces of information (such as drug names, symptoms, ...);
- a robust parser able to process long and complex sentences in order to identify key dependencies between these meaningful pieces of information.

The following paper presents such a system in the following sections, applied to monitoring of hospital acquired infections through information extraction in patient discharge summaries.

2. HOSPITAL ACQUIRED INFECTIONS

2.1. Definition

A Hospital Acquired Infection can be defined as: *An infection occurring in a patient in a hospital or other health care facility in whom the infection was not present or incubating at the time of admission. This includes infections acquired in the hospital but appearing after discharge, and also occupational infections among staff of the facility. If the exact status of the patient is not clearly known when he*

first came in a medical unit, a period of 48 hours (or superior to the incubation period if it is known) is considered to separate HAI from other kinds of infections coming from outside. As for infections related to surgery a period of 30 days is considered and extended to 12 months in case of implanted device. [5]

2.2. Burden of disease

Studies [6] show that in Europe, the frequencies of HAI hit 5 to 10% of hospitalized patients. In the extended European Union, there are approximately 3 million identified cases and 50,000 related deaths per year. Mortality related to HAI was estimated in 2005 to 4000 deaths per year in France and 20 to 30 per cent of these deaths are estimated avoidable by adapted prevention guidelines.

HAI related costs are greatly dependent on the type of infection and on the patient's risk factors. The costs associated to HAI ranged from 500 € for a urinary infection up to 40 000 € for a serious bacteraemia in Intensive Care Units (ICU) [7]. In France, estimating that the number of HAI by year is 750 000 and that the preventable part of these cases is 30%, the overcharge for the health system may represent a total cost between 0.11 and 9 billion €

2.3. Monitoring Systems

Automated surveillance is defined by Wright et al. [8] as a process of obtaining useful information from infection control data through the systematic application of medical informatics and computer science technologies. This definition recovers very different ways of processes.

The first way of process is based on the combination of different hospital databases (bacteriological data, antibiotic exposure, claim data...). Different studies have demonstrated the efficiency of systems based on the combination of bacteriological data, antibiotic exposure or discharge diagnoses [9]. In France, Bouam and al [10] evaluated the sensitivity of automated nosocomial infections detection based on bacteriological databases to 59% and the specificity was 91% compared to manual detection. A Danish study [11] showed that the sensitivity of nosocomial infections detection was higher by combining different infection parameters (microbiology, antibiotic treatment, leucocytes counts, C-reactive protein concentrations) (94%) than by using each infection parameter separately (61% to 82%). However the specificity was lower (47% for combined parameters vs. 53% to 70% for each parameter used separately).

A second way of automated process is based on using natural language processing of discharge summaries. There is no national guideline or obligation whatsoever to standardize recording systems neither in private clinics nor in public hospitals. However information exists and is organized in a way or another. Main

clinical events occurring during a patient's hospitalization are recorded in the patient's record by the medical staff. A summary is written at the hospital discharge which is the main source of communication between the various medical units. But today there is no real standardized summary of this information and medical staff can write this discharge report on the way they want. It stresses therefore the need for an automated tool able to scan daily reports, to mine information written in these texts in order to detect potential risk patterns and to send alerts to appropriate people.

Very few experiences were already performed [3]. Melton and al. used for instance the MedLEE natural language processor for the detection of adverse events, comprising nosocomial infections. The sensitivity to detect adverse events was evaluated to 28% (IC95% = 17-42) and the specificity to 98,5% (IC95% = 98,4 – 98,6) .We can hypothesize that the low sensitivity of this tool is linked to the broad type of adverse events searched (venous thrombosis, post-operative wound, perioperative myocardial infarction, falls...). The medical language being very complex, the use of natural language tools for the detection of adverse events should be developed by specific adverse events topics (nosocomial infections, therapeutic adverse events, ...).

3. A STRATEGY FOR RISK ASSESSMENT USING NATURAL LANGUAGE TECHNOLOGY ON PATIENT DISCHARGE SUMMARY

3.1. First step: develop interoperable information extraction systems

It appears that information recording system in hospitals, is not always the top priority with respect to investments. Even if the current tendency is to computerize all data to make them available to electronic databases and for automated processes some hospitals are still relying most of paper based documents as some other have make the move to the digital world. There is not national obligation to normalize these systems. Each hospital, and some time each department, can decide which equipment to adopt and deploy. The result of this is a complete mess of heterogeneous systems not really interoperable where information is duplicated and not easily available for global analysis. Interoperability is still an important issue as stressed by EU reports [12].

The HAI surveillance system must not be specific to a given hospital or department but should be operational at a National level. This step implies developing processes and ad-hoc methods to gather required documents from local databases.

3.2. Second step: anonymization of patient's records

Medical data are highly sensitive. At that point one important task is to *anonymize* these data. This means removing all personal information related to people or places (in order to protect patients' privacy). This means that not only people names should be removed but also any information that could lead to an identification such as personal address, phone number, social security number, and so on. This is often required by specific national regulation. While at the moment anonymisation of patient records is done manually, natural language techniques specifically designed to detect this kind of Named Entities [13] can be applied here to perform this task.

In order to do so and also for the remaining text processing steps we are using the Xerox Incremental Parser [14] that combines five linguistic processing layers which are: pre-processing (tokenization, morphological analyzer and part of speech tagging); named entities; chunking; dependency extractions between words on the basis of sub-tree patterns over chunk sequences and finally a combination of those dependencies with boolean operators to generate new dependencies or to modify or delete existing dependencies. XIP comprises an engine and a meta-language that allows users to write grammar rules or add words in the lexicon. XIP integrates also a Named Entity Recognition module.

4. THE DIFFERENT LINGUISTIC STEPS TO BE ACHIEVED FOR HAI SURVEILLANCE

4.1. Entity detection

Once documents have been normalized and *anonymised* they can be processed by a Terminology Server in order to identify and locate all Named Entities that will be useful for the remaining decision process (e.g. drug names, symptoms, processes, dates). The goal of this step is twofold: to perform a Part of Speech analysis to allow a further computation of syntactic dependencies, and then to detect and disambiguate at a semantic level all key entities that will be involved in the risk pattern detection step.

Furthermore in the context of information extraction and risk analysis a proper recognition of specific vocabulary allows also to add a semantic tag to some words or multi-word expressions that can be involved in the description of an adverse event. This step will be performed by the Named Entity Recognition module enriched with some specialized terminology contained in medical structured terminologies. Indeed, terminological resources, and to be even more specific Taxonomies such as SNOMED¹ for instance, are very important at this step. They allow a system to identify these entities with respect to their definition in these dictionaries. These entities can be composed of several words (or tokens). In this

¹ <http://www.snomed.org/>

case specific detection rules apply to regroup all these words under a same semantic tag. In the context of HAI, these entities, taken to the largest extend, can be drug names (e.g. “Tienam”), disease name (e.g. “Surgical site infection”), exam (e.g. “abdominal ultrasound”), symptoms (e.g. “abdominal pain”), etc.

Applying now the XIP parser to texts enables the system to detect chunks of related words. Coupling this general tagger, the XIP chunker with a medical terminology infrastructure like for instance SNOMED enables the system to semantically tag the different concepts.

“The postoperative consequences were marked by abdominal pain and fever due to multiple intra-peritoneal abscesses and peritonitis without anastomotic dehiscence that required a peritoneal toilet on September 29th of this year. It was an infection with Klebsiella only sensitive to Tienam which was probably facilitated by the preoperative biliary drainage and the splenectomy. The evolution was finally favorable.”

Detected Entities:

SYMPTOM(postoperative consequences)
SYMPTOM(abdominal pain)
SYMPTOM(fever)
DIAGNOSIS (multiple intra-peritoneal abscesses)
DIAGNOSIS(peritonitis)
DIAGNOSIS(infection)
PROCEDURE(peritoneal toilet)
TREATMENT(Tienam)
BACTERIA(Klebsiella)

Figure 1: Named Entity Detection

4.2. Risk pattern detection

Once a patient Discharge Summary has been processed to assign POS tags and identify Named Entities, the next step is to detect some typical combinations of named entities that may be involved in the description of an adverse event.

Characterizing these events is not simply identifying some keywords inside texts; it is about finding special relations between these keywords. Therefore a syntactic analysis is required to detect the potential links, and more specifically the order of these relations. This can be simply summarized by trying to found: What produces what to whom when and how.

The first step consists in processing each sentence of a report to compute all syntactic dependencies. This is done thanks to a set of grammar rules designed for common language. XIP provides already grammar rules for almost 10 different languages including for example French and English. This rules applies on the POS tags assigned to each words (or tokens) at preceding step. Syntactic dependencies extraction does not need to be customized for a specific domain provided that it has been done for the lexical level.

What need to be customized for the domain is the rules to characterize key information searched inside texts. For the detection of HAI it includes the detection of various types of information such as:

- where does the situation takes place
- who is the patient (male, female, young, old)
- what are the treatment or drug involved
- what symptoms are detected
- are characteristic adverse events terms appearing inside the text (e.g name of a virulent bacteria)

The connection between these elements is important because according to their order it may characterize an HAI or just a normal case. In order to detect these elements specific rules have to be designed that takes into account both the semantic tags assigned to words or multi-words expression thanks to domain specific terminologies that help identify symptoms and drug names for example, and the detected syntactic relation between these entities.

These rules have to be defined by experts from the domain. In this case this is experts from surveillance groups that already spent time reading report to find potential indication about HAI case. They must formalize what are the criteria they use to say whether or not if there is a potential HAI case emerging from a report. Once these rules are formalized, then linguist can convert them into parsing rules than can be processed by the text parser.

The result of such analysis can be illustrated by the following example that characterize key information element that will be used when trying to find a match between what is extracted from the text and potential HAI scenario.

“The postoperative consequences were marked by abdominal pain and fever due to multiple intra-peritoneal abscesses and peritonitis without anastomotic dehiscence that required a peritoneal toilet on September 29th of this year. It was an infection with Klebsiella only sensitive to Tienam which was probably facilitated by the preoperative biliary drainage and the splenectomy. The evolution was finally favorable.”

Dependencies between pertinent entities and events

- Symptom (*pain, without, dehiscence*)
- Preliminary_Condition (*yes, pain*)
- Preliminary_Condition (*No, dehiscence*)
- Detailed_Symptom (*abdominal, pain*)
- Location (*abdominal*)
- Prescribed_AntiBio (*Tienam*)

Figure 2 : Detected Syntactic dependencies

4.3. Risk assessment

Once some key entities and specific links among them have been detected inside a text, the next step is to evaluate a potential match with predefined scenarios characterizing HAIs.

In order to do so, these scenarios which detail all the criteria that are taken into account to define one specific HAI, have to be formalized by experts. They must details both all the elements (symptoms, drugs, ...) that can be involved in a case definition and also the various types of links that should exist among them.

At this level several strategies are possible. One might consist in simply remaining at the sentence level to find direct ordered syntactic links between key elements. This can be illustrated for example by the detection in a single sentence of a symptom that is the consequence of a new treatment (or drug prescription that must belong to a specific category of drugs such as Anti-Biotic) which produces the following effect (or symptoms).

However, HAI are complex to characterize and generally involve various different elements that must occur in a specific order. This is why all the needed information to decide whether or not we face an HAI is generally not contained in one single sentence. Other information than just dependencies between Events are therefore part of the reasoning process. According to input documents metadata and localization (sections where the information is detected) can also be taken into account to make the decision. Paragraphs ordering as well as dates inside these paragraphs provide useful data to build a timeline that helps classifying information between actions and consequences. Therefore it is important to build a complete discourse analysis to take into account all these elements. This requires some kind of discourse representation mechanisms, and to do some decision support systems designed to formalized medical knowledge can be well adapted to do so.

Ontologies are important at this level because it allows to formalize a scenario at an abstracted level which reduce the number of cases the knowledge expert has to take into account to cover all possible combinations of keywords that may be involved into the definition of one single case. Ontologies provide hierarchies of terms (Drug names, symptoms, ...) this allows for example to state simply in one scenario that if a specific type of anti-biotic is detected inside a text in combination of a specific type of bacteria then this related to an HAI. There will be therefore a link made between the semantic tag assigned to the elements detected inside text and the abstracted concepts used inside the scenarios thought the use of such Ontologies that will provide the link between these two elements.

One last element that should be taken into account for HAI detection is flexibility and this because most of the time HAI are not clearly indicated inside text. There could be pieces of evidences but not a clear statement because for example the case has not been detected by the medical staff as so, and therefore not detailed explicitly. This means that several levels of HAI detection confidence should be taken into validating a detection. Some elements can be very characteristic such as the name of a given bacteria (e.g. “**infection with *Klebsiella***”), some strong candidate such as the use of a specific type of antibiotic drug in specific department (e.g. “**tienam**” and

“Intensive Care Unit”) and some require a combination with various other elements to truly characterize an HAI. The alert mechanism must therefore be able to compute the level of HAI likelihood according to the elements extracted from text that match a given scenario.

5. PLAN AND NEXT STEPS

A Patient Discharge Summary Analysis Strategy is currently investigated in the context of a project that is starting up between the Lyon University Hospital and the Xerox Research Centre Europe to apply state of the art Computational Intelligence Techniques to address a problem jeopardizing public health. This project will be developed in close collaboration between HAI surveillance experts and Linguistic and Knowledge Management experts in order to design the necessary set of rules to identify HAI from medical reports.

On a first hand only some departments and infections will be targeted. These are those related to the highest risks and that have the highest impact on human health. These departments are: *Intensive Care Unit* and *Surgery*.

A consistent set of reports containing HAI cases will be selected and highlighted by experts for system design and validation. We are currently building the first version of the system architecture focusing on the detection of the following events:

- Context (e.g. “re-entry”, “Surgery”, “HAI”, ...)
- Clinical Parameters (e.g. “fever”, “inflammatory trace”, “pulmonary secretion”, “cough”, ...)
- Biological Parameters (e.g. “bacteriological exam”, ...)
- Biochemical Parameters (e.g. “white-cell > 10 000/mm³”, ...)
- Treatments (e.g. “specific anti-biotic drugs”, ...)

The real impact of such a project remains to be determined by experiments. In particular, we need to evaluate how much relevant information is present in the patient record.

6. CONCLUSION

Hospital Acquired Infection is a major issue that has a very important impact both for the patient and for added medical cost. Providing tools to shorten the time and effort necessary to discover and react against HAI is crucial to reduce this impact.

In this paper we presented a strategy that aims at applying Natural Language Processing techniques to mine patient Discharge Summaries in order to identify HAI. This strategy implies a strong collaboration with HAI surveillance experts in order to formalize the detection rules and linguist to convert these rules into appropriate

grammars for advanced parser and decision mechanisms to trigger alerts. This will be made in the context of a collaboration started between XRCE and HAI surveillance experts from the Lyon University Hospital to design a control and prevention system able to analyze medical reports for HAI detection.

REFERENCE

1. P. Michel, J. Quenon, A. Djihoud, S. Tricaud-Vialle et al. Les événements indésirables graves liés aux soins observés dans les établissements de santé : premiers résultats d'une étude nationale. Etudes et résultats DRESS 2005(n°398).
2. Emori T, Culver D, Horan T, Jarvis W, White J, Olson D. *National Nosocomial Infections Surveillance System (NNIS) : description of surveillance methods*. American Journal of Infection Control 1991;19(1):19-35.
3. Murff HJ, Forster AJ, Peterson JF, Fiskio JM, Heiman HL, Bates DW. *Electronically screening discharge summaries for adverse medical events*. J Am Med Inform Assoc 2003;10(4):339-50.
5. Prevention of hospital-acquired infections: A practical guide. 2nd edition. World Health Organization.
http://www.who.int/csr/resources/publications/drugresist/WHO_CDS_CSR_EPH_2002_12/en/
6. H, S. E. Humphreys. "Prevalence surveys of healthcare-associated infections : what do they tell us, if anything?" Clin Microbiol Infect 2006; 12: 2-4.
7. S. D. Bärowolf, C. Geffers, C. Brandt, R.P. Vonberg, et al. "Reduction of surgical site infections after caesarean delivery using surveillance." Journal of Hospital Infection 64: (2006) pp. 156-161
8. Wright M. *Automated surveillance and infection control:toward a better tomorrow*. Am J Infect Control 2008;36:S1-6.
9. Bellini C, Petignat C, Francioli P, Wenger A, Bille J, Klopotov A, Vallet Y, Patthey R, Zanetti G. Comparison of automated strategies for surveillance of nosocomial bacteremia. Infect Control Hosp Epidemiol 2007;28(9):1030-5.
10. S. Bouam, E. Girou, et al. "An intranet-based automated system for the surveillance of nosocomial infections: prospective validation compared with physicians' self-reports." Infection Control and Hospital Epidemiology (2003) 24(1): pp. 51-55.
11. Leth, R.A. and J.K. Moller, *Surveillance of hospital-acquired infections based on electronic hospital registries*. J Hosp Infect, 2006. 62(1): p. 71-9.
12. eHealth for Safety. Impact of ICT on Patient Safety and Risk Management. European Commission. Information Society and Media. ISBN 13 978 92 79 06841-6.
13. C. Brun, C. Hagège. Intertwining deep syntactic processing and named entity detection. ESTAL 2004, Alicante, Spain, October 20-22, 2004.
14. S. Ait-Mokhtar, J.P. Chanod, (1997) *Incremental Finite-State Parsing*. In Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97), Washington March 31st to April 3rd, 1997, pp.72-79.