Separate the grain from the chaff: make the best use of language and knowledge technologies to model textual medical data extracted from electronic health records

Abstract

Electronic Health Records (EHRs) contain information that is crucial for biomedical research studies. In recent years, there has been an exponential increase in scientific publications about using textual processing of medical data in fields as diverse as medical decision support, epidemiological studies and data and semantic mining. While the use of semantic technologies in this context demonstrates promising results, a first experience with such an approach, shed light on some challenges among which the need for smooth integration of specific terminologies and ontologies into the linguistic processing modules as well as independency of linguistic and expert knowledge. Our works lies at the cross road between natural language processing, knowledge representation and reasoning and aims at providing a truly generic system to support extraction and structuration of medical information contained in EHRs. This paper presents an approach which combines sophisticated linguistic processing with a multi-terminology server and an expert knowledge server focusing on independency of linguistic and expert rules.

Keywords: NLP, ontologies, biomedical terminologies, expert systems, knowledge representation, reasoning, Semantic Web technologies, epidemiology.

1. Introduction

With the development of Electronic Health Records (EHRs), the linguistic analysis of textual data in the medical sector is receiving increasing interest. However, existing approaches vary in terms of the granularity and sophistication of linguistic processing carried out. For example, the MedLEE system processes medical text based on the recognition of named entities, without taking into account the relationships between these entities (Friedman et al., 1996). Others use more sophisticated linguistic approaches relying on syntactic and semantic analysis. Among these systems, some implement rule-based approaches, (Wang, 2007) (Ben Abacha & Zweigenbaum, 2011) while others employ statistical-based methods (Ehrentraut et al., 2012), or combine an expert-based system with a machine learning system (Zweigenbaum et al., 2013). Most of these combine their linguistic approach with the use of medical terminologies or ontologies.

In a previous project, we demonstrated the feasibility and good performance of linguistic processing via the development of a semantic analysis tool to detect Hospital Acquired Infections. However, a number of scientific and technical challenges also came out of this first experience: the need for deep temporal analysis of events in the medical domain, smooth integration of terminologies and ontologies into the linguistic processing modules, independency of linguistic and expert knowledge rules in order to provide a truly generic system to support medical studies and decision making.

In this context, we present the development of a generic solution that extracts semantic information from medical data and organizes this medical information in such a way that it can be used to support epidemiological studies or medical decision-making. In this generic solution, medical staff will be able to write their own expert rules, exploiting language intelligence, independently of their domains of specialty and their knowledge of linguistics.

From a scientific and technological standpoint, the project objectives are: to develop fine-grained linguistic rules to extract temporal expressions, to interface a semantic analyser with a multi-terminology server upstream during the extraction phase, and to interface a linguistic engine and knowledge representation module. For the generation of expert knowledge rules, we propose a general modular architecture that clearly distinguishes between linguistic rules and expert knowledge rules in such a way as to allow medical users to generate their own decision rules and enable semantic queries on extracted information. The outcome will be an operational system that integrates the various technological modules described in the next section.

After briefly presenting the general modules of the system, we focus on the approach we adopt for the expert decision rules that encode medical knowledge and in particular on their links with the linguistic processing component.

2. The four main components of the general architecture

Our objective is to develop an approach that combines sophisticated linguistic processing with a multi terminology server and a knowledge server.

In such a context, the quality and the quantity of information of different types coming from the components described below (, multi-terminology server, linguistic server and knowledge server) determine the performance of the overall system. Furthermore, the fact that the three components are independent and interact together to populate the knowledge base, ensures that the system is generic and adaptable to different medical sub-domains. As a consequence the resulting system will enable medical users to write their own expert knowledge rules, taking advantage of refined linguistic technologies, without necessarily being a linguist. To achieve this goal, we propose to provide medical staff with a system based on linguistic and semantic web technologies that support the building of a knowledge database. This knowledge base is then used to analyse extracted data in the context of epidemiological studies or assisted medical decisions.

The whole system consists of four main components:

- The multi-terminology server, which provides all processing modules with relevant lexical-semantic information in the medical domain.
- The linguistic server, which analyses textual medical input in order to provide a semantically enriched document to the next component.
- The knowledge server, which extracts high level knowledge using both outputs of the terminology and the linguistic servers.

The general planning component, which calls the different modules and provides results, either directly through the general user interface, or to each component that might need input from another component. The originality of our approach relies on domain adaptability. This adaptability is achieved by combining three distinct components a linguistic server, a terminology server and a knowledge server. The three main components are used to enrich a facts database, or knowledge base, which medical users can query according to different facets.

Besides all these components, in order to respect privacy policies, we developed a de-identifier which masks any information in the document that could help to identify either the patient or the medical staff. In some cases, for medical decision making, it is important that, at the end of the semantic processing, the authorized medical staff have access to patients 'identity to make the most appropriate decision for them. The de-identification tool thus also provides a re-identification facility.

The general architecture of the system is presented in Figure 1:

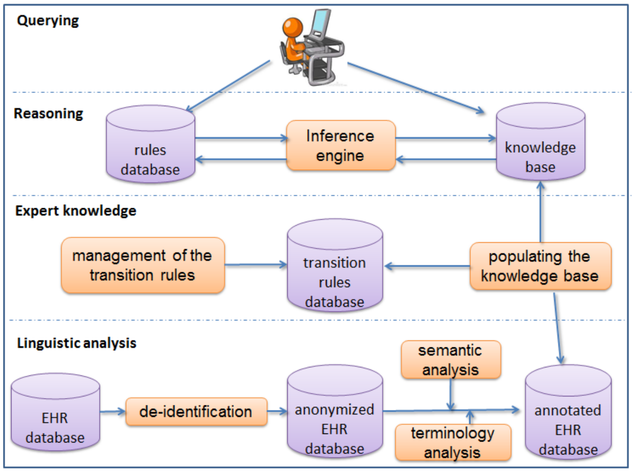

Figure 1: General system architecture

In the next section we describe the linguistic processing of the medical textual document that will serve as input to the inference engine.

3. Linguistic processing, medical terminologies

The linguistic processing module takes EHRs as input and enriches them with linguistic information. Linguistic processing is carried out by an NLP pipeline based on a combination of symbolic and statistical methods. The output is passed as input to the expert knowledge server in order to populate the knowledge base.

There are three main linguistic processing steps involved in the complete chain:

1. Sentence detection, tokenization, morpho-syntactic tagging and lemmatization.
2. Syntactic analysis (dependency parsing).
3. Semantic analysis.

The output of step 2 is a graph in which the nodes represent the words of the sentence, along with their morpho-syntactic features, and the arcs represent the syntactic dependencies between words.

Figure 2 shows an example for the French sentence *Nous avons découvert un abcès pulmonaire chez le patient en 2001 (We discovered a lung abscess in the patient in 2001).*

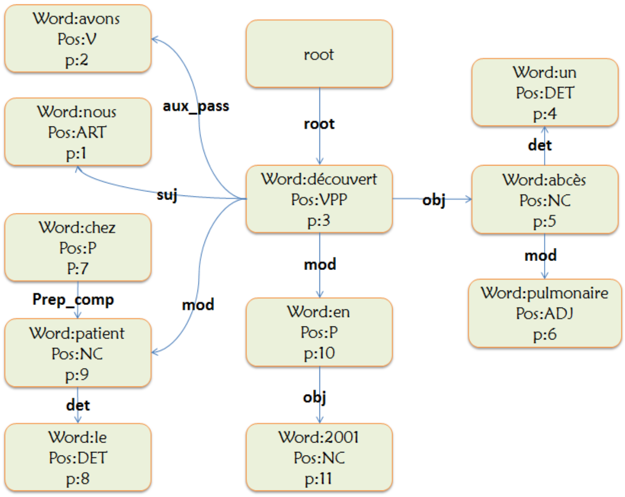

Figure 2: Graphical representation of syntactic dependencies.

Using the output of the ***syntactic*** analysis, combined with external general knowledge and knowledge coming from the multi-terminology terminology server, both medical and general, the ***semantic*** analysis step aims at representing the meaning of elements relevant to the medical domain. For example, a medical event related to a date, the presence or absence of certain characteristic symptoms, a dosage of medication, etc. Again, the semantic representation is a graph in which the nodes represent **entities** such as objects, persons, medications, and locations, while arcs represent the relations among these entities. These relations are **thematic roles** such as *agent*, *patient*, *goal*, *theme*, *cause*, *instrument*, and *beneficiary*.

The graphical representation of the semantic analysis of the previous sentence is shown in Figure 3.

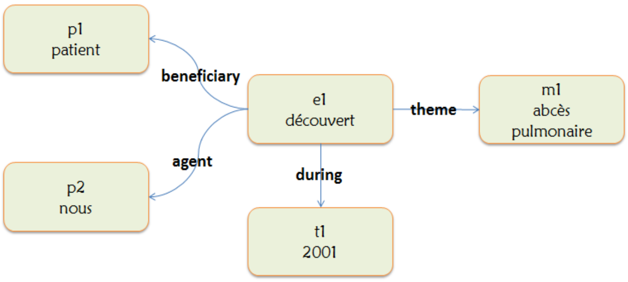

Figure 3: Graphical semantic representation

In the Figure 3, the variable e1 refers to an event (the verb *découvert, discovered*), p1 and p2 refer to a person (*le patient, the patient* and *nous, we*), m1 refers to the disease (*abcès pulmonaire, lung abscess*), and t1 refers to the time period *(2001*).

Four types of relations link these entities: BENEFICIARY which states that the one who receives the outcome of the event e1 is p1; AGENT, which states that the one who performs the outcome of the event e1 is p2; THEME, which states that the one that undergoes e1 is m1, and DURING, which states that e1 happens at the time t1.

Entities' nodes can be enriched with information coming from medical terminologies and ontologies, enriching the semantic analysis. For instance, entities such as *pulmonary abscess*, *cataract surgery* or *chronic tobacco smoking* can be associated to medical codes stored in several medical terminologies and/or ontologies such as ICD-10, SNOMED CT, MedDRA accessible through the UMLS knowledge browser (Bodenreider, 2006), BioPortal (Whetzel et al., 2011) or the HeTOP (Soualmia et al., 2011). Thanks to these terminology servers, for *pulmonary abscess* one can find that: it is a *disorder* (MeSH), it is a synonym of *lung abscess* (MedDRA), *abscess of the lung* (SNOMED CT), it may be related to *abscess of the lung with pneumonia* (ICD-10) and *abscess of the lung without pneumonia* (ICD-10). More relations can also be found: it is a *disorder of the lung* (SNOMED CT), an *inflammatory disorder of the respiratory tract* (SNOMED CT) that has one *finding site* the *lung structure* (body structure) and two *associated morphologic relations* the *abscess* (as disorder) and *abscess* (as morphologic abnormality).

4. Knowledge representation and reasoning

In the case of epidemiological studies, linguistic analysis is not enough as it does not retrieve information such as antecedent, chronic disease, lifestyle, etc. For example, in the previous sentence (section 3), the linguistic analysis detects that a lung abscess was discovered in 2001, but it does not detect that the lung abscess is an antecedent, which is exactly what the expert is interested in finding out.

We investigate two radically different approaches to building the expert server: the first one is based on well-established Business Rules Management Systems (BRMSs) and the second one, more research-oriented, based on Semantic Web technologies. We concentrate below on the second approach. At a high level, the different tasks to be performed are:

- Choosing the information model for the knowledge base.
- Populating the knowledge base.
- Reasoning on the knowledge base.
- Querying the knowledge base.

In what follows, after an overview of the information model we chose to build the knowledge base that will store the extracted facts from EHRs (section 4.1). We present what we call "Transition Rules" that are based on linguistic analysis or expert knowledge and serve as a means of populating the facts database (section 4.2).

4.1 Choosing the information model

Before going any further, we need to decide how we will represent the knowledge extracted by the system. This is very important as the adopted model will impact on the type of reasoning that will be possible over the extracted information and, as a consequence, on the new facts that may be discovered.

Until recently, the model most widely used to represent and store information has been the relational model. However, with the advent of the Semantic Web, the RDF *model* (*Resource Description Framework*), introduced by the W3C[1], has become the model of data exchange *par excellence* as it offers a powerful representation of graphs. It allows for explicit expression of relationships between two resources. Knowledge is expressed as *subject – predicate – object* triples, where *subject* is the resource to describe, *predicate* is the relationship between the subject and object, and *object* is the value of the predicate. Figure 5 shows an example graph where patient 22146[2] has a lung abscess (*abcès pulmonaire*) as antecedent at time T-12A (12 years previously).

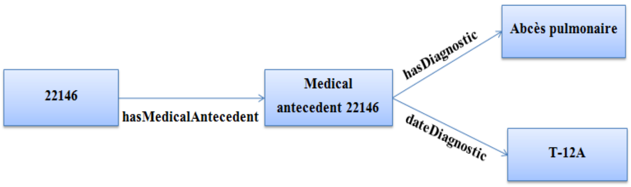

Figure 5: RDF graph representation

While RDF provides a way to model individuals (*22146*, *medical history* 22146, pulmonary abscess and T-12A), it does not define the semantics of an application domain and its power of reasoning is very limited. *RDFS* (*Resource Description Framework Schema*) is an *RDF*

[1] http://www.w3.org/

[2] Each patient has a unique identifier.

extension; it defines groups of similar *concepts*, a hierarchy over these concepts, relationships between concepts (*properties*), a hierarchy of properties and *instances* of concepts. However, *RDFS* has limited expressive power because some characteristics of the properties, such as *transitivity*, *inverse properties*, *etc.* are not supported, and it does not allow *value restrictions* or *cardinality*. OWL (*Web Ontology Language*) provides a more expressive knowledge representation language that allows the specification of ontologies. In addition to the ability to describe concepts and properties, *OWL* also allows cardinality. For instance, a patient must have a unique identifier (*22146*). Figure 6 models our example with OWL using OWL-DL (Description Logics) because it provides maximum expressiveness while ensuring the completeness of reasoning (all inferences are computable) and ensure computability (all calculations end in a finite time).

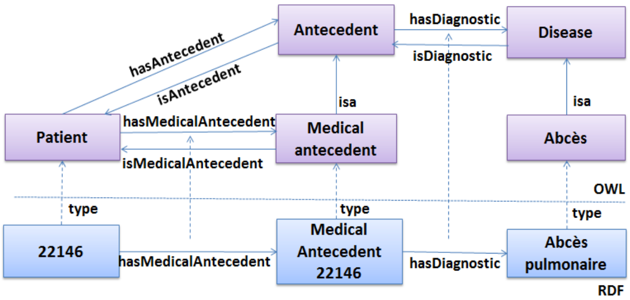

Figure 6: OWL representation

4.2 Populating the knowledge base

The second task is to populate the knowledge base with the facts that are relevant to medical staff. As mentioned before, linguistic analysis generates information that will help to populate the knowledge base, but this information is not always sufficient. The originality of our approach lies in our use of Transition Rules that combine both linguistic information and expert knowledge. Transition Rules, which are written by an expert and stored in the system's general planning component, convert the output of the semantic processing (the semantic graph) into facts that populate the knowledge base.

We distinguish two types of Transition Rules: (1) *linguistic-based Transition Rules* and (2) *expert-based Transition Rules*.

For the sentence, *Nous avons découvert un abcès pulmonaire chez le patient en 2001(“we discovered a lung abscess in this patient in 2001”)* the linguistic server provides as output: m1 (*m1, abcès pulmonaire, lung abscess*), TIME (*t1, T-12A*) and DURING (*t1, m1*) that allows one to locate the pathology in time. The linguistic-based Transition Rules convert information into the explicit semantic relation *hasDiagnosticDate* which can then be used to write the following type of rules

- `If THEME(e1,m1) and DURING(e1,t1) Then (m1-hasDiagnosticDate-t1)`

However, the linguistic server fails to provide detailed information about event e1, such as the fact that event e1 is a medical antecedent. The expert-based Transition Rules are there to enrich information extracted by the linguistic Transition Rules with expert knowledge, for example:

- `If THEME(e1,m1) and DURING(e1,t1) and t1 > d Then (m1-isa-Antecedent)`

In the previous example of expert Transition Rules, the variable *d* is the minimum interval of elapsed time required for pathology to be considered an antecedent and its value is determined by the expert. For instance, if *d=T-2A* then (abcès pulmonaire-isa-Antecedent) because *t1=T-12A*. The granularity of our model of representation (Figure 6) requires us to know the nature of the antecedent (medical or surgical), and a condition is then added to the previous rule:

- `If THEME(e1,m1) and DURING(e1,t1) and t1 > d and (m1-isa-DISEASE) then (m1-isa-MedicalAntecedent)`
- `For the same rule, if (m1-isa-SURGICAL PROCEDURE) Then (m1-isa-SurgicalAntecedent)`

The value of the variable m1is obtained thanks to a call to the terminology server.

The next step will consist in reasoning on the knowledge base and in the process of introducing new knowledge.

5. Conclusion

In this paper we propose to combine terminology, natural language processing, knowledge representation and Semantic Web technologies in order to analyze patient records in the context of epidemiological studies. The originality of our approach lies in the notion of Transitions Rules that combine, in separate modules, linguistic and expert knowledge.

Over the course of this three-year project we will perform different types of evaluations for the different linguistic services as well as for its different applications.

The project still being under development, we do not yet have any evaluation results to present. However, we shortly describe below the different evaluation steps we plan to carry out, together with the approach we plan to follow.

There will be two types of evaluations: evaluation in vitro and evaluation in vivo.

The first type of evaluation, *in vitro*, will consist in evaluating the performance of the different components, for instance, evaluating the coverage of the multi-terminology server for the two medical sub-domains, evaluating precision and recall of the syntactic and semantic components of the linguistic module, and evaluating the precision and recall of the expert rules. The results of these evaluations will be compared with existing systems and approaches described in the literature and, when possible, we will participate in academic evaluation campaigns.

The second type of evaluation, *in vivo*, will consist in evaluating the quality of extracted information in the context of epidemiological studies. System performance will be evaluated in two domains: Hospital-Acquired Infections and cancer.

These evaluations will ensure that obtained results are:

- consistent with the end user's needs (requirements and evaluation),
- technically and scientifically sound (quality assurance).

We also plan to compare results obtained using the Semantic Web-based expert module described in this paper with a module, also developed in parallel in this project, based on an existing open source Business Rules Management System (BRMS), from the field of expert systems.

References

Ben Abacha A & Zweigenbaum P. (2011) *Automatic extraction of semantic relations between medical entities: a rule based approach*. Journal of Biomedical Semantics, 2(Suppl 5):S4.

Boussaid O, Messaoud R, Choquet R and Anthoard S (2006). *XWarehousing: An XML-Based Approach for Warehousing Complex Data*. In the Proceedings of the 10 th East-European Conference on Advances Databases. pp??

Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, Cooper J, Guan W, de Groen PC (2009). *Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model*. J Biomed Inform; 42(5):937-49

Ehrentraut, C, H. Tanushi, H. Dalianis and J. Tiedemann. (2012) *Detection of Hospital Acquired Infections in sparse and noisy Swedish patient records. A machine learning approach using Naïve Bayes, Support Vector Machines and C4.5.* In the Proceedings of the Sixth Workshop on Analytics for Noisy Unstructured Text Data, AND, held in conjunction with Coling 2012, Bombay. pp??

Friedman C, Shagina L, Socratous SA, and Zeng X. (1996) *A WEB-based version of MedLEE: A medical language extraction and encoding system*, In Cimino JJ, ed. Proceeding of the fall 1996 AMIA Conference, p. 938.

Gicquel Q, Dini L, Kergourlay I, Arnod-Prin P, Chariout S, Bittar A, Soualmia L, Guedez P, Segond F, Ruhlmann M, Darmoni S, Metzger MH (2013), SYNODOS SYstème de Normalisation et d'Organisation de Données médicales textuelles pour l'Observation en Santé, Medinfo, FRSIGIMIA, Copenhague - Danemark, August 2013. Pp??

Inmon W (1995). *What is a data warehouse?* volume 1. Prism Tech Topic.

Jensen PB, Jensen LJ, Brunak S (2012). *Mining electronic health records: towards better research applications and clinical care*, Nat Rev Genet; 13(6): 395-405

Kimball R and Strehlo K (1995). *Why decision support fails and how to fixit.* ACM SIGMOD . pp??

Proux D, Hagège C, Gicquel Q, Pereira S, Darmoni S, Segond F. Metzger MH. (2011) *Architecture and Systems for Monitoring Hospital Acquired Infections inside Hospital Information Workflows.* Proceedings of the Second Workshop on Biomedical Natural Language, Bulgaria, pp 43-48.

Soualmia LF, Griffon N, Grosjean J, Darmoni SJ. (2011) *Improving Information Retrieval by Meta-Modelling Medical Terminologies*. 13th conference on Artificial Intelligence in Medicine (AIME): Springer, Heidelberg; 215-219.

Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, Vongsangnak W, Shen B. (2013) *Biomedical text mining and its application in cancer research.* J Biomed Inform; 46:200-211

Zweigenbaum P, Lavergne T, Grabar N, Hamon T, Rosset S, Grouin C. (2013) *Combining an Expert-Based Medical Entity Recognizer to a Machine-Learning System: Methods and a Case-Study*. In: Biomed Inform Insights, N°6 (Suppl 1)p.51–62.