TIDA: A Spanish EHR semantic search engine

Roberto Costumero[1], Consuelo Gonzalo[1], and Ernestina Menasalvas[1]

Universidad Politécnica de Madrid - Centro de Tecnología Biomedica, Madrid, Spain
roberto.costumero, consuelo.Gonzalo, ernestina.menasalvas@upm.es

Abstract. Electronic Health Records (EHR) and the constant adoption of Information Technologies in healthcare have dramatically increased the amount of unstructured data stored. The extraction of key information from this data will bring better caregivers decisions and an improvement in patients' treatments. With more than 495 million people talking Spanish, the need to adapt algorithms and technologies used in EHR knowledge extraction in English speaking countries, leads to the development of different frameworks. Thus, we present TIDA, a Spanish EHR semantic search engine, to give support to Spanish speaking medical centers and hospitals to convert pure raw data into information understandable for cognitive systems. This paper presents the results of TIDA's Spanish EHR free-text treatment component with the adaptation of negation and context detection algorithms applied in a semantic search engine with a database with more than 30,000 clinical notes.

Keywords: Natural Language Processing, Electronic Health Records, Negation Detection

1 Introduction

The adoption and effective use of information technologies in healthcare (in particularly the Electronic Health Record - EHR) has been hindered for decades. The leading English speaking countries have brought new technologies to build health decision-making systems in the last years such as Natural Language Processing (NLP), Negation and context detection algorithms, frameworks like Mayo's cTAKES [1], terminology systems as the Unified Medical Language System (UMLS) [2] or actual systems such as IBM Watson™[3]. Despite the interest shown in the adoption of such technologies in healthcare and having more than 495 million people talking Spanish throughout the world [4] there has not been great efforts in the translation and adoption in the Spanish environment.

In this paper we present the design of TIDA (Texts and Images Data Analyzer), the architecture of a Spanish EHR semantic search engine that makes it possible to: I) process clinical notes and radiological information identifying negation and context of medical terms; II) process images and other media to identify patterns of abnormality and key factors physicians need in order to complete a patient's study and III) extract knowledge using Big Data Analytics using the sources aforementioned and patient's structured data to build intelligent end-user applications. TIDA's components and their functionality are presented over the different sections, though, in this paper, we concentrate on TIDA's component responsible for medical natural text processing.

The main contributions of the paper are: I) to present the architecture to homogenize the structure of the data as the basis for the analytic system; II) to introduce the problems regarding the health related texts in Spanish (clinical notes and imaging reporting) and III) to adapt NegEx and ConText algorithms to be able to deal with Spanish texts, integrating with the output of existing indexing tools so they can be used as the input for the adapted ConText algorithm for Spanish towards semantic enrichment of queries.

The rest of the paper has been organized as follows: Section 2 presents the related works with emphasis on natural language processing . In Section 3 we present the architecture of the system, the algorithms and technologies used. Section 4 presents preliminary results. To end with, conclusion and future works are presented in section 5.

2 Related Work

In Natural Language Processing (NLP) the input of natural language is transformed in several steps of a pipeline to get computers to understand it. In the treatment of free-text NLP, the text serves as an input to the system which will lead to several structured components with semantic meaning so the information can be manipulated knowing the importance of the different parts of the speech. To do the proper training for NLP algorithms to learn, a properly annotated corpus is needed.

These components learn from the training data included in the corpus to analyze future input sentences. Although there are several studies introducing different English corpora to train models to get NLP process working, we focus here on the ones which have been used in the healthcare domain. As cited in Savova et al. [1], there are no community resources such as annotated medical corpus in the clinical domain, so in the evaluation on tools like cTAKES, own corpus has been developed. Using the gold standard linguistic annotations of Penn TreeBank (PTB) [10] and GENIA corpus [11] together with their own Mayo Clinic EMR corpus, cTAKES models were trained. The lack of corpus in Spanish language for the healthcare domain makes such training difficult these days.

The usual NLP pipeline components are a *Sentence Detector*, a *Tokenizer*, a *Part of Speech Tagger*, a *Chunker or Shallow Parser* and a *Named Entity Recognition*. In Figure 2 the process we have implemented is illustrated. Another module that is useful once terms are found is the negation detector that identifies when a particular term or expression appears negated in a text. Depending on the domain context and hypothetical flavor of an expression may also be required.

Natural Language Processing is being extensively used in fields like medicine with many different framework approaches such as cTAKES [1]. It is an open source framework for information extraction from EHR's free-text. It has been developed by the Mayo clinic and open sourced by the Apache Foundation. It is built upon IBM's UIMA (Unstructured Information Management Architecture) [7] and Apache's OpenNLP [8]. cTAKES also includes several components that can be reused independently or executed together in a pipelined process whose output is a structured representation of the unstructured free-text.

cTAKES processes clinical notes written in English identifying different medical concepts from several dictionaries, included own developed ones, but also Unified Medical Language System (UMLS)[2] For this purpose, the algorithm NegEx [9] is used.

Once preprocessed, clinical notes can be searched to find relevant information for clinical and research purposes. In this process two important aspects require special attention: I) negation of terms and II) context of terms. Negation can invert the sense of a term consequently yielding numerous false-positive matches in a search. On the other hand, the context of a term identifies the subject it refers to (also known as the experiencer). This is specially important in clinical notes so to be able to identify the subject of a symptom or disease.

NegEx [9] is a simple algorithm for identifying negated findings and diseases in clinical notes based on the appearance of key negation phrases on the text, limiting the scope of the negation trigger. It is a well known algorithm for negation detection in the healthcare domain and the terms and the algorithm have already been translated and tested in other languages such as Swedish, German and French. In the original paper [9] the algorithm was tested with 1235 findings and diseases in 1000 sentences from discharge summaries. It gave an specificity of 94.5% and a positive predictive value of 84.5% with a sensitivity of 77.8%.

ConText [13] is an extension of the NegEx algorithm that not only identifies negation but also identifies an hypothetical status. It is intended to work moderately at finding the person experiencing the symptoms or whether the condition occurred historically. ConText algorithm [13] was tested with 4654 annotations combining the development and test set of 240 reports (2377 annotated conditions in the development set and 2277 annotated conditions in the test set). Overall negation detection has a F-Measure of 0.93. Overall historical detection has a F-Measure of 0.76 while the overall hypothetical detection has a F-Measure of 0.86.

3 TIDA Spanish EHR semantic search engine

3.1 Introduction

The complexity of healthcare information management is not only due to the amount of data generated but also by its diversity and the challenges of extracting knowledge from unstructured data. Solutions proposed until now have been focused on different aspects of the information process, ranging from unstructured text analysis from discharge summaries and radiology reports, to analysis of PET imaging or Rx, but none of them giving an integrated solution to process and mine obtaining information from all sources.

3.2 TIDA's architecture design

We present TIDA (Text, Image and Data Analytics) a Spanish EHR semantic search engine. TIDA makes it possible to address the problem of Spanish text indexing in the healthcare domain by adapting different techniques and technologies which are explained in this section.

TIDA is designed to build flexible applications over a typical data storage system. TIDA's architecture mainly relies on the information obtained from healthcare databases, which has previously been gathered together into a common storage, so different components get the information from a common warehouse. In order to fulfill this requirements, the architecture (see Figure 1) presents the following components: a **DB** as common data storage system with all hospital's data, from reports to images and patient's structured information which will serve information to the immediate upper level of components; the **Mayo/Apache cTAKES** as free-text analysis system built upon Mayo's cTAKES framework. This framework relies on information gathered through UMLS from different dictionaries for diseases, drugs or laboratory test classifications such as ICD, RxNorm, LOINC or SNOMED CT; an **Image transformation framework** including a set of own developed applications to determine anomalies and automatically annotate medical images using the IBM UIMA architecture which cTAKES is built upon; the **IBM/Apache UIMA** component to gather the two previous components' output in order to get a structured view of the unstructured data; the **Patient's structured data**; a **Structured data, images and text annotator** in charge of annotating text and images supported by UIMA and the structured information; An instance of **Apache Lucene** which indexes all the previously annotated data to serve different kinds of applications; **Apache Solr** to bring quick, reliable semantic search into the picture; An **API** powered by Solr's output, to bring more functionality and link end-user web applications to the whole set of the architecture; and finally, the **End-user web application** to serve end-user applications to give different functionalities on the same data structures.

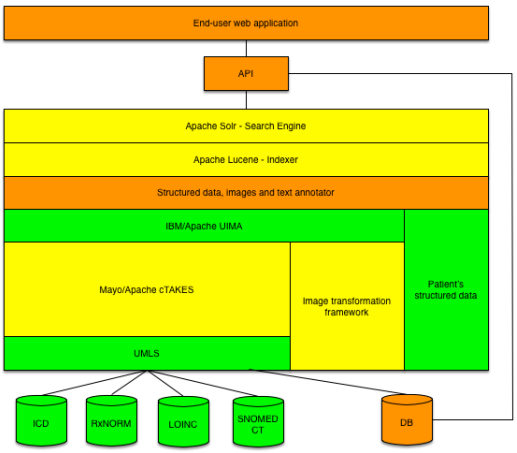

Fig. 1. Principal components of TIDA's architecture

TIDA's architecture is designed to be language-independent, so the platform principles stay the same no matter which language texts are written in. Language-dependent components UMLS and cTAKES can be adapted so the architecture works as expected.

In this paper we present the Spanish adapted version of TIDA, which has been built incorporating the components in Spanish.

3.3 Text analysis in TIDA

It should be noted that, though this architecture is presented to get the whole picture of the work that needs to be done, this paper concentrates in the advances obtained in the free-text analysis. Figure 2 shows the process followed in the free-text analysis done in this paper.

Due to the lack of health related corpora in Spanish, we decided to use a general domain annotated corpus, so at least the models can be trained to be used with Spanish words. AnCora [14] is one of the fully annotated Spanish corpus containing more than 500,000 words.

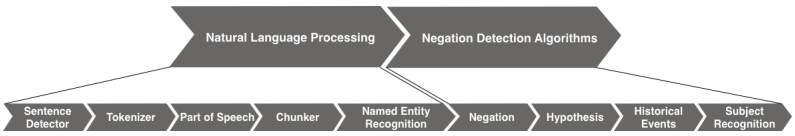

Fig. 2. TIDA's text analysis process

There should be considered two main processes in the text analysis:

1. **Indexation**, which is a heavy process due to a complete text analysis and which runs in a pipeline the different processes presented in Figure 2, starting with Natural Language Processing and following with the Negation detection algorithms.
2. **Searching**, which thanks to indexation is a very lightweight process.

TIDA's text analysis relies on cTAKES [1], which has been briefly introduced in Section 2. cTAKES is an English-centric development, which means that, although it has a very good design and is very modular, it has been developed towards an English comprehensive system. This introduces a challenge to make it suitable to use in other languages. It is not on the scope of this paper to introduce cTAKES architecture, but we will introduce key components to develop our work.

cTAKES uses a common NLP pipeline including a Sentence Detector, a Tokenizer, a Part of Speech Tagger, a Chunker and a Dictionary Lookup for Named Entity Recognition. Although this components should be modified, using the newly trained models to work in Spanish, we are focusing on this paper in the Assertion module, which has been developed outside the scope of cTAKES in a prototype to prove its functionality prior its integration in cTAKES pipeline. cTAKES assertion module, the one involved in negation detection, is the one replicating the functionality on the analysis of the negation, hypothesis and the experiencer on a particular condition.

We are working in line with the NegEx algorithm, which relies on several keywords acting as "negation triggers". The translation of the triggering terms helps us to detect

Trigger in English	Trigger in Spanish
can be ruled out can rule him out	se puede descartar
no, not	no
no evidence	sin evidencia, no evidencia
no suspicion of	ninguna sospecha de

Table 1. Example of NegEx triggers' translation

in a moderated way the detection of the negation in Spanish. An example of such terms can be seen in Table 1.

When analyzing the experiencer for a certain condition in ConText, triggering terms must be used. The results obtained when querying for the condition must have the actual condition but associated to an experiencer distinct from the patient. Examples of different historical and hypothetical triggering terms can be found in Table 2.

Historical in English	Historical in Spanish	Hypothesis in English	Hypothesis in Spanish
again noted	observado de nuevo	could	podría
change in	cambio en	likely	probablemente
chronic	crónico	looks like	parece
clinical history	historia clínica de	not certain	sin certeza

Table 2. Example of ConText triggers' translation

4 Results

TIDA's text analysis component has been trained with AnCora [14] corpus, which has more than 500,000 words. Around 500 words have been translated for the negation detection algorithms. In order to test the system clinical notes and reports (> 30,000) from different services of several public hospitals in Spain have been provided to test the system doing the proper indexation and then querying the system with some medical terms. In this section, we show some results of the application of the process both indexation and searching through the different processes of free-text analysis. The software has been executed in a machine with a Quad-core 2.83GHz CPU with 4GB of RAM.

Figure 3 shows a particular case used to test the NLP component before doing the indexation of the clinical notes with Apache Lucene [15]. This sentence has been extracted from a clinical note. This same figure serves to see what the actual output of the Sentence Detector would be. Tokenization is shown below the Sentence Detection, where 11 are returned, one for each of the words in the sentence. Afterwards, the process analyzes the part of speech for each token. Note that *NN* stands for a Noun, *JJ* stands for an Adjective, *IN* stands for a preposition and *NNS* stands for a Noun in plural, following Penn Treebank's PoS tags [10], which are widely understood in English

speaking environment. Finally, the chunker output is found. PoS tokens are grouped into: Noun Phrases (NP), Verb Phrases (VP), Prepositional Phrases (PP) or Others (O).

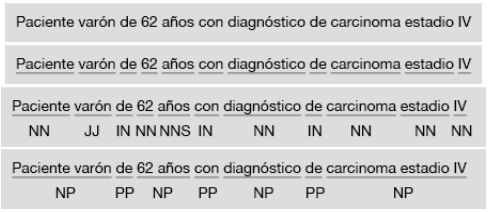

Fig. 3. NLP process example for a given sentence

Testing of the searching system has been done with several queries using Apache Solr [16] to get the user input and the results displayed in a user-friendly environment. A particular case is the use of the metastasis medical term. Figure 4 shows an example of the query results for "metástasis" (Spanish spelling of metastasis) in which the patient actually has metastasis and below the result where the patient does not. The system returned a total of 6,835 results found in 7ms. In the latter query, the system returned a total of 6,296 results found in 8ms. As indexation is done separately and before the process of the actual search, the performance of the searching process is very good, with response in the range of milliseconds. However, the analysis of the performance of the system is left for future work.

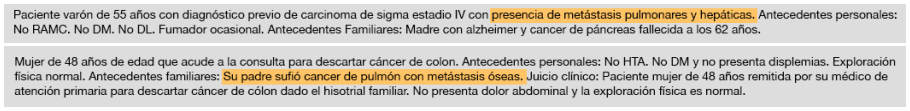

Fig. 4. Metastasis positive and negative query result

5 Conclusions and future work

The increase of data generated and the adoption of IT in healthcare have motivated the development of systems such as IBM Watson and frameworks like cTAKES (integrated exclusively in an English speaking environment). The huge amount of Spanish speaking people leads to the development of new solutions adapted to Spanish. Thus, in this paper we have presented TIDA, an approach which focuses into the generation of a complete EHR semantic search engine which brings an integrated solution to medical experts to analyze and identify markers in a system that brings together text, images and structured data analysis.

Results presented demonstrate that the adaptation of existing algorithms and technologies to the Spanish environment is currently working and that cognitive systems

can be built to work in the health domain. The proper treatment of Spanish texts together with the correct adaptation of ConText algorithm lead to the correct indexation for better queries.

Future work leads to the process on integrating the newly trained models and the Spanish adapted algorithms into the cTAKES framework as well as validating the whole system's performance is left to future work. Also, we are leaving for future work the application of data mining techniques so the indexing component boosts medical related terms and the search engine returns better results.

References

1. SAVOVA, Guergana K., et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association, 2010, vol. 17, no 5, p. 507-513.
2. BODENREIDER, Olivier. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 2004, vol. 32, no suppl 1, p. D267-D270.
3. IBM Watson. http://www.ibm.com/watson (December 13th, 2013).
4. Cervantes Institute. *El español: una lengua viva. Informe 2012*. [on-line] http://cvc.cervantes.es/lengua/anuario/anuario_12/i_cervantes/p01.htm (January 8th, 2014).
5. Logica and Nordic Healthcare group. *Results from a survey conducted by Logica and Nordic Healthcare Group January 2012*. [PDF] Available at: http://www.logica-group.com/we-are-logica/media-centre/thought-pieces/2012/market-study-of-electronic-medical-record-emr-systems-in-europe/~/media/Global%20site/Media%20Centre%20Items/Thought%20pieces/2012/WPPSEMRJLv16LR.ashx (January 7th, 2014)
6. HM Hospitales. *Estadísticas y Resultados Sanitarios 2012*. [on-line] http://www.hmhospitales.com/grupohm/Estadisticas/Paginas/Estadisticas-Generales.aspx (January 7th, 2014)
7. FERRUCCI, David; LALLY, Adam. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering, 2004, vol. 10, no 3-4, p. 327-348.
8. Apache OpenNLP. http://opennlp.apache.org (November 21st, 2013)
9. CHAPMAN, Wendy W., et al. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics, 2001, vol. 34, no 5, p. 301-310.
10. MARCUS, Mitchell P.; MARCINKIEWICZ, Mary Ann; SANTORINI, Beatrice. Building a large annotated corpus of English: The Penn Treebank. Computational linguistics, 1993, vol. 19, no 2, p. 313-330.
11. KIM, J.-D., et al. GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics, 2003, vol. 19, no suppl 1, p. i180-i182.
12. ILSP. *Hellenic National Corpus*. [on-line] http://hnc.ilsp.gr/en/ (January 9th, 2014)
13. HARKEMA, Henk, et al. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of biomedical informatics, 2009, vol. 42, no 5, p. 839-851.
14. TAULÉ, Mariona; MARTÍ, Maria Antònia; RECASENS, Marta. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. En LREC. 2008.
15. Apache Lucene. https://lucene.apache.org/core/ (November 21st, 2013)
16. Apache Solr. http://lucene.apache.org/solr/ (November 21st, 2013)