Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 259-262 recibido 26-06-10 revisado 03-07-10 aceptado 06-07-10

First Steps in The Manual and Automatic Annotation of Clinical Notes in Spanish

Primeros pasos en la anotación tanto manual como automática de informes clínicos en Español

M. Oronoz
IXA Group
U. of the Basque Country
maite.oronoz@ehu.es

A. Diaz de Ilarraza
IXA Group
U. of the Basque Country
a.diazdeilarraza@ehu.es

O. Torices
IXA Group
U. of the Basque Country
otorices@ikasle.ehu.es

Resumen: Este artículo presenta la anotación de un corpus de informes clínicos de pacientes de ginecología y obstetricia, así como el desarrollo de un esquema de anotación para su etiquetado manual. Centramos nuestra descripción en el etiquetado manual de los informes y en la adaptación de la herramienta para el Procesamiento del Lenguaje Natural Freeling al dominio médico.
Palabras clave: Corpus, Informes Clínicos, Procesamiento del Lenguaje Natural

Abstract: This paper presents the annotation of a corpus of gynaecology and obstetrics patient records and the development of an annotation scheme for its hand tagging. We focus our description in the manual annotation of the clinical notes and in the adaptation of the Natural Language Processing analyzer Freeling to the medical domain.
Keywords: Corpus, Clinical Notes, Natural Language Processing

1 Introduction

The generation of a text corpus of clinical notes is the basis for the development of many tools and applications for medicine that could doubtlessly improve the quality of care. Although the amount of available biomedical corpora is large (GENIA, PennBioIE, ...), there is not any available annotated corpus in the clinical domain (Roberts et al., 2009) in Spanish.

This paper describes the first steps in the development of a corpus of clinical notes and is organized as follows: In section 2 we explain the way we analyze automatically the structure of the notes and the creation of the annotated corpus together with the initial guidelines. In section 3 we describe the work accomplished in the use of Natural Language Processing (NLP) tools for the automatic tagging of the corpus. We integrate a medical abbreviation and acronym dictionary for Spanish (Yetano y Alberola, 2003) in FreeLing[1]. This initial adaptation has some consequences in the performance of Freeling that are listed. Besides the work done at morphological level, we explain the experiment designed for the extraction of semantic information from the corpus. Finally, we present some conclusions and future work.

2 Manual Annotation

We received 400 semi-structured discharging notes (years 2000-2007) of the gynecology and obstetric services of the Cruces Hospital[2] with a total of 63125 words. These documents comprise data describing subjects that go from child birth to breast cancer operations. The steps followed to process the texts before manual annotation are: i) automatic analysis of the structure, ii) definition of the elements to tag and selection of the annotation tool, and iii) design of the initial guidelines.

2.1 Automatic Analysis of the Structure

We write a set of regular expressions to standardized the structure of the documents obtaining XML files together with the corresponding DTD and XML-Schema. Figure 1 shows the result of the structural analysis. The structural information indicates aspects

[1] FreeLing [http://www.lsi.upc.edu/ nlp/freeling/]

[2] Cruces Hospital [http://www.hospitalcruces.com]

ISSN 1135-5948

such as "Family Medical History", "treatment" and so on. This information will be very useful in the identification of clinical concepts because it could be inferred that each concept type usually appears in some specific parts of the notes. For example, "Body Parts" and "Procedures" usually appear in the operation section.

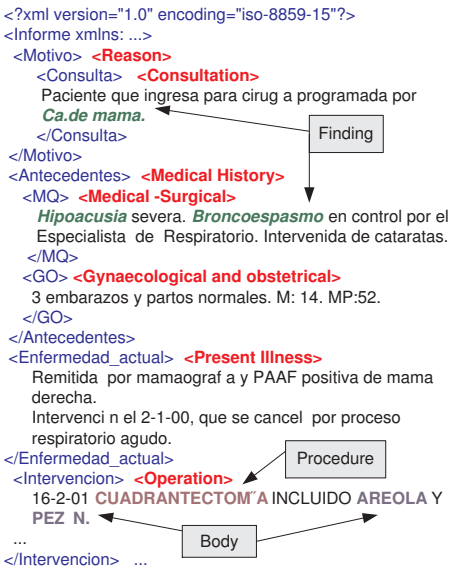

Figure 1: Output of the Structural Analysis.

2.2 Annotation Process

For the selection of the annotation tool we examine different options: WordFreak[3], Callisto[4], iAnnotate[5], GATE[6] and UIMA(Ferrucci y Lally, 2004). We selected Knowtator[7] because it is open-source, well documented, and it uses stand-off annotation.

We randomly chose a subset of 10 documents and two linguists jointly annotated them in order to produce the initial guidelines. During the annotation process, the class hierarchy was updated and redefined to represent the phenomena found in the corpus.

2.3 Initial Guidelines

Table 1 describes the categories defined and their frequency of appearance. We identified 6 main categories in the annotation schema: i) entities based on the SNOMED CT[8] concept categories, ii) phrase types, iii) misspellings, iv) abbreviations, v) dates and vi) measures.

Main Category	Sub-Category	Number of tags	% of Tags
BioEntity	Finding	126	15.29
	Procedure	82	9.95
	Substance	27	3.28
	Body	41	4.98
	Occupation	12	1.46
	Observable	23	2.79
	Social Context	20	2.43
	Qualifier	102	12.38
	Physical obj.	1	0.12
	Environment	19	2.31
Phrase Type	Negation	13	1.58
	Concessive	1	0.12
	Condition	2	0.24
	Ellipsis	1	0.12
	Cause	2	0.24
Misspelling		70	8.5
Abbreviation		133	16.14
Date		18	2.18
Measure	Atomic	39	4.73
	Range	4	0.49
	Percent	6	0.73
	Dose	17	2.06
	Size	17	0.26
	Time period	36	4.37
	Weight	12	1.46
Total		824	

Table 1: Annotation Schema.

As expected we found several difficulties when annotating the texts.

- When treating **BioEntity** class, several doubts arise relating to the classification of a concept as belonging to "observable" or "finding" classes. In these cases we take the category defined in SNOMED CT for this concept.
- The subclasses in the **Phrase-Type** class are syntactic (negation, condition...) while the other classes contain basically semantic and conceptual information. In our opinion the identification of, for instance, expressions indicating the lack of a concept (e.g. *sin alergias medicamentosas'* 'without drug allergies') could be as important as the appearance of the concept itself.
- We defined subclasses for the **Measure** class to identify among others ranges, doses ...The great variety of ways for giving information about doses can be a clear source for disagreement.
- Under the class of **Abbreviation** we include acronyms and abbreviations.

[3]http://bioie.ldc.upenn.edu/wiki/index.php/Main_Page
[4]http://callisto.mitre.org/
[5]http://www.dbmi.columbia.edu/ cop7001/iAnnotateTab/
[6]http://www.gate.ac.uk/projects.html
[7]http://knowtator.sourceforge.net
[8]SNOMED CT [http://www.ihtsdo.org/snomed-ct/]

- Misspelling's category has been one of the most frequent and, problematic.

The guidelines gather annotation decisions taken to solve detected inconsistencies. As we use standard SNOMED concepts and general linguistic concepts (misspellings, abbreviations...) we think that the annotation schema is quite general and not specific for this corpus. The inter-annotator agreement is of 91.88% for class matching and 83.93% in class/span matching.

3 Use of NLP tools

3.1 Morphological Analysis

"Freeling is an open-source multilingual language processing library providing a wide range of language analyzers for several languages" (Padró et al., 2010). We use the tools for morphological analysis in Spanish provided by Freeling. The linguistic resources (lexicons, grammars...) in Freeling could be modified, so we take advantage of this by extending some of the linguistic data containing files with medical abbreviations. There are only a few analyzers adapted to the clinical domain, e.g. the GENIA tagger (Tsuruoka et al., 2005), but no one for Spanish. This is the first step in the adaptation of Freeling to this domain.

Regarding the integration of abbreviations in Freeling we must distinguish those ending with a full stop (e.g. ca.= *'carcinoma'*) from those without it (e.g. de= *'disfunción erectil'*, erectil disfunction). In the last case, the abbreviation is added only to the dictionary, while those ending with a full stop are integrated also in the file related to tokenization. Some other not medical abbreviations appear very often in clinical records, (e.g. dra=*'doctora'*, female doctor...) but are not properly analyzed in Freeling.

A fact to be checked when adding medical abbreviations to the analyzer, is the increment of the morphological ambiguity. For example, in Freeling it has increased from 1.374 analysis per word to 1.721. This asks for the retraining of Freeling's statistical disambiguator to adapt it to the medical domain. Let us see some examples of the newly generated ambiguity:

- *Words having both, medical and not medical analysis.* In the expression *'durante la lactancia'* (while breastfeeding) the word *'la'* (which correct analysis is definite article) has 4 more meanings when being analyzed as a medical abbreviation (see table 2). Freeling selects the appropriate analisy with a probability of 0.972094 out of 1.
- *Words having all the analysis related to the medical domain.* In the sentence *'G^o Rh: B'*(Rh G^o: B), the abbreviation 'Rh' has three possible analysis; all of them belonging to the clinical domain[9]. Freeling analyses the abbreviation giving the same probability, 0.33, to all the analysis, instead of choosing 'the rhesus factor'. This is a normal behaviour as the disambiguator has not been trained for this domain.

Regarding to the analysis of misspellings. The module in Freeling that assigns the mentioned probabilities, also works as an unknown word guesser. Being misspellings unknown words, Freeling usually gives them a lemma and tries to guess a POS. For example, in the expression *'Pópidos cervicales'* the analysis assigned to the incorrect word *'Pópido'* (instead of *Pólipo*, polyp) is `Pópipos-pópipos-CNMP`[10]`-1.0`. In this way, we can not differentiate a correct word from a misspelled one. In the manual annotation, misspellings are one of the most frequently annotated categories, so their identification is very important. For this reason, we need to apply an independent speller before the analysis or to change the word guesser in Freeling.

In summary, the inclusion of medical abbreviations in any tagger usually increases its ambiguity. This is a general problem for everyone that wants to adapt a general purpose analyzer. The problem with the misspellings in Freeling is very specific of this tool.

3.2 Kyoto

One of the aims of the Kyoto[11] project is to allow people to extract knowledge and facts from texts. With this purpose the Kybot (*Knowledge yielding robot*) technology was defined. Kybots are programs that use concepts already connected to ontologies to detect actual concept instances and relations in

[9] *'receptores hormonales'* (hormone receptors), *'factor rhesus'* (rhesus factor) and *'símbolo del rodio'* (rhodium symbol)

[10] CNMP: common noun masculine plural.

[11] See the http://www.kyoto-project.eu/ webpage.

Form	Lemma or expanded form	POS	Probability
la	*'lo'*	Personal feminine pronoun	0.027710
la	*'la'*	Common noun (musical note)	0.000039
la	*'lactancia artificial'* (artificial breast feeding)	Medical abbreviation	0.000039
la	*'linfadenectomía axilar'* (axillary lymphadenectomy)	Medical abbreviation	0.000039
la	*'líquido amniótico'* (amniotic fluid)	Medical abbreviation	0.000039
la	*'líquido ascítico'* (ascitic fluid)	Medical abbreviation	0.000039
la	*'el'*	Definite feminine article	**0.972094**

Table 2: Analysis of the word *'la'* (definite article).

text".

The Kyoto project usually works in the environmental domain, but its technology can be easily adapted to the medical domain. For example, we could define Kybots to establish relationships between clinical procedures and the body parts they affect (see figure 2). For example, it would be possible to extract documents containing information related to the following request: *"Casos clínicos en los que se ha realizado una* ***cuadrantectomía*** *para el tratamiento del carcinoma de* ***mama****."*[12]

After applying the Kybot in figure 2 the document in figure 1 is retrieved.

```
Process:  'cuadrantectomía' quadrantectomy
Involves:  'areola', 'pezón' areola, nipple
When:  16-01-0
```

Figure 2: BioKybot.

4 Conclusions and Future Work

In this paper we have presented the first steps in the annotation of discharging records in Spanish in two ways: i) manually by means of a graphical annotation tool with the aim of obtaining the initial guidelines for the annotation of the whole corpus and, ii) automatically, using a first adaptation to the medical domain of the Freeling analyzer for Spanish. In the manual annotation, misspellings and abbreviations are two of the categories that appear most frequently. Regarding NLP processing, we want to face the problem of the misspellings by identifying and correcting them as in (Patrick et al., 2010). We have improved the understanding of the abbreviations by their inclusion into the dicctionaries of Freeling. In the future we must work in the disambiguation of these abbreviations.

[12]Clinical cases in which it has made a **quadrantectomy** for the treatment of **breast** carcinoma

The adaptation of Freeling to the medical domain requires some changes: i) the tokenizer has to identify tokens such as doses and measures, ranges, periods of time... ii) it is important to identify clearly misspellings from unknown words and, iii) the process of disambiguation of abbreviations must be integrated.

Bibliografía

Ferrucci, David y Adam Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4).

Padró, L., S. Reese, E. Agirre, y A. Soroa. 2010. Semantic Services in Freeling 2.1: WordNet and UKB. En *Global Wordnet Conference*, Mumbai, India.

Patrick, J., M. Sabbagh, S. Jain, y H. Zheng. 2010. Spelling Correction in Clinical Notes with Emphasis on First Suggestion Accuracy. En *LREC, 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, Valletta, Malta.

Roberts, A., R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, y A. Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, 42.

Tsuruoka, Y., Y. Tateishi, J. Kim, T. Ohta, J. McNaught, S. Ananiadou, y J. Tsujii. 2005. Developing a Robust Part-of-Speech Tagger for Biomedical Text. En *10th Panhellenic Conference on Informatics*, Genoa, Italy.

Yetano, J. y V. Alberola. 2003. *Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias.* Ministerio de Sanidad y Consumo.