Annotating clinical notes in Spanish for NLP based text mining

Text mining in biomedical text corpora could doubtlessly improve the quality of care by the development of decision support systems or evidence-based medicine. Although the amount of available biomedical corpora is large (GENIA, PennBioIE …), there is not any available semantically annotated corpus in the clinical domain [1] and in Spanish. This abstract presents the preprocessing of a corpus of gynaecology and obstetrics patient records and the development of an annotation schema for its hand tagging (problems and doubts generated and obtained inter-annotator agreement). We will focus our description in the manual annotation of the clinical notes, but we have already accomplished some work in the use of Natural Language Processing tools for the automatic tagging of this corpus. We integrated a Spanish medical abbreviation dictionary [2] in FreeLing[a] and we designed some experiments for adapting Kybot technology from the Kyoto[b] project to extract the medical semantic information in the corpus, especially that related to clinical procedures.

Regarding manual annotation, we received 400 semi-structured and de-identified clinical notes (years 2000-2007) from the Cruces Hospital[c]. The structure of these notes was automatically annotated and standardized in XML. The structural information indicates, the "Family Medical History" of the patients, the "treatment" and so on. In our opinion, this structural information will be very useful in the identification of clinical concepts as each concept type usually appears in some specific parts. We randomly chose a subset of 10 documents and two human annotators jointly annotated them by means of the Knowtator[d] tool to produce the initial guidelines. Table 1 describes the categories indicated by the annotators and their frequency of appearance. We identified 6 main categories in the annotation schema: i) entities based on the SNOMED CT[e] concept categories, ii) phrase types, iii) misspellings, iv) abbreviations, v) dates and vi) measures. Misspelling's category has been one of the most frequent and at the same time, problematic. The guidelines created so far gather concrete annotation decisions taken to solve, among others, these problems. Knowtator's functionality for measuring inter-annotator agreement establishes a value of 91.88% for class matching and 83.93% for class/span matching.

---

[a] FreeLing
[http://www.lsi.upc.edu/~nlp/freeling/]. FreeLing is an open source language analysis tool suite that includes from a lemmatizer to a semantic disambiguation tool.
[b] Kyoto Project
[http://www.kyoto-project.eu/]
[c] Cruces Hospital
[http://www.hospitalcruces.com/elHospitalPresentacion.asp?lng=en]
[d] Knowtator
[http://knowtator.sourceforge.net/]
[e] SNOMED CT
[http://www.ihtsdo.org/snomed-ct/]

Table 1: Annotation Schema

Main Category	Subcategory	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 object	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
	Period of Time	36	4.37
	Weight	12	1.46
Elements in total		**824**	

References

1. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts and A. Setzer. **Building a semantically annotated corpus of clinical texts**. Journal of Biomedical Informatics 42, pages 950-966. 2009

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