Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names

Maite Oronoz[1], *Arantza Casillas*[2], *Koldo Gojenola*[1], *and Alicia Perez*[1]

[1] Departamento de Lenguajes y Sistemas Informáticos. IXA taldea. UPV-EHU
[2] Departamento de Electricidad y Electrónica. IXA taldea. UPV-EHU
maite.oronoz@ehu.es, arantza.casillas@ehu.es, koldo.gojenola@ehu.es, alicia.perez@ehu.es

Abstract. This paper presents an annotation tool that detects entities in the biomedical domain. By enriching the lexica of the Freeling analyzer with bio-medical terms extracted from dictionaries and ontologies as SNOMED CT, the system is able to automatically detect medical terms in texts. An evaluation has been performed against a manually tagged corpus focusing on entities referring to pharmaceutical drug-names, substances and diseases. The obtained results show that a good annotation tool would help to leverage subsequent processes as data mining or pattern recognition tasks in the biomedical domain.

Index Terms: development of linguistic tools, annotation, medical domain.

1 Introduction

Syntactic and semantic annotation has been used in many applications such as data mining and pattern recognition. There are a variety of supervised and semi-supervised training algorithms that require to be boosted from annotated data sets. The aim of this paper is to automatically annotate different types of entities in the biomedical domain.

Over the last years Spanish health care services are storing most of the information concerning patients in electronic medical records. These clinical texts constitute a rich source of information about diseases, allergies, and any information that the sanitary personnel is interested in. Access to this information is of great interest and value for clinical research. Many current methods for accessing information are based on statistical and machine learning methods, that need annotated data. However, the annotation process is time-consuming and expensive to be performed manually. Biomedicine is an area where the corpora have a confidential nature, hence, open resources are scarce and when comparing it to other fields it does not seem an eligible task for exploiting publicly available resources such as the semantic web, althouh there are some publicly available resources such as parallel corpora in various languages [1, 2]. Besides, the annotators' expertise is crucial, and thus, it is not an option for crowd sourcing or

social annotation as it was done in other tasks like language model adaptation [3]. Making this an automatic process would allow to save work and money.

The Pharmaceutical Service of the Galdakao's Hospital performs the task of manually detecting Adverse Drug Reactions (ADRs). The aim of the tool presented in this paper is to automatically annotate medical texts with brand-name drugs, disease names and substance names, opening the way for the future automatic detection of ADRs.

Figure 1 shows a fragment of a clinical note with annotations for diseases, substances and drugs as well as allergies and adverse drug effects, obtained by means of Brat [4], a tool for text annotation. This tool allows not only to highlight such events but also to detail cause-effect relations. Note that while the figure shows a manual annotation provided by medical experts, the aim of this work is to produce the annotation automatically. As a result, reading a clinical note (or conversely, supervising a dictated note) would be easier, since this tool would allow to draw the attention to specific items.

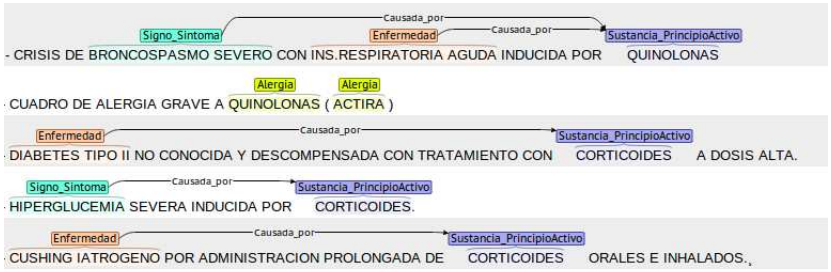

Fig. 1. Medical record manually annotated with the Brat toolkit.

The core element of the proposed automatic annotation toolkit lies in creating a syntactic and semantic analyzer for Spanish in the specific domain of biomedicine. In this paper, we will focus on the description of the adaptation of the linguistic analyzer Freeling [5] to the domain of medicine. The annotations provided by the presented domain-adapted analyzer will be evaluated with respect to annotations provided by human experts. The benefits of having an automatic analyzer are twofold: (1) automatic annotation is much faster and cheaper (2) the annotated data will serve for developing advanced information extraction and data mining systems.

There are only a few publically accesible analyzers adapted to the clinical domain in the Spanish language. For English, the GENIA tagger [6] is specifically tuned for biomedical texts. Patrick et al. [7] introduce a new method to automatically identify medical concepts from the Systematized Nomenclature of Medicine-Clinical Texts (SNOMED CT) in English free text. MetaMap Transfer (MMTx) [8] is a program to map biomedical text to the UMLS[3] Metathesaurus or, equivalently, to discover concepts from the Metathesaurus in texts. In [9] the

[3] http://www.nlm.nih.gov/research/umls/

authors present a first simple approach to the Spanish MetaMap, using Google Translator to obtain an English version of the text and then applying English MMTx to extract the concepts. In [10] a system for the automated identification of biomedical concepts in Spanish-language clinical notes is presented.

The rest of the paper is arranged as follows: Section 2 delves into the adaptation and enrichment of the linguistic analyzer. Section 3 is devoted to the experimental evaluation of the tool against a set of manually annotated texts. Finally, conclusions and future work are given in section 4.

2 Automatic Analysis of Electronic Medical Records

For the initial processing of medical records, we have made use of a basic Natural Language Processing toolkit, *Freeling*[4], together with several available medical ontologies and dictionaries. Freeling is an open-source multilingual language processing library providing a wide range of language analyzers for several languages [5]. In this work, we used the tools for Spanish morphological analysis provided by Freeling. The linguistic resources (lexica, grammars, ...) in Freeling can be modified, so we took advantage of this flexibility by extending the linguistic data files with large-scale resources containing medical information.

As it is a standard approach in Natural Language Processing, where there is a distinction between morphology and syntax on one side and semantics on the other, we will distinguish two levels of processing. In our case, during morphosyntactic processing, our system will only categorize word-forms using their basic part-of-speech (POS) categories (explained in section 2.1), while the semantic distinctions will be dealt with in a second stage (see section 2.2). Following this approach, if the term that we want to insert already existed in Freeling's standard Spanish dictionary, e.g. *bar* as common noun (bar or pub), the entries with medical meanings will not be added to the lexicon, e.g. *bar* or *bacilo acidorresistente* (acid-fast rod) because this term also corresponds to a common noun. The medical meanings are added in a later semantic tagging phase. This solution helps to avoid an explosion of ambiguity in the morphosyntactic analysis and enables a clear separation between morphosyntax and semantics.

2.1 Enriching dictionaries in Freeling

In order to extend the standard Freeling analyzer in Spanish to the medical domain, we enriched two dictionaries: a basic dictionary of terms consisting of a unique word, and a multiword-term dictionary. The former should be enriched with terms such as *enteroplastia* (repair of intestine), and the latter with composed terms as, for example, *canal vertebral lumbar* (lumbar spinal canal).

As we previously explained, to keep the distinction between morphosyntactic and semantic ambiguity in the lexica is essential for us. We decided to add a term to the files with POS information in Freeling only if it did not exist before.

[4] http://nlp.lsi.upc.edu/freeling/

For example, *xilosuria* (xylosuria) or *Zofenil* (pharmaceutical product) will be new entries applying this principle. The first row in Table 1 shows the number of entries of the standard lexica for Spanish within the Freeling 2.2. standard package. The medical resources used to enhance them are the following:

Medical Abbreviations. Yetano and Alberola [11] gathered the abbreviations used in some hospitals to develop a dictionary of medical abbreviations and acronyms for Spanish. After a manual examination, we obtained a list of 3,196 entries. Some of them were ambiguous, e.g. *ADR* meaning *adrenalina* (adrenalin), or *adriamicina* (adriamycin), while others were not, e.g. *HTA* (*Hipertensión arterial* for "high blood pressure"). Table 1 shows the number of abbreviations already contained in the standard lexica (first row in Table 1), and the number of new abbreviations. The majority of the abbreviations are new entries in the Freeling lexica because they correspond specifically to the medical language (e.g. *vvz* extended *virus varicela zoster*). All the abbreviated chemical elements (e.g. *as*, *bi*), measure units (e.g. *kg*, *cm* ...) were already in the lexica.

SNOMED CT Terms. SNOMED CT is a comprehensive clinical terminology that provides clinical content and expressivity for clinical documentation and reporting. SNOMED CT is based on concepts, that is, units of thought or clinical ideas, coded by means of alphanumeric identifiers (e.g. 106190000 refers to *allergy*). Concept-descriptions are classified into *Fully Specified Name* in which the hierarchy the term belongs to is indicated (body part, procedure...), *Preferred Terms* and *Synonyms*. We have added the preferred terms and the synonyms of the 31th of October 2011 release to the lexica in Freeling.

The Unified Medical Language System (UMLS), is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems. SNOMED CT is part of the Metathesaurus knowledge source in UMLS. We tagged the terms in Spanish with their corresponding SNOMED CT identifiers but also with their UMLS identifiers (see figure 2). In this way we will have the option of accesing the other ontologies in UMLS and of getting additional medical information.

Table 1 shows that 94.1% of the terms from SNOMED CT have more than one word and 94% of them were new in the multiword-term file. This fact gives an idea of the complexity of the terms used in SNOMED CT. In proportion, the number of single word terms already in the dictionaries, 9,302 out of 23,399, is relatively high, compared to the number of locutions or multiword terms.

Bot PLUS. Bot PLUS is a database of sanitary knowledge distributed by the General Council of Spanish Pharmacologists[5]. Bot PLUS stores the names of all the medicines that are commercialized in Spain. The knowledge stored in the Bot PLUS database makes up for the lack of this kind of information in SNOMED CT. For the work presented in this paper, we have obtained the following lists:

[5] http://www.portalfarma.com

i) brand names or pharmaceutical drug names and ii) substances. Table 1 shows the lexical entries incorporated to Freeling, having Bot PLUS as a basis.

Regarding the insertion of medicine brand-names in the lexica, it is worth remarking that from 9,984 names, 9,902 entries are new in the lexica and only 82 existed already (e.g. *rizan* from the verb "to curl"). In the case of substance-names with a unique token, there are more terms already in the dictionaries (1,590) than those entered as new ones (1,406) because they have their place in SNOMED CT, and they were already in the lexica.

Table 1. Number of entries in the lexica of Freeling and added resources.

		Unique word terms	**Multiword terms**	**Total**
FreeLing	Standard	556,212	1,480	557,692
Abbreviations	In dictionary	369	4	373
	New	2,654	169	2,823
	Total	3,023	173	3,196
SNOMED CT	In dictionary	9,302	125	9,427
	New	23,399	521,973	545,372
	Total	32,701	522,098	554,799
Bot PLUS	**Medicine brand-names**			
	In dictionary	61	21	82
	New	3,746	6,156	9,902
	Subtotal	3,807	6,177	9,984
	Substances			
	In dictionary	1,590	158	1,748
	New	1,406	1,072	2,478
	Subtotal	2,996	1,230	4,226
	Total	6,803	7,380	14,210
ICD-9	In dictionary	530	1,029	1,559
	New	268	17,950	18,218
	Total	798	18,979	19,777

ICD-9. The International Statistical Classification of Diseases is a medical classification list compiled by the World Health Organization (WHO). All the medical records from the Basque Health System should be tagged with a code indicating the medical diagnosis of the patient, following the 9th version of this classification (ICD-9). Table 1 shows the data about the integration of these terms in Freeling's lexica and the complexity of the terms in ICD-9.

The four lexica have been integrated in Freeling in their order of appearance in the paper, that is, abbreviations first, and then SNOMED CT, Bot PLUS and ICD-9. We decided to give priority to SNOMED CT against Bot PLUS and ICD-9, because it is a well structured and extensive clinical terminology. The expansion of the abbreviations first is essential if we want to add meanings, e.g. from SNOMED CT, to the expanded lemmas.

2.2 Semantic Postprocess

With the augmented lexica, Freeling performs tokenization, morphological analysis, POS tagging, lemmatization, shallow parsing and dependency parsing. The medical records are analyzed with linguistic information at all these levels but at

the present work we will make use of information about terms, which gives access to all information levels except syntactic dependencies. All the entries described in section 2 have been inserted as nouns in the lexica, but also indicating the source of information of each entry.

In case of an ambiguity of meanings, that ambiguity would correspond to the semantic level. Being this the case, we insert this medical information and, in consequence, ambiguity in the analysis. The example in figure 2 shows that the word *Estreptomicina* was already in Freeling as a common noun feminine singular (tag NCFS000). For medical information extraction tasks, it is important to know that this is a *substance* or *product*, so we will insert this information as an *External Reference* (*extRef*). In the *extRef* we include information about the *resource* (Snomed CT in Spanish version of the date 31 October 2011), the SNOMED CT *Concept Identifier* in the *reference* attribute and the *reftype*, in our case corresponding to the semantic tag of the term in SNOMED CT (*product* and *substance*). For future works we aim to access the entire UMLS, this is why we have also inserted the *UMLS's Concept Unique Identifier* in the analysis.

Overall, the enhancement process of the lexical resources adds 47,132 standard entries and 554,807 locutions, taking an outstanding step ahead in text processing of the biomedical domain.

```
<term lemma="estreptomicina" pos="N.NCFS000" tid="t56" >
  <extRefs>
    <extRef resource="SCT_20111031" reference="40877002" reftype="producto">
      <extRef resource="UMLS-2010AB" reference="C0038425"/>
    </extRef>
    <extRef resource="SCT_20111031" reference="387223008" reftype="sustancia">
      <extRef resource="UMLS-2010AB" reference="C0038425"/>
    </extRef>
  </extRefs>
</term>
```

Fig. 2. Analysis with augmented information.

3 Evaluation

Although our adapted linguistic analyzer is able to detect terms from the 19 content hierarchies of SNOMED CT (i.e. organisms, procedures,...), one of the first uses of the analyzer will be to detect adverse drug events. This is the reason for focusing our first evaluation in the detection of drug-names, diseases and substances. We distinguish between brand-name drugs (e.g. *Nolotil*) and substances that could be active ingredients (e.g. *Metamizol*) or any substance that could create an adverse drug reaction (e.g. *polen* meaning pollen).

We did not found any publicly avalaible corpus composed of electronic medical records in Spanish, so after several meetings with the legal advice services of the University and the Hospital, and after signing the corresponding confidentiality agreement, we obtained a corpus of patient records. Having a "private" corpus, our results are not comparable to others, as in other related works [10].

A corpus of 100 medical records was collected from the outpatient consultations of the Galdakao Hospital and it was manually tagged by doctors and

pharmacologists. The corpus is composed of 51,061 words and the experts have manually tagged 690 drug names, 891 diseases and 735 substances. The performance of the analyzer was assessed using the manually tagged corpus. These data samples were shuffled and randomly split into three disjoint sets for training (60 documents), development (20 documents) and test purposes (20 documents).

The system is assessed by means of the F-Measure that compares the human annotation with the output of the analyzer by combining precision and recall. In order to set out if two elements are equal, an approximate correctness criteria was applied: two elements are considered to be equivalent if an element given by the system is entirely contained within an extension of a manually tagged element by six positions both to the left and to the right. This follows the standard approach of allowing an approximate boundary matching, as in the BioNLP Shared Task [12]. Table 2 shows the number of drugs, substances and diseases in the test set, also presenting the number of True Positives (TP), False Negatives (FN) and False Positives (FP) returned by the system for each category of elements. Precision (PR), recall (RE) and F-Measure (F-M) are calculated for each type of element. The results are encouraging, with an F-Measure of 0.90, and imply that the designed analyzer can automatically generate reliable annotated corpus with morphosyntatic and medical-concept tags.

Table 2. Results achieved by the automatic tagger on the test set.

	Manual	TP	FN	FP	PR	RE	F-M
Diseases	211	354	88	12	0.97	0.80	0.88
Drugs	180	175	8	0	1.00	0.96	0.98
Substances	184	357	27	65	0.84	0.92	0.88
Total	575	886	123	77	0.92	0.88	0.90

4 Conclusions

The goal of this work was to create an analyzer for clinical texts in Spanish that identifies medical entities. To attain this goal we have added medical information to a standard linguistic analyzer for Spanish. The incorporated information was extracted from different sources such as ontologies, a medical abbreviation dictionary and a pharmaceutical drug element database. The system is robust enough to deal with electronic medical records in which abbreviations and errors are very common. We think that in the same way, it is able to analyze other types of texts within the medical domain (journal papers, books...).

The contributions of this work are threefold: 1) the enhancement of standard Spanish dictionaries for the biomedical domain in the FreeLing toolkit; 2) the development of a system based on FreeLing to automatically annotate medical records providing an F-Measure of 0.90; 3) the compilation of a corpus of medical documents tagged with medical concepts in Spanish.

In the near future, we aim to improve the system by adding, as external references, the missing information about abbreviations, drug names from Bot PLUS and diseases from ICD. This will produce an increase in the semantic ambiguity of the terms. For that reason, we want to use UKB [13], a tool for graph-based word sense disambiguation to select the adequate medical sense.

Acknowledgements

We would like to thank the Pharmacy Service of the Hospital Galdakao-Usansolo and the Pharmacovigilance Service of the Basque Government. This work was supported by the Department of Industry of the Basque Government (IT344-10, S-PE12UN114), the University of the Basque Country (GIU09/19), the Spanish Ministry of Science and Innovation (MICINN, TIN2010-20218).

References

1. Jimeno-Yepes, A., Prieur-Gaston, É., Névéol, A.: Combining medline and publisher data to create parallel corpora for the automatic translation of biomedical text. BMC Bioinformatics **14** (2013) 146
2. Tiedemann, J.: Parallel data, tools and interfaces in opus. Proc. Language Resources and Evaluation (LREC), 2012.
3. Wu, Y., Abe, K., Dixon, P.R., Hori, C., Kashioka, H.: Leveraging Social Annotation for Topic Language Model Adaptation. Proc. International Speech Communication Association (INTERSPEECH), 2012.
4. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: Brat: A web-based tool for nlp-assisted text annotation. Proc. EACL 2012.
5. Padró, L., Reese, S., Agirre, E., Soroa, A.: Semantic Services in Freeling 2.1: WordNet and UKB. In: Global Wordnet Conference, Mumbai, India (2010)
6. Tsuruoka, Y., Tateishi, Y., Kim, J., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a Robust Part-of-Speech Tagger for Biomedical Text. In: 10th Panhellenic Conference on Informatics (2005)
7. Patrick, J., Wang, Y., Budd, P.: An Automated System for Conversion of Clinical Notes into SNOMED Clinical Terminology. In: Proc. Australasian symposium on ACSW frontiers. Volume 68 of ACSW '07. (2007) 219–226
8. Aronson, A.: Effective Mapping of Biomedical Text to the UMLS Metathesaurus: the MetaMap program. In: Proc. of AMIAS. (2001) 17–21
9. Carrero, F.M., Cortizo, J.C., Gómez, J.M., de Buenaga, M.: In the Development of a Spanish Metamap. In: Proc. of the 17th ACM conference on Information and Knowledge Management. (2008) 1465–1466
10. Castro, E., Iglesias, A., Martínez, P., no, L.C.: Automatic Identification of Biomedical Concepts in Spanish-Language Unstructured Clinical Texts. In: Proc. of the 1st ACM International Health Informatics Symposium. IHI '10 (2010) 751–757
11. Yetano, J., Alberola, V.: Diccionario de Siglas Médicas y Otras Abreviaturas, Epónimos y Términos Médicos Relacionados con la Codificación de las Altas Hospitalarias. Ministerio de Sanidad y Consumo (2003)
12. Kim, J.D., Pysalo, S., Ohta, T., Bossy, R., Nguyen, N., Tsujii, J.: Overview of BioNLP Shared Task 2011. In: Proc. of BioNLP Shared Task 2011, ACL (2011)
13. Agirre, E., Soroa, A., Stevenson, M.: Graph-based word sense disambiguation of biomedical documents. Bioinformatics **26** (2010) 2889–2896