Journal of Biomedical Informatics 56 (2015) 318–332

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions

Maite Oronoz, Koldo Gojenola, Alicia Pérez, Arantza Díaz de Ilarraza, Arantza Casillas *

IXA Group, University of the Basque Country (UPV-EHU), Computer Engineering Faculty, P. Manuel Lardizabal, 1, 20018 Donostia-San Sebastián, Spain[1]

ARTICLE INFO

Article history:
Received 24 October 2014
Revised 21 May 2015
Accepted 23 June 2015
Available online 30 June 2015

Keywords:
Gold standard
Clinical text
Adverse drug reaction
Text mining

ABSTRACT

The advances achieved in Natural Language Processing make it possible to automatically mine information from electronically created documents. Many Natural Language Processing methods that extract information from texts make use of annotated corpora, but these are scarce in the clinical domain due to legal and ethical issues. In this paper we present the creation of the IxaMed-GS gold standard composed of real electronic health records written in Spanish and manually annotated by experts in pharmacology and pharmacovigilance. The experts mainly annotated entities related to diseases and drugs, but also relationships between entities indicating adverse drug reaction events. To help the experts in the annotation task, we adapted a general corpus linguistic analyzer to the medical domain. The quality of the annotation process in the IxaMed-GS corpus has been assessed by measuring the inter-annotator agreement, which was 90.53% for entities and 82.86% for events. In addition, the corpus has been used for the automatic extraction of adverse drug reaction events using machine learning.

1. Introduction

Advances in Natural Language Processing (NLP) and machine learning enable the automatic extraction of information from electronically created documents, including the detection of cause-effect events and of entities involved in such events. Training a machine learner usually relies on the existence of annotated corpora, and thus corpus creation is a significant part of the development of information mining techniques. This paper describes the creation of a corpus annotated with medical concepts and relationships between them.

As Cohen and Demner-Fushman [1] note, while research in genomic NLP has benefited from a growing number of corpora and document collections consisting of scientific publications, research in the clinical domain has been hampered by the legal and ethical issues associated with corpora of clinical documents like electronic health records (EHRs). Furthermore, most NLP work in biomedicine has concerned English (although there are recent efforts that incorporate other languages, such as Spanish [2], French [3,4], Swedish [5], and Finnish [6]).

The goal of this work is to address both the scarcity of clinical corpora and the heavy focus on English by developing a corpus consisting of collections of EHRs, annotated with medical entities and events in Spanish, a major world language. This corpus, which we call IxaMed-GS, consists of syntactically and semantically annotated discharge summaries written in spontaneous Spanish by doctors. The aim is to use this corpus to develop tools for automatic annotation and, hence, make it easier for doctors to recover information from EHRs. The data—electronic health records gathered in the Galdakao-Usansolo Hospital—was subject to an agreement between the Basque Health Service and the University of the Basque Country, in which the Health Service provided a corpus completely divested of identifying personal information and authorized its use exclusively for research purposes.

Generating reliably annotated corpora is typically dependent on human experts. Since annotation work is time-consuming and monotonous, it would be desirable to facilitate it with automated tools. Therefore, in this paper we present a process in which a seed of manual annotations served to develop an automatic annotation tool, FreeLing-Med [7]. This tool annotated the documents and, in a second step, experts revised the tags to check the validity of the annotations. We describe the annotation procedure and assess it with the aid of results obtained by different experts, also discussing expert disagreement. We show that by the end of the process a reliable annotated corpus can be obtained by coordinating automatic methods and manual work by experts.

The corpus focuses on Adverse Effects (AEs), which the ENEAS report [8]—a national study on Adverse Effects associated with

[1] http://ixa.si.ehu.eus.

* Corresponding author. Tel.: +34 946015321; fax: +34 946013071.
E-mail address: arantza.casillas@ehu.eus (A. Casillas).

http://dx.doi.org/10.1016/j.jbi.2015.06.016
1532-0464/© 2015 Elsevier Inc. All rights reserved.

hospitalization published by the Spanish Ministry of Health and Consumption—defines as follows: an accident or incident that either injured or may have injured the patient during treatment. Among the different AEs distinguished in the report, the large majority are related to one of the following three causes: (i) drug prescriptions (37.4% of all AEs); (ii) nosocomial infections (25.3% of all AEs); and (iii) procedures (25.0% of all AEs). Our work particularly concentrates on adverse drug reactions (ADRs), defined as unavoidable or difficult-to-avoid disorders, with or without injury, produced when drugs are used in an appropriate way [8]. The results and conclusions from the ENEAS report were presented in [9]: twenty-four Spanish hospitals were studied to determine the impact and preventability of AEs. Of the AE events associated with medication-use (37.4% of all AEs), 34.8% were classified as preventable.

A key issue in preventive medicine is the documentation of ADRs and related cause-effect events such as drug-disease, substance-allergy, drug-drug interactions, or disease-symptoms. While no estimate of the monetary cost associated with ADRs was available for Spain, Harpaz et al. [10] report an estimated cost associated with ADRs of $75 billion annually in the United States. Technological aids like machine learning, if they succeed in reducing the incidence of ADRs, can therefore potentially both improve health outcomes and eliminate unnecessary expenses.

The contribution of this paper is threefold:

- Generation of a corpus based on real EHRs in Spanish, enhanced with syntactic and semantic information: IxaMed-GS.
- Definition of annotation guidelines for manual annotation by experts, annotation assessment and consensus agreement.
- Development of tools for medical entity annotation (FreeLing-Med) as well as AE event annotation.

We also explore a use case of the IxaMed-GS corpus: detection of AE events, with ADRs being the main category of such events, and describe preliminary experiments that provide evidence of the usability of the IxaMed-GS corpus.

The remainder of this article is arranged as follows: Section 2 provides a summary of the state of the art in corpus development and ADR extraction; Section 3 presents the methodology followed to create the IxaMed-GS corpus from the manual annotation guidelines to the computer-assisted annotation framework based on FreeLing-Med, a linguistic analyzer for the biomedical domain. After annotating entities such as drugs, substances, active ingredients, procedures, diseases, and allergies, we explore the performance of the corpus in the development of an automatic ADR annotation system. Section 4 is devoted to the quantitative assessment of the aforementioned materials and methods, among them the consensus achieved by the annotators and the performance of the automatic event annotation system. Next, in Section 5, the results presented are discussed; finally, Section 6 presents the main conclusions and contributions of this work.

2. Related work

In the past few years, several annotated corpora have been created for a number of medical domains [1]. The nature of the annotated concepts depends on the purpose for which the corpus is developed: corpora are used in a variety of tasks, of which extracting relationships like drug to drug interactions and adverse drug reactions are only one. Below, we review some of the most important corpora annotated with drugs and diseases used in biomedical language processing research, with a special emphasis on the annotation of ADRs. Table 1 summarizes the information about the described corpora.

Table 1
Medical domain corpora and their characteristics.

Corpus	Size	Annotation	Number of annotators
CLEF	20,234 Clinical documents	Semantic in Inter/intra sentence	2
BioText	100 titles 40 abstracts From Medline	Treatments, Disorders, Relations in sentence	1
Arizona	794 PubMed abstracs	Diseases,	2 after
Disease	2775 sentences	Symptom in sentence	Automatic annotation
EU-ADR	300 Medline Abstracts	Drugs, disease Relations in sentence	3 after Automatic annotation
ADE	2972 Medline Case reports	Drugs, disease Relations in sentence	3
DailyStrength	10,617 comments Social media	Adverse effects For specific drugs in sentence	2
DDI	792 text from Drug Bank 233 Medline abstract	drugs, brands, substances, groups of drugs, D–D interactions in sentence	2
i2b2	1243 Discharge summaries	Information Related to medications in sentence	3

One of the most renowned corpora in this context is the CLEF corpus (CLinical E-Science Framework) [11], a semantically annotated corpus of 20,234 clinical documents (structured records and free text). The free texts are of three types: clinical narratives; histopathology reports; and imaging reports. Each document is independently annotated by two annotators.

In the BioText corpus [12], relationships between disorders and treatments are annotated. Treatments comprise both drugs and medical treatments. The corpus is composed of a set of 100 titles and 40 abstracts extracted from Medline, and the annotations are performed at the sentence level. All the documents were annotated manually by one expert.

The Arizona Disease Corpus [13,14] includes 794 PubMed abstracts with diseases and symptoms annotated. The documents do not contain tags for AEs but the locations of mentions of a disease or a symptom are marked and mapped to Unified Medical Language System's (UMLS) Concept Unique Identifiers (CUIs). After the automatic annotation of the concepts, two domain experts revised the results. The corpus has 3,206 diseases annotated, mapped to UMLS CUIs, and distributed in 2775 sentences.

The EDGAR system [15] was built to extract medical information such as drugs, genes and relations from the medical literature. It counts on the Medline database of biomedical citations and abstracts and the UMLS.

The EU-ADR corpus [16] is an annotated corpus of 300 Medline abstracts where drugs, diseases, targets, and their relationships are marked. The annotation process was performed by three annotators and divided into two steps: (i) a named-entity recognition system produced a first annotation and (ii) annotators revised this annotation using a web-based interface. The corpus contains the drug-disease relation annotated, with an indication of whether a particular drug may produce an adverse effect. The corpus and the annotation tool are available.

The ADE corpus [17] is comprised of a subset of 2972 Medline case reports that were manually annotated by three annotators and subsequently harmonized. It contains annotations of 5063 drugs, 5776 conditions (e.g. diseases, signs, symptoms, dosages)