Expert Systems With Applications 61 (2016) 235–245

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Learning to extract adverse drug reaction events from electronic health records in Spanish

Arantza Casillas[a], Alicia Pérez[b,*], Maite Oronoz[c], Koldo Gojenola[b], Sara Santiso[b]

[a] Dep. Electricity and Electronics. Faculty of Science and Technology, (UPV-EHU), Leioa, Spain
[b] Dep. Languages and Computer Systems. School of Engineering of Bilbao (UPV-EHU), Bilbao, Spain
[c] Dep. Languages and Computer Systems. Faculty of Computer Science (UPV-EHU), San Sebastian, Spain

ARTICLE INFO

Article history:
Received 30 December 2015
Revised 24 May 2016
Accepted 25 May 2016
Available online 26 May 2016

Keywords:
Text mining
Unbalanced classification problem
Medical event extraction

ABSTRACT

Objective: To tackle the extraction of adverse drug reaction events in electronic health records. The challenge stands in inferring a robust prediction model from highly unbalanced data. According to our manually annotated corpus, only 6% of the drug-disease entity pairs trigger a positive adverse drug reaction event and this low ratio makes machine learning tough.

Method: We present a hybrid system utilising a self-developed morpho-syntactic and semantic analyser for medical texts in Spanish. It performs named entity recognition of drugs and diseases and adverse drug reaction event extraction. The event extraction stage operates using rule-based and machine learning techniques.

Results: We assess both the base classifiers, namely a knowledge-based model and an inferred classifier, and also the resulting hybrid system. Moreover, for the machine learning approach, an analysis of each particular bio-cause triggering the adverse drug reaction is carried out.

Conclusions: One of the contributions of the machine learning based system is its ability to deal with both intra-sentence and inter-sentence events in a highly skewed classification environment. Moreover, the knowledge-based and the inferred model are complementary in terms of precision and recall. While the former provides high precision and low recall, the latter is the other way around. As a result, an appropriate hybrid approach seems to be able to benefit from both approaches and also improve them. This is the underlying motivation for selecting the hybrid approach. In addition, this is the first system dealing with real electronic health records in Spanish.

1. Introduction

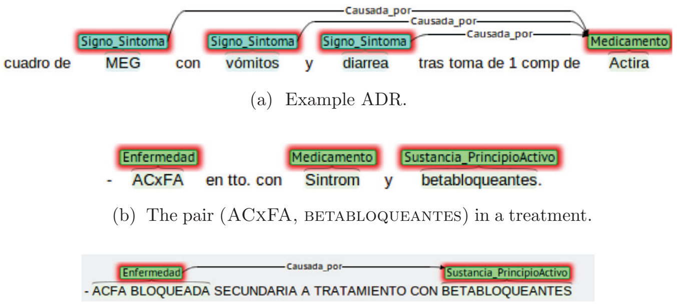
Text mining in the clinical domain has emerged as a field of interest during the last decade with several attempts in the literature that aim at easing the reading of reports in English (Chaplin et al., 2015; Rink, Harabagiu, & Roberts, 2011; Toldo, Bhattacharya, & Gurulingappa, 2012). There is a great variety in the type of biomedical events aimed at, such as binary protein-protein interaction (Wong, 2001), biomolecular event extraction (Jin-Dong et al., 2011), drug-drug interaction extraction (Segura-Bedmar, Martínez, & Herrero-Zazo, 2013), and cause-effect event extraction (Björne, Ginter, Pyysalo, Tsujii, & Salakoski, 2010; Mihaila, Ohta, Pyysalo, & Ananiadou, 2013).

We focus on the detection of Adverse Drug Reactions (ADRs): when a drug prescribed to combat a disease can be the cause of other new diseases. In Fig. 1a some adverse reactions originally written in Spanish are shown (*"MEG"*, *"vómitos"*, *"diarrea"*) caused by a drug (*"Actira"*). In this particular case of ADR detection, there are solutions based on rules, machine learning techniques or combinations, with encouraging results in different contexts:

- **Medical literature** (SriJyothsna, Aditya, Saipradeep, Govindakrishnan, & Rajgopal, 2014; Xu & Wang, 2013): e.g. scientific journals. These texts tend to be grammatically correct and without misspellings.
- **Social media** (Nikfarjam & Gonzalez, 2011; SriJyothsna et al., 2014): e.g. blogs and tweets related to health usually written by non-experts.

[*] Corresponding author.
E-mail addresses: arantza.casillas@ehu.eus (A. Casillas), alicia.perez@ehu.eus (A. Pérez), maite.oronoz@ehu.eus (M. Oronoz), koldo.gojenola@ehu.eus (K. Gojenola), sara.santiso@ehu.eus (S. Santiso).

http://dx.doi.org/10.1016/j.eswa.2016.05.034
0957-4174/© 2016 Elsevier Ltd. All rights reserved.



cuadro de MEG con vómitos y diarrea tras toma de 1 comp de Actira

(a) Example ADR.

- ACxFA en tto. con Sintrom y betabloqueantes.

(b) The pair (ACxFA, BETABLOQUEANTES) in a treatment.

- ACFA BLOQUEADA SECUNDARIA A TRATAMIENTO CON BETABLOQUEANTES

(c) The pair (ACxFA, BETABLOQUEANTES) in an ADR.

Fig. 1. Example of EHRs in the Brat framework. The tags `Enfermedad` and `Signo_Sintoma` stand for "disease" and "sign or symptom", belonging to the DISEASE family; the tag `Medicamento` and `Sustancia_PrincipioActivo` stand for "drug" and "substance or active ingredient" (they belong to the DRUG family), and the tag `Causada_por` stands for "caused by".

- **Electronic health records (EHRs)** (Karlsson, Zhao, Asker, & Boström, 2013; Sohn, Kocher, Chute, & Savova, 2011): e.g. clinical reports. They do not use either a fully formal linguistic register or a lay register. They can contain abbreviations, typos, or grammatically incorrect sentences.

The aim of this work is to automatically highlight ADRs in EHRs in order to alleviate the work-load of various services within a hospital (pharmacy, documentation, etc.) that have to read these reports. The processing of these texts presents a real challenge to the rapid detection of ADRs and, in consequence, to the safety of the patient.

We present a hybrid ADR extraction system to cope with this task. It entails two stages in sequence: 1) the first one carries out, among others, the annotation of entities such as drugs and substances (from now onwards both of them shall be referred to as DRUG), and also disorders and findings (referred to as DISEASE); 2) the second stage determines whether a given (DRUG, DISEASE) pair of entities represents an ADR event. Note that we are interested in highlighting events involving (DRUG, DISEASE) pairs where the DRUG caused the DISEASE. The final system should present the ADRs marked in a friendly front-end. To this end, we will represent the text in the framework provided by *Brat rapid annotation tool* (Stenetorp et al., 2012). Fig. 1 shows examples, represented in Brat, of some cause-effect events manually tagged by experts.

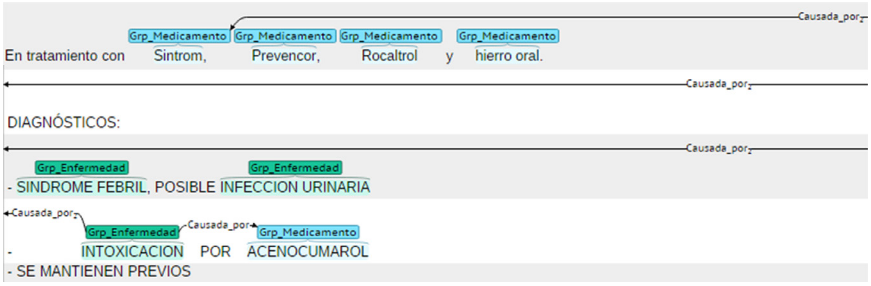
ADRs differ from medication errors where drugs are used in an inappropriate way and, in consequence, are preventable situations, while ADRs are hardly preventable (MSC, 2006). Thus, a drug is prescribed to combat a disease but, in some situations, it could cause unexpected side effects on specific patients. In this work we aim to differentiate between (DRUG, DISEASE) pairs causing an ADR and those pairs with positive or neutral consequences for patients. The task of detecting ADRs in EHRs is tough since the same pair might represent both ADR and non-ADR events. As an example, Fig. 1 presents the pair (AC × FA, BETABLOQUEANTES) twice in the same document. AC × FA is an abbreviation that indicates *"Arritmia cardíaca por fibrilación auricular"*, meaning "atrial fibrillation", and *"betabloqueantes"* are "beta blockers", a drug family. That pair, in one case (depicted in Fig. 1b), represents a treatment; however, in the same document the same pair is found but as an ADR (see Fig. 1c). 6% of the drug-disease entity pairs trigger an adverse drug reaction, and these results are in accord with similar estimates for other health systems. For example, Henriksson, Kvist, Dalianis, and Duneld (2015) state that ADRs are responsible for approximately 3−5% of hospital admissions world-wide, and they suffer heavily from under-reporting. The ENEAS report, written by the Spanish Ministry of Health (MSC, 2006) examined twenty-four hospitals in Spain to determine the impact and preventability of ADRs, concluding that 42% of the adverse affects are avoidable.

The personnel at the hospital uses prescription management systems that help to avoid the most typical and frequent drug-disease pairs causing ADRs. However, these lists are not specially useful in the context of analyzing EHRs because the ADRs that we intend to discover are those that do not belong to the list of typical ADRs, which are filtered before the prescription of the drug, according to each patient's characteristics. As Henriksson et al. (2015) pointed out, ADRs are heavily under-reported in EHRs most of the times, except in the few cases when they are the main cause of disease. Additionally, the main difficulty of the present task is to discover ADRs which are specified in multiple ways in EHR texts, potentially with big differences with respect to the standard naming of drugs, diseases and ADR triggers. We are dealing with discharge EHRs written by around 400 different doctors. These records are not written with the aid of a template, thus, they do not follow a pre-determined structure, and this, by itself, entails a challenge. Most of the recent works cope with event extraction within the same sentence, that is, intra-sentence events. By contrast, in this work we have realised that around 22% of the ADR events in our EHRs occur between medical entities that are in different sentences, and some of them are far from each other.

Fig. 2 presents an example of an inter-sentence event, the disease *"INTOXICACION"* (intoxication) is caused by the drug "Sintrom" and each entity is in a different sentence. The ADRs between entities that are in different sentences are going to be explained in depth in Section 4.1. The pair (*"INTOXICACION"*, *"ACENOCUMAROL"*) corresponds to an ADR where the drug and the disease are in the same sentence. The pair ("INTOXICACION","Sintrom") is an ADR with the drug and the disease in different sentences. Note that *Acenocumarol* (in English, Acenocoumarol) is the active ingredient of the pharmacological product called *Sintrom*.

2. Related work

Since 2008 various European projects have tackled the problem of the early detection of adverse drug reactions, some examples



En tratamiento con Sintrom, Prevencor, Rocaltrol y hierro oral.

DIAGNÓSTICOS:

- SINDROME FEBRIL, POSIBLE INFECCION URINARIA

- INTOXICACION POR ACENOCUMAROL

- SE MANTIENEN PREVIOS

Fig. 2. Example of intra-sentence and inter-sentence ADRs.