Adverse Drug Event prediction combining shallow analysis and machine learning

Sara Santiso
Alicia Pérez
Koldo Gojenola
IXA Taldea (UPV-EHU)

Arantza Casillas
Maite Oronoz
IXA Taldea (UPV-EHU)
http://ixa.si.ehu.es

Abstract

The aim of this work is to infer a model able to extract cause-effect relations between drugs and diseases. A two-level system is proposed. The first level carries out a shallow analysis of Electronic Health Records (EHRs) in order to identify medical concepts such as drug brandnames, substances, diseases, etc. Next, all the combination pairs formed by a concept from the group of drugs (drug and substances) and the group of diseases (diseases and symptoms) are characterised through a set of 57 features. A supervised classifier inferred on those features is in charge of deciding whether that pair represents a cause-effect type of event.

One of the challenges of this work is the fact that the system explores the entire document. The contributions of this paper stand on the use of real EHRs to discover adverse drug reaction events even in different sentences. Besides, the work focuses on Spanish language.

1 Introduction

This work deals with semantic data mining within the clinical domain. The aim is to automatically highlight the Adverse Drug Reactions (ADRs) in EHRs in order to alleviate the work-load to several services within a hospital (pharmacy service, documentation service,...) that have to read these reports. Event detection was thoroughly tackled in the Natural Language Processing for Clinical Data 2010 Challenge. Since then, cause-effect event extraction has emerged as a field of interest in the Biomedical domain (Björne et al., 2010; Mihaila et al., 2013). The motivation is, above all, practical. Electronic Health Records (EHRs) are studied by several services in the hospital, not only by the doctor in charge of the patient but also by the pharmacy and documentation services, amongst others. There are some attempts in the literature that aim to make the reading of the reports in English easier and less time-consuming by means of an automatic annotation toolkit (Rink et al., 2011; Botsis et al., 2011; Toldo et al., 2012). This work is a first approach on automatic learning of relations between drugs causing diseases in Spanish EHRs.

This work presents a system that entails two stages in cascade: 1) the first one carries out the annotation of drugs or substances (from now onwards both of them shall be referred to as DRUG) and diseases or symptoms (referred to as DISEASE); 2) the second one determines whether a given (DRUG, DISEASE) pair of concepts represents a cause-effect reaction. Note that we are interested in highlighting events involving (DRUG, DISEASE) pairs where the drug caused an adverse reaction or a disease. By contrast, often, (DRUG, DISEASE) pairs would entail a drug prescribed to combat a disease, but these correspond to a different kind of events (indeed, diametrically opposed). Besides, (DRUG, DISEASE) pairs might represent other sort of events or they might even be unrelated at all. Finally, the system should present the ADRs marked in a friendly front-end. To this end, the aim is to represent the text in the framework provided by Brat (Stenetorp et al., 2012). Figure 1 shows an example, represented in Brat, of some cause-effect events manually tagged by experts.

There are related works in this field aiming at a variety of biomedical event extraction, such as binary protein-protein interaction (Wong, 2001), biomolecular event extraction (Kim et al., 2011), and drug-drug interaction extraction (Segura-Bedmar et al., 2013). We are focusing on a variety of interaction extraction: drugs causing diseases. There are previous works in the literature that try to warn whether a document contains or not this type of events. There are more recent works that

Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014, pages 85–89,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

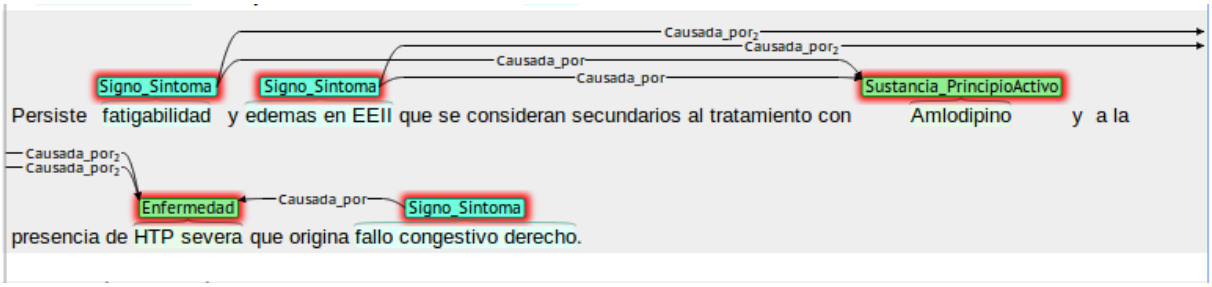

Figure 1: Some cause-effect events manually annotated in the Brat framework.

cope with event extraction within the same sentence, that is, intra-sentence events. By contrast, in this work we have realised that around 26% of the events occur between concepts that are in different sentences. Moreover, some of them are at very long distance. Hence, our method aims at providing all the (DRUG, DISEASE) concepts within the document that represent a cause-effect relation.

We cope with real discharge EHRs written by around 400 different doctors. These records are not written in a template, that is, the EHRs do not follow a pre-determined structure, and this, by itself entails a challenge. The EHRs we are dealing with are written in a free structure using natural language, non-standard abbreviations etc. Moreover, we tackle Spanish language, for which little work has been carried out. In addition, we do not only aim at single concept-words but also at concepts based on multi-word terms.

2 System overview

The system, as depicted in Figure 2 entails two stages.

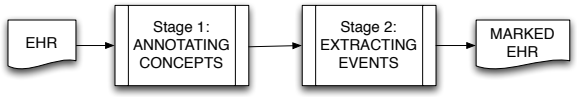

Figure 2: The ADR event extraction system.

In the first stage, relevant pairs of concepts have to be identified within an EHR. Concept annotation is accomplished by means of a shallow analyser system (described in section 2.1). Once the analyser has detected (DRUG, DISEASE) pairs in a document, all the pairs will be examined by an inferred supervised classifier (described in section 2.2).

2.1 Annotating concepts by shallow analysis

The first stage of the system has to detect and annotate two types of semantic concepts: drugs and diseases. Each concept, as requested by the pharmacy service, should gather several sub-concepts stated as follows:

1. DRUG concept:
   (a) Generic names for pharmaceutical drugs: e.g. corticoids;
   (b) Brand-names for pharmaceutical drugs: e.g. Aspirin;
   (c) Active ingredients: e.g. vancomycin;
   (d) Substances: e.g. dust, rubber;
2. DISEASE concept:
   (a) Diseases
   (b) Signs
   (c) Symptoms

These concepts were identified by means of a general purpose analyser available for Spanish, called FreeLing (Padró et al., 2010), that had been enhanced with medical ontologies and dictionaries, such as SNOMED-CT, BotPLUS, ICD-9-CM, etc. (Oronoz et al., 2013). This toolkit is able to identify multi-word context-terms, lemmas and also POS tags. An example of the morphological, semantic and syntactic analysis, provided by this parser is given in Figure 3. In the figure two pieces of information can be distinguished: for example, given the word "secundarios" (meaning secondaries) 1) the POS tag provided is AQOM corresponding to Qualificative Adjective Ordinal Masculine Singular; and 2) the provided lemma is "secundario" (secondary). Besides, in a third layer, the semantic tag is given, that is, the tag "ENFERMEDAD" (meaning disease) involves the multi-word concept "HTP severa" (severe pulmonary hypertension).

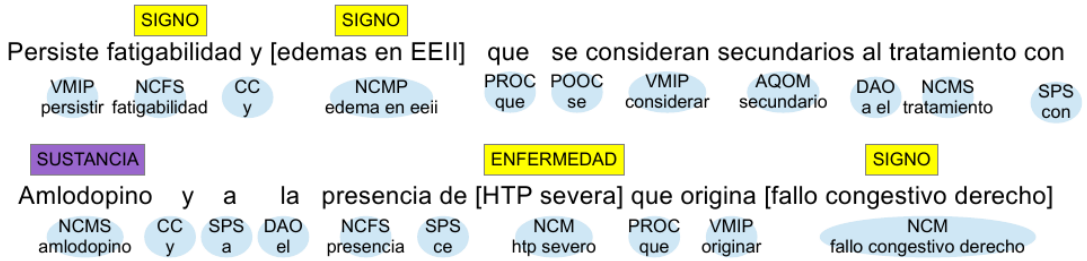

Figure 3: Lemmas, POS-tags and semantic tags are identified by the clinic domain analyser (diseases in yellow and drugs or substances in violet).

2.2 Extracting adverse drug reaction events using inferred classifiers

The goal of the second stage is to determine if a given (DRUG, DISEASE) pair represents an ADR event or not. On account of this, we resorted to supervised classification models. These models can be automatically inferred from a set of documents in which the target concepts had been previously annotated. Hence, first of all, a set of annotated data representative for the task is required. To this end, our starting point is a manually annotated corpus (presented in section 2.2.1). Besides, in order to automatically learn the classifier, the (DRUG, DISEASE) pairs have to be described in an operative way, that is, in terms of a finite-set of features (see section 2.2.2). The supervised classification model selected was a type of ensemble classifier: Random Forests (for further details turn to section 2.2.3).

2.2.1 Producing an annotated set

A supervised classifier was inferred from annotated real EHRs. The annotation was carried out by doctors from the same hospital that produced the EHRs. Given the text with the concepts marked on the first stage (turn to section 2.1) and represented within the framework provided by Brat[1], around 4 doctors from the same hospital annotated the events. This annotated set would work as a source of data to get instances that would serve to train supervised classification models, as the one referred in section 2.2.

2.2.2 Operational description of events

As it is well-known, the success of the techniques based on Machine Learning relies upon the features used to describe the instances. Hence, we selected the following features that eventually have proven useful to capture the semantic relations between ADRs. The features can be organised in the following sets:

- Concept-words and context-words: to be precise, we make use of entire terms including both single-words and multi-words.
  - DRUG concept-word together with left and right context words (a context up to 3, yielding, thus, 7 features).
  - DISEASE concept-word together with left and right context words (7 features).
- Concept-lemmas and context-lemmas for both drug and disease (14 features overall)
- Concept-POS and context-POS for both drug and disease (14 features)
- Negation and speculation: these are binary valued features to determine whether the concept words or their context was either negated or speculated (2 features).
- Presence/absence of other drugs in the context of the target drug and disease (12 features)
- Distance: the number of characters from the DRUG concept to the DISEASE concept (1 feature).

2.2.3 Inferring a supervised classifier

Given the operational description of a set of (DRUG, DISEASE) pairs, this stage has to deter-

[1]Brat is the framework a priori selected as the output front-end shown in Figure 1

mine if there exists an ADR event (that is, a cause-effect relation) or not. To do so, we resorted to Random Forests (RFs), a variety of ensemble models. RFs combine a number of decision trees being each tree built on the basis of the C4.5 algorithm (Quinlan, 1993) but with a distinctive characteristic: some randomness is introduced in the order in which the nodes are generated. Particularly, each time a node is generated in the tree, instead of chosing the attribute that maximizes the Information Gain, the attribute is randomly selected amongst the k best options. We made use of the implementation of this algorithm available in Weka-6.9 (Hall et al., 2009). Ensemble models were proved useful on drug-drug interaction extraction tasks (Thomas et al., 2011).

3 Experimental results

We count on data consisting of discharge summaries from Galdakao-Usansolo Hospital. The records are semi-structured in the sense that there are two main fields: the first one for personal data of the patient (age, dates relating to admittance) that were not provided by the hospital for privacy issues; and the second one, our target, a single field that contains the antecedents, treatment, clinical analysis, etc. This second field is an unstructured section (some hospitals rely upon templates that divide this field into several subfields, providing it with further structure). The discharge notes describe a chronological development of the patient's condition, the undergone treatments, and also the clinical tests that were carried out.

Given the entire set of manually annotated documents, 34% were randomly selected without replacement to produce the evaluation set. The resulting partition is presented in Table 1 (where the train and evaluation sets are referred to as Train and Eval respectivelly).

	Documents	**Concepts**	**Relations**
Train	144	6,105	4,675
Eval	50	2,206	1,598

Table 1: Quantitative description of the data.

All together, there are 194 EHRs manually tagged with more than 8,000 concepts (entailing diseases, symptoms, drugs, substances and procedures). From these EHRs all the (DRUG,DISEASE) pairs are taken into account as event candidates, and these are referred to as relations in Table 1.

The system was assessed using per-class averaged precision, recall and f1-measure as presented in Table 2.

Precision	**Recall**	**F1-measure**
0.932	0.849	0.883

Table 2: Experimental results.

Semantic knowledge and contextual features have proven very relevant to detect cause-effect relations. Particularly, those used to detect the concepts and also negation or speculation of the context in which the concept appear.

A manual inspection was carried out on both the false positives and false negative predictions and the following conclusions were drawn:

- The majority of false positives were caused by i) pairs of concepts at a very long distance; ii) pairs where one of the elements is related to past-events undergone while the other element is in the current treatment prescribed (e.g. the disease is in the antecedents and the drug in the current diagnostics).
- The vast majority of false negatives were due to concepts in the same sentence where the context-words are irrelevant (e.g. filler words, determiners, etc.).

4 Concluding Remarks and Future Work

This work presents a system that first identifies relevant pairs of concepts in EHRs by means of a shallow analysis and next examines all the pairs by an inferred supervised classifier to determine if a given pair represents a cause-effect event. A relevant contribution of this work is that we extract events occurring between concepts that are in different sentences. In addition, this is one of the first works on medical event extraction for Spanish.

Our aim for future work is to determine whether the (DRUG, DISEASE) pair represents either a relation where 1) the drug is to overcome the disease; 2) the drug causes the disease; 3) there is no relationship between the drug and the disease.

The aim of context features is to capture characteristics of the text surrounding the relevant concepts that trigger a relation. More features could also be explored such as trigger words, regular patterns, n-grams, etc.

Acknowledgments

The authors would like to thank the Pharmacy and Pharmacovigilance services of Galdakao-Usansolo Hospital.

This work was partially supported by the European Commission (325099 and SEP-210087649), the Spanish Ministry of Science and Innovation (TIN2012-38584-C06-02) and the Industry of the Basque Government (IT344-10).

References

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. Complex event extraction at pubmed scale. Bioinformatics [ISMB], 26(12):382–390.

Taxiarchis Botsis, Michael D. Nguyen, Emily Jane Woo, Marianthi Markatou, and Robert Ball. 2011. Text mining for the vaccine adverse event reporting system: medical text classification using informative feature selection. JAMIA, 18(5):631–638.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. SIGKDD Explorations, 11(1):10–18.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of bionlp shared task 2011. In Proceedings of the BioNLP Shared Task 2011 Workshop, pages 1–6. Association for Computational Linguistics.

Claudiu Mihaila, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. Biocause: Annotating and analysing causality in the biomedical domain. BMC Bioinformatics, 14:2.

Maite Oronoz, Arantza Casillas, Koldo Gojenola, and Alicia Perez. 2013. Automatic annotation of medical records in Spanish with disease, drug and substance names. In Lecture Notes in Computer Science, volume 8259, pages 536–547. Springer-Verlag.

Lluis Padró, S. Reese, Eneko Agirre, and Aitor Soroa. 2010. Semantic Services in Freeling 2.1: WordNet and UKB. In Global Wordnet Conference, Mumbai, India.

Ross Quinlan. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

Bryan Rink, Sanda Harabagiu, and Kirk Roberts. 2011. Automatic extraction of relations between medical concepts in clinical texts. JAMIA, 18:594–600.

Isabel Segura-Bedmar, P Martínez, and Maria Herrero-Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Proceedings of Semeval, pages 341–350.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: A web-based tool for nlp-assisted text annotation. In In Proceedings of the Demonstrations Session at EACL 2012.

Philippe Thomas, Mariana Neves, Illés Solt, Domonkos Tikk, and Ulf Leser. 2011. Relation extraction for drug-drug interactions using ensemble learning. 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011), pages 11–18.

Luca Toldo, Sanmitra Bhattacharya, and Harsha Gurulingappa. 2012. Automated identification of adverse events from case reports using machine learning. In Workshop on Computational Methods in Pharmacovigilance.

Limsoon Wong. 2001. A protein interaction extraction system. In Pacific Symposium on Biocomputing, volume 6, pages 520–531. Citeseer.