Journal of Biomedical Informatics 45 (2012) 231–239

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

Summary of Product Characteristics content extraction for a safe drugs usage

S. Rubrichi *, S. Quaglini

Laboratory for Biomedical Informatics, "Mario Stefanelli", Department of Computers and Systems Science, University of Pavia, Pavia, Italy

ARTICLE INFO

Article history:
Received 27 March 2011
Accepted 30 October 2011
Available online 10 November 2011

Keywords:
Information extraction
Conditional random fields
Support vector machines
Adverse drug events
Medication errors
Summary of Product Characteristics

ABSTRACT

The use of medications has a central role in health care provision, yet on occasion, it may injure the person taking them as result of adverse drug events. A correct drug choice must be modulated to acknowledge both patients' status and drug-specific information. However, this information is locked in free-text and, as such, cannot be actively accessed and elaborated by computerized applications. The goal of this work lies in extracting content (active ingredient, interaction effects, etc.) from the Summary of Product Characteristics, focusing mainly on drug-related interactions, following a machine learning based approach. We compare two state of the art classifiers: conditional random fields with support vector machines. To this end, we introduce a corpus of 100 interaction sections, hand annotated with 13 labels that have been derived from a previously developed conceptual model. The results of our empirical analysis demonstrate that the two models perform well. They exhibit similar overall performance, with an overall accuracy of about 91%.

1. Introduction

Adverse drug events (ADEs) have been defined as injuries resulting from medical intervention related to a drug [1] that endanger patient's safety and account for increased heath care costs. Examples include injuries (e.g., rash, confusion, or loss of function) caused by incorrect dosage as well as allergic reactions occurring in a patient not known to be allergic to a given medication and so forth. Many of these injuries suffered by patients are inevitable but at least a quarter may be secondary to medication errors [2], as errors in medication-management process are generally called [3]. These damages are not unavoidable and can be prevented. According to one estimate, medication errors occur most frequently at the prescribing and administration stages [2]. The rate of medication errors and preventable ADEs represents a serious cause for concern. The Institute of Medicine (IOM) committee, in its report Preventing Medication Error, estimates that more than 1.5 million ADEs are preventable each year in the US alone [2]. This report outlines the main priorities for research on safe medication use that will address this problem. It proposes electronic prescribing, by computerized provider order entry (CPOE) systems, as one of the most highly effective error prevention strategies.

CPOE systems are computer applications that allow direct, electronic entry of orders for medications, laboratory, radiology, referral, and procedures [4]. Several systematic reviews have shown the benefits of CPOE systems [5–10] resulting from their ability to detect unsafe and potentially fatal medication orders. Computerization in fact enables the delivery of clinical decision support [11], alerts to guide ordering, allowing for checks for allergies, drug–drug interactions, clinical conditions. One of the factors, which mainly influences the overall performance of such a system, is the quality and validity of the knowledge base underlying the system. Safe medication use requires that providers and consumers synthesize several types of information, including knowledge of the medication itself, as well as understanding of how it may interact with coexisting illnesses and medications, and how its use might be monitored. This information is constantly changing, and while most of the necessary updated knowledge is available somewhere, it is not always readily accessible, creating a situation in which it is almost impossible for health care providers to have current knowledge of every medication they prescribe.

In this work, we consider the problem of automatic extraction of drug information conveyed in the Summary of Product Characteristics (SPC), focusing on a specific section concerned with drug-related interactions. Our contributions are as follows:

1. We formulate the problem in a machine learning framework, in which we seek to assign the correct semantic label, such as *InteractionEffect* or *ActiveDrugIngredient*, to each word, or segment of sentence, of the text. We employ two state of the art classifiers: linear-chain conditional random fields (CRFs) and structural support vector machines (SVMs). These classifiers discriminate between semantically interesting and uninteresting content through the automatic adaptation of hundreds of engineered text characteristics, taking into account the properties of a document, on both a local (word) and global (sentence) level.

* Corresponding author. Fax: +39 0382 525638.
E-mail address: stefania.rubrichi@unipv.it (S. Rubrichi).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.10.012

2. We introduce a corpus of 100 interactions sections in Italian language that have been annotated with 13 distinct semantic labels, with respect to a previously implemented ontology.
3. We apply the CRFs and the SVMs to our data set and evaluate their overall and individual label results. Both the classifiers achieve a micro-averaged F_1-measure (see Section 4.2) greater than 90%, which is promising for real-world applications.

In a next step, the extracted information will be used to populate the ontology, in order to carry out automated reasoning on data. This reasoning process could help, for example, to determine whether it is possible for a particular drug to have any interaction with a particular active ingredient or diagnostic test, to find the effect of a particular interaction and so on. Moreover, ontology population can compensate for the possible lack of completeness and/or congruence among different SPCs. Like this, such knowledge model can be made available to specific prescription applications, such as CPOE, for integrating the underlying base of knowledge, thus improving the prescription process.

2. Background

SPCs represent a vast source of information for health professionals on how to use medicines safely and effectively. It forms an intrinsic and integral part of marketing authorization. In order to obtain an authorization to place a medicinal product on the market, a SPC shall be included in the application made to the competent authority. A SPC lays out the agreed position (results of physico-chemical, biological or microbiological tests, toxicological and pharmacological tests, clinical trials, etc.) on the medicinal product as collected during the course of the assessment process. Its content is regulated by Article 11 of Directive 2001/83/EC. Accordingly, SPCs of specialty medicines for human use are organized into 12 sections: name, therapeutic categories, active ingredient, excipients, indications, contraindication/side effects, undesired effects, posology, storage precautions, warnings, interactions, and use in case of pregnancy and nursing.

All this information is locked in free-text, the most convenient and natural way to convey medical knowledge for human communication [12]. This narrative form, however, cannot be actively used by health information systems. Reliable access to this comprehensive information, by Natural Language Processing (NLP) systems, can provide a wide range of coded data [12,13]. Such data will then be available for new or enriched clinical applications, thus facilitating and improving the prescription process. It is therefore an important aspect for improving quality of care and preventing medication errors.

Among NLP techniques, information extraction (IE) methods have been largely employed in the biomedical domain to extract facts from free-text [14–18] and to make them available for subsequent tasks such as case finding, summarization, decision support, or statistical analysis.

In this work, we therefore propose to extract drug-related interaction information reported in SPC, following a named entity recognition (NER) approach. NER is an important step in an integral IE task and aims at identifying words or phrases in natural language text belonging to certain classes of interest (i.e., Named Entities), such as diseases or drugs, and labeling them with their appropriate type [19]. In NER, an attempt will be made to associate each token with a label that indicates its appropriate domain-specific category. Approaches to NER span a broad range, from rule-based systems to machine learning. Rule-based systems make decisions on sets of hand-written disambiguation rules that play an important role in discovering a named entity, which specify, for example, that an ambiguous word belongs to a particular named entity rather than to another one if it follows another specific named entity.

On the other hand, a typical application of machine learning works to classify a novel instance x as belonging to a particular class y. In the field of label sequence problems such techniques aim at identifying the most likely sequence of labels for the words in any given sentences. These methods generally resolve tagging ambiguities by using training corpus to compute the probability of a given word having a given tag in a given context. Then, they automatically tune their own parameters to maximize their performance on the training corpus. The machine then generalizes from these samples.

Rule-based NER systems can be very effective but require some manual effort. Machine learning approaches can successfully extract named entities but require large annotated training data. Advantages of machine learning approaches are that they do not require human intuition, they are general and clearly separate the algorithm from the data, so that it is easier to apply them to any domain by simply retraining without reprogramming.

2.1. Related work

Several studies have addressed the issue of IE in medication domain. Some approaches concentrated their analysis on the extraction of drug names. Levin et al. [20] implemented a system based on lexicon (RxNorm) and regular expressions (Hints List) to extract and normalize drug names from an anesthesia electronic health record, into a standardized terminology. RxNorm and Hints List concepts were used in the mapping module as references for drug names, and medical abbreviations and jargon, respectively. In another study, Sirohi and Peissig [21] performed a dictionary-based NLP study to determine the effects of using varying lexicon to extract drug names from electronic medical records. These authors have shown how the accuracy of results can be enhanced by refining the drug lexicon.

Other studies focus on extracting more specific drug features, such as drug names and dosage. In one example, Evans et al. [22] reported a method of extracting drug and dosage data from a collection of discharge summaries. They first draw a conceptual model of drug-dosage information and then identified this information using a semantically driven extraction module. This module combines readily available NLP facilities from the Clarit system with newly created resources, including a set of pattern rules and a lexicon. A study by Shah and Martinez [23] derived numerical information about daily dosage from unstructured dosage instructions from a patient records database, using a dictionary to standardize words and phrases. Then, they converted the extracted information into structured fields.

Lately, more studies have been geared toward the extraction of a more complete set of drug characteristics. In particular, Gold et al. [24] built Merki, a parser that can extract drug names and other relevant information from discharge summaries using a lexicon and a set of parsing rules. Evaluation showed that the system identified drug names, but other information such as dose and frequency had lower precisions. Similarly, Xu et al. [25] implemented a NLP system, MedEx, which extracts medication information from clinical notes. Relying on a more detailed medication representation model, they integrated a semantic tagger and Chart parser to capture drug names and signature information from clinical narratives and then to map it onto structured representation.

The Third i2b2 NLP Challenge [26] focused on the extraction of medication information, in particular medication names, dosages, routes of administration, frequencies, durations, and reasons for administration, from discharge summaries. Different approaches have been proposed to address this task: rule-based, machine learning, and hybrid systems. Among others, Doan et al. [27] integrated