Journal of Biomedical Informatics 57 (2015) 333–349

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Identifying adverse drug event information in clinical notes with distributional semantic representations of context

Aron Henriksson [a,*], Maria Kvist [a,b], Hercules Dalianis [a], Martin Duneld [a]

[a] Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden
[b] Department of Learning, Informatics, Management and Ethics (LIME), Karolinska Institutet, Sweden

ARTICLE INFO

Article history:
Received 5 May 2015
Revised 19 July 2015
Accepted 10 August 2015
Available online 17 August 2015

Keywords:
Adverse drug events
Electronic health records
Corpus annotation
Machine learning
Distributional semantics
Relation extraction

ABSTRACT

For the purpose of post-marketing drug safety surveillance, which has traditionally relied on the voluntary reporting of individual cases of adverse drug events (ADEs), other sources of information are now being explored, including electronic health records (EHRs), which give us access to enormous amounts of longitudinal observations of the treatment of patients and their drug use. Adverse drug events, which can be encoded in EHRs with certain diagnosis codes, are, however, heavily underreported. It is therefore important to develop capabilities to process, by means of computational methods, the more unstructured EHR data in the form of clinical notes, where clinicians may describe and reason around suspected ADEs. In this study, we report on the creation of an annotated corpus of Swedish health records for the purpose of learning to identify information pertaining to ADEs present in clinical notes. To this end, three key tasks are tackled: recognizing relevant named entities (disorders, symptoms, drugs), labeling attributes of the recognized entities (negation, speculation, temporality), and relationships between them (indication, adverse drug event). For each of the three tasks, leveraging models of distributional semantics – i.e., unsupervised methods that exploit co-occurrence information to model, typically in vector space, the meaning of words – and, in particular, combinations of such models, is shown to improve the predictive performance. The ability to make use of such unsupervised methods is critical when faced with large amounts of sparse and high-dimensional data, especially in domains where annotated resources are scarce.

1. Introduction

The digitization of healthcare data, as a result of the increasingly widespread adoption of electronic health records (EHRs), has rendered its analysis possible on a large and unprecedented scale. However, despite the widely acknowledged transformative potential of exploiting EHR data for secondary use in the endeavor of improving healthcare and supporting public health activities, it remains a largely underutilized resource [1], partly as a result both of technical challenges and possible application areas being underexplored. To ensure, then, that this valuable resource is more widely tapped and its potential fully realized, computational methods need to be developed for this particular domain.

A nascent line of research concerns the application of machine learning algorithms to EHR data for the construction of predictive models that can be employed in a wide range of tasks. There are, however, many challenges involved in learning high-performing predictive models from EHR data, such as the high dimensionality caused by the large number of variables that can be used to describe a given set of observations, as well as the typically accompanying sparsity. There is also an inherent heterogeneity in EHR data, entailing that the various data types cannot be handled in an identical fashion. The majority of EHR data is, for instance, expressed in natural language, albeit in a form that is greatly specialized and domain-dependent: clinical text typically does not conform to standard grammar rules and is often littered with shorthand and misspellings [2,3], further exacerbating the aforementioned dimensionality and sparsity issues. There is perhaps, then, a particular need to adapt natural language processing (NLP) techniques to the genre of clinical text, lest the potentially valuable information contained therein should be ignored. It is moreover critical that research is conducted on languages other than English.

One public health activity that may be supported through secondary use of EHR data is pharmacovigilance, i.e. post-marketing drug safety surveillance, as alternatives to spontaneous reporting

* Corresponding author.
E-mail addresses: aronhen@dsv.su.se (A. Henriksson), maria.kvist@karolinska.se (M. Kvist), hercules@dsv.su.se (H. Dalianis), xmartin@dsv.su.se (M. Duneld).

http://dx.doi.org/10.1016/j.jbi.2015.08.013
1532-0464/

systems for information on adverse drug events (ADEs) are currently being explored, not least in order to address the gross underreporting of ADEs that create obstacles in obtaining reliable incidence estimates. In comparison to spontaneous case reports, EHRs have several advantages, such as providing longitudinal observations of patient treatment, including drug prescriptions and administration. Unfortunately, ADEs are also heavily underreported in EHRs, where they can be encoded by a, albeit rather limited, set of diagnosis codes. It is therefore important to develop capabilities to process clinical notes, where clinicians may describe and reason around suspected ADEs.

Extracting information pertaining to ADEs in clinical notes requires a number of key components: (1) named entity recognition, i.e. being able to detect mentions of, for instance, drugs, symptoms and disorders; (2) concept attribute labeling, e.g. being able to determine if named entity mentions are expressed with negation, speculation or a non-current temporality (past/future events); and (3) relation extraction, i.e. being able to detect and classify relations that may hold between pairs of named entity mentions. Machine learning can be leveraged to construct predictive models to perform such tasks automatically. Doing so, however, requires access to substantial amounts of labeled data, which is typically not readily available and, to create for every problem, domain and language, is prohibitively expensive, particularly when medical experts are required to provide the annotations.

In this paper, we describe the creation of such an annotated resource – comprising clinical notes written by physicians in Swedish – that is then used to construct predictive models, which, in turn, are used to identify information pertaining to ADEs in clinical notes. To address the aforementioned challenges – high dimensionality and sparsity, on the one hand, and limited availability of annotated resources in the clinical domain, on the other – we investigate how models of distributional semantics can be leveraged to obtain enhanced predictive performance on the three identified tasks. Distributional semantics essentially allow word representations, typically in vector space, to be obtained in a wholly unsupervised manner. These can be used to generate (semantic) features that can subsequently be exploited by a learning algorithm when constructing predictive models. In a series of experiments, such representations of the data are shown to be more conducive to learning high-performing predictive models in comparison to the commonly employed bag-of-words (BOW) approach. The ability to exploit large amounts of unlabeled data is critical when faced with volumes of EHR data that are approaching "big data".

2. Background

Pharmacovigilance is carried out throughout the life-cycle of a drug in order to inform decisions on its initial and sustained use in the treatment of patients. The need to monitor the safety of drugs post marketing is caused by the inherent limitations of clinical trails in terms of sample size and study duration, making it particularly difficult to identify rare and long-latency ADEs. There are, in fact, several cases in which drugs have been discovered to cause severe, even fatal, ADEs, resulting in their withdrawal from the market [4,5]. Moreover, ADEs have been estimated to be responsible for approximately 3–5% of hospital admissions worldwide [6,7], causing suffering and inflated healthcare costs, often unnecessarily so, as ADEs are in many cases preventable: according to one meta-analysis, around 50% of adverse drug reactions are preventable [8]. Post-marketing surveillance of drug safety has primarily relied on case reports that are reported voluntarily by clinicians and drug users in so-called spontaneous reporting systems, such as the US Food and Drug Administration's Adverse Event Reporting System, the Yellow Card Scheme in the UK and the World Health Organization's Global Individual Case Safety Reporting Database: Vigibase. Relying solely on spontaneous reports has, however, proven to be insufficient. In addition to several limitations inherent in collecting information in this way, such as selective reporting, incomplete patient information and indeterminate population information [9], spontaneous reporting systems suffer heavily from underreporting: according to one estimate, more than 94% of ADEs are not reported in such systems [10].

As a result, alternative – and complementary – sources of information for pharmacovigilance are being explored, including the biomedical literature [11], user-generated data in social media [12] and, as previously mentioned, EHRs. The latter has the distinct advantage of containing data collected from the clinical setting, thereby providing access to longitudinal observations of patients, their medical condition and drug use. Health records contain various types of data, which can crudely be categorized into structured and unstructured: the structured data includes, e.g., diagnosis and drug codes, clinical measurements and lab tests, while clinical notes written in free-text make up the more unstructured parts. Although ADEs signals can, to some extent, be detected from the structured EHR data – and this constitutes an ongoing line of research [13–17] – a substantial amount of information pertaining to ADEs is expressed only in clinical notes, where clinicians may describe and reason around potential ADEs. Methods that can identify information pertaining to ADEs in clinical notes would therefore be very valuable.

2.1. Detecting adverse drug events in clinical notes

In recent years, there have been a few studies investigating the possibility of detecting and extracting various types of ADE information from clinical notes. Some of the methods that have been developed are based on hand-crafted rules, which tend to rely heavily on the existence of extensive dictionaries in the target language and domain. One such rule- and dictionary-based approach was developed for Danish clinical notes and evaluated on around six thousand health records of psychiatric patients [18]. The system identified a large number of potential ADEs of various kinds with, according to an evaluation through manual inspection, high precision (0.89) and moderate recall (0.75). This approach has later been employed in conjunction with temporal data mining techniques to allow for the identification of dose-specific ADEs [19]. Rule-based approaches often tend to perform fairly well; however, they are also known not to generalize well to other domains and over time, while being cumbersome and expensive to create.

Another type of approach to the exploitation of clinical notes for pharmacovigilance is primarily based on statistical methods. An early attempt in this vein used an NLP system, MedLEE, to extract relevant clinical events from discharge summaries, for which co-occurrence statistics were calculated in order to detect drug-ADE associations [20]. A small set of drugs with known ADEs were selected to evaluate the system, yielding a precision of 0.31 and a recall of 0.75. It was also shown that this method could be used to detect novel ADEs. A similar approach is to extract events or concepts from a large number of clinical notes – as many as fifty million in one study – and then to apply disproportionality methods to detect drug-ADE signals, as well as ADEs caused by drug-drug interactions [21–23]. The authors demonstrate the ability of their methods to flag for ADEs, in some cases before an official alert is made.

A third approach is to employ (supervised) machine learning to build predictive models that can identify potential relations, including ones that indicate an ADE, between drugs and medical problems. In one study, 435 Japanese discharge summaries were