Journal of Biomedical Informatics 45 (2012) 885–892

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports

Harsha Gurulingappa [a,b,*], Abdul Mateen Rajput [c], Angus Roberts [d], Juliane Fluck [a], Martin Hofmann-Apitius [a,b], Luca Toldo [c]

[a] Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany
[b] Bonn-Aachen International Center for Information Technology (B-IT), Dahlmannstraβe 2, 53115 Bonn, Germany
[c] Department of Knowledge Management, Merck KGaA, Frankfurterstraβe 250, 64293 Darmstadt, Germany
[d] Department of Computer Science, University of Sheffield, Sheffield S1 4DP, United Kingdom

ARTICLE INFO

Article history:
Received 15 March 2011
Accepted 11 April 2012
Available online 25 April 2012

Keywords:
Adverse drug effect
Benchmark corpus
Annotation
Harmonization
Sentence classification

ABSTRACT

A significant amount of information about drug-related safety issues such as adverse effects are published in medical case reports that can only be explored by human readers due to their unstructured nature. The work presented here aims at generating a systematically annotated corpus that can support the development and validation of methods for the automatic extraction of drug-related adverse effects from medical case reports. The documents are systematically double annotated in various rounds to ensure consistent annotations. The annotated documents are finally harmonized to generate representative consensus annotations. In order to demonstrate an example use case scenario, the corpus was employed to train and validate models for the classification of informative against the non-informative sentences. A Maximum Entropy classifier trained with simple features and evaluated by 10-fold cross-validation resulted in the F_1 score of 0.70 indicating a potential useful application of the corpus.

1. Introduction

Adverse drug effect is a response of a drug which is noxious and unintended, and which occurs at doses normally used in humans for the prophylaxis, diagnosis, therapy of disease, or for the modification of physiological function [1]. Most information about the drug's efficacy and adverse effects are obtained during clinical trials and post-marketing surveillance [2]. Organizations like the World Health Organization (WHO), the Food and Drug Administration (FDA), the European Medicines Agency (EMEA), and the Medicines and Healthcare products Regulatory Agency (MHRA) maintain a reporting system that enables individuals to spontaneously report the experienced adverse effects related to the use of medicines or healthcare products. A large portion of information that includes public as well as proprietary resources are carefully monitored by the drug manufacturers and the drug regulatory agencies where the medical complications are brought into public notice through data sources such as RXList,[1] Drug Information Portal,[2] or PharmaPendium.[3] Adverse effects present major ethical and legal issues for the pharmaceutical and health care industries. Although discretely visible drug-related information is publicly available in a semi-structured manner, a substantial amount of information remains uncovered in the textual form. This includes the electronic patient health records, hospital discharge summaries, medical case reports, full text research articles, blogs [4], and news reports [5].

With the growing amount of unstructured textual data, information extraction (IE) technologies [3,6] have gained popularity over more than a decade. The aim of information extraction is to automatically extract useful facets of information from the huge volumes of unstructured textual data. In the context of medical sciences, such processing may involve identifying the names of medical entities, the relationships between various entities and the events associated with them. Information extraction has immense potential in the medical domain [7]. A typical example of a medical information extraction system is the MedLEE [9] system that has found various applications in the medical scenarios [8]. Examples of EU-sponsored projects that have aimed at systematic

* Corresponding author at: Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany.
E-mail addresses: harsha.gurulingappa@scai-extern.fraunhofer.de (H. Gurulingappa), abdul-mateen.rajput@external.merckgroup.com (A.M. Rajput), a.roberts@dcs.shef.ac.uk (A. Roberts), juliane.fluck@scai.fraunhofer.de (J. Fluck), martin.hofmann-apitius@scai.fraunhofer.de (M. Hofmann-Apitius), luca.toldo@merckgroup.com (L. Toldo).

[1] http://www.rxlist.com/script/main/hp.asp.
[2] http://druginfo.nlm.nih.gov/drugportal/drugportal.jsp.
[3] https://www.pharmapendium.com.

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.04.008

exploration of information in text include EU-ADR,[4] EU-PSIP,[5] and IMI-EHR4CR.[6] Although there has been a significant progress in the information extraction research, a precise practical task would still require the availability of manually annotated corpora. A manually annotated corpus serves multiple purposes. First, it provides the necessary data for developing or optimizing the system irrespective of the underlying methodology (i.e. statistical, rule-based or machine learning-based). It serves as a gold standard (often referred to as ground truth) against which the performance of automatic systems can be compared. Furthermore, annotated corpora can also be used as curated dataset for the construction of literature-based knowledgebase (such as MetaCore,[7] and BRENDA[8]).

In the biological domain, efforts have been made to generate semantically annotated corpora like GENIA [10], BioCreative[9] and BioNLP.[10] However, these bio-corpora are restricted to the entities and events of biological interest such as gene names, protein names, cellular location or cellular events such as protein–protein interactions. In comparison to the biology domain, the availability of annotated corpora in the medical domain is limited. This is partially due to the proprietary nature of the existing data as well as ethical issues. In recent years, collaborative efforts such as CLEF [11] have been investing efforts to generate semantically annotated medical corpora for information extraction. The medical NLP challenge I2B2 [12] provides de-identified and annotated patient discharge summaries as well as a platform for common evaluation of information extraction techniques. There is a limited availability of task-specific corpora such as the AZDC corpus [13] annotated with the disease names or the Chem corpus [14] annotated with the chemical names that can be applied for specific named entity recognition tasks. The *DISAE* corpus [15] contains 400 MEDLINE articles annotated with the names of diseases and adverse effects without information about the drugs. Nevertheless, there is no annotated corpus that is publicly available (to the best of author's knowledge) that can be used for training, optimization or evaluation of the techniques for the identification of drug-related adverse effects from free text.

This paper reports on the construction of a gold standard corpus in which MEDLINE case reports[11] have been annotated for the mentions of drugs, adverse effects, dosages as well as the relationships between them. The entities and the relationships are annotated systematically to ensure that the quality of data is reliable enough to support information extraction research. Finally, as an example with an application point of view, the usability of the corpus is demonstrated by developing and validating a sentence classification model that can discriminate between informative sentences against the non-informative ones. The corpus is named as the *ADE* (adverse drug effect) corpus and annotations over the corpus are made freely available online at https://sites.google.com/site/adecorpus/.

2. Methods

2.1. The ADE corpus characteristics

During the development of a benchmark corpus, several characteristics have to be considered. Amongst them, two important ones are the domain suitability of the corpus and the target user group. Considering the domain suitability, medical case reports were of the first choice since they provide important and detailed information about symptoms, signs, diagnosis, treatment, and follow-up of individual patients. More importantly, case reports can serve as an early warning signal for the under-reported or unusual adverse effects of medications [16]. Since the goal of this work is to generate a corpus for public usability, MEDLINE articles were used due to their nature of free public availability. Therefore, the *ADE* corpus constitutes a subset of MEDLINE case reports.

2.2. Document sampling

Currently, MEDLINE contains more than 1.5 million medical case reports. In order to restrict the scope of the corpus to drug-related adverse events, a PubMed[12] search with *drug therapy* and *adverse effect* as MeSH [17] terms was performed limiting the language to *English*. The text option was chosen to be *abstract* in order to eliminate the documents with only title and no abstract text. A precise PubMed query performed on 2010/10/07 is as follows:

"adverse effects"[sh] AND (hasabstract[text] AND Case Reports[ptyp]) AND "drug therapy"[sh] AND English[lang] AND (Case Reports[ptyp] AND ("1"[PDAT]: "2010/10/07" [PDAT]))

This process retrieved nearly 30,000 documents from PubMed out of which 3000 documents (referred to as *ADE* corpus) were randomly selected for the annotation and benchmarking purposes. A corpus of 3000 annotated documents is believed to be substantially large to support the development and validation of information extraction systems.

An additional set of 100 non-overlapping documents (referred to as *ADE-seed* corpus) were selected in order to be used by the annotators for practicing the annotation task as well as for the annotation guideline refinement and stabilization. KNIME[13] was used for document sampling and dataset generation for the annotation task. KNIME is an open source workflow management system that provides graphically viewable data manipulation and processing environment. KNIME-based workflows are easily reproducible and minimize data handling errors.

2.3. Annotation guidelines

A critical issue that reflects the quality of an annotated corpus is consistency [19]. In order to generate an annotated corpus for information extraction modeling or performance benchmarking, consistent and uniform annotation across all the documents is essential. To ensure the consistency, a set of draft guidelines was developed and provided to all annotators. The guidelines provide rules that annotators should follow when working on documents. Draft guidelines were periodically revised before beginning the annotation of *ADE* corpus (see Section 2.4 for details). Important components of the annotation guidelines are as follows:

2.3.1. Drug

Names of drugs and chemicals that include brand names, trivial names, abbreviations and systematic names were annotated. Mentions of drugs or chemicals should strictly be in a therapeutic context. This category does not include the names of metabolites, reaction byproducts, or hospital chemicals (e.g. surgical equipment disinfectants).

2.3.2. Adverse effect

Mentions of adverse effects include signs, symptoms, diseases, disorders, acquired abnormalities, deficiencies, organ damage or death that strictly occur as a consequence of drug intake.

[4] http://www.alert-project.org/.
[5] http://www.psip-project.eu/.
[6] http://www.ehr4cr.eu/.
[7] http://www.genego.com/metacore.php.
[8] http://www.brenda-enzymes.org/.
[9] http://www.biocreative.org/news/corpora/biocreative-iii-corpus/.
[10] http://bionlp-corpora.sourceforge.net/.
[11] http://www.nlm.nih.gov/bsd/indexing/training/PUB_050.htm.
[12] http://www.ncbi.nlm.nih.gov/pubmed/.
[13] http://www.knime.org/.