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/.

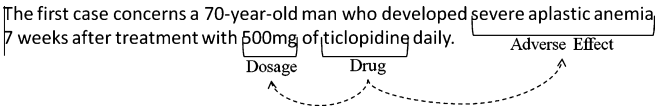

Fig. 1. Example of a sentence annotated with drug, dosage, adverse effect and the relationships between them.

2.3.3. Dosage

Dosage information that include the quantitative measurements (e.g. 0.1 mg/kg/day) as well the frequency mentions (e.g. two tablets twice daily) were annotated.

2.3.4. Relationship

The scope of a relationship was defined and restricted to the sentence level. There should be a clear mention of a drug/chemical resulting in an adverse effect defined within the context of a sentence. Mentions of drugs, disorders or dosages that do not fit into a relation were not annotated. Relationships were annotated between the drugs and adverse effects as well as between the drugs and dosages. Fig. 1 shows an illustration of a sentence annotated with the entities and relationship between them.

2.4. Annotation methodology

2.4.1. Annotation participants

Altogether, five individuals participated in the generation and revision of the annotation guidelines. Amongst them, three individuals were involved in the annotation task. All the annotators possess a minimum qualification of M.Sc. degree with the background related to Biomedicine. Two annotators have substantial experience working in the biomedical text mining domain whereas the third annotator has comparatively little practical experience working with text mining-related topics.

2.4.2. Annotation workflow

The annotation workflow follows the standards established by the CLEF framework [19]. Knowtator [21] *version 1.9 beta 2* was the tool used for annotation. The CLEF framework provides an easily configurable text annotation environment plugged into the knowtator toolkit. Fig. 2 shows the workflow adapted for the annotation task.

An individually single annotated document (i.e. a document annotated by only one annotator) can reflect several problems. They include idiosyncratic errors made by the annotators, missing annotations or the consistent under-performance of the individuals. In order to overcome these problems, a strategy of triple annotation [20] was applied. During the process of triple annotation, each document is independently annotated by three annotators and the sets of annotations are compared thereafter for quality assurance.

The annotation task started with applying the draft guidelines for annotating the *ADE-seed* corpus. First, the *ADE-seed* corpus of 100 documents was divided into *ADE-seed-set1* and *ADE-seed-set2* with each comprising 50 non-overlapping documents. As indicated in Fig. 2, initially the *ADE-seed-set1* sub-corpus was annotated by all three annotators by strictly applying the draft guidelines provided. The agreement between the annotators was determined using the Inter-Annotator Agreement (IAA) scores (see Section 3.1). The IAA scores were determined for the entities as well as for the relationships (see Section 3.2). The level of agreement was determined for all the documents and the under-performing documents were manually reviewed to check for disagreeing instances. An underperforming document is one that contains at least one disagreeing annotation. Depending on the necessity, changes were made to the annotation guidelines that were used.

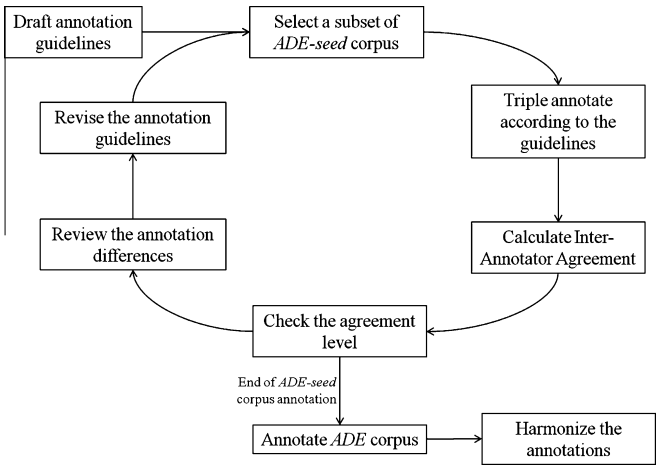

Fig. 2. The workflow employed for the annotation task (adapted from [19]).

The process was repeated for the *ADE-seed-set2*. Counts of the annotated entities and relationships over the *ADE-seed* corpus for two preliminary rounds of annotation are provided in Tables 1 and 2.

Before starting the annotation of the *ADE* corpus, an interactive stabilization of the annotation guidelines was performed based on the experiences gained during the annotation of *ADE-seed* corpus. An interactive stabilization of annotation guidelines involved mutual discussion between the annotators concerning any inconsistencies or conflicts in understanding the guidelines followed by subsequent refinement of the annotation rules. The *ADE* corpus of 3000 documents was divided into *ADE-set1*, *ADE-set2*, and *ADE-set3* subsets with each comprising 1000 non-overlapping documents. Each annotator processed two corpus subsets. With this strategy, every document was annotated by two annotators and the total number of documents that each annotator has to read was reduced by one-third. Fig. 3 shows the distribution of the subsets of *ADE* corpus among the different annotators. Table 3 shows the counts of the annotated entities and relationships over the *ADE* corpus.

2.4.3. Annotation harmonization

During the harmonization process, the double annotated documents were subjected to a review by the respective annotators in order to resolve the conflicting annotations and to improve the overall quality of the annotated corpus. Each document was reviewed by only two annotators who annotated the respective document earlier. The aim of annotation harmonization is to focus on the differences in annotator's interpretation of the guidelines and the differences in their interpretation of the documents. During harmonization, the documents were not completely read. Only those instances that were marked by at least one annotator were investigated for correctness. Documents that do not contain any annotation from both the annotators or documents where both the annotators agree completely were not reviewed. Documents that contain at least one conflicting annotation were subjected to the review process by the respective annotators. The following precautions were taken during the harmonization process.

1. No entirely new annotations were added if they were not marked earlier by either of the annotators.
2. No annotations were removed if they were marked earlier by both the annotators.
3. Annotations were added or removed if they were marked by any one of the annotators and provided they both agree on the decision thereafter.

Table 1
Counts of the annotated entities and relations in the *ADE-seed-set1* corpus. Numbers within the brackets indicate the unique number of sentences that contain at least one relation. Enumerations related to dosages are zeroes since no dosage information was annotated during this round.

	Entity counts			Relation counts	
	Drug	Adverse effect	Dosage	Drug-adverse effect	Drug-dosage
Annotator-1	116	139	0	166 (90)	0 (0)
Annotator-2	120	159	0	177 (84)	0 (0)
Annotator-3	57	132	0	52 (26)	0 (0)

Table 2
Counts of the annotated entities and relations in the *ADE-seed-set2* corpus. Numbers within the brackets indicate the unique number of sentences that contain at least one relation.

	Entity counts			Relation counts	
	Drug	Adverse effect	Dosage	Drug-adverse effect	Drug-dosage
Annotator-1	91	83	4	110 (68)	4 (4)
Annotator-2	86	77	3	95 (65)	3 (3)
Annotator-3	54	60	0	59 (46)	0 (0)

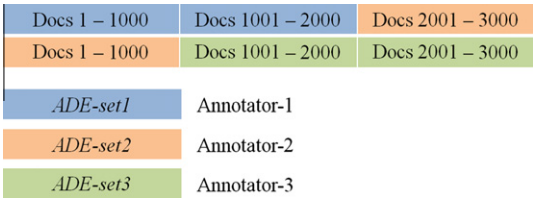

Fig. 3. Distribution of the subsets of *ADE* corpus among the different annotators. Each subset contains 1000 non-overlapping documents.

Table 3
Counts of the annotated entities and relations in the *ADE* corpus. Numbers within the brackets indicate the unique number of sentences that contain at least one relation. Each annotator handles only 2000 documents that are distributed according to Fig. 3.

	Entity counts			Relation counts	
	Drug	Adverse effect	Dosage	Drug-adverse effect	Drug-dosage
Annotator-1	2391	3330	129	3995 (2490)	140 (111)
Annotator-2	3097	3464	69	4028 (2681)	71 (60)
Annotator-3	3999	4604	77	5489 (3404)	83 (77)

4. In case of partially overlapping annotations, only the conflicting parts were resolved. For instance, Annotator-1 marks *acute lymphoblastic leukemia* whereas the Annotator-2 marks *lymphoblastic leukemia*, then the decision will be made to resolve the annotation of the word *acute*.

The harmonization was performed over the complete *ADE* corpus in the presence of annotators for both the entities as well as the relationships. Table 4 shows the counts of the annotated entities and relationships over the *ADE* corpus after the harmonization procedure. Twenty-eight documents were removed from the *ADE* corpus due to errors induced by the annotation software as well as manual handling errors (such as missing annotations and annotation offset shifts). At the end of harmonization, the *ADE* corpus contains 2972 documents having 420,515 tokens within 20,967 sentences out of which 4272 sentences have been annotated with names and relationships between drugs, adverse effects and dosages. The sentences with drug-dosage relationships (i.e. 213 sentences) constitute a subset of 4272 sentences that contain drug-adverse effect relationships.

Table 4
Counts of the annotated entities and relations in the *ADE* corpus after harmonization. Numbers within the brackets indicate the unique number of sentences that contain at least one relation.

Entity counts	
Drug	5063
Adverse effect	5776
Dosage	231
Relation counts	
Drug-adverse effect	6821 (4272)
Drug-dosage	279 (213)

2.5. Modeling a sentence classifier

In order to demonstrate an example scenario where the *ADE* corpus can be applied, models for sentence classification were developed and validated. Although the scope of this work does not aim to extensively build a sentence classification model, with an application point of view the usability of the corpus has been shown. Additionally, the experiments performed with sentence classification do not implicitly restrict the scope of the corpus to sentence classification task.

A sentence classification framework helps in the extraction of sentences that contain statements of hypotheses or declarations about the adverse effects related to certain drugs. It helps in retrieving the documents or sentences that contain potential signals indicating the drug-related adverse effects that can help in safety monitoring of the drugs as well as to develop new hypothesis based on existing evidences.

The *ADE* corpus (after harmonization containing 2972 documents) was transformed into sentences using the Genia Sentence Splitter [23]. The sentence splitting was performed after the complete annotation and harmonization of the *ADE* corpus. As a result of sentence splitting, the *ADE* corpus of 2972 documents generated 20,967 sentences. The corpus of sentences was divided into 4272 sentences containing at least one drug-related adverse effect mention that were labeled as *Positive*. The remaining 16,695 sentences that contain no information about the adverse drug effects were separated and labeled as *Negative*. Therefore, the sentences corpus contains overall 20,967 instances labeled as either *Positive* or *Negative* that can be subjected to the classification task. Fig. 4 illustrates the sentence classification framework applied for training and validating a model for classifying the informative and non-informative sentences.

For the purpose of training and validation, the Naïve Bayes classifier [24] and the Maximum Entropy classifier [25] implemented within the MALLET toolkit [26] were applied with the default

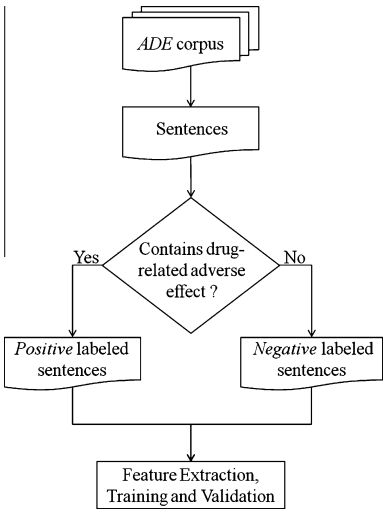

Fig. 4. Sentence classification framework applied for training and validating a model for classifying the informative (*Positive* labeled) and non-informative sentences (*Negative* labeled).

parameter settings using all words occurring in sentences as features.

3. Results and discussion

3.1. Inter-annotator agreement metrics

Over the *ADE-seed* as well as the *ADE* corpora, the double annotated documents were used for the determination of Inter-Annotator Agreement (IAA) scores. The IAA scores were calculated using the F_1 score as a criterion [22]. The F_1 score measures the harmonic mean of precision and recall between the annotators using one annotator as a standard and the other as a reference. The reason for applying F_1 score compared to conventional measures such as Kappa [34] is that F_1 scores are computed at the level of entities. Since the annotation task requires finding the correct boundaries of entities in addition to label assignment, chance agreement is effectively zero, and therefore F_1 score is equivalent to kappa. This offers an advantage of understanding the named entities on which annotators completely agree, partially agree, or completely disagree. In addition, the F_1 score provides an easy and effective way for calculating the level of agreement on relationships. The IAA scores were determined for both the entities as well as for the relationships. GATE [18] framework was used for the determination of IAA scores.

For the entities, IAA scores were determined using the *exact match* and *partial match* as criteria. *Exact match* is a situation where both the annotations should completely overlap whereas *partial match* is a situation where the annotations may partially or completely overlap. Using complete and partial match as criteria helps in understanding the entities over which the annotators disagree completely and those instances where annotators agree at least partially. For the relationships, two types of evaluations were applied. They are the *exact entity match with exact relation* and *partial entity match with exact relation*. The *exact entity match with exact relation* requires that the annotations of the entities overlap completely and the relationship is correctly annotated. In case of *partial entity match with exact relation*, a relationship that links two partially or completely matching entity spans is considered to be correct.

3.2. Inter-annotator agreement calculation

The IAA scores between the annotators were determined over the *ADE-seed* corpus during two preliminary rounds of annotation. Whereas, the IAA scores over the *ADE* corpus were determined before the final harmonization was performed. The agreement levels were determined for the entities as well as for the relationships. Tables 5 and 6 show the IAA scores over the *ADE-seed-set1* and *ADE-seed-set2* corpora respectively. The *ADE-seed-set1* corpus did not contain any mentions of dosages that fit into a pre-defined relationship with drugs. Therefore, the IAA scores for dosages were enumerated as zero for the entity mentions as well as for the relationships with drugs. During the preliminary annotation rounds, the level of agreement between Annotator-1 and Annotator-2 remained consistent for drug names. A potential reason is that drug names often occur as one word entities (e.g. minocycline) and they hardly suffer from boundary mismatch problems. However, the agreement levels for the exact name matches of adverse effects and dosages were unsatisfactory. The names of adverse effects often occur as descriptive multiword terms and deciding the correct term boundaries was a major problem. For instance, Annotator-1 marked *non-metastatic gestational trophoblastic tumor* whereas the Annotator-2 marked the same instance as *gestational trophoblastic tumor*. Nevertheless, the partial name matches for adverse effects had substantial level of agreement. Dosage information faced severe annotation problems. Mentions such as *low-dose* were often misinterpreted or overseen by the annotators and were not annotated. Such instances represent the contemporary errors induced during the annotation process that were improved later on. Typical examples of relationship annotation errors include one in the sentence *the patient developed monoarthritis 2 weeks after initiation of IFN-beta, which persisted during 14 months of therapy and resolved with discontinuation of the medication [PMID: 16393774]*. A relationship between *monoarthritis* and *IFN-beta* exists which was correctly annotated by one annotator whereas overseen by the other annotator. Such instances were exemplified in the annotation guidelines and thoroughly discussed before the annotation of main corpus was performed. Annotator-3 having minimum experience with text annotation exercises often achieved lower agreement scores with rest of the annotators. However, after the annotation of *ADE-seed-set2*, the agreements between Annotator-3 and rest of the annotators improved and reached until 60% for adverse effects and their relationships to drugs. Since the final goal was to harmonize the complete annotations, the *ADE* corpus was annotated by three annotators with no further revisions of the annotation guidelines.

Table 7 shows the IAA scores between the annotators over the large *ADE* corpus that contains 3000 documents. The *ADE* corpus was strategically divided and annotated by three annotators. Therefore, the IAA scores were determined over the sets of 1000 documents that were commonly annotated by two annotators. Based on the experiences gained during the preliminary annotation rounds, all the three annotators were able to consistently annotate the drug names. Although the names of adverse effects underwent frequent boundary problems, the results of partial name matches were consistent amongst all three annotators. The dosage information being the poorest annotated entity class was strictly resolved during the harmonization process. Therefore, the authors believe that dosage annotations in spite of having poor agreement between annotators are still valuable due to the thorough harmonization later on. All the annotated entities and relationships were subjected to the harmonization procedure after the complete annotation of *ADE* corpus in the presence of respective annotators in order to achieve a consistent final annotation. Randomly selected 200 annotated sentences from the *ADE* corpus after harmonization were reviewed by two drug safety experts in order to judge the quality of annotations where both experts agreed on all annotations. This illustrates the suitability of annotated corpus for modeling real world information extraction challenges related to drug safety.

Table 5
IAA scores between the annotators over the *ADE-seed-set1* corpus containing 50 documents. Enumerations related to dosages are zeroes since no dosage information was annotated during this round.

Annotators	Entity (*exact match*)			Entity (*partial match*)		
	Drug	Adverse effect	Dosage	Drug	Adverse effect	Dosage
1 and 2	0.76	0.66	0.00	0.82	0.86	0.00
1 and 3	0.28	0.43	0.00	0.38	0.55	0.00
2 and 3	0.29	0.40	0.00	0.38	0.51	0.00

Annotators	Relation (*exact entity match with exact relation*)		Relation (*partial entity match with exact relation*)	
	Drug-adverse effect	Drug-dosage	Drug-adverse effect	Drug-dosage
1 and 2	0.64	0.00	0.79	0.00
1 and 3	0.14	0.00	0.37	0.00
2 and 3	0.10	0.00	0.37	0.00

Table 6
IAA scores between the annotators over the *ADE-seed-set2* corpus containing 50 documents.

Annotators	Entity (*exact match*)			Entity (*partial match*)		
	Drug	Adverse effect	Dosage	Drug	Adverse effect	Dosage
1 and 2	0.73	0.88	0.29	0.90	0.88	0.86
1 and 3	0.63	0.65	0.00	0.77	0.66	0.00
2 and 3	0.57	0.66	0.00	0.76	0.67	0.00

Annotators	Relation (*exact entity match with exact relation*)		Relation (*partial entity match with exact relation*)	
	Drug-adverse effect	Drug-dosage	Drug-adverse effect	Drug-dosage
1 and 2	0.69	0.28	0.87	0.85
1 and 3	0.51	0.00	0.65	0.00
2 and 3	0.46	0.00	0.66	0.00

Table 7
IAA scores between the annotators over the *ADE* corpus containing 3000 documents. Each annotator handles only 2000 documents that are distributed according to Fig. 3. IAA scores are calculated over the sets of 1000 documents that are commonly annotated by two annotators.

Annotators	Entity (*exact match*)			Entity (*partial match*)		
	Drug	Adverse effect	Dosage	Drug	Adverse effect	Dosage
1 and 2	0.80	0.72	0.26	0.82	0.80	0.43
1 and 3	0.75	0.68	0.05	0.77	0.77	0.37
2 and 3	0.76	0.63	0.03	0.78	0.77	0.09

Annotators	Relation (*exact entity match with exact relation*)		Relation (*partial entity match with exact relation*)	
	Drug-adverse effect	Drug-dosage	Drug-adverse effect	Drug-dosage
1 and 2	0.68	0.17	0.78	0.26
1 and 3	0.63	0.14	0.74	0.18
2 and 3	0.60	0.12	0.75	0.15

3.3. Semantic corpus analysis

After the harmonization procedure, in order to analyze the semantic distribution of entities in the *ADE* corpus, the annotated names of drugs and adverse effects were mapped to standard ontologies using the ProMiner [28] system. The drug names were mapped to the Anatomical Therapeutic Chemical (ATC) classification system [29] using the DrugBank [31] dictionary. The ATC hierarchically classifies several drugs according to their pharmacotherapeutic properties. Since ATC is hierarchical, its level two classes were used for the analysis. The names of adverse effects were mapped to the MedDRA [30] classification system. MedDRA contains a hierarchically organized medical terminology and it been widely applied for pharmacovigilance and drug regulatory affairs. Similar to ATC, the level two MedDRA classes were used for analysis. Out of 5063 annotated drug names, 4205 could be normalized to the ATC (i.e. 83%) whereas for the adverse effects, 4356 out of 5776 names (i.e. 75%) could be mapped to the MedDRA. Table 8 shows top five ATC classes to which frequently occurring drugs belong. Table 9 shows top five MedDRA classes to which frequently occurring adverse effects belong.

Table 8
Top five ATC classes to which the frequently occurring drugs belong.

ATC class	% of drugs
Antineoplastic agents	22
Ophthalmologicals	11
Antibacterial agents	11
Immunosuppressants	9
Antiepileptics	8

Table 9
Top five MedDRA classes to which the frequently occurring adverse effects (AE) belong.

MedDRA class	% of AE
Cardiac arrhythmias	12
General system disorders	11
Epidermal and dermal conditions	9
Allergic conditions	9
Hepatic and hepatobiliary disorders	8

Table 10
Performances of sentence classifiers validated by 10-fold cross-validation of the training data. Precision, Recall, and F_1 score for the class positive are reported.

	Precision	Recall	F_1 score
Naïve Bayes	0.91	0.08	0.14
MaxEnt	0.75	0.64	0.70

3.4. Performance measure of sentence classifiers

The performance of the sentence classification framework was evaluated using 10-fold cross-validation [27]. During the cross-validation experiments, the discriminative capability of the classifier to correctly distinguish between the informative and non-informative sentences was measured using *Precision*, *Recall*, and *F_1 score* over the class *Positive* as a criterion.

3.5. Results of sentence classification

The Naïve Bayes and the Maximum Entropy (MaxEnt) classifiers were applied for training and validating models for sentence classification that can discriminate the potentially informative against the non-informative sentences (see Section 2.5 for experimental details). The sentences that contain at least one drug-related adverse effect formed the informative (*Positive* labeled) dataset whereas the remaining sentences formed the non-informative (*Negative* labeled) dataset. Upon evaluation by 10-fold cross-validation, the performances of Naïve Bayes and MaxEnt classifiers are shown in Table 10.

A MaxEnt model trained with simple features such as words in the sentence resulted in a good classification performance. Removal of stop-words degraded the performance of classification. There has been a significant amount of work dedicated to the sentence classification task in the medical text [32]. A variety of feature sets and techniques such as feature boosting have been proposed for improving the performance of classification. However, since the scope of this work does not extensively aim at building a sentence classifier, the performance of classification has been shown as an example scenario where the corpus could already find a direct application. A detailed framework on application of the corpus for adverse effect sentence classification using shallow linguistic features as well as its ability to support the identification of novel drug-adverse effect associations has been demonstrated by Gurulingappa et al. [33].

4. Conclusion

A semantically annotated corpus designed to support the extraction of information about drug-related adverse effects from medical case reports has been presented. The corpus is intended to facilitate the development and validation of automated systems for evidence-based pharmacovigilance and hence overcome the manual reading task that has been performed in many pharmaceutical and healthcare companies. The corpus has been systematically double annotated using a well-defined annotation schema in order to ensure the maximum consistency. The annotations were performed in different rounds with intermediate quality control checks using the inter-annotator agreement scores. The process of harmonization of double annotated documents in order to generate a representative consensus annotation is discussed as well. Finally, an example application scenario has been demonstrated by applying the corpus for training and validating sentence classifiers that can discriminate the informative sentences against the non-informative ones.

The corpus has been made publicly accessible in order to encourage the research in the direction of evidence-based pharmacovigilance and drug safety aspects. Currently, a significant amount of work has been going on in order to develop a system for automatic extraction of drug-related adverse effects. As an immediate following step, a second version of the corpus thoroughly checked by the clinical experts will be released. Our future work intends to evaluate the performance of couple of commercial and freely available tools for named entity recognition and relationship extraction. We also plan to extend our work to include blogs and news reports for the analysis of drug-related adverse events.

Acknowledgments

This work has been funded in part by a "B-IT Research School" scholarship Grant from the state of NorthRhineWestfalia to Harsha Gurulingappa. We also thank drug safety experts at Merck, Dr. Maren Rohrbacher (MD PhD) and Dr. Yorki Tayrouz (MD PhD) for reviewing a subset of annotations.

References

[1] World Health Organization (WHO) glossary of terms used in Pharmacovigilance. <http://www.who-umc.org/graphics/15338.pdf>.
[2] Ahmad SR. Adverse drug event monitoring at the Food and Drug Administration. J Gen Intern Med 2003;18(1):57–60.
[3] Erhardt RA, Schneider R, Blaschke C. Status of text-mining techniques applied to biomedical text. Drug Discov Today 2006;11(7–8):315–25.
[4] Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In: Proceedings of the 2010 workshop on biomedical natural language processing; 2010. p. 117–25.
[5] Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, Tateno Y, et al. BioCaster: detecting public health rumors with a Web-based text mining system. Bioinformatics 2008;24(24):2940–1.
[6] Jagannathan V, Mullett CJ, Arbogast JG, Halbritter KA, Yellapragada D, Regulapati S, et al. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. Int J Med Inform 2009;78(4):284–91.
[7] Rodriguez-Esteban R. Biomedical text mining and its applications. PLoS Comput Biol 2009;5(12):e1000597.
[8] Hyun S, Johnson SB, Bakken S. Exploring the ability of natural language processing to extract data from nursing narratives. Comput Inform Nurs 2009;27(4):215–23.
[9] Chen L, Friedman C. Extracting phenotypic information from the literature via natural language processing. Stud Health Technol Inform 2004;107:758–62.
[10] Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus – semantically annotated corpus for bio-textmining. Bioinformatics 2003;19(Suppl. 1):i180–2.
[11] Roberts A, Gaizauskas R, Hepple M, Davis N, Demetriou G, Guo Y, et al. The CLEF corpus: semantic annotation of clinical text. AMIA Annu Symp Proc 2007:625–9.
[12] Uzuner O, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. J Am Med Inform Assoc 2010;17(5):519–23.
[13] Leaman R, Miller C, Gonzalez G. Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. In: Proceedings of the 3rd international symposium on languages in biology and medicine; 2009. p. 82–9.
[14] Kolarik C, Klinger R, Friedrich CM, Hofmann-Apitius M, Fluck J. Chemical names: terminological resources and corpora annotation. In: The Proceedings of BioTxtM workshop (6th edition of the language resources and evaluation conference; 2008.
[15] Gurulingappa H, Klinger R, Hofmann-Apitius M, Fluck J. An empirical evaluation of resources for the identification of diseases and adverse effects in biomedical literature. In: 2nd workshop on building and evaluating

resources for biomedical text mining (7th edition of the language resources and evaluation conference), Valetta, Malta; May 2010.
[16] Kidd M, Hubbard C. Introducing journal of medical case reports. J Med Case Rep 2007;1:1.
[17] Lowe HJ, Barnett GO. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. JAMA 1994;271(14): 1103–8.
[18] Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: a framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th anniversary meeting of the association for computational linguistics (ACL'02), Philadelphia; July 2002.
[19] Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Roberts I, et al. Building a semantically annotated corpus of clinical texts. J Biomed Inform 2009;42(5):950–66.
[20] Wilbur WJ, Rzhetsky A, Shatkay H. New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinformatics 2006;7:356.
[21] Ogren PV. Knowtator: a plug-in for creating training and evaluation data sets for Biomedical Natural Language systems. In: Proc 9th internat protege conf; 2006.
[22] Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and *F*-score, with implication for evaluation. Adv Inform Retrieval Lect Notes Comput Sci 2005;3408:345–59.
[23] Rune S, Yoshida K, Yakushiji A, Miyao Y, Matsubayashi Y, Ohta T. AKANE system: protein–protein interaction pairs in BioCreAtIvE2 challenge, PPI-IPS subtask. In: Proceedings of the second biocreative challenge evaluation, workshop; April 2007. p. 209–12.
[24] Rish I. An empirical study of the naive Bayes classifier. In: Proceedings of IJCAI-01 workshop on empirical methods in artificial intelligence; 2001.
[25] Nigam K. Using maximum entropy for text classification. In: Proceedings of IJCAI-99 workshop on machine learning for information filtering; 1999. p. 61–7.
[26] McCallum AK. Mallet: a machine learning for language toolkit; 2002. <http://mallet.cs.umass.edu>.
[27] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI'95 Proceedings of the 14th international joint conference on artificial intelligence; 1995.
[28] Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J. ProMiner: rule-based protein and gene entity recognition. BMC Bioinformatics 2005;6(Suppl. 1):S14.
[29] Miller GC, Britt H. A new drug classification for computer systems: the ATC extension code. Int J Biomed Comput 1995;40(2):121–4.
[30] Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). Drug Saf 1999;20(2):109–17.
[31] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucl Acids Res 2006;34(Database issue):D668–72.
[32] McKnight L, Srinivasan P. Categorization of sentence types in medical abstracts. AMIA Annu Symp Proc 2003:440–4.
[33] Gurulingappa H, Fluck J, Hofmann-Apitius M, Toldo L. Identification of adverse drug event assertive sentences in medical case reports. In: First international workshop on knowledge discovery and health care management (KD-HCM), European conference on machine learning and principles and practice of knowledge discovery in databases (ECML PKDD); 2011.
[34] Blackman NJM, Koval JJ. Interval estimation for Cohen's kappa as a measure of agreement. Stat Med 2000;19(5):723–41.