



The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships

Erik M. van Mulligen^{a,*}, Annie Fourier-Reglat^b, David Gurwitz^c, Mariam Molokhia^d, Ainhoa Nieto^e, Gianluca Trifiro^{a,f}, Jan A. Kors^a, Laura I. Furlong^g

^a Dept. of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

^b Université de Bordeaux, U657, F-33000 Bordeaux, France

^c Tel-Aviv University, Tel Aviv, Israel

^d Kings College London, London, United Kingdom

^e University of Santiago de Compostela, Santiago de Compostela, Spain

^f University of Messina, Messina, Italy

^g Research Programme on Biomedical Informatics (GRIB), IMIM (Hospital del Mar Research Institute), Universitat Pompeu Fabra, Barcelona, Spain

ARTICLE INFO

Article history:

Received 2 May 2011

Accepted 11 April 2012

Available online 25 April 2012

Keywords:

Text mining

Corpus development

Machine learning

Adverse drug reactions

ABSTRACT

Corpora with specific entities and relationships annotated are essential to train and evaluate text-mining systems that are developed to extract specific structured information from a large corpus. In this paper we describe an approach where a named-entity recognition system produces a first annotation and annotators revise this annotation using a web-based interface. The agreement figures achieved show that the inter-annotator agreement is much better than the agreement with the system provided annotations. The corpus has been annotated for drugs, disorders, genes and their inter-relationships. For each of the drug–disorder, drug–target, and target–disorder relations three experts have annotated a set of 100 abstracts. These annotated relationships will be used to train and evaluate text-mining software to capture these relationships in texts.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Systems for automated extraction of information from large corpora of biomedical documents hold promise as an adjunct to human information processing capabilities. For example, information extraction systems have applications in the support of database curation [1], interpretation of high-throughput experiments [2], and discovery of novel protein–protein interactions [3].

To develop and evaluate such systems annotated corpora are necessary. Most currently available corpora have annotations for only one type of entity, and only very few annotate relationships between entities. The entities most frequently annotated are proteins and genes, with relationship annotation focusing on protein–protein interactions. However, relationships involving other entities, such as drugs and diseases, are increasingly being studied. In the EU-ADR project, for example, data mining of massive amounts of electronic health records generates many potential signals (drug–disease combinations in which the disease can be considered as an adverse event due to the drug) [4]. If a drug–disease association is reported in literature, the signal can be discarded; else drug–target (proteins, genes, and gene variants) and target–disease relationships are sought to substantiate the signal,

i.e., to find a plausible biological mechanism that can explain the signal. Systems to automatically retrieve these relationships from scientific literature and databases would be very helpful, but annotated corpora to train and test these systems are currently not available.

Another limitation of most existing corpora is that they do not provide information about the agreement between the corpus annotators. This makes it difficult to compare the performance of an automated system with that of an individual human annotator.

In this study we describe the construction of a corpus, dubbed the EU-ADR corpus, which contains annotations of multiple entities (drugs, diseases, and targets) and relationships between these entities. Inter-annotator agreement scores are also provided.

2. Background

Annotated corpora have been developed for a number of biomedical domains, mostly focusing on gene and protein annotations. The Genia corpus v3.0 consists of 2000 Medline abstracts that have been annotated for a subset of the substances and the biological locations involved in reactions of proteins [5,6]. Several corpora were constructed as part of the BioCreative challenges [7]. One corpus consisted of 20,000 sentences from Medline abstracts, which were annotated for genes and proteins [8]. In another corpus, consisting of 358 full-text documents curated by

* Corresponding author. Fax: +31 10 7044722.

E-mail address: e.vanmulligen@erasmusmc.nl (E.M. van Mulligen).

MINT or IntAct, protein–protein interactions were annotated only at the abstract level, i.e., the annotators did not indicate the location of the interacting proteins in the text. Another example of an annotated corpus is PennBioIE [9]. It provides annotations of proteins and malignancies (cancer) for 2258 PubMed abstracts. In many of these corpora apart from the medical entities also lexical information (paragraphs, sentences, and part of speech) are provided.

Few corpora have been annotated for drugs and diseases. A subset of 1731 texts has been extracted from the PharmGKB database and annotated for drugs, diseases, and genes, together with their relationships [10,11]. The annotations only indicate terms and relationships that are mentioned in an abstract, not their actual location in the text. The corpus is not publicly available.

In the BioText corpus [12] several relationships between disorders and treatments have been annotated. It is composed of 100 titles and 40 abstracts from Medline. The treatments comprise both drugs and medical treatments. The annotations are performed at the sentence level. Both positive and negative relationships have been annotated.

For testing chemical dictionaries a manually annotated set of 100 PubMed abstracts has been used for the SCAI corpus [13]. Based on a complexity indicator for chemical terms a set of abstracts was selected and annotated by two independent annotators.

Leaman developed a manually annotated disease and symptoms corpus – known as the Arizona Disease Corpus – existing of 794 PubMed abstracts [14]. This set does however not contain adverse effects mentions. The corpus contains the location of the mentions of a disease or symptom and the mapping to an equivalent UMLS concept. This corpus is freely downloadable.

Some corpora provide annotations that were generated by automatic means. EDGAR is a corpus of 103 PubMed abstracts about cancer, which has been annotated for drugs, genes and the relationships between these entities by the EDGAR program [15]. The Wisconsin corpus [16] is composed of 1,529,731 sentences that were annotated automatically for three types of relationships: gene–disorder, protein–protein interaction and protein–subcellular localization. The locations of the entities are not provided and human experts did not review the annotations in these corpora for their quality.

Almost all these corpora are used either for training and evaluating text mining systems and typically one finds precision, recall and f-measures describing the quality of the system under evaluation against the gold standard. However, only little data is provided by the constructors of these corpora on the inter-annotator agreement (IAA). Sometimes, a description of the annotation guidelines is given with a process to refine the annotations into a certain consensus. Roberts presented in his paper on the CLEF corpus how inter-annotator agreement was computed (with a relaxed and corrected IAA) [17].

Summarizing, existing corpora often focus on the annotation of single entities, provide only limited relationship annotation, if any, and do not provide inter-annotator agreement scores. A corpus with annotations for drug–disease, drug–target, and target–disease relationships is not available. Here we report on the construction of such a corpus.

3. Material and methods

3.1. Strategy for corpus development

The development of the EU-ADR corpus consisted of the following phases:

- Definition of the scope of the corpus and first version of the annotation guidelines.

- Selection of the expert annotators within the EU-ADR project. Five experts volunteered to participate in the annotation project. Their area of expertise covers molecular biology, pharmacology, pharmacogenomics, and pharmacovigilance. Based on the individual expertise of each annotator, each of them was assigned to different annotation tasks.
- Pilot annotation experiment. A pilot annotation exercise was conducted on a small set of documents in order to train the annotators, learn how to use the annotation tool, refine the annotation guidelines, and evaluate the time and effort required for annotation.
- Document selection from Medline.
- Start of annotation project.
- Evaluation of the annotations. Analysis of the inter-annotator agreement, development of the final, consensus EU-ADR corpus.

The annotation was performed on Medline abstracts. Text from the title and the abstract was annotated at the sentence level. An important consideration in development of a corpus is the way the annotations are encoded in the text. It is not enough to say that an article describes an association between drug X and target Y, but pointers to the text fragments that denote the drug X and the target Y have to be provided.

3.2. Annotation guidelines

The annotation guidelines were developed after manual inspection of abstracts concerning the relationships of interest to the project and our previous survey on biomedical corpora. In addition, the GENIA project for annotation of biomedical events [18] and the BioIE project [20] were used as guides for the preparation of the guidelines. Here we only provide the general guidelines used in the annotation of the EU-ADR corpus:

1. Associate all the annotations with actual expressions in the text. This is known as “Text-bound annotations”, as described in [18].
2. The annotations were performed on single sentences. All the evidence for a relationship should come from the same sentence. The context (the rest of the abstract) can be used for disambiguation.
3. Annotate only relationships between the entities drug, target, and disease. Other types of relationships should not be annotated.
4. Annotate relationships according to the provided categories. Relationships that do not belong to these categories should not be annotated.
5. Entities will be already annotated in the text by means of an automatic NER system (Peregrine) [19]. Annotators will have to check if the annotation of entities is correct, and perform additional annotation of entities if they find an entity that was missed by the NER system.

3.3. Entity annotation

This task involves the identification of text spans that correspond to the biomedical entities of our interest: drug, targets, and diseases. The annotation tool provided the pre-annotation of entities. Thus, the annotators did not have to annotate entities from scratch, unless the NER system Peregrine [21] missed entities or annotated them incorrectly. Peregrine uses a thesaurus that contains per concept all term variations, its definition, external identifiers/codes, and the semantic type and hierarchical relations. It will use the terms to detect in the text occurrences of the concept (maximum term matching) and attach the concept identifier to

the text. However, the annotators had to check that the annotation of entities is correct, and perform additional annotation of entities if they found an entity that was missed or incorrectly annotated by Peregrine.

The entities on which we focus have been defined as follows:

- **Target:** genes, proteins and sequence variants of genes and proteins. More specifically:
 - Gene/RNA: a gene or RNA element, and families of genes.
 - Protein: gene products, protein complexes and protein families.
 - Variation: any type of sequence variation within the population as well as mutations.

This corresponds to the types Genes & Molecular sequences and SNP & Sequence variants in the annotation tool.

- **Disease:** disease phenotypes of the adverse drug reactions. In EU-ADR we focus on ADRs, most of which will classify as disorders according to the UMLS Metathesaurus. Thus, disease phenotypes in general are annotated, which correspond to the type “Diseases & Disorders” in the annotation tool.
- **Drug:** biologically active chemicals, marketed drugs and drug metabolites. Drug classes are also allowed. This corresponds to the type “Chemicals & Drugs” in the annotation tool.

3.4. Relationship annotation

In this annotation project we considered only three relationships between entities. In addition, for each relationship, we specified the type of the relationship according to the level of certainty. The relationships considered are the following:

- **Target–disease:** this relationship indicates if the target is associated with the disease, for instance, if the target plays a role in the mechanism underlying the disease, or is a marker of the disease, or if there is any evidence that is associated to the disease.
- **Target–drug:** this relationship indicates if the drug binds to a protein target, but also if the drug affects gene expression or modifies in some way the gene or the protein function.
- **Drug–disease:** this relation indicates if the drug is associated with the disease, for instance, if the drug may produce an adverse effect.

Once we have identified a relationship between two entities, we have to specify the type of the relationship. The relationship types are defined with respect to the level of certainty used to express a relationship in the text at the sentence level:

- **Positive association (PA):** the sentence clearly states that there is an association between the entities.
- **Negative association (NA):** the sentence clearly states that there is no association between the entities.
- **Speculative association (SA):** the sentence describes a putative relationship between the target and the disease. This might be confirmed or refuted later in the abstract, but in the sentence under study the relationship is presented as a speculation.

We illustrate these types using the target–disease relationship in which the targets are SNPs.

- **Positive association:** the sentence states that the SNP is associated with the disease. It can be a causative association (the SNP introduces a change that causes the disease), a marker

association (the SNP is in linkage disequilibrium with the “real” causative variation, which is not known, thus the SNP can be used as a disease marker).

Examples (entities shown in bold):

“We report on 3 sisters with severe **hyperhomocysteinemia** due to homozygosity for the **CBS 833T** → **C** mutation”.

“A point mutation was observed in an individual diagnosed with **HEP**, resulting in an **alanine to glycine change at amino acid position 80** and was present on both alleles”.

“Our data provide the first evidence that the **Cdc6 G1321A** polymorphism is associated with **decreased risk of cancer**”.

- **Negative association:** the sentence states that there is no association between the SNP and the disease phenotype. With this category we consider negative findings in the literature regarding relationships.

Example:

“The frequency of **Lys-251** was 10-fold greater in African-Americans than in Caucasians, but was not associated with **essential hypertension**”.

- **Speculative association:** the association is expressed as a hypothesis, can be identified by the use of expressions such as “might be”, “suggest”, etc.

Example:

“The **Met98Lys** change may be associated with a fraction of normal-tension **glaucoma** in patients of Japanese ethnicity”.

In summary, the relationship types apply to each relationship (target–disease, target–drug, drug–disease), and indicate the level of certainty used to express a fact in text.

3.5. Document selection

Three sets of 100 Medline abstracts were obtained (from 2007 until 2008) using PubMed with four different queries (Table 1). For the drug–disorder and drug–target relation 100 abstracts were randomly selected from the PubMed result. The target–disorder set was composed of 50 randomly selected abstracts for gene–disorder and 50 from the SNP–disorder relation. Only for the drug–disorder query the “adverse effects” subheading is applicable and added. There were no other similar subheadings available for the other relations. Three experts independently annotated the entities and their inter-relations for each abstract in each of the three sets of 100 abstracts according to the annotation guidelines described below. Some experts annotated more than one set.

3.6. Annotation tool

A web-based annotation tool has been developed that assisted the experts in the annotation tasks (see Fig. 1). The main features of the annotation tool are: it is a web-based system and the annotations are stored server-side per annotator; entities are pre-annotated automatically, the annotator has to correct these annotations and provide new ones if required; relation annotations are automatically generated based on co occurrence of entities in one sentence; the system provides standoff annotations. The tool is based on software from Knewco [21] that shows in-text highlights of terms recognized by the named entity recognition (NER) system Peregrine. The relations between the terms are automatically derived based on co occurrence and limited to the semantic types of the selected annotation relationship – so for the drug–disorder relationship only co occurrences between a drug and a disorder are considered – and proposed to the annotator. The annotator can revise the entities marked up and the relations found indicating the presence of the relationship and its type. In addition, the annotators can add new entities if required.

In order to minimize the burden on the annotators we integrated a NER system – Peregrine – to pre-annotate a PubMed

Table 1
PubMed queries used to construct the different annotation sets.

Relation	Query
Drug-disorder	(“Inorganic Chemicals”[Mesh] OR “Organic Chemicals”[Mesh] OR “Heterocyclic Compounds”[Mesh] OR “Polycyclic Compounds”[Mesh] OR “Hormones, Hormone Substitutes, and Hormone Antagonists”[Mesh] OR “Carbohydrates”[Mesh] OR “Lipids”[Mesh] OR “Amino Acids, Peptides, and Proteins”[Mesh] OR “Nucleic Acids, Nucleotides, and Nucleosides”[Mesh] OR “Biological Factors”[Mesh] OR “Biomedical and Dental Materials”[Mesh] OR “Pharmacologic Actions”[Mesh]) AND “Diseases Category”[Mesh] AND hasabstract[text] AND (“2007”[PDAT]: “2008”[PDAT]) AND “humans”[MeSH Terms] AND English[lang] AND “adverse effects”[Subheading]
Drug-target	(“Inorganic Chemicals”[Mesh] OR “Organic Chemicals”[Mesh] OR “Heterocyclic Compounds”[Mesh] OR “Polycyclic Compounds”[Mesh] OR “Hormones, Hormone Substitutes, and Hormone Antagonists”[Mesh] OR “Carbohydrates”[Mesh] OR “Lipids”[Mesh] OR “Amino Acids, Peptides, and Proteins”[Mesh] OR “Nucleic Acids, Nucleotides, and Nucleosides”[Mesh] OR “Biological Factors”[Mesh] OR “Biomedical and Dental Materials”[Mesh] OR “Pharmacologic Actions”[Mesh]) AND “Proteins”[Mesh] AND hasabstract[text] AND (“2007”[PDAT]: “2008”[PDAT]) AND “humans”[MeSH Terms] AND English[lang] AND (“agonists”[Subheading] OR “antagonists and inhibitors”[Subheading] OR “Gene Expression Regulation/ drug effects”[Mesh])
Gene-disorder	(“Diseases Category”[Mesh] AND “genetics”[Subheading] AND (hasabstract[text] AND (“2007”[PDAT]: “2008”[PDAT]) AND “humans”[MeSH Terms] AND English[lang]))
SNP-disorder	(“Polymorphism, Single Nucleotide”[Mesh] AND (hasabstract[text] AND (“2007”[PDAT]: “2008”[PDAT]) AND “humans”[MeSH Terms] AND English[lang]) AND “Diseases Category”[Mesh])

abstract with the entities under consideration. When the annotator opens the abstract for the first time the entities as identified by the NER system are already marked in the text. The annotator can modify or delete system annotations very easily. After the annotation is saved the system annotation is not available anymore. The web interface automatically computes per sentence a set of relations that according to the relation type can be derived. Again, the annotator can modify or delete these suggested relations. As soon as the user marks a new entity, the set of relations is updated.

3.7. Inter-annotation agreement

Since the annotations are performed based on background knowledge and also on the particular context of the text, measures to control for the variability of the annotations have to be considered. Clear and detailed annotation guidelines and the use of “text bound” annotation can reduce, but not eliminate, the variability between annotators. A measure of the inter-annotator agreement was used to estimate how the annotations differed between annotators and to look for strategies to get a consensus annotation. In order to maximize the agreement on the annotation of relationships, which is the main goal of the project, we allowed some flexibility on the annotation of entities (for instance by allowing variable boundaries on the annotation of entities).

To harmonize the different annotations, a simple majority-voting scheme was applied: if two of the three annotators agreed on a given annotation, it became part of the final EU-ADR corpus. If only one annotator gave an annotation, it was discarded. To allow for slightly differing entities marked by different annotators, we applied a matching scheme that regarded terms as a match if they shared at least one word rather than a scheme that requires terms to match exactly. For example, one annotator may have marked “diabetes” whereas the other marked “severe diabetes”. With exact match, this would be considered a disagreement, whereas one may argue that the annotators agreed on the term “diabetes”. A relationship was only included in the corpus if there was agreement both on the entities involved and on the type of the relationship (positive, negative, and speculative). Preliminary analysis of the inter-annotator agreement on the relationships

Fig. 1. An interactive web-based application allows the annotator to add new annotations or modify annotations suggested by the NER system in an easy way. The colored items indicate the entities that have been annotated, with each color representing a different type of entity. The annotations are saved at the server side per annotator when ‘save’ is selected and can be pulled up anytime. The relationships can be annotated as well with different association types. Note: the data here shown only serves as an example. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

indicated that there was not a clear decision regarding positive and speculative associations among the annotators. Thus, speculative and positive relationships were considered equivalent when assessing inter-annotator agreement. If relations were typified as speculative they were always annotated by other annotators as either speculative or positive. We assumed therefore that negative speculative relations are unlikely to occur in our corpus. Therefore we treated the speculative and positive relations as similar for computing agreement. As an example, the relation between the drug “RTX” and the disorder “pSS” in the following sentence “Controlled trials should be performed to confirm the efficacy of RTX in pSS. [PubMed, PMID: 16950808]” was annotated both as speculative and as positive.

4. Results

Based on all annotations of entities and relationships we can analyze the number of entities and relations for whom a majority exists and in consequence were included in the final EU-ADR corpus (see Table 2). Although the agreement on the relationships is low we have to take into account that a relation can only be included if both entities for which the relationship holds are agreed upon. The second percentage under the relationships agreement shows the agreement given that the annotators agreed on the two entities. Given that correction, it is evident that the clarity and complexity of how a relation is expressed depends on the type of entities involved in the relation and as a consequence impacts the agreement. Agreement has been computed as the fraction of all annotations on which the annotator is in agreement with the majority.

In order to test the agreement of each annotator with the EU-ADR corpus we computed both the agreement statistics for both the entities (Table 3) and the relations (Table 4). The agreement figures show a good correspondence between the different annotations. From the results we can see that apart from annotator A4 all annotators show a good agreement with the EU-ADR corpus.

In addition to comparing the annotations against the annotated corpus we also computed the inter-annotator agreement for each relationship (Table 5).

5. Discussion

The agreement statistics are comparable with what has been shown in other annotation efforts [13,17]. The agreement on the entity annotation is a little higher than on the relationships. One reason for this is that it may be difficult for annotators to distinguish between a relationship being described in the text and the relationship actually being true. During the training that preceded the annotation effort it was clear that some annotators had difficulty to distinguish these differences, however further clarifications were made to the guidelines following consensual

Table 3

Agreement between the annotators (A1–A5) and the automatic tool against the EU-ADR corpus for the annotated entities.

Relationship type	A1	A2	A3	A4	A5	NER system
Drug–disorder	0.83			0.78	0.87	0.73
Target–disorder	0.80	0.84	0.88			0.63
Target–drug	0.82	0.83	0.87			0.67
Overall	0.82	0.83	0.88	0.78	0.87	0.67

Table 4

Agreement between the annotators (A1–A5) and the automatic tool against the EU-ADR corpus for the annotated relationships.

Relationship type	A1	A2	A3	A4	A5	NER system
Drug–disorder	0.75			0.51	0.83	0.69
Target–disorder	0.85	0.72	0.71			0.57
Target–drug	0.77	0.79	0.50			0.79
Overall	0.80	0.75	0.62	0.51	0.83	0.66

Table 5

Per relationship type the inter-annotator (and NER system) agreement statistics.

	A1	A4	A5	NER system
<i>Drug–disorder</i>				
A1	1.00	0.78	0.72	0.59
A4	0.78	1.00	0.70	0.64
A5	0.72	0.70	1.00	0.56
NER system	0.59	0.64	0.64	1.00
<i>Target–disorder</i>				
A1	1.00	0.73	0.74	0.46
A2	0.73	1.00	0.75	0.49
A3	0.74	0.75	1.00	0.58
NER system	0.46	0.49	0.58	1.00
<i>Target–drug</i>				
A1	1.00	0.78	0.75	0.49
A2	0.78	1.00	0.74	0.52
A3	0.75	0.74	1.00	0.58
NER system	0.49	0.52	0.58	1.00

discussions. It can also be related to the lack of information provided in the abstract to assess the relationships.

Even though a named entity recognition system has been used to suggest annotations to the annotators the agreement between this system and the annotators is lower than the inter-annotator agreement. The inter-agreement is the fraction of all annotations where two annotators agree. This means that the annotators modified the suggested annotations and were consistent on suggestions for change. Nevertheless, in our experience the use of NER system is highly recommended to facilitate the annotation.

The web-based interface made it possible for the annotators to easily make annotations without the need to install any software.

Table 2

Number of annotated entities and relationships and their agreement in the EU-ADR corpus. For the relationships agreement the second percentage shows the agreement given agreement on the entities.

Relationship type	Entities	Total	Agreement entities	Relations	Agreement relations
Drug–disorder	Drugs	929	739 (79.5%)	668	300 (44.9%, 71.5%)
	Disorders	1029	812 (78.9%)		
	Total	1958	1551 (79.2%)		
Target–disorder	Targets	1664	1177 (70.7%)	941	424 (45.1%, 86.3%)
	Disorders	964	724 (75.1%)		
	Total	2628	1901 (72.3%)		
Target–drug	Target	1601	1241 (77.5%)	827	313 (37.8%, 66.7%)
	Drugs	824	587 (71.2%)		
	Total	2425	1828 (75.3%)		
Overall		7011	5280 (75.3%)	2436	1037 (42.5%, 74.7%)

The web-based interface kept track of the progress and stored the annotations at a secure central server based at the Erasmus University Medical Center.

6. Conclusion

We have developed a corpus of 300 abstracts with drugs, disorders and targets and their inter-relationships annotated. The annotators achieved a consistently high agreement and improved on the computer annotation that was provided as a first annotation. We foresee that the availability of such a corpus will be useful for the development and testing of text-mining tools.

For evaluating and training text-mining systems on an annotated corpus it is essential to know the agreement on the annotations. It is therefore essential to provide with each annotated corpus data on the inter-annotator agreement.

7. Availability

The EU-ADR corpus can be downloaded from: http://euadr.erasmusmc.nl/sda/euadr_corpus.tgz

The annotation tool is available online at the following URL: <http://euadr.erasmusmc.nl/sda/annotate.py>

Acknowledgments

This research received funding from the European Union Community in the framework of the FP7/2007–2013 convention-governing subsidy no. 215847 – the EU-ADR project, the Innovative Medicines Initiative [eTOX,115002], and the Instituto de Salud Carlos III FEDER (CP10/00524). The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB) and a member of the COMBIOMED network.

References

- [1] Yeh AS, Hirschman L, Morgan AA. Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics* 2003;19(1):331–9.
- [2] Loging W, Lee Harland L, Williams-Jones B. High-throughput electronic biology: mining information for drug discovery. *Nat Rev Drug Discov* 2007;6:220–30.
- [3] van Haagen HHHBM, 't Hoen PAC, Botelho Bovo A, de Morrée A, van Mulligen EM, et al. Novel protein–protein interactions inferred from literature context. *PLoS One* 2009;4(11):e7894.
- [4] Coloma PM, Schuemie MJ, Trifirò G, Gini R, Herings R, Hippisley-Cox J, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Saf* 2011;20(1):1–11.
- [5] Tateisi Y, Yuka, Yakushiji A, Ohta R, Tsujii J. Syntax annotation for the GENIA corpus. In: *Proceedings of the IJCNLP 2005, Companion volume*. Jeju Island, Korea; October, 2005. p. 222–7.
- [6] Kim J-D, Ohta T, Tateisi Y, Tsujii J. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics* 2003;19(1):180–2.
- [7] Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreative IV: critical assessment of information extraction for biology. *BMC Bioinformatics* 2005;6(1).
- [8] Smith L et al. Overview of BioCreative II gene mention recognition. *Genome Biol* 2008;9:S2.
- [9] Mandel MA. Integrated annotation of biomedical text: creating the PennBioE corpus. In: *Proceedings of the workshop on text mining, ontologies and natural language processing in biomedicine*, 20–21 March, 2006, Manchester, UK; 2006.
- [10] Garten Y, Coulet A, Altman RB. Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics* 2010;11(10):1467–89.
- [11] Garten Y, Tatonetti NP, Altman RB. Improving the prediction of pharmacogenes using text-derived drug–gene relationships. *Pac Symp Biocomput* 2010;15:305–14.
- [12] Rosario B, Hearst MA. Classifying semantic relations in bioscience text. In: *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL 2004)*, Barcelona; July, 2004.
- [13] Kolářik C, Klinger R, Friedrich CM, Hofmann-Apitius M, Fluck J. Chemical names: terminological resources and corpora annotation. In: *Proceedings of the LREC 2008 workshop on building and evaluating resources for biomedical text mining*; 2008.
- [14] Leaman R, Miller C, Gonzalez G. Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. In: *Handbook of the 3rd international symposium on languages in biology and medicine*; 2009.
- [15] Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac Symp Biocomput* 2000:517–28.
- [16] Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. In: *Proceedings of the seventh international conference on intelligent systems for molecular biology*. Heidelberg (Germany): AAAI Press; 1999. p. 77–86.
- [17] Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Setzer A. Semantic annotation of clinical text: the CLEF corpus. In: *Proceedings of the LREC 2008 workshop on building and evaluating resources for biomedical text mining*; 2008. p. 19–26.
- [18] Kim J-D, Ohta T, Tsujii J. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* 2008;9:10.
- [19] Schuemie M, Jelier R, Kors JA. Peregrine: lightweight gene name normalization by dictionary lookup. In: *Proceedings of the Biocreative 2 workshop 2007*, April 23–25, Madrid; 2007. p. 131–40.
- [20] http://bioie ldc.upenn.edu/wiki/index.php/Main_Page.
- [21] <http://www.knewco.com>.