REVIEW

An analysis on the entity annotations in biological corpora [v1; ref status: indexed, http://f1000r.es/2o0]

Mariana Neves

Hasso-Plattner-Institut, Potsdam Universität, Potsdam, Germany

v1 **First published:** 25 Apr 2014, **3**:96 (doi: 10.12688/f1000research.3216.1)
Latest published: 25 Apr 2014, **3**:96 (doi: 10.12688/f1000research.3216.1)

Abstract

Collection of documents annotated with semantic entities and relationships are crucial resources to support development and evaluation of text mining solutions for the biomedical domain. Here I present an overview of 36 corpora and show an analysis on the semantic annotations they contain. Annotations for entity types were classified into six semantic groups and an overview on the semantic entities which can be found in each corpus is shown. Results show that while some semantic entities, such as genes, proteins and chemicals are consistently annotated in many collections, corpora available for diseases, variations and mutations are still few, in spite of their importance in the biological domain.

Open Peer Review

Invited Referee Responses

	1	2
version 1 published 25 Apr 2014	☑ report	☑ report

1 **Roman Klinger**, Bielefeld University Germany

2 **Paloma Martínez**, Charles III University of Madrid Spain

Latest Comments

No Comments Yet

Corresponding author: Mariana Neves (marianalaraneves@gmail.com)

How to cite this article: Neves M. **An analysis on the entity annotations in biological corpora [v1; ref status: indexed, http://f1000r.es/2o0]** *F1000Research* 2014, **3**:96 (doi: 10.12688/f1000research.3216.1)

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing interests: The author declares no competing interests.

First published: 25 Apr 2014, **3**:96 (doi: 10.12688/f1000research.3216.1)
First indexed: 14 Jul 2014, **3**:96 (doi: 10.12688/f1000research.3216.1)

Introduction

Annotated collections of documents are crucial components for developing new methods in text mining, such as extraction of named entities and relationships from the scientific literature. This lies in the fact that supervised learning systems need to rely on annotated documents to train the algorithms, and therefore, "learn" how to efficiently perform a certain task. Additionally, use of a standard collection of documents is practically the only way of performing an unbiased comparison between different systems for a particular task.

In natural language processing (NLP), a corpus can be defined as a collection of documents which usually belongs to a particular topic and that has been annotated according to a pre-defined schema. When annotating semantic information, a schema is usually composed of some entities (e.g., genes, proteins), and optionally, relationships (e.g., protein-protein interactions, gene-disease relationships). The number of documents may vary from a couple of full text documents[1,2], to hundreds of abstracts[3] or thousands of sentences[4].

A schema can be composed of an arbitrary list of annotation types or based on terms pertaining to one or more ontologies. For example, the Variome[1] and the CellFinder[2] corpora contain annotations for a pre-defined list of entities, such as genes/proteins, cell lines, diseases and mutations. On the other hand, the CRAFT corpus[5] includes annotations according to concepts in seven ontologies and terminologies to allow a better identification of the annotations and their interoperability with other biomedical resources. The annotation schema is usually part of a comprehensive guideline document in which more details of the annotation process are described, such as an overview of the concepts, the provenance of the documents and examples of situations where the annotation should (or should not) be carried out.

Corpora are usually constructed for training or evaluation purposes during the development of a particular system (e.g., Gerner *et al.*[6]) but are often also created in the context of a challenge or shared task (e.g., Krallinger *et al.*[7]) to foster improvements on a particular task and allow comparison between different solutions. Corpora are usually manually annotated by human experts in a particular field or automatically derived using NLP techniques. When manually constructed by one or more annotators, it receives the denomination of a gold-standard corpus. In this process, annotators are required to carefully read the texts and manually annotate the text according to the pre-defined schema. This annotation process is usually supported by an annotation tool, such as Brat[8] or Knowtator[9], which provides a nice graphical user interface and ways to previously configure the annotation schema. A comprehensive survey of the annotation tools for the biomedical domain can be found in Neves *et al.*[10]. A good approach on corpus construction should include training for the particular annotation tool and the guidelines, an inter-annotator agreement and the construction of a consensus corpus derived from the later.

Frequently, manual annotation is supported by text mining by providing automatically extracted annotations which are later validated by the annotators. This validation process should not only include checking the annotations which were automatically extracted by the text mining tools but also carefully reading the text to identify missing ones. Hybrid corpora in which part of the annotations correspond to non-validated automatic annotations and then manually annotated with others, such as relationships, can also be found. For instance, for the Drug-Drug Interaction corpus[11], drugs were automatically annotated using the Metamap tool[12] followed by the manual annotation of relationships by experts.

Finally, corpora can also be completely derived from automatized methods and never manually validated by experts, the so-called silver-standard corpora. Despite the undeniable presence of wrong annotations and the absence of many others, previous works have demonstrated that these corpora can support development of semi-supervised or distant supervised systems for named-entity[13] and relationship extraction[14]. As manual annotation or validation is not required in this case, such corpora tends to be much larger than the gold-standard ones. CALBC[15] is an example of a silver-standard corpus derived from a community-based project which intended to automatically harmonize annotations generated from a variety of named-entity recognition tools.

In this work, I present a review on 36 corpora which are available for the biomedical natural language processing (BioNLP) domain and perform an analysis on the semantic types which they include. The motivation for this work is to provide the first comprehensive overview on BioNLP corpora and thus support choosing the most appropriate collection whenever necessary. Additionally, I show the impact of each corpora in the field and give insights for the construction of new corpora or for the extension of existing ones.

Corpora and semantic types

List of corpora

Here I present a comprehensive study on the semantic entities included in the gold-standard corpora which have been annotated for the named-entity recognition (NER), relationship extraction and event extraction tasks. Although there are corpora available for other BioNLP tasks, such as text classification[16] and question answering[17], these are not covered in this survey. I focus on gold-standard corpora which contain annotations for entity types, such as genes/proteins, chemicals and species. Thus, I also did not include corpora which have only text span annotations not related to a particular semantic entity, such as the Data Deposition corpus[18] which contains annotations on data deposition statements. Given the focus on Biology, I did not consider corpora which were built with the medical domain in mind, such as BioText[19] and Variome[1]. Silver-standard corpora, such as CALBC[15], were also not included here. I also do not cover corpora which focused on the linguistic aspects rather on semantic annotations, such as the BioScope corpus[20], which contains annotations for negations and speculations statements. Finally, only corpora which are still available for download were included.

In this section, I give an overview of 36 corpora made available on the BioNLP domain and describe how the semantic analysis of the corpora has taken place.

List of biological corpora

Here I present the list of 36 corpora which have been considered in this study. For each of them, I include a brief description of its origin, which may include the type of documents it contains (abstracts and full texts), its annotation schema, tools which have been based on it, further extensions it has received and the number of citations its publications have received. Table 1 shows a summary of these corpora, including their first publications, year of release (according to the main publication), number of citations according to Google Scholar (as December-2013) and the corresponding URL. Some of the corpora I present here are included in the WBI repository (http://corpora.informatik.hu-berlin.de), which provides their full visualization using the Stav/Brat annotation tool[8]. The collections are presented in the alphabetical order.

Table 1. Overview of the corpora: main publication, year of publication, citations in Google Scholar (as December-2013) and the URL are shown for each corpus.

Corpus	Ref.	Year	Cit.	URL
AIMed	[21]	2005	270	ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/
AnEM	[22]	2012	9	http://www.nactem.ac.uk/anatomy/
AZDC	[23]	2009	19	http://diego.asu.edu/downloads/AZDC
Bact. Gene Int.	[24]	2012	11	https://sites.google.com/site/bionlpst/home/bacteria-gene-interactions
BioCreative GM	[4]	2008	126	http://biocreative.sourceforge.net/biocreative_2_gm.html
BioInfer	[25]	2007	246	http://mars.cs.utu.fi/BioInfer
CellFinder	[2]	2012	5	http://cellfinder.de/about/annotation/
CG	[26]	2013	3	http://2013.bionlp-st.org/tasks/cancer-genetics
CHEMDNER	[7]	2013	7	http://www.biocreative.org/tasks/biocreative-iv/chemdner/
CRAFT	[5]	2012	17	http://bionlp-corpora.sourceforge.net/CRAFT/
Craven	[27]	1999	374	http://www.biostat.wisc.edu/~craven/ie/
DDI	[28]	2013	0	http://labda.inf.uc3m.es/ddicorpus
EBI Disease	[29]	2008	66	ftp://ftp.ebi.ac.uk/pub/software/textmining/corpora/diseases
EDGAR	[30]	2000	395	ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe/EDGAR_GS.txt
EPI	[31]	2012	14	https://sites.google.com/site/bionlpst/home/epigenetics-and-post-translational-modifications
EU-ADR	[32]	2012	4	http://euadr.erasmusmc.nl/sda/euadr_corpus.tgz
GeneReg	[33]	2010	11	http://www.julielab.de/Resources/GeneReg.html
Genia	[3]	2003	575	http://www.nactem.ac.uk/aNT/genia.html
Genia Ev. Extr.	[34]	2008	236	http://bionlp.dbcls.jp/redmine/projects/bionlp-st-ge-2013/wiki/Wiki
GETM	[35]	2010	13	http://getm-project.sourceforge.net/
GREC	[36]	2009	53	http://www.nactem.ac.uk/GREC/
HPRD50	[37]	2007	268	http://www2.bio.ifi.lmu.de/publications/RelEx/
ID	[31]	2012	14	https://sites.google.com/site/bionlpst/home/infectious-diseases
IEPA	[38]	2002	208	http://orbit.nlm.nih.gov/resource/iepa-corpus
Linnaeus	[6]	2010	79	http://linnaeus.sourceforge.net/
LLL	[39]	2005	163	http://genome.jouy.inra.fr/texte/LLLchallenge/
Metab. Enzym.	[40]	2011	14	http://www.nactem.ac.uk/metabolite-corpus/
MutationFinder	[41]	2007	83	http://mutationfinder.sourceforge.net/
Nagel	[42]	2009	12	http://sourceforge.net/projects/bionlp-corpora/files/ProteinResidue/
NCBI Disease	[43]	2012	10	http://www.ncbi.nlm.nih.gov/CBBresearch/Fellows/Dogan/disease.html
OSIRIS	[44]	2008	20	https://sites.google.com/site/laurafurlongweb/databases-and-tools/corpora
PC	[45]	2013	4	http://2013.bionlp-st.org/tasks/pathway-curation
PICAD	[46]	2011	1	http://stat.fsu.edu/~jinfeng/resources/PICAD.txt
SCAI	[47]	2008	57	http://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/research-development/information-extraction-semantic-text-analysis/named-entity-recognition/chem-corpora.html
SNPCorpus	[48]	2011	3	http://www.scai.fraunhofer.de/snp-normalization-corpus.html
Species	[49]	2013	1	http://species.jensenlab.org/

AIMed. The AIMed corpus[21] contains annotation on proteins and protein-protein interactions (PPI) for 200 abstracts, which were selected from the documents for which curated annotations were found in the Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu/dip/Main.cgi). The corpus is one of the five corpora widely used for the development of PPI extraction methods[50] and thus, has been used for the development of a variety of PPI tools[51].

AnEM. The recently published AnEM corpus[22] contains a total of 500 documents which contains annotations on the following anatomical entity types: organism subdivision, anatomical system, organ, multi-tissue structure, tissue, cell, developing anatomical structure, cellular component, organism substance, immaterial anatomical entity and pathological formation. It is probably the largest manually annotated corpus on anatomical entities and has been used for the development of the AnatomyTagger tool[52].

AZDC. The AZDC corpus[23] contains almost 800 abstracts which includes the ones available in the EBI disease corpus (cf. below) and some from the Craven corpus (cf. also below). It contains annotations for diseases and normalization to UMLS unique concepts for some semantic subtypes and was used for the development of named-entity recognition tools for disease names, such as the recent DNorm system[53].

Bacteria Gene Interaction. The Bacteria Gene Interaction (BGI) corpus[24] was developed in the scope of the BioNLP Event Extraction Shared Tasks 2011 for assessing the extraction of genetic processes in *Bacillus subtilis*. It is derived from the LLL corpus (cf. below) for PPIs. This corpus has been extended for the Gene Regulation Network (GRN) task[54] in the 2013 edition of the same challenge.

BioCreative 2 Gene Mention. The BioCreative 2 Gene Mention[4] corpus has been used in two editions of the BioCreative challenges (http://www.biocreative.org/) to foster improvements for gene/protein extraction. It is composed of sentences, opposed to documents, which were derived from Medline documents and contains annotation on gene and protein, though it does not make distinction between them. Given that it has been used in one of the most popular challenges in the BioNLP community, several studies have used this corpus for the development of gene/protein extraction systems, such as BANNER[55].

BioInfer. BioInfer[25] is also one of the five popular corpora available for PPI[50]. It contains sentences derived from more than 800 documents and annotations are available for genes, DNA families or groups, proteins, protein complexes and protein families and groups. Just as the other five PPI corpora, the BioInfer corpus has been used for training and evaluation of several tools[51].

CellFinder. The CellFinder corpus[2] was developed in the scope of the CellFinder database (http://cellfinder.de/) and includes annotations for six entity types (anatomical parts, cell lines, cell types, species and cell components) for 10 full text documents in the stem cell research field. This corpus has been mainly used for the evaluation of named-entity recognition approaches for the above entity types in Neves *et al.*[2,56].

Cancer Genetics. The Cancer Genetics (CG) corpus[26] was constructed for the Cancer Genetics task in the BioNLP Event Extraction Shared Task in 2013 and includes annotations on the development and progress of cancer. The corpus is composed of 600 abstracts split into three datasets and events are composed of anatomical and molecular entities, as well as annotations for organisms.

CHEMDNER. The CHEMDNER corpus[7] has been recently created in the scope of the CHEMDNER task in BioCreative IV for assessing performance of named-entity recognition tools for chemical compounds. It contains 10,000 abstracts split into training, development and test datasets and annotations for chemicals are classified in eight categories, such as systematic, formula or abbreviation.

CRAFT. The CRAFT corpus[5] is a recent and very comprehensive collection of 97 full text documents which has been annotated with concepts, such as gene/proteins, species, cells and chemicals, from nine ontologies and terminologies. The authors have reserved 30 of the full texts for a text mining challenge that is going to be carried out in the near future.

Craven. The so-called Craven corpus[27] is in fact a collection of three corpora which contains annotations on sub-cellular locations, PPIs and gene-disease associations. These corpora have been used for the development of methods for extracting the above binary relationships and support construction of knowledge bases.

Drug-Drug Interaction. The Drug-Drug Interaction (DDI) corpus[28] includes more than 700 documents derived from Medline and DrugBank, and includes annotations for drugs and binary relationships between them. It has been already evaluated on two shared tasks[11,57] and thus, has been extensively used for both training and evaluation for NER and relatiosnhip extarction tasks.

EBI Disease. The EBI Disease corpus[29] is composed of 600 sentences selected from the Craven corpus (cf. above) which have been extended with associations to unique concepts in the UMLS terminologies.

EDGAR. The EDGAR corpus[30] contains annotations for genes, drugs and cells, including binary relationships between genes and drugs, genes and cells, and drugs and cells.

Epigenetics and Post-translational Modifications. The Epigenetics and Post-translational Modifications (EPI) corpus[31] was developed for the BioNLP Event Extraction Shared Task 2011 and contains 1,200 abstracts annotated with events related to epigenetic changes. Just like the Genia Event Extraction corpus (cf. below), it contains annotations for genes/proteins and annotations identified as “Entity” which might refer to a variety of entity types, such as cell locations or small molecules.

EU-ADR. The EU-ADR corpus[32] was constructed in the scope of the EU-ADR project, which aimed to automatically process health records. The corpus contains a total of 300 abstracts which are split into three groups, each containing annotations for two entity types and binary relationships: drug-target, drug-disease and target-disease.

GeneReg. The GeneReg corpus[33] is composed of 314 abstracts related to *Escherichia coli* and contains annotations of events for gene expression regulation. It has been created in order to allow its interoperability with the Genia corpus (cf. below) and other lexical resources, such as WordNet and the Specialist lexicon.

Genia. The Genia corpus[3] is probably one of the most popular corpora in the biomedical domain and has been used for the development of many named-entity tools, such as ABNER[58], and also to assess systems in a shared task[59]. It contains 2,000 Medline abstracts with annotations based on the Genia ontology for DNA, RNA, proteins, lipids, cells, tissues, body parts and cell lines, among others.

Genia Event Extraction corpora. The Genia Event Extraction (Genia EE) corpus[34] has started from the annotation of 1,000 abstracts, half of the Genia corpus (cf. above), and was annotated with genes/proteins and biological events, such as gene expression and gene regulations. This version of the corpus was used for the BioNLP Event Extraction Shared Task which took place in 2009[60] and then extended with 15 full texts for the following edition of the challenge that took place in 2011[61]. A new corpus composed of 34 full texts was constructed for the third edition of the shared task that took place in 2013[62]. The corpora have been used for the development and comparison of a variety of systems for extracting events.

GETM. The GETM corpus[35] is composed of 150 abstracts derived from the development dataset of the Genia Event Extraction corpus (cf. above). Relationships were annotated between the gene expression events and the annotations for cells and anatomical locations which were present in the original corpus. It was used for the evaluation of a rule-based relationship extraction system on gene expression events in cell locations.

GREC. The GREC corpus[36] contains annotations for 240 Medline abstracts for events on gene regulation and expression related to ontologies, such as Gene Ontology and Sequence Ontology.

HPRD50. The HPRD50 corpus[37] has been created in the scope of the RelEx system and contains 50 abstracts and annotations for PPIs. The corpus is also one of the five PPI corpora[50] and has been used for the development of a variety of PPI tools[51].

ID. The ID corpus[31] was developed for the BioNLP Event Extraction Shared Task 2011 and contains 30 full text documents annotated with biomolecular mechanisms of infectious diseases. The corpus is split into three datasets (training, development and testing) and events are related to annotations of proteins, chemicals and organisms.

IEPA. The IEPA corpus[38] is composed of more than 200 sentences extracted from Medline abstracts and is annotated with binary relationships between proteins. It is also one of the five popular corpora available for PPI[50].

Linnaeus. The Linnaeus corpus[6] contains 100 full text documents annotated with annotations for organisms, all linked to identifiers in NCBI taxonomy (http://www.ncbi.nlm.nih.gov/taxonomy). It was built for the development of the Linnaeus system, one of the state-of-art tools for the annotation of organism names.

LLL. The LLL corpus[39] for PPI in *Bacillus subtilis* was release in the scope of the Learning Language in Logic (LLL) shared task and was later also included in the package of the five popular corpora available for PPIs[50]. The proteins are identified as agent or target in the relationships.

Metabolites and Enzymes. The Metabolites and Enzymes corpus[40] contains annotations for metabolites and enzymes names in almost 300 abstracts and was used for the evaluation of dictionary-based approaches for the recognition of these entity types.

MutationFinder. The MutationFinder corpus[41] is composed of 508 Medline abstracts annotated with mutations and it was used for the evaluation of the homonymous tool based on regular expression techniques.

Nagel. The Nagel corpus[42] contains annotations for protein residues, species and mutations in 100 Medline abstracts which have been used for the evaluation of a system developed for the extraction of these triplets.

NCBI Disease. The NCBI Disease corpus[43] is composed of almost 800 abstracts derived from the AZDC corpus (cf. above) split into three datasets for training, development and blind testing. Annotations are classified into categories, such as modifier and specific disease, and it has been used for the development of the DNorm tool[53].

OSIRIS. The OSIRIS corpus[44] contains abstracts annotated with genes and sequence variants and was used for the evaluation of a dictionary-based system developed for the extraction of the later. Annotations for genes are normalized to identifiers from the NCBI EntrezGene database (http://www.ncbi.nlm.nih.gov/gene).

Pathway Curation. The Pathway Curation (PC) corpus[45] was created for the homonymous task in the BioNLP Event Extraction Shared Task 2013 in which participants were required to extract biomolecular events to support curation of pathways. It includes a total of 525 abstracts annotated with events which contain chemicals, gene, proteins, complexes and cellular components as arguments.

PICAD. The PICAD corpus[46] is another less popular PPI corpus composed of more that 1,000 sentences which were assembled in the scope of the development of a tool for this purpose.

SCAI. The SCAI corpus[47] includes 100 abstracts with annotations for chemicals and training and test datasets for the recognition of IUPAC names. This has been one of the most popular corpora for chemical named-entity recognition and has been used for the development of many tools, such as ChemSpot[63].

SNPCorpus. The SNPCorpus[48] contains almost 300 abstracts and annotations for protein sequence and nucleotide sequence mutations and it has been used by the authors for extraction of these mentions from the text and their association to identifiers in biological databases.

Species. The Species corpus[49] has been recently built as an alternative to Linnaeus (cf. above). Instead of using full text documents, it aimed at providing more variability on the species names by using eight groups of 100 abstracts on the following categories: bacteriology, botany, entomology, medicine, mycology, protistology, virology, and zoology.

Semantic analysis of corpora

In this section, I show an analysis of the semantic types of the annotations present in the corpora discussed above. This analysis has been carried out based on the publications associated with the corpora and sometimes by checking the annotation types for the corpora which are available at the WBI Corpora repository. Here I only consider those annotations which are meaningful enough to be associated with one of the pre-defined semantic types under consideration (cf. below). For instance, I do not consider the "Entity" annotations in the Genia Event Extraction corpus[60].

Six top level semantic types were decided based on the annotations available in the corpora and on the UMLS semantic types (http://semanticnetwork.nlm.nih.gov/SemGroups/SemGroups.txt). The following are the types along with their mapping to the UMLS sematic groups and types:

- gene/protein: semantic group "Genes & Molecular Sequences" (GENE), as well as the types T116 (Amino Acid, Peptide, or Protein) and T114 (Nucleic Acid, Nucleoside, or Nucleotide);
- variant/mutation: semantic type T045 (Genetic Function);
- drug/chemical: semantic group "Chemicals & Drugs" (CHEM), except for the types T116 and T114 which were considered gene/proteins (cf. above);
- cell/anatomy: sematic group "Anatomy" (ANAT);
- disease: semantic group "Disorders" (DISO);
- organisms: semantic group "Living beings" (LIVB).

The gene/protein category covers a wide range of small molecules and includes gene, proteins, protein complexes, gene complexes, protein families/groups, RNA, DNA families/groups, regulons, etc. Most of the corpora which include these entities do not make a distinction between them, such as the BioCreative Gene Mention[4]. In the cell/anatomy semantic type, I include all kinds of cellular and anatomical locations, whether *in vivo* or *in vitro*, as follows: cell lines, cell types, cell components, sub-cellular locations, developing anatomical structures, anatomical systems, organs and tissues. Drugs and chemicals were put together in the same group as some corpora include both of them, although these are sometimes classified into categories. Variants and mutations were assembled in a single group and, finally, one category for diseases and one for species, which are more homogeneous groups and whose annotations are not usually classified in distinct categories in corpora.

Comparison and discussion

In this section I present an analysis of the semantic types for the named-entities present in 36 corpora. Figure 1 shows an overview of which annotations are available for each corpora, as well as which corpora contain annotations for a particular semantic type. It also gives an idea of the similarities between corpora in terms of the entity types they share.

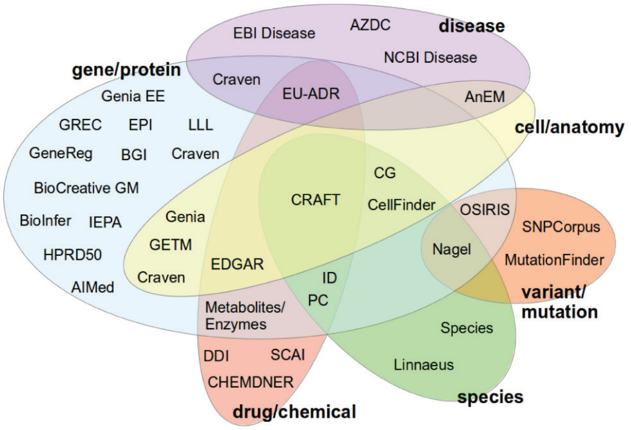

Figure 1. Classification of the corpora according to the semantic annotations they contain.

The closer a corpus is to the center of the figure, more distinct semantic types it contains. The CRAFT corpus is the collection which contains the higher number of semantic entities, namely, gene/proteins, species, chemicals and cells, but still lacks annotations for disease, variants/mutations and anatomical parts. The Cancer Genetics (CG), CellFinder, EDGAR, EU-ADR, Infectious Disease (ID), Pathway Curation (PC) and Nagel corpora are the ones which come closer to the CRAFT corpus, each containing annotations for three different types, but with a great variability on which of these three types are considered.

On the other hand, the farther a corpus is to the center of the figure, less distinct semantic types it contains and most of the corpora fall into this situation. There are 12 corpora which contain only annotations for gene/proteins, three for diseases, two for variants or mutations, two for species or organisms and three for chemical or drugs. Curiously, no corpus contains annotations only for cell anatomical entities, except the AnEM corpus, which was also placed in the disease semantic types because it contains annotations on pathological formations.

Genes and proteins are the most popular entities in biomedical corpora: in a total of 26 collections. However, these have different purpose and number of documents. Early initiatives, such as the BioCreative Gene Mention corpus, were based on sentences instead of documents, but following corpora have annotated the abstracts instead. Recently developed corpora include annotation of full text documents, such as CellFinder, CRAFT and the Genia Event Extraction, in order to allow systems to make use of the complexities of the languages which can only be found in the full text but not in the abstracts[64]. Most of the corpora classified in this group make no distinction between genes, proteins, complexes, or families, except for Genia and the Bacteria Gene Interaction corpora. Corpora whose annotations are mapped to identifiers in a database, e.g., EntrezGene, such as CRAFT and OSIRIS, allow their use for the development of gene/protein normalization tools[65]. Finally, the high number of corpora available for gene/protein corpora is due to the importance of these entities for the molecular biology domain and to the research in the last years on PPIs and biological events.

Corpora which contained annotations for chemicals and drugs were few until the release of the SCAI corpus, which focused initially on the IUPAC nomenclature. But this has become a hot topic in the last couple of years and following corpora have provided annotations also for drugs and their interactions (DDI corpus), as well as anotations on full text documents (CRAFT corpus). The CHEMDNER corpus classifies chemicals in some predefined categories and was used in the one of the shared tasks in last BioCreative challenge, which attracted the participation of many teams. Relationships of chemical compounds with other semantic entities can be found in the EU-ADR and also for more complex events, such as in the shared tasks of Cancer Genetics and the Infectious Disease in the BioNLP Event Extraction Shared Tasks.

During many years, the Linneaus corpus and tool have been the state-of-art resources for benchmarking and extraction of species annotations, respectively. The simplicity of the nomenclature and the high performance of Linnaeus has not encouraged further research in this line. However, the release of the Species corpus some months ago aims to provide more variety on the annotations for organism, by choosing a higher number of abstracts, as opposed to few full text documents in the Linnaeus corpus. Additionally, abstracts are grouped on eight categories of organisms (bacteriology, botany, entomology, medicine, mycology, protistology, virology, and zoology), thus, ensuring the diversity of annotations. Other recent full text corpora which contains annotations for organisms are the CellFinder and CRAFT corpora.

Annotations for cell and anatomical parts have since many years been limited to the cell lines and cell types in the Genia and EDGAR corpora. However, the recent release of the AnEM corpus, which include a careful classification of these entities based on many ontologies, along with the AnatomyTagger tool[52], will certainly encourage new solutions in this area. Other recent corpora for cell annotations are the full text documents of the CRAFT corpus, including mapping to the Cell Ontology, as well as the annotations for cell lines, cells types and anatomical parts in the CellFinder corpus.

Most corpora which contain annotations for diseases exclusively are somehow related to each other as all of them contain documents which have been selected from the AZDC and the Craven corpora. The recent release NCBI Disease corpora aims to improve research in this field by classifying mentions based on some pre-defined categories, followed by the release of the DNorm tool[53]. Associations of diseases with other entity types are still scarce and only present in the EU-ADR and Craven corpora.

Finally, variations and mutations have also received little attention from the BioNLP community, and the four available corpora are composed only of abstracts. Co-occurrence of these entities in the text is available for genes in the OSIRIS corpus, however, no explicit relationships was annotated between them. Such relationships are only available with genes and species in the Nagel corpus, but its small size (100 abstracts) hinders text mining solutions based on machine learning methods, being only suitable for evaluation purposes.

From Figure 1, it is straightforward to observe which corpora are available according to the entity types of interest. The aim of this study is to encourage the use of less popular corpora which are already available and whose suitability for the text mining tasks has been scientifically evaluated. However, when choosing to use more than one corpora, the text miners will probably need to deal with more than one format for the documents and annotations, and write specific parsers for each of them. This is a problem that the BioC initiative[66] is aiming to solve with the recent introduction of the BioC XML format. Indeed, many of the corpora shown here have already been converted to this format using the Brat2BioC tool[67] and made available in the WBI Corpora repository. Given that most of the corpora are available under a flexible license, this review will also serve as a starting point for further updates on the repository and allow not only their availability for visualization but also for download in the BioC format.

Conclusions

In this survey I presented an overview on the semantic entity types available for 36 corpora in the biomedical domain. The annotations were classified in six categories (gene/protein, drug/chemical, cell/anatomy, variant/mutation, species and disease) and an overview on which corpora contain each of these semantic types has been shown. I hope that this review can be of help when choosing the best corpora for developing a named entity recognition tool and also to encourage re-use (re-annotation) of existing corpora instead of building a new one.

Competing interests

The author declares no competing interests.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgements

I would like to thank Prof. Ulf Leser for hosting the WBI Corpora repository and Dr. Antonio Jimeno Yepes for discussions on the UMLS semantic types.

References

1. Verspoor K, Jimeno Yepes A, Cavedon L, *et al.*: **Annotating the biomedical literature for the human variome.** *Database (Oxford).* 2013; **2013**: bat019.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Neves M, Damaschun A, Kurtz A, *et al.*: **Annotating and evaluating text for stem cell research.** In *Proceedings of the Third Workshop on Building and Evaluation Resources for Biomedical Text Mining (BioTxtM 2012) at Language Resources and Evaluation (LREC),* Istanbul, Turkey. 2012; 16–23.
Reference Source

3. Kim JD, Ohta T, Tateisi Y, *et al.*: **GENIA corpus--semantically annotated corpus for bio-textmining.** *Bioinformatics.* 2003; **19**(Suppl 1): i180–2.
PubMed Abstract | Publisher Full Text

4. Smith L, Tanabe LK, Ando RJ, *et al.*: **Overview of BioCreative II gene mention recognition.** *Genome Biol.* 2008; **9**(Suppl 2): S2.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Bada M, Eckert M, Evans D, *et al.*: **Concept annotation in the CRAFT corpus.** *BMC Bioinformatics.* 2012; **13**: 161.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Gerner M, Nenadic G, Bergman CM: **LINNAEUS: a species name identification system for biomedical literature.** *BMC Bioinformatics.* 2010; **11**: 85.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Krallinger M, Leitner F, Rabal O, *et al.*: **Overview of the chemical compound and drug name recognition (chemdner) task.** In *BioCreative IV workshop.* 2013; **2**: 2–33.
Reference Source

8. Stenetorp P, Pyysalo S, Topić G, *et al.*: **brat: a webbased tool for nlp-assisted text annotation.** In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France. Association for Computational Linguistics. April 2012. 102–107.
Reference Source

9. Ogren PV: **Knowtator: a protégé plug-in for annotated corpus construction.** In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Morristown, NJ, USA.* 2006; 273–275.
Publisher Full Text

10. Neves M, Leser U: **A survey on annotation tools for the biomedical literature.** *Brief Bioinform.* 2014; **15**(2): 327–40.
PubMed Abstract | Publisher Full Text

11. Segura-Bedmar I, Martinez P, Sanchez-Cisneros D: **The 1st ddiextraction-2011 challenge task: Extraction of drug drug interactions from biomedical texts.** In *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction 2011.* 2011; 1–9.
Reference Source

12. Aronson AR, Lang FM: **An overview of MetaMap: historical perspective and recent advances.** *J Am Med Inform Assoc.* 2010; **17**(3): 229–236.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Ando RK: **Biocreative ii gene mention tagging system at ibm watson.** In *BioCreative 2 workshop.* 2007.
Reference Source

14. Thomas P, Bobic T, Leser U, *et al.*: **Weakly labeled corpora as silver standard for drug-drug and protein-protein interaction.** In *Proceedings of the Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM) on Language Resources and Evaluation Conference (LREC).* Instanbul, Turkey, May . European Language Resources Association (ELRA). 63–70.
Reference Source

15. Rebholz-Schuhmann D, Jimeno Yepes AJ, Van Mulligen EM: **CALBC silver standard corpus.** *J Bioinform Comput Biol.* 2010; **8**(1): 163–179.
PubMed Abstract | Publisher Full Text

16. Krallinger M, Leitner F, Rodriguez-Penagos C, *et al.*: **Overview of the protein-protein interaction annotation extraction task of BioCreative II.** *Genome Biol.* 2008; **9**(Suppl 2): S4.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Morante R, Krallinger M, Valencia A, *et al.*: **Machine reading of biomedical texts about alzheimer's disease.** In *CLEF (Online Working Notes/Labs/Workshop).* 2012.
Reference Source

18. Névéol A, Wilbur WJ, Lu Z: **Extraction of data deposition statements from the literature: a method for automatically tracking research results.** *Bioinformatics.* 2011; **27**(23): 3306–3312.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Rosario B, Hearst MA**: Classifying semantic relations in bioscience texts.** In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.* 2004.
Publisher Full Text

20. Vincze V, Szarvas G, Farkas R, *et al.*: **The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes.** *BMC Bioinformatics.* 2008; **9**(Suppl 11): S9.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Bunescu R, Ge R, Kate RJ, *et al.*: **Comparative experiments on learning information extractors for proteins and their interactions.** *Artif Intell Med.* 2005; **33**(2): 139–55.
PubMed Abstract | Publisher Full Text

22. Ohta T, Pyysalo S, Tsujii J, *et al.*: **Open-domain anatomical entity mention detection.** In *Proceedings of ACL 2012 Workshop on Detecting Structure in Scholarly Discourse (DSSD).* 2012; 27–36.
Reference Source

23. Leaman R, Miller C, Gonzalez G: **Enabling recognition of diseases in biomedical text with machine learning: Corpus and benchmark.** In *Proceedings of the Symposium on Languages in Biology and Medicine.* 2009; 82–89.

24. Bossy R, Jourde J, Manine AP, *et al.*: **BioNLP Shared Task--The Bacteria Track.** *BMC Bioinformatics.* 2012; **13**(Suppl 11): S3.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Pyysalo S, Ginter F, Heimonen J, *et al.*: **BioInfer: a corpus for information extraction in the biomedical domain.** *BMC Bioinformatics.* 2007; **8**(1): 50.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Pyysalo S, Ohta T, Ananiadou S: **Overview of the cancer genetics (cg) task of bionlp shared task 2013.** In *Proceedings of the BioNLP Shared Task 2013 Workshop,* Sofia, Bulgaria, August 2013. Association for Computational Linguistics. 58–66.
Reference Source

27. 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,* AAAI Press, 1999. 77–86.
Reference Source

28. Herrero-Zazo M, Segura-Bedmara I, Martínez P, *et al.*: **The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions.** *J Biomed Inform.* 2013; **46**(5): 914–20.
PubMed Abstract | Publisher Full Text

29. Jimeno A, Jimenez-Ruiz E, Lee V, *et al.*: **Assessment of disease named entity recognition on a corpus of annotated sentences.** *BMC Bioinformatics.* 2008; **9**(Suppl 3): S3.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Rindflesch TC, Tanabe L, Weinstein JN, *et al.*: **EDGAR: extraction of drugs, genes and relations from the biomedical literature.** *Pac Symp Biocomput.* 2000; 517–528.
PubMed Abstract | Free Full Text

31. Pyysalo S, Ohta T, Rak R, *et al.*: **Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011.** *BMC Bioinformatics.* 2012; **13**(Suppl 11): S2.
PubMed Abstract | Publisher Full Text | Free Full Text

32. van Mulligen EM, Fourrier-Reglat A, Gurwitz D, *et al.*: **The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships.** *J Biomed Inform.* 2012; **45**(5): 879–884.
PubMed Abstract | Publisher Full Text

33. Buyko E, Beisswanger E, Hahn U: **The genereg corpus for gene expression regulation events an overview of the corpus and its in-domain and out-of-domain interoperability.** In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10),* Valletta, Malta, European Language Resources Association (ELRA). 2010.
Reference Source

34. Kim JD, Ohta T, Tsujii J: **Corpus annotation for mining biomedical events from literature.** *BMC Bioinformatics.* 2008; **9**: 10.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Gerner M, Nenadic G, Bergman CM: **An exploration of mining gene expression mentions and their anatomical locations from biomedical text.** In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing,* BioNLP '10, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 2010; 72–80.
Reference Source

36. Thompson P, Iqbal SA, McNaught J, *et al.*: **Construction of an annotated corpus to support biomedical information extraction.** *BMC Bioinformatics.* 2009; **10**: 349.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Fundel K, Küffner R, Zimmer R: **RelEx--relation extraction using dependency parse trees.** *Bioinformatics.* 2007; **23**(3): 365–371.
PubMed Abstract | Publisher Full Text

38. Ding J, Berleant D, Nettleton D, *et al.*: **Mining MEDLINE: abstracts, sentences, or phrases?** *Pac Symp Biocomput.* 2002; 326–37.
PubMed Abstract

39. Nédellec C: **Learning language in logic -genic interaction extraction challenge.** In *Proceedings of the Learning Language in Logic 2005 Workshop at the International Conference on Machine Learning.* 2005.
Reference Source

40. Nobata C, Dobson PD, Iqbal SA, *et al.*: **Mining metabolites: extracting the yeast metabolome from the literature.** *Metabolomics.* 2011; **7**(1): 94–101.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Caporaso JG, Baumgartner WA Jr, Randolph DA, *et al.*: **MutationFinder: a high-performance system for extracting point mutation mentions from text.** *Bioinformatics.* 2007; **23**(14): 1862–5.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Nagel K, Jimeno-Yepes A, Rebholz-Schuhmann D: **Annotation of protein residues based on a literature analysis: cross-validation against UniProtKb.** *BMC Bioinformatics.* 2009; **10**(Suppl 8): S4.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Doğan RI, Lu Z: **An improved corpus of disease mentions in pubmed citations.** In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing,* BioNLP '12, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. 2012; 91–99.
Reference Source

44. Furlong LI, Dach H, Hofmann-Apitius M, *et al.*: **OSIRISv1.2: A named entity recognition system for sequence variants of genes in biomedical literature.** *BMC Bioinformatics.* 2008; **9**(1): 84.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

45. Ohta T, Pyysalo S, Rak R, *et al.*: **Overview of the pathway curation (pc) task of bionlp shared task 2013.** In *Proceedings of the BioNLP Shared Task 2013 Workshop,* Sofia, Bulgaria, August 2013. Association for Computational Linguistics. 67–75.
Reference Source

46. Bell L, Zhang J, Niu X: **Mixture of logistic models and an ensemble approach for protein-protein interaction extraction.** In *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine.* BCB '11, New York, NY, USA, ACM. 2011; 371–375.
Publisher Full Text

47. Kolárik C, Klinger R, Friedrich CM, *et al.*: **Chemical names: Terminological resources and corpora annotation.**. In *Proc. of the Workshop on Building and Evaluating Resources for Biomedical Text Mining.* 2008; 51–58.
Reference Source

48. Thomas PE, Klinger R, Furlong LI, *et al.*: **Challenges in the association of human single nucleotide polymorphism mentions with unique database identifiers.** *BMC Bioinformatics.* 2011; **12**(Suppl 4): S4.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

49. Pafilis E, Frankild SP, Fanini L, *et al.*: **The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text.** *PLoS One.* 2013; **8**(6): e65390.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

50. Pyysalo S, Airola A, Heimonen J, *et al.*: **Comparative analysis of five protein-protein interaction corpora.** *BMC Bioinformatics.* 2008; **9**(Suppl 3): S6.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

51. Tikk D, Thomas P, Palaga P, *et al.*: **A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature.** *PLoS Comput Biol.* 2010; **6**: e1000837.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

52. Pyysalo S, Ananiadou S: **Anatomical entity mention recognition at literature scale.** *Bioinformatics.* 2014; **30**(6): 868–75.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

53. Leaman R, Islamaj Dogan R, Lu Z: **DNorm: disease name normalization with pairwise learning to rank.** *Bioinformatics.* 2013; **29**(22): 2909–17.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

54. Bossy R, Bessières P, Nédellec C: **Bionlp shared task 2013 – an overview of the genic regulation network task.** In *Proceedings of the BioNLP Shared Task 2013 Workshop,* Sofia, Bulgaria, August 2013. Association for Computational Linguistics. 153–160.
Reference Source

55. Leaman R, Gonzalez G: **BANNER: An executable survey of advances in biomedical named entity recognition.** *Pacific Symposium of Biocomputing.* 2008; 652–63.
PubMed Abstract

56. Neves M, Damaschun A, Mah N, *et al.*: **Preliminary evaluation of the CellFinder literature curation pipeline for gene expression in kidney cells and anatomical parts.** *Database (Oxford).* 2013; **2013**: bat020.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

57. Segura-Bedmar I, Martínez P, Zazo MH: **Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013).** In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013),* Atlanta, Georgia, USA Association for Computational Linguistics. 2013; 341–350.
Reference Source

58. Settles B: **Abner: an open source tool for automatically tagging genes, proteins and other entity names in text.** *Bioinformatics.* 2005; **21**(14): 3191–2.
PubMed Abstract | **Publisher Full Text**

59. Kim JD, Ohta T, Tsuruoka Y, *et al.*: **Introduction to the bio-entity recognition task at jnlpba.** In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications,* JNLPBA '04, Stroudsburg, PA USA, Association for Computational Linguistics. 2004; 70–75.
Reference Source

60. Kim JD, Ohta T, Pyysalo S, *et al.*: **Extracting bio-molecular events from literature — the bionlp'09 shared task.** *Computational Intelligence.* 2011; **27**(4): 513–540.
Publisher Full Text

61. Kim JD, Nguyen N, Wang Y, *et al.*: **The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011.** *BMC Bioinformatics.* 2012; **13**(Suppl 11): S1.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

62. Kim JD, Wang Y, Yasunori Y: **The genia event extraction shared task, 2013 edition overview.** In *Proceedings of the BioNLP Shared Task 2013 Workshop,* Sofia, Bulgaria August 2013, Association for Computational Linguistics. 8–15.
Reference Source

63. Rocktäschel T, Weidlich M, Leser U: **ChemSpot: A hybrid system for chemical named entity recognition.** *Bioinformatics.* 2012; **28**(12): 1633–40.
PubMed Abstract | **Publisher Full Text**

64. Bretonnel Cohen K, Johnson H, Verspoor K, *et al.*: **The structural and content aspects of abstracts versus bodies of full text journal articles are different.** *BMC Bioinformatics.* 2010; **11**(1): 492.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

65. Morgan AA, Lu Z, Wang X, *et al.*: **Overview of BioCreative II gene normalization.** *Genome Biol.* 2008; **9**(Suppl 2): S3.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

66. Comeau DC, Doğan R, Ciccarese P, *et al.*: **Bioc: a minimalist approach to interoperability for biomedical text processing.** *Database (Oxford).* 2013; **2013**: bat064.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

67. Yepes AJ, Neves M, Verspoor K: **Brat2bioc: conversion tool between brat and bioc.** In *BioCreative IV.* 2013.
Reference Source

Open Peer Review

Current Referee Status: ☑ ☑

Referee Responses for Version 1

☑ **Paloma Martínez**
Department of Computer Science, Charles III University of Madrid, Madrid, Spain

Approved: 14 July 2014

Referee Report: 14 July 2014
doi:10.5256/f1000research.3456.r5114

The article shows a review of 36 publicly annotated corpora in the biological domain. The analysis is performed taking into account different features such as type of text (abstracts, full text publications, etc.), types of entities, types of relationships (if any), number of documents, if automatic or manual annotation, if annotations are related to other resources (such as ontolologies), etc.It is an interesting and useful work that can help researchers to find a corpus to perform evaluations of their machine learning methods. As the author describes in the introduction sections, annotated corpora are valuable resources that allow to train and test algorithms and also to compare state-of-the art works each other.I think that the paper is a good contribution to the journal. Below I give several suggestions in order to improve the article as well as some data to correct a couple of mistakes.

- In the Introduction section, it would be beneficial to introduce to the reader the "*inter-annotator agreement*" as well as the main ways to calculate it.
- Before introducing the annotation schema it is necessary to give a definition of "*Schema*" (is it a conceptual schema? What are the elements a schema should contain)
- In the paragraph about DDI corpus, some remarks should be included: in the first version of the corpus used in the DDIExtraction 2011 task the drugs were automatically annotated by Metamap tool[1] but in the new version (used in DDI Extraction 2013 shared tasks), every annotation was manually revised by two pharmacists)
- At the end of this section, I would like to see something about the precision experts have annotating entities and relationships. This would also help to know what is the limit systems are able to manage recognizing entities and relations.
- In section "*Corpora and semantic types*" it is a good idea to give the number of citations from Google Scholar, although the most recent corpora have almost no citations. The DDI corpus has 0 citations because the reference is from October 2013; Segura-Bedmar *et al.* (2011)[1]is another reference that the author could include in the article corresponding to the first version of the corpus.
- Concerning the figures of DDI corpus, it consists of 792 texts selected from the DrugBank database (DDI-DrugBank dataset) and other 233 Medline abstracts (DDI-MedLine dataset) on the

subject of DDIs. The corpus was manually annotated with a total of 18,502 pharmacological substances and 5028 DDIs, including both pharmacokinetic (PK) as well as pharmacodynamic (PD) interactions.

- Concerning drugs, there are two corpora that do not appear in the article: (a) PK Corpus and PF DDI Corpus[2] with approx. 600 abstracts about clinical pharmacokinetics and pharmacogenetics, in-vitro and in-vivo drug-drug interactions. (b) PK DDI corpus[3] consisting of 64 FDA drug labels with annotations for drugs with their precipitant or object roles in drug-drug interaction.

- Instead of having a paragraph for each corpus, I suggest including a tabular representation of corpora and characteristics. For instance, with a column for each feature (document type, annotation tool, categories of entities, number of mentions, format, availability, etc.). This representation would help to compare different corpora.

- I also suggest mentioning that linguistic phenomena such as co-reference resolution are also required in annotation task, especially in the detection of entities. For instance, how are words such as “drug”, “disease”, “medication” and others annotated in these corpora?

- Some typos: In the section: ***List of biological corpora - Drug-Drug Interaction:***

  *“...has been extensively used for both training and evaluation for NER and **relatiosnhip extarction** tasks.”*

References

1. Segura-Bedmar I, Martínez P, Sanchez-Cisneros D: The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts. *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction.*2011; **761**: 1-9 Reference Source

2. Wu HY, Karnik S, Subhadarshini A, Wang Z, Philips S, Han X, Chiang C, Liu L, Boustani M, Rocha LM, Quinney SK, Flockhart D, Li L: An integrated pharmacokinetics ontology and corpus for text mining.*BMC Bioinformatics*. 2013; **14** (35). PubMed Abstract | Free Full Text | Publisher Full Text

3. Boyce R, Gardner G, Harkema H: Using natural language processing to identify pharmacokinetic drug-drug interactions described in drug package inserts. *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*. 2012. 206-213 Reference Source

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

☑

Roman Klinger
Cognitive Interaction Technology – Center of Excellence (CIT-EC), Bielefeld University, Bielefeld, Germany

Approved: 06 May 2014

Referee Report: 06 May 2014

doi:10.5256/f1000research.3456.r4561

The paper *"An analysis on the entity annotations in biological corpora"* is a review of several corpora which are available to the research community and which consist of textual data with annotation. These annotations are either on an entity level or with relations between entities.

The author makes very clear what this paper is about, which corpora are discussed and where the border is drawn to resources not discussed here. The comprehensive list of corpora is enriched with citation counts to make clear which corpora are already commonly used and which might be under-explored by the community. She makes clear that she would like to motivate researches to make use of these corpora as well.

In general, I think that this paper is a very valuable review of resources available to the community. The author does not define strictly what the motivation for this paper is, what the audience should be and who can specifically benefit from it. I could think of at least two scenarios:

1. A researcher would like to improve or develop a method for a specific domain/entity class (like drug-drug-interaction or recognition of chemical names, for instance). Then the author can get an overview of available corpora and be quite sure that she or he does not miss a corpus of the specific domain when the class is mentioned in this review.

2. A researcher would like to evaluate a method and needs resources, he or she does not necessarily care so much about the specific domain and can select from the variety of discussed resources in this review.

I propose that such or similar motivations are added to the introduction.

Title: I am not sure if a corpus can be "*biological*" — or dealing with classes from the bio(medical) domain.

Abstract: Should indicate for whom this review might be of value. Some examples of cases for use would be great, I think.

Introduction:

- A schema does not consist of entities only, but also of entity classes.
- Maybe it would be interesting to discuss the issue of having a bias in the annotation towards the automatic tool when annotations are only validated.
- "*Such corpora tends…"* —> "*corpora tend*" or "*such corpus tends*"
- "*I show the impact of each corpora*" —> "*each corpus*"
- The author cites the paper on the 5 reviews commonly used to study PPI. Are there no other such reviews?

Corpora and semantic types:

List of corpora:

- "*I also did not include corpora which have only text span annotations not related to a particular semantic entity…"* — I do not understand that. Several of the corpora the author is discussing do not have links to database or ontology IDs. This should be made clear. (examples are the BC2 GM corpus, the SCAI corpus, I think AZDC as well.)
- Abner is at least evaluated on BioCreative data as well. I am not sure it has only been trained on GENIA.

Semantic Analysis of Corpora

- "*I only consider those annotations which are meaningful enough to be associated with one of the pre-defined semantic types under consideration*" I think that formulation could be improved to make clearer what '*meaningful'* means.

Comparison and Discussion

- "*On the other hand*" — without "*On the one hand*"
- "*different number of documents*" — "*numbers*"?
- "*relationships was*" -> "*relationships were*"
- "*corpora which contains*" -> "*contain*"

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.