Linking Biomedical Data to the Cloud

Stefan Zwicklbauer(✉), Christin Seifert, and Michael Granitzer

University of Passau, Innstraße 33, 94032 Passau, Germany
{Stefan.Zwicklbauer,Christin.Seifert,Michael.Granitzer}@uni-passau.de

Abstract. The application of Knowledge Discovery and Data Mining approaches forms the basis of realizing the vision of Smart Hospitals. For instance, the automated creation of high-quality knowledge bases from clinical reports is important to facilitate decision making processes for clinical doctors. A subtask of creating such structured knowledge is entity disambiguation that establishes links by identifying the correct semantic meaning from a set of candidate meanings to a text fragment. This paper provides a short, concise overview of entity disambiguation in the biomedical domain, with a focus on annotated corpora (e.g. CalbC), term disambiguation algorithms (e.g. abbreviation disambiguation) as well as gene and protein disambiguation algorithms (e.g. inter-species gene name disambiguation). Finally, we provide some open problems and future challenges that we expect future research will take into account.

Keywords: Linked data cloud · Entity disambiguation · Text annotation · Natural language processing · Knowledge bases

1 Introduction

The amount of digital data, also called the digital universe, grows rapidly, amounting to 4.4 Zetabytes in 2013[1]. Thus, medical doctors and biomedical researchers of today are confronted with increasingly large volumes of high-dimensional, heterogeneous and complex data from various sources, which pose substantial challenges to the computational sciences [1]. Overall, the majority of such information (e.g. medical reports) is transmitted through unstructured documents [2], more suitably defined as non-standardized data [3]. The task of Knowledge Discovery is to extract implicit, previously unknown, and potentially useful information from such unstructured data [4].

The application of Knowledge Discovery and Data Mining approaches forms the basis of realizing the vision of *Smart Hospitals* [1,5]. A prominent example is the (automated) creation of high-quality knowledge bases (KB) from clinical reports. The Comparative Toxicogenomics Database (CTD) [6], for instance, is a high-quality data base for researching the influence of chemicals on human health, but is manually curated and therefore restricted in its coverage of the

[1] The digital universe of opportunities http://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf.

A. Holzinger et al. (Eds.): Smart Health, LNCS 8700, pp. 209–235, 2015.
DOI: 10.1007/978-3-319-16226-3_9

documents annotated by experts. Providing high-quality, automatic methods for populating the KB from clinical reports would facilitate decision making processes for clinical doctors [1]. The demand for automatic methods is also reflected in the natural language processing challenges posed by various initiatives, like the BioCreative initiative[2] and the BioNLP shared tasks [7]. For instance, in the domain of biomedical research, the understanding of two-component regulatory systems (TCSs), a mechanism widely used by bacteria to sense and respond to the environment, can be facilitated [8]. TCSs are of particular interest for infectious disease researchers including virulence, response to antibiotics, quorum sensing and bacterial cell attachment [9].

For these purposes, the recognition and assignment of symptoms, chemicals, genes, proteins etc. to a unique identifier in a KB is an important subtask. This chapter gives an overview of the state-of-the art of linking unstructured biomedical data to the Linked Data Cloud, with a special emphasis on biomedical entity disambiguation.

The remainder of the chapter is structured as follows: Sect. 2 defines the technical terms required for understanding the chapter. Section 3 gives a clear definition of the problem that should be solved and illustrates why linking biomedical entities to the cloud is a challenging task by examples. Section 4 then provides the foundations for understanding the reviewed algorithms by exemplifying the data structures used by disambiguation methods. The state-of-the-art review in Sect. 5 is divided into four subsections:

- The state of the biomedical Linked Data Cloud is described in Sect. 5.1,
- Section 5.2 presents annotated corpora for training linking algorithms,
- Algorithms for biomedical term disambiguation are reviewed in Sect. 5.3,
- Algorithms for gene and protein disambiguation are presented in Sect. 5.4.

The chapter concludes with an overview of open problems in Sect. 6 and an outlook on future work is given in Sect. 7.

2 Glossary and Key Terms

Automatic Term Recognition (ATR) Recognition and linking of terms to domain specific data bases [10], synonym to ↑ NED.

Disambiguation The process of linking a ↑ surface form to a ↑ URI.

Entity A modeled abstract or concrete object of the real world, for example a specific gene. In the context of ↑ disambiguation also called label [11].

Knowledge Base (KB) describes a knowledge repository that stores facts about the world. Knowledge bases can be coarsely classified into structured and unstructured knowledge bases depending on the form of the data representation. An orthogonal classification is specific for general-purpose knowledge bases, depending on the type of knowledge stored.

[2] http://www.biocreative.org.

Linked (Open) Data describes the concept of providing semantic information for data sets. The goal is to support automatic sharing and linking pieces of the data on a semantic level. The basic technologies for Linked Data are ↑ URIs and ↑ RDF. Linked Open Data (LOD) encompasses the idea that these data sets should be openly accessible.

Linked (Open) Data Cloud subsumes the (openly accessible) data sets represented as ↑ Linked Data.

Named Entity A modeled, concrete object of the real world, referenced by proper names or acronyms in the text. Originally introduced in the Message Understanding Conference (MUC) Challenges, the commonly agreed types were person, location and organization, later date and time, measures and email addresses were added [12]. Depending on the application domain, other domain-specific named entities exist. These are for instance names of drugs or proteins in the biomedical domain.

Named Entity Recognition (NER) The process of identifying a ↑ named entity, i.e. identifying that a surface form represents a named entity (but not yet knowing, which entity exactly).

Named Entity Disambiguation (NED) The process of linking a ↑ surface form representing a ↑ named entity to a unique meaning [13].

Resource Description Framework (RDF) is a general concept for the semantic description of resources. The building blocks of RDF are triplets consisting of subject (the thing that is described), the object (to which it is related) and a relation (specifying the relationship between subject and object). Relations are unidirectional. All parts of a triplet are uniquely identifiable by the means of ↑ URIs.

Surface Form refers to the piece of textual information (words or phrases) that should be linked to a semantic entity [14,15]. Also called mention, entity mention, mention occurrence, spot [11], or lemma [16].

Uniform Resource Identifier (URI) is a string of characters identifying a resource. The most prominent example is the Uniform Resource Locator (URL) used in the World Wide Web.

Word Sense Disambiguation (WSD) The process of linking a ↑ surface form to a unique entry in a dictionary. In general, the linked ↑ surface forms are not ↑ entities. Consider for instance the different meanings of the word "mind" (depending on the context it could be used as verb or noun and may have different meanings in each grammatical form.).

3 Problem Statement

Entity annotators undertake a crucial processing step in producing structured knowledge. They "ground" the underlying texts with respect to an adequate semantic representation. The entity annotation task can be subdivided into the following two sub steps:

- **Entity Recognition:** The identification of short-and-meaningful sequences of terms, also called surface forms, which can be linked to entities in a catalog.
- **Entity Disambiguation:** The annotation of surface forms with unambiguous identifiers (entities) drawn from a catalog.

Entity Recognition. Entity recognition forms the first step of creating entity annotations. It identifies proper nouns that can be linked to a semantic meaning. Proper nouns often exhibit *structural ambiguity* that complicates the correct identification. For example, the components of "Victoria and Albert Museum and IBM and Bell Laboratories" look identical. The term "and" is part of the name of the museum in the first example, but a conjunction joining two computer company names in the second [17]. The task of named entity recognition (NER) focuses on identifying surface forms in a text which are the names of things, such as person, organization, gene or protein names. Overall, (named) entity recognition is a well studied research topic. State-of-the-art algorithms for generic knowledge entities score ≈90 % of F-measure [18], while accuracy of biomedical NER strongly depends on the entities' types (e.g. proteins, genes, diseases) [19].

Entity Disambiguation. The task of entity disambiguation establishes links between identified surface forms and entities within a catalog (KB) and faces the problem of *semantic ambiguity* [17]. Formally, entity disambiguation inherently involves resolving many-to-many relationships. Multiple distinct surface forms may refer to the same entity. Simultaneously, multiple identical surface forms may refer to distinct entities [20]. Figure 1 shows a specific example of this relationship. We assume a sentence containing the surface forms "Ford" and "CART" (depicted in the yellow rectangle). Both surface forms may refer to different entities, e.g. Ford by itself could be an actor (Harrison Ford), the

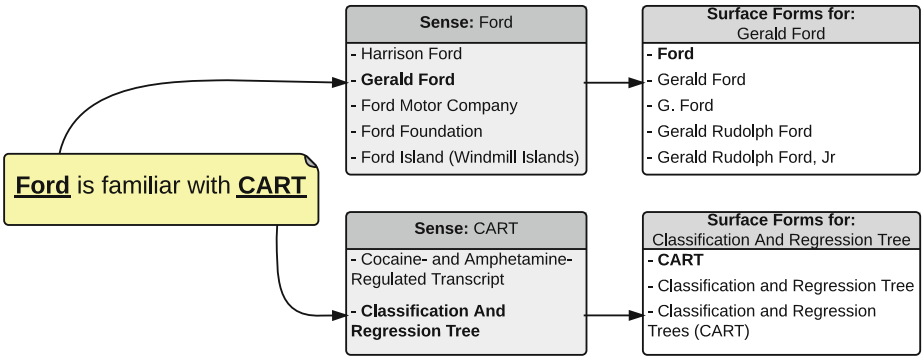

Fig. 1. Surface forms (bold) within a sentence (yellow rectangle) may refer to different entities (rectangles in the middle) depending on the context. Additionally, an entity may be addressed by various surface forms (rectangles on the right) (Colour figure online).

38th President of the United States (Gerald Ford), an organization (Ford Motor Company) or a place (Ford Island). In our context, we assume "Gerald Ford" to be the correct entity, which may be expressed in several ways, e.g. "Gerald Rudolph Ford, Jr.". However, similar to NER, the task of named entity disambiguation (NED) focuses on surface forms constituting the names of special entity classes. The ever-increasing publication rate of biomedical documents now means that entity disambiguation in the biomedical domain is becoming more and more important. Biomedical NED is constrained to biomedical entities only, but is extremely challenging [21] since a surface form

1. could refer to another type of biomedical entity, such as a protein or phenotype, e.g. the mouse gene "hair loss".
2. could be other types of concepts in closely related domains, such as the clinical field, e.g. the mouse gene "diabetes".
3. could be the same as common English words, e.g. fly genes "can" and "lie".
4. could refer to several, different genetic entities, either from the same or from other species, e.g. cow or chicken.

In biomedical entity disambiguation, genes and gene products (i.e. proteins) form an important class of entities. To map surface forms of these entity classes to an entity within a KB, it is important to identify what organisms (species) the genes and proteins belong to, and on what species the experiments are carried out to understand particular biological phenomena. There are dozens of species commonly used in biological studies, such as Escherichia coli, Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens and hundreds more are frequently mentioned in biological research papers. For example, without context, "tumor protein p53" may associate to over 100 proteins across 23 species[3]. To identify the proteins (i.e. the underlined terms) in the following sentence, knowing the "focus" species of the article is not sufficient, as they belong to three different species: human, mouse and rat.

The amounts of human and mouse CD200R-CD4d3+4 and rCD4d3+4 protein on the microarray spots were similar ...

The authors of [21] investigated the extent of the ambiguity problem in the biomedical domain. They obtained genes from 21 species and quantified naming ambiguities within and across species, with English words and with medical terms. The results revealed that official gene symbols display negligible ambiguity within a specific species (0.02 % regarding uppercase letters) and a high ambiguity across-species (14.20 %). Additionally, the results showed a moderate ambiguity rate with general English words (0.57 %) and medical terms (1.01 %). The analysis of correct gene disambiguation results within abstracts of biomedical research paper also showed a very high number of ambiguous genes across species [21] (85.1 %).

Overall biomedical NED is a challenging task and thus has attained much attention in research in the last decade.

[3] Querying RefSeq database (http://www.ncbi.nlm.nih.gov/refseq/). The number of species was manually counted.

4 Entity Representation

A crucial factor for creating a disambiguation system is the way entities are represented within a KB. Generally an entity can be defined intensionally, i.e. through a set of describing properties, or extensionally, i.e. through instances and usage in documents [22]. In the following we differentiate more precisely between these representations and give examples of how entities might be represented within disambiguation KBs in practice.

4.1 Intensional Description

An intensional definition of an entity can be understood as a thesaurus or logical representation, as it is provided by Linked Open Data repositories. In the context of entity disambiguation, KBs comprising intensionally defined entities are referred to as *entity-centric KBs* [23]. Formally, an entity-centric KB can be described as

$$Kb_{\text{ent}} = \{e_0, ..., e_n | e_i \in E, n \in \mathbb{N}\} \tag{1}$$

The set of all entities available in the entity-centric KB Kb_{ent} is denoted as E, with e_i being a single entity [23]. All entities $e_i \in \text{Kb}_{\text{ent}}$ usually provide a unique primary key ID which combines the name of the knowledge source as well as its identifier in the knowledge source. Additionally, a variable number of fields k contain domain-independent attributes, e.g. descriptions, and domain-dependent information, e.g. the sequence length of genes. Formally, such an entity can be denoted as

$$e_{\text{i}} = (ID, Field_1, ..., Field_k) \tag{2}$$

Table 1 shows a specific example of how the entity "Phenylalanyl-tRNA–protein transferase" might be represented in an entity-centric KB. The entity contains standard attributes, i.e., name, synonyms, description, link to web resource, type, as well as occurrence information. More specifically, all referenced surface forms for this entity and the respective amount of occurrences with this surface form are stored in *Occurrences*. The field *Cooccurrences* contains surface forms of entities that appeared near the described entity in any text and the amount of appearances of the respective surface form in the context range (i.e. 300 words).

4.2 Extensional Description

An extensional entity definition resembles information on the usage context of an entity. For instance, natural language text documents annotated with entities can be used as such usage context. KBs containing extensional entity definition are referred to as *document-centric KBs* [23]. Formally, a document-centric KB is defined as

$$Kb_{\text{doc}} = \{d_0, ..., d_n | d_i \in D, n \in \mathbb{N}\} \tag{3}$$

An entry d_i in a document-centric KB Kb_{doc} consists of the document content representing a text string and a list of annotations of surface forms $t^l_{e_i}$, with

Table 1. Example of an entity-centric KB entry

Field	Content
ID	UNQ9A741
Name	Phenylalanyl-tRNA–protein transferase
Synonyms	Leucyltransferase
Description	Functions in the N-end rule pathway of protein degradation where it conjugates Leu, Phe and, less efficiently, Met from aminoacyl-tRNAs to the N-termini of proteins
Mainlink	http://www.uniprot.org/uniprot/Q9A741
Type	Caulobacter
Occurrences	aat:::3
Co-Occurrences	substrate:::3, Leu:::6, Phe:::6

l denoting the l^{th} annotation in the document. Annotated surface forms are described by their position in the document and a list of their entity references. An entry in a document-centric KB is denoted as

$$d_{\mathrm{i}} = (Document, \{(Start, End, \{ID\}), ...\}) \tag{4}$$

Table 2 shows an example of a biomedical document containing the surface form "'Myeloma"' in a document-centric KB. The document's content is subdivided in *title* and *titleandtext*, which is a concatenation of the document's title and main content. Furthermore, all available annotations (and its respective properties) are stored in the field *Annotations*. The field ID depicts a unique document identifier.

Table 2. Example of a document-centric KB entry

Field	Content
ID	174996
Title	Antibody therapy for treatment of multiple myeloma
Abstract	Monoclonal antibody therapy antibody therapy has emerged as a viable treatment option for patients with lymphoma and some leukemias. It is now beginning to be...
TitleAndAbs	Antibody therapy for treatment of multiple myeloma. Monoclonal antibody therapy antibody therapy has emerged as a viable treatment option for patients with...
Keywords	Myeloma::43::50::diso:umls:C0026764:T191:diso

5 State-of-the-Art

In this section the state-of-the-art is reviewed along three dimensions. First, we review the state of the biomedical linked data cloud in Sect. 5.1. Second, we describe available annotated corpora for training algorithms in Sect. 5.2. Third, we review the algorithms for for biomedical term disambiguation in Sect. 5.3 and for gene and protein disambiguation in Sect. 5.4. We note that we do not describe and review text (pre-)processing steps (e.g. tokenization, normalization, stemming) which are necessary for entity recognition and disambiguation. An overview of relevant steps for text processing in the biomedical domain can be found in [1].

5.1 The Biomedical Linked Data Cloud

According to the "State of the LOD Cloud 2014"[4] the Linked Open Data cloud comprises 1014 data sets, 83 (8.19 %) belong to the life sciences domain as of April 2014. Data sets use different vocabularies, proprietary or non-proprietary. Proprietary vocabularies are only used by one data set and thus are not useful for interlinking differently linked data repositories. Non-proprietary vocabularies are used by at least two data sets and comprise only 41.76 % of all encountered 649 vocabularies. In terms of data sets, 23.17 % (241) data sets use proprietary vocabularies, but also nearly all of the data sets (99.87 %) use non-proprietary vocabularies. In the life sciences this amount is slightly higher. 35 different proprietary vocabularies are used in 26 data sets (these amount to 29.21 % of all life sciences data sets). Only 28.57 % of these data sets are fully linkable to other data sets, i.e. can be fully interpreted by automatic mechanisms. 65.71 % of these data sets are not linkable at all.

5.2 Annotated Corpora

This section presents an overview of annotated corpora for biomedical entity disambiguation. We omitted corpora that were not, or are no longer publicly available.

GENIA Corpus

The GENIA corpus [24], released in 2003, contains ≈2000 MEDLINE abstracts from the domain of molecular biology. The corpus is freely available for download[5]. The MEDLINE abstracts were collected by querying PubMed for the three MeSH terms "human", "blood cells", and "transcription factors". They were syntactically and semantically annotated, resulting in six different sub-corpora corresponding to the specific annotations:

[4] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/.
[5] http://www.nactem.ac.uk/genia/genia-corpus.

Table 3. Statistics of the GENIA corpus (term annotations)

	GENIA
Documents	2,000
Document Type	MEDLINE abstract
Surface Forms	89,862
Release Date	2003 (version 3.0)

- Part-of-Speech annotation subcorpus,
- Constituency (phrase structure) syntactic annotation subcorpus,
- Term annotation subcorpus,
- Event annotation subcorpus,
- Relation annotation subcorpus,
- Coreference annotation subcorpus.

Linguistic structures are annotated with biological terms from the GENIA ontology in the term annotation subcorpus, which represents the corpus for entity disambiguation. Table 3 provides an overview of the GENIE term annotation subcorpus.

BioCreative Corpora

The BioCreative (Critical Assessment of Information Extraction in Biology) community has released various annotated corpora since 2004. The data sets are freely available for non-commercial purposes[6].

GM Corpus (BioCreative I and II): The BioCreative I data set [25] for the Gene Mention (GM) task was released in 2005 and consists of sentences from MEDLINE abstracts annotated with gene mentions. The provided sentences have already been tokenized. The BioCreative II data set [26] is an extended and refined version of the BioCreative I data set and was released in 2008. The changes include an addition of 5000 sentences, a review of the annotations with ≈13 % changes and linkage of the gene mentions to either the GENE or ALTGENE KB. Further, in the BioCreative II data set the sentences were not tokenized a-priori. An overview of the basic statistics for the BioCreative I+II data sets can be found in Table 4.

Table 4. Statistics of the GM I and II corpus (aggregated training, test and development set)

	GM I	GM II
Documents	1,500	2,000
Document Type	MEDLINE abstract	MEDLINE abstract
Surface Forms	1,800	44,500
Release Date	2005	2008

[6] http://www.biocreative.org/resources/.

Table 5. Statistics of the ChemDNER corpus (aggregated training, test and development set)

	ChemDNER
Documents	10,000
Document Type	PubMed abstract
Surface Forms	84,355
Entities	19,805
Release Date	2013

ChemDNER Corpus (BioCreative IV): The ChemDNER (Chemical and Drug Named Entity Recognition) corpus [27], released by the BioCreative community in 2013 (part of BioCreative IV), contains PubMed abstracts manually annotated with chemical compounds and drugs. Each abstract was annoated by at least two experts with an overall inter-annotater agreement of 91 %, thus the corpus can be considered a gold standard for chemical NER. Table 5 provides a summary statistics of the corpus with all values aggregated over training, test and development set. More details on corpus construction and statistic can be found in [27].

BC4GO Corpus (BioCreative IV): The Gene Ontology (GO) corpus [28] was released by the BioCreative community in 2013 as part of the BioCreative IV challenge. The corpus consists of 200 annotated full-text articles from PMC. The task associated with this corpus involves extracting gene function terms and the associated evidence sentences. Table 6 provides an overview of the corpus.

CalbC Corpus

The CalbC (Collaborative Annotation of a Large Biomedical Corpus) corpus is a very large, community-wide shared text corpus annotated with biomedical entity references [29]. CalbC represents a silver standard corpus which results from the

Table 6. Statistics of the BC4GO corpus (aggregated training, test and development set)

	BC4GO
Documents	200
Document Type	PMC full-texts
Gene mentions	5,162
Entities (Genes)	665
GO term mentions	5,275
Entities (GO terms)	1,311
Release Date	2013

harmonization of automatically generated annotations and is freely accessible[7]. The data set is released in 3 different sizes: small (CalbCSmall), big (CalbCBig) and pilot, with the former two being the most widely used. Table 7 provides an overview of the basic properties of CalbCSmall and CalbCBig. A comparison regarding the overlap of entities within both corpora shows that a very high percentage of entities occurs in both data sets. Hence, there are few entities which occur in CalbCBig but are not present in the small corpus. In contrast to other disambiguation corpora like Dbpedia, a surface form may be linked to more than one entity resource per annotation. Due to a comprehensive taxonomy and classification system a surface form provides 9 entity annotations on average. Figure 2 presents an overview of the distribution of surface forms and their corresponding entities. The histogram axis showing the number of entities is truncated at 40 entities due to very few existing surface forms which contain a lot of different meanings (maximum 9895). Nearly half of all surface forms may attain between 2 and 7 different entities. The other half of surface forms attains up to 9895 different entity meanings. Figure 3 shows an overview of the distribution of surface forms over entities. More than 10,000 different surface forms address general entities like "kinase" or "protein".

Table 7. Statistics of the CalbCSmall and CalbCBig corpora

	CalbCSmall	CalbCBig
Documents	174,999	714,282
Document Type	MEDLINE abstract	MEDLINE abstract
Surface Forms	2,548,900	10,304,172
Unique Surface Forms	50,725	101,439
Entities	37,309,221	96,526,575
Unique Entities	453,352	308,644
Used Unique Entities	265,532	228,744
Namespaces	14	16
Release Date	2011	2011

CRAFT Corpus

The CRAFT (Colorado richly annotated full text) corpus [30] is an annotated corpus consisting of 67 full-text journal articles from the biomedical domain. The corpus contains ≈100,000 annotations from the biomedical domain, linking it to 7 different repositories (Chemical Entities of Biological Interest, Cell Ontology, Entrez Gene, Gene Ontology, NCBI Taxonomy, Protein Ontology and Sequence Ontology). Table 8 provides an overview of the data set. The corpus is licenced under the Creative Commons Attribution 3.0 license (CC BY) and is available online[8].

[7] http://www.calbc.eu/.
[8] http://bionlp-corpora.sourceforge.net/CRAFT/.

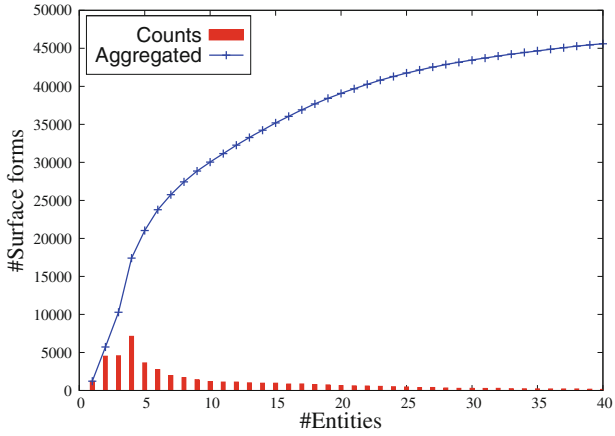

Fig. 2. Distribution of surface forms and their corresponding entities

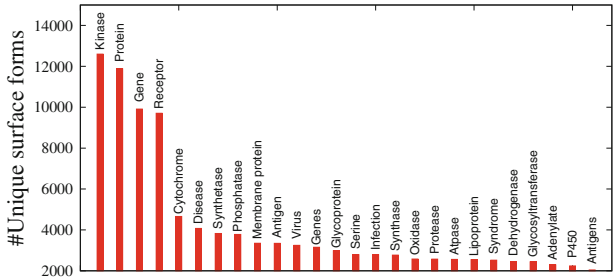

Fig. 3. Number of entity annotations (Only entities annotated for more than 2000 different entities are shown.)

BioNLP Shared Tasks Corpera

The BioNLP Shared Tasks corpora originates from the GENIA corpus (see above). In 2004, 2009 and 2011, the initiative covering different natural language tasks for the biomedical domain released several corpora. The data sets are available online[9]. Here, we describe subcorpora from the release in 2011 [8], which is also publicly available[10].

EPI Corpus: The EPI corpus (Epigenetics and Post-translational Modifications) was crafted to research automatic extraction of events related to epigenetic changes. The corpus consists of 1,200 MEDLINE abstracts, annotated with entities representing proteins or genes. Additional annotations are made for events (e.g. hydroxylation, DNA methylation), and event modifications (e.g.

[9] http://www.nactem.ac.uk/genia/shared-tasks.
[10] http://2011.bionlp-st.org/.

Table 8. Statistics of the CRAFT corpus

	CRAFT
Documents	67
Document Type	PubMed full-texts
Surface Forms	≈100,000
Namespaces	7
Unique Entities	4,319
Release Date	2012

catalysis, positive regulation, negation or speculation). An overview of the EPI corpus is presented in Table 9.

Table 9. Statistics of the EPI corpus from the BioNLP Shared Task (aggregated training, test and development set)

	EPI
Documents	1,200
Document Type	PubMed abstract
Surface Forms (Protein, Gene)	15,190
Surface Forms (Event)	3,714
Surface Form (Modification)	369
Release Date	2011

ID Corpus: The ID (infectious diseases) corpus was designed to study the molecular mechanism of infectious diseases. It consists of 30 full-text documents from the PMC data base. The documents are annotated with five types of entities (protein, two-component system, regulon-operon, chemical and organism), event types (e.g. for example gene expression, binding, regulation) and modifications. The latter indicates whether a statement is a speculation or a negation. Table 10 provides an overview of the ID corpus.

CDT Corpus

The Comparative Toxicogenomic Database (CTD) [6] is a publicly available database[11] containing the following types of manually curated annotations:

- Chemical-gene interactions,
- Chemical-disease associations,

[11] http://ctdbase.org/.

Table 10. Statistics of the ID corpus from the BioNLP shared Task (aggregated training, test and development set)

	ID
Documents	30
Document Type	PMC full-texts
Surface Forms (Entity)	12,740
Surface Forms (Event)	3,714
Surface Form (Modification)	369
Release Date	2011

– Gene-disease associations,
– Chemical-phenotype associations.

The manual data collection started in 2004 and is constantly updated. An overview of the data sets as of July 2014 can be found in Table 11.

Table 11. Statistics of the CDT corpus (figures correspond to the version from July 2014)

	CDT
Documents	109,701
Document Type	PubMed full-texts
Chemicals	13,446
Diseases	6,347
Genes	36,393
Release Date	Silent releases, constantly updated

5.3 Biomedical Term Disambiguation

Biomedical term disambiguation focuses on disambiguating all classes of biomedical entities (e.g. medical terms, abbreviations, genes, chemicals). Official biomedical symbols display only a moderate degree of ambiguities with general English words, medical terms and concepts [21]. Thus, the number of works resolving these ambiguities is limited.

String Matching Algorithms. String Matching algorithms are able to map case-sensitive surface forms to the respective KB entries. The work by Tsuruoka et al. [31] focused on learning a string similarity measure from a dictionary with logistic regression. The experiments were conducted on several large-scale gene and protein name dictionaries. Results showed that a logistic regression-based similarity measure outperforms existing similarity measures like Hidden Markov

Model [32], SoftTFIDF [33], Jaro-Winkler [34] and Levenshtein in dictionary look-up tasks.

Another work from Rudniy et al. [35] describes the problem of mapping entities in biomedical data to the UMLS Metathesaurus. The work introduces the Longest Approximately Common Prefix (LACP) method as an algorithm for approximate string matching that runs in linear time. The authors compare the LACP method to nine other well-known string matching algorithms (e.g. TF-IDF [36], Jaro-Winkler [34], Needleman-Wunsch [37]) in terms of precision and performance. As a result, LACP outperforms all nine string similarity methods in both disciplines, performance and accuracy. It attains the best F1 values (up to 92 %) when evaluated on three out of the four data sets.

A major disadvantage of these approaches is the non-availability of disambiguation techniques. In other words, if surface forms are ambiguous these algorithms are hardly able to determine the potential entity candidate.

Abbreviation Disambiguation. There are a number of systems that have been developed to map biomedical abbreviations to appropriate entities. Methods for mapping abbreviations to full forms fall into two broad categories [38]: abbreviations are linked to entities with the help of pattern or rules when the entities' full forms appear nearby in the text [39,40], or statistical disambiguation methods choose entities for an abbreviation based on the context the abbreviation occurs in [38,41].

The intention of the AbbRe system [39] (Abbreviation Recognition and Extraction) was to map abbreviations to entities when the entities' full forms are explicitly defined in biomedical full-text articles. AbbRE operates through a set of manually annotated rules assigning matches between letters in the abbreviations and words in the full form. AbbRE was evaluated in full-text biomedical articles and found to have 70 % recall and 95 % precision.

Yu et al. [38] proposed the first model that resolves the problem of abbreviation ambiguity in full-text journal articles. The approach is built upon the earlier work AbbRe and presents a semi-supervised method that applies MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. The authors trained supervised learning algorithms (i.e. Naive Bayes and Support Vector Machines) on 11 million MEDLINE abstracts which were annotated with AbbRe first.

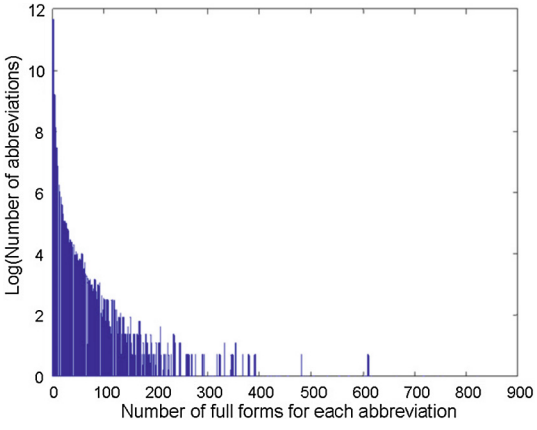

Fig. 4. Distribution (from eleven million MEDLINE records) of the numbers of abbreviations paired with different numbers of full forms [38].

Figure 4 shows the distribution of the numbers of abbreviations paired with different numbers of full forms occurring in the annotated MEDLINE abstracts. The abbreviations "or" and "ca" correspond to the largest numbers of different full forms. Overall, the authors report up to 92 % precision when disambiguating biomedical abbreviations.

General Biomedical Term Disambiguation. Few works focused on general biomedical term disambiguation, which comprises all kinds of biomedical surface forms that can be linked to an entity (i.e. medical terms, gene names, abbreviations).

The work of Chen et al. [42] presents a simple method for biomedical term disambiguation, which can be viewed as a context-based classification approach. Instead of directly using all of a words' surrounding words, the authors only select certain words with high "discriminating" capabilities as features. By using this method, unimportant surrounding words are discarded to improve disambiguation quality. The top-n influential context terms are used as feature vector. These feature vectors serve as input to a classification method for creating classifiers (i.e. Support Vector Machine, Naive Bayes, Ripper and C4.5), which map each surface form to an entity in the KB. A major contribution of this method is its unique way of selecting the features of the ambiguous terms and building feature vectors.

Zwicklbauer et al. [23] investigated biomedical entity disambiguation with entity- and document-centric KBs. The authors state that document-centric KBs outperform laboriously constructed entity-centric KBs if an adequate amount of annotations is available. In this context, they investigated to which degree disambiguation results depend on the quality of entity repositories [23]. They showed that the quality of disambiguation results with an entity-centric KB is distinguished from the use of different repositories and biomedical subdomains (e.g. UMLS, Uniprot, Entrez Gene). A major limitation is the non-use of machine learning algorithms. Instead, the authors apply standard approaches like the Vector Space Model [43] with TF-IDF [36] and BM-25 [44].

5.4 Gene and Protein Disambiguation

A bulk of works specialized on disambiguating genes and proteins, which constitutes a challenging task due to a high degree of ambiguous gene/protein mentions across species [21]. The goal of gene and protein disambiguation, a subtask of the Gene Normalization (GN) process (also comprises gene and protein recognition [45]), is to determine the unique identifiers of genes and proteins mentioned in scientific literature. A unique identifier comprises a unique species id as well as a unique id for the respective gene or protein. Basically, the gene and protein disambiguation (in the following denoted as gene disambiguation) faces the following ambiguity problems:

1. Gene-Protein name ambiguity: a surface form may refer either to a gene or a protein, but is unambiguous within the set of all genes or proteins across all species.

2. Intra-species gene name ambiguity: a surface form could be the identifier of several genes or proteins belonging to a specific species when the species identifier is provided.
3. Inter-species gene name ambiguity: a surface form could be the identifier of several genes or proteins across species.

In the following we describe the most important works addressing the respective ambiguities.

5.4.1 Gene-Protein Name Ambiguity

The simplest form of ambiguity occurs if a surface form either refers to a gene or a protein while being unambiguous within the set of all genes or proteins across all species. This assumption can be modeled as a binary classification problem which classifies the surface form into the gene or protein class.

While recent work do not *explicitly* distinguish between both classes, the authors of [46] conducted experiments on how standard classification approaches like Naive Bayes and C4.5 [47] perform on this disambiguation task. When Naive Bayes was combined with a well-chosen smoothing function, it attained ≈80 % accuracy in the classification task on different data sets. Ginter et al. [48] introduced a new classifier based on ordering and weighting the feature vectors obtained from word counts and work co-occurrence in the text. An additional improvement was attained after weighting by positions of the words in the context of annotated article abstracts downloaded from the PubMed [49] database. Pahikkala et al. [50] further improved accuracy by incorporating a weighting scheme based on distances of context words into a conventional Support Vector Machine.

Overall gene-protein classification is quite simple and thus attains accuracy values between 85 % and 90 % with standard approaches.

5.4.2 Intra-species Gene Name Ambiguity

It is more likely that a surface form could be the identifier of several genes or proteins belonging to a specific species when the species identifier is provided. Algorithms that resolve an intra-species gene name ambiguity do not *explicitly* distinguish between the gene and protein class. The BioCreative I and II challenges [45] were conducted to map genes from the EntrezGene KB when specific sets of species are provided. Focusing on gene recognition in text and gene disambiguation (and also on protein-protein interactions), the BioCreative II dataset is commonly used for evaluation purpose of intra-species gene-protein name ambiguity. However, by also including the gene recognition task, the result values of the evaluated systems are not applicable to the disambiguation task in general.

Semantic Approaches Xu et al. [51] proposed a gene profile-based approach which examines gene name disambiguation under several idealistic assumptions: 1. Perfect gene mentions are assumed with most being restricted to short string

gene symbols, and 2. among the possible gene candidates in their disambiguation task one candidate is always the correct answer, which ignores the fact that an apparent gene mention in a text may not denote a gene at all [52]. However, in their approach, they extract a profile with different types of information (e.g. context terms, context ontological semantic concepts) from each gene from already annotated knowledge sources. Their disambiguation approach describes an information retrieval approach which ranks the similarity scores between the context of the surface form and the candidate gene profiles. A look at their results, however, reveals that a plain bag-of-words approach performs almost equally well.

A complex semantic disambiguation approach was introduced by Hakenberg et al. [53,54]. They identify genes by using background knowledge from EntrezGene, UniProt and GeneOnthology (GO). For each candidate ID that is assigned to a gene surface form and thus to a text, the approach tries to find all information in the text and picks the ID with the highest likelihood. To calculate the similarity based on GO terms, GO terms in the surface form context are compared with gene candidate GO terms. For each potential tuple taken from the two sets, the system calculates a distance of the terms in the ontology tree. These distances yield a similarity measure for two terms, even if they do not belong to the same sub-branch or are immediate parents/children of each other. The distance takes the shortest path via the lowest common ancestors into account, as well as the depth of this lowest common ancestor in the overall hierarchy. The distances for the closest terms from each set then define a similarity between the gene and the text [54]. The approach currently achieves an F-measure of 86.4 % on the BioCreative II gene normalization data and, thus, belongs to the best intra-species gene name disambiguation systems.

Machine Learning Approaches. There are also a few machine-learning approaches for intra-species gene ambiguity. One system is Azure, which is able to automatically assign gene names to their LocusLink[12] ID in previously unseen MEDLINE abstracts [55]. Azure contains a supervised learning approach that covers tens of thousands of genes and proteins. Apparently, it is possible to achieve high quality gene disambiguation using scalable automated techniques. Wermter et al. [52] developed GeNo, a highly competitive system for gene name normalization. The authors apply a Maximum Entropy string similarity measure for candidate retrieval and calculate a semantic similarity score for checking semantic matches. Additionally, the authors show that (i) machine learning methods perform superiorly when integrated with publicly available training data in a well-designed manner and (ii) a simple bag-of-words semantic approach to biological background knowledge performs as well as more complex semantic disambiguation [52].

Major disadvantages for machine learning and profile-based approaches are: As new biological entities are discovered very quickly, there may be no mention in the previous existing literature for that sense or for that symbol. A partial

[12] http://www.ncbi.nlm.nih.gov/Web/Newsltr/Summer99/locus.html.

solution is to perform updates to the profiles and machine learning models regularly.

5.4.3 Inter-species Gene Name Ambiguity

In inter-species gene name ambiguity tasks the species information for genes is not provided. Hence, a surface form could be the identifier of several genes or proteins across species. The disambiguation task requires the disambiguation of species first, and the resolution of the intra species gene name ambiguity in the second step (cf. Sect. 5.4.2). Species disambiguation faces the problem that multiple species assignments may be correct and that therefore multiple correct entities may exist. Hence, determining the parameter of how many results should be retrieved for each disambiguation task is a challenge. If not explicitly mentioned the proposed algorithms return a single species with the most likelihood.

Rule-Based Approaches. A simple approach to link surface forms to a species is by looking for species words in the context. More specifically, several works use one of the following rules as a baseline system [56]:

1. Previous species word: if the word preceding an entity is a species word, assign the species ID indicated by that word to the entity.
2. Species word in the same sentence: if a species word and an entity appear in the same sentence, assign its species ID to the entity. When more than one species word co-occurs in the sentence, priority is given to the species word to the entity's left with the smallest distance. If all species words occur to the right of the entity, take the nearest one.
3. Majority vote: assign the most frequently occurring species ID in the document to all entity mentions.

A well-known system to detect the species of genes in scientific publications is GNAT and was proposed by Hakenberg et al. [54]. Their approach relies on a multi-stage procedure with descending reliability to assign species to genes. For instance, a gene and a species could occur in the same phrase, including enumerations: "rat and murine Eif4g1". If no rule can be applied, the approach checks the abstract for general mentions of kingdoms, classes, etc. The system obtained one of the best performance for the Gene Normalization task in BioCreative II.

A recent approach [57] defines a three-step species disambiguation system. First, a preprocessing step including tokenization and cue word extraction for each gene surface form is performed. Second, the algorithm estimates focus species with the proposed EF-AISF coefficient, the entity frequency-augmented invert species frequency, to calculate the relevance between the cue words of a surface form and species. The species with the highest correlation coefficient is chosen as the probable focus species. Third, an appropriate species is assigned to each gene surface form with the help of the introduced *Relational Guide Factor* which enhances the capability of species assignment. An evaluation shows that

the usage of EF-AISF may significantly outperform other (machine-learning) approaches like SVMs in the task of entity species disambiguation.

Wang et al. [58] introduced and compared a number of rule-based and machine-learning based approaches to resolve species ambiguity in mentions of biomedical named entities, and demonstrated that a hybrid method achieves the best overall accuracy at 71.7 %, as tested on the gold-standard ITI-TXM corpora [59]. The authors performed multiple species assignments and investigated the average rank of the first correct species annotation.

They also introduced a hybrid species information tagging system (a combination between rule-based and machine learning approach), which improved the rule-based term identification system by up to 10 % [58].

Machine Learning Approaches. The authors of [60] describe a generic approach to disambiguate specific entity classes (e.g. species). Instead of classifying each individual occurrence of an entity, it classifies pair-wise relations between the surface form in question and the cue words in its adjacent context, where each cue word is assumed to bear a semantic class (e.g. a specific species). If a cue word features a "positive" relation with the surface form, the corresponding semantic tag of the cue word is assigned to the surface form. While an individual surface form may belong to a large number of semantic classes, a relation can only take one of two values: positive or negative, hence transforming a complex multi-classification problem into a less complicated binary classification task. The binary classification problem was solved with Support Vector Machines. One drawback of the relation classification systems is that they cannot cover all surface forms but only the ones with informative keywords co- occurring in the same sentence. The authors overcame that drawback by using spreading rules [60].

The approach by Harmston et al. [61] transforms a MEDLINE record into a mixture of adjacency matrices. By performing a random walk over the resulting graph, the authors are able to perform multi-class supervised classification, allowing the assignment of taxonomy identifiers to individual gene mentions. This method does not require training data for all potential classes in order to achieve high performance and does not only perform classification but also provides a probability, which serves to quantify the certainty attached to a classification. This species disambiguation approach shows significant improvements over the relation method proposed by Wang et al. [60]. Once the reliable corpora are in place, the approach can be applied in an automatic fashion without any user intervention, which will aid its employment in the context of novel organisms [61].

Wang et al. [56] compared a parser-based (e.g. Stanford parser), a supervised multi-classification [58] and a relation-based [60] species disambiguation approach. Promising results are obtained by training a machine learning model on syntactic parse trees, which is then used to decide whether an entity belongs to the model organism denoted by a neighboring species-indicating word (e.g. yeast). The parser-based approaches are also compared with a supervised classification method and results indicate that the former are a favorable choice when

domain portability is of concern. The best overall performance was obtained by combining the strengths of a syntactic parser (i.e. ENJU-Genia), a relation classification model, and a supervised classification model. Their method does not function well if no species term co-occurs with the gene mentions in a sentence. Similarly, the method cannot handle the articles that lack species mentions.

A comparison between rule-based and machine learning approaches shows that machine learning approaches attain satisfying results. However, the availability of training data is often limited, and the available data sets tend to be imbalanced and, in some cases, heterogeneous.

6 Open Problems

This chapter lists the open problems for linking biomedical data to the cloud, categorized into problems with the data (Sect. 6.1) and algorithm-related problems (Sect. 6.2).

6.1 Dataset Related Problems

Annotated corpora for training linking algorithms contain surface forms linked to entities from different KBs and namespaces (e.g. Uniprot, UMLS, SnomedCT). This implies that algorithms trained on one specific corpus with its respective KBs are only able to link to these KBs. Depending on the application scenario, however, references to different KBs might be required. Although the Semantic Web standard accounts for connections between two repositories in the Linked Data Cloud by special types of relations, e.g. the `owl:sameAs` relation, the majority of the biomedical linked data repositories (65.71 %) is not linkable to other repositories (see Sect. 5.1). Thus, an open problem is the missing links between the various available repositories, also termed ontology alignment. High-quality automatic ontology alignment is still an open problem, while semi-automatic approaches seem to yield promising results [62], but require considerable human effort.

Further problems root in the missing provenance and licensing information of the Linked Data Cloud repositories. As described in Sect. 5.1 for the life sciences domain, only 3.37 % of the data sets provide licensing information in RDF and pose a challenge for fully automatic exploitation of these KBs. Applications using Linked Data repositories rely on the actuality and correctness of the represented knowledge, but only the minority of the life sciences data sets in the Linked Data Cloud (23.60 %) contain provenance information.

6.2 Algorithm Related Problems

Analyzing available disambiguation algorithms shows three major, important and open problems which have been addressed insufficiently so far.

Inter-domain Entity Disambiguation. Scientific literature is being published in various domains (e.g. biomedical, computer science domain). Consequently, these documents comprise entities from different domains. Generally, existing disambiguation systems are able to disambiguate entities belonging to a specific domain, either generic entities as available in Wikipedia or special knowledge entities (e.g. biomedical entities). Zwicklbauer et al. [23] showed that large-scale and heterogeneous entity KBs may mitigate disambiguation results significantly. An open problem is how different entity repositories from different domains can be combined while providing reliable disambiguation results.

Supervised or Unsupervised Classification. Disambiguation tasks (i.e. intra-species and inter-species gene name ambiguity) may be interpreted as classification tasks. Thus, many approaches rely on supervised classification, which needs a non-negligible amount of training data. The availability of training data is often limited, and the available data sets tend to be imbalanced and, in some cases, heterogeneous [60]. Another problem of making extensive use of training data is that new biological entities are discovered very quickly. There may be no surface form in the previous existing literature for that sense or for that symbol [51]. Unsupervised or rule-based algorithms are either not available or do not provide similar results as supervised algorithms [58]. The question remains how algorithms provide reliable results despite requiring less or no training data.

Multiple Species Assignments. As shown in Sect. 3, a surface form of genes or proteins may belong to several different species (e.g. the proteins in sentence "human and mouse CD200R-CD4d3+4 and rCD4d3+4 protein" belong to the species human, mouse and rat). Hence, these surface forms refer to multiple entities. Existing algorithms usually extract the corresponding species providing the highest score. Furthermore, a static threshold often denotes the top-n relevant species to be extracted. However, existing approaches lack algorithms to investigate how many and which species belong to surface forms of genes or proteins.

7 Conclusion and Outlook on Future Work

Biomedical entity disambiguation has benefited from substantial interest from researchers and from practical needs of several domains (e.g. smart hospitals, infectious disease researchers), especially in the last ten years. In this work we provide an overview of biomedical entity disambiguation, with a special focus on annotated corpora, term disambiguation algorithms as well as gene and protein disambiguation algorithms.

As stated in the section above, there is a need for disambiguation systems for entities across several domains (e.g. entities from computer science and biomedical domain). A first important step would be to investigate how to combine two KBs, comprising entities from different domains, without mitigating disambiguation results due to an increase of heterogeneity and quantity.

Another important direction to add more flexibility to disambiguation systems is in reducing the necessity of training data by intelligent algorithm design and data exploitation. Most works are built upon supervised algorithms and need a huge amount of annotated data sets. Promising approaches avoid using expensive manually annotated data for each new domain and thus achieve better portability, e.g. [60].

With the entity linking approaches becoming more and more sophisticated, the application tasks shift to more complex recognition tasks. This shift can for instance, be observed in the community challenges issued by the BioNLP consortium. Starting with 2011, the event detection task additionally involved co-reference resolution and relation identification, and assumed a correct entity disambiguation system as prerequisite [8].

Acknowledgments. The presented work was developed within the EEXCESS project funded by the European Union Seventh Framework Programme FP7/2007–2013 under grant agreement number 600601.

References

1. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., Verspoor, K.: Biomedical text mining: state-of-the-art, open problems and future challenges. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. LNCS, vol. 8401, pp. 271–300. Springer, Heidelberg (2014)
2. Gantz, J., Reinsel, D.: Extracting value from chaos. Technical report. IDC iview (2011)
3. Holzinger, A.: On Knowledge Discovery and Interactive Intelligent Visualization of Biomedical Data - Challenges in Human-Computer Interaction and Biomedical Informatics. INSTICC, Rome (2012)
4. Piateski, G., Frawley, W.: Knowledge Discovery in Databases. MIT press, Cambridge (1991)
5. Holzinger, A., Jurisica, I.: Knowledge discovery and data mining in biomedical informatics: the future is in integrative, interactive machine learning solutions. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. LNCS, vol. 8401, pp. 1–18. Springer, Heidelberg (2014)
6. Davis, A.P., Grondin, C.J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B.L., Wiegers, T.C., Mattingly, C.J.: The comparative toxicogenomics database's 10th year anniversary: update 2015. Nucleic acids research (2014)
7. Kim, J.D., Pyysalo, S.: Bionlp shared task. In: Dubitzky, W., Wolkenhauer, O., Cho, K.H., Yokota, H. (eds.) Encyclopedia of Systems Biology, pp. 138–141. Springer, New York (2013)
8. Pyysalo, S., Ohta, T., Rak, R., Sullivan, D., Mao, C., Wang, C., Sobral, B., Tsujii, J., Ananiadou, S.: Overview of the ID, EPI and REL tasks of BioNLP shared task 2011. BMC Bioinform. **13**(Suppl 11), S2 (2012)
9. Krell, T., Lacal, J., Busch, A., Silva-Jiménez, H., Guazzaroni, M.E., Ramos, J.L.: Bacterial sensor kinases: diversity in the recognition of environmental signals. Annu. Rev. Microbiol. **64**, 539–559 (2010)
10. Krauthammer, M., Nenadic, G.: Term identification in the biomedical literature. J. Biomed. Inform. **37**(6), 512–526 (2004). Named Entity Recognition in Biomedicine

11. Kulkarni, S., Singh, A., Ramakrishnan, G., Chakrabarti, S.: Collective annotation of wikipedia entities in web text. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD 2009, pp. 457–466. ACM, New York, NY, USA (2009)
12. Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: Proceedings of the 16th Conference on Computational Linguistics, COLING 1996, vol. 1, pp. 466–471. Association for Computational Linguistics, Stroudsburg, PA, USA (1996)
13. Gentile, A.L., Zhang, Z., Xia, L., Iria, J.: Semantic relatedness approach for named entity disambiguation. In: Agosti, M., Esposito, F., Thanos, C. (eds.) IRCDL 2010. CCIS, vol. 91, pp. 137–148. Springer, Heidelberg (2010)
14. Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 708–716. Association for Computational Linguistics, Prague, Czech Republic (2007)
15. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, CIKM 2007, pp. 233–242. ACM, New York, NY, USA (2007)
16. Limaye, G., Sarawagi, S., Chakrabarti, S.: Annotating and searching web tables using entities, types and relationships. Proc. VLDB Endow. **3**(1–2), 1338–1347 (2010)
17. Wacholder, N., Ravin, Y., Choi, M.: Disambiguation of proper names in text. In: Proceedings of the Fifth Conference on Applied Natural Language Processing, ANLC 1997, pp. 202–208. Association for Computational Linguistics, Stroudsburg, PA, USA (1997)
18. Marsh, E., Perzanowski, D.: Muc-7 evaluation of ie technology: overview of results. In: Proceedings of the Seventh Message Understanding Conference (MUC-7) (1998)
19. Campos, D.: Srgio Matos. Theory and Applications for Advanced Text Mining, J.L.O. (2012)
20. Bagga, A., Baldwin, B.: Entity-based cross-document coreferencing using the vector space model. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL 1998, vol. 1, pp. 79–85. Association for Computational Linguistics, Stroudsburg, PA, USA (1998)
21. Chen, L., Liu, H., Friedman, C.: Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics **21**(2), 248–256 (2005)
22. Ogden, C., Richards, I.A.: The Meaning of Meaning: a Study of the Influence of Language Upon Thought and of the Science of Symbolism, 8th edn. Harcourt Brace Jovanovich, New York (1923). Reprint
23. Zwicklbauer, S., Seifert, C., Granitzer, M.: Do we need entity-centric knowledge bases for entity disambiguation? In: Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies. i-Know 2013, pp. 4:1–4:8. ACM, New York, NY, USA (2013)
24. Kim, J.D., Ohta, T., Tateisi, Y., Tsujii, J.: Genia corpusa semantically annotated corpus for bio-textmining. Bioinformatics **19**(suppl 1), i180–i182 (2003)
25. Yeh, A., Morgan, A., Colosimo, M., Hirschman, L.: Biocreative task 1a: gene mention finding evaluation. BMC Bioinform. **6**(Suppl 1), S16 (2005)

26. Smith, L., Tanabe, L., Johnson nee Ando, R., Kuo, C.J., Chung, I.F., Hsu, C.N., Lin, Y.S., Klinger, R., Friedrich, C., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C., Povinelli, R., Vlachos, A., Baumgartner, W.A., Hunter, L., Carpenter, B., Tzong-Han Tsai, R., Dai, H.J., Liu, F., Chen, Y., Sun, C., Katrenko, S., Adriaans, P., Blaschke, C., Torres, R., Neves, M., Nakov, P., Divoli, A., Maa-Lpez, M., Mata, J., Wilbur, W.: Overview of biocreative II gene mention recognition. Genome Biol. **9**(Suppl 2), S2 (2008)
27. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A.: Overview of the chemical compound and drug name recognition (chemdner) task. In: BioCreative Challenge Evaluation Workshop, vol. 2. (2013)
28. Van Auken, K., Schaeffer, M.L., McQuilton, P., Laulederkind, S.J., Li, D., Wang, S.J., Hayman, G.T., Tweedie, S., Arighi, C.N., Done, J. et al.: Corpus construction for the biocreative IV go task. In: Proceedings of the BioCreative IV workshop, Bethesda, MD, USA (2013)
29. Rebholz-Schuhmann, D., Yepes, A.J.J., Van Mulligen, E.M., Kors, J., Milward, D., Corbett, P., Buyko, E., Beisswanger, E., Hahn, U.: Calbc silver standard corpus. J. Bioinform. Comput. Biol. **8**(01), 163–179 (2010)
30. Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W.A., Cohen, K., Verspoor, K., Blake, J., Hunter, L.: Concept annotation in the craft corpus. BMC Bioinform. **13**(1), 161 (2012)
31. Tsuruoka, Y., McNaught, J., Tsujii, J., Ananiadou, S.: Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. Bioinformatics **23**(20), 2768–2774 (2007)
32. Smith, L.H., Yeganova, L., Wilbur, W.J.: Hidden markov models and optimized sequence alignments. Comput. Biol. Chem. **27**(1), 77–84 (2003)
33. Cohen, W., Minkov, E.: A graph-search framework for associating gene identifiers with documents. BMC Bioinform. **7**(1), 440 (2006)
34. Winkler, W.E.: String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In: Proceedings of the Section on Survey Research, pp. 354–359 (1990)
35. Rudniy, A., Song, M., Geller, J.: Mapping biological entities using the longest approximately common prefix method. BMC Bioinform. **15**, 187 (2014)
36. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. **24**(5), 513–523 (1988)
37. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**(3), 443–453 (1970)
38. Yu, H., Kim, W., Hatzivassiloglou, V., Wilbur, W.J.: Using medline as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. J. Biomed. Inform. **40**(2), 150–159 (2007)
39. Yu, H., Hripcsak, G., Friedman, C.: Mapping abbreviations to full forms in biomedical articles. JAMIA **9**(3), 262–272 (2002)
40. Pustejovsky, J., Castaño, J., Saurí, R., Rumshinsky, A., Zhang, J., Luo, W.: Medstract: Creating large-scale information servers for biomedical libraries. In: Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain, BioMed 2002, vol. 3, pp. 85–92. Association for Computational Linguistics, Stroudsburg, PA, USA (2002)
41. Pakhomov, S.: Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL 2002, pp. 160–167. Association for Computational Linguistics, Stroudsburg, PA, USA (2002)

42. Chen, P., Al-Mubaid, H.: Context-based term disambiguation in biomedical literature. In: Proceedings of the 19th International FLAIRS conference FLAIRS Conference, pp. 62–67 (2006)
43. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
44. Spärk Jones, K., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments. Inf. Process. Manage. **36**(6), 493–502 (2000)
45. Morgan, A.A., Lu, Z., Wang, X., Cohen, A., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H.H., Torres, R., Krauthammer, M., Lau, W., Liu, H., Hsu, C.N., Schuemie, M., Cohen, K.B.: Overview of biocreative ii gene normalization. Genome Biol. **9**(Suppl 2), S13 (2008)
46. Hatzivassiloglou, V., Dubou, P.A., Rzhetsky, A.: Disambiguating proteins, genes, and RNA in text: a machine learning approach. In: ISMB (Supplement of Bioinformatics), pp. 97–106 (2001)
47. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
48. Ginter, F., Boberg, J., Järvinen, J., Salakoski, T.: New techniques for disambiguation in natural language and their application to biological text. J. Mach. Learn. Res. **5**, 605–621 (2004)
49. McEntyre, J., Lipman, D.: PubMed: bridging the information gap. CMAJ Can. Med. Assoc. J. (journal de l'Association medicale canadienne) **164**(9), 1317–1319 (2001)
50. Pahikkala, T.: Filip Ginter, J.B.: Contextual weighting for support vector machines in literature mining: an application to gene versus protein name disambiguation. BMC Bioinform. **6**(1), 157 (2005)
51. Xu, H., Fan, J.W., Hripcsak, G., Mendonça, E.A., Markatou, M., Friedman, C.: Gene symbol disambiguation using knowledge-based profiles. Bioinformatics **23**(8), 1015–1022 (2007)
52. Wermter, J., Tomanek, K., Hahn, U.: High-performance gene name normalization with geno. Bioinformatics **25**(6), 815–821 (2009)
53. Hakenberg, J., Plake, C., Royer, L., Strobelt, H., Leser, U., Schroeder, M.: Gene mention normalization and interaction extraction with context models and sentence motifs. Genome Biol. **9**(Suppl 2), S14 (2008)
54. Hakenberg, J., Plake, C., Leaman, R., Schroeder, M., Gonzalez, G.: Inter-species normalization of gene mentions with GNAT. In: ECCB, pp. 126–132 (2008)
55. Podowski, R.M., Cleary, J.G., Goncharoff, N.T., Amoutzias, G., Hayes, W.S.: Azure, a scalable system for automated term disambiguation of gene and protein names. In: CSB, pp. 415–424. IEEE Computer Society (2004)
56. Wang, X., Tsujii, J., Ananiadou, S.: Disambiguating the species of biomedical named entities using natural language parsers. Bioinformatics **26**(5), 661–667 (2010)
57. Hsiao, J.C., Wei, C.H., Kao, H.Y.: Gene name disambiguation using multi-scope species detection. IEEE/ACM Trans. Comput. Biol. Bioinform. **11**(1), 55–62 (2014)
58. Wang, X., Matthews, M.: Distinguishing the species of biomedical named entities for term identification. BMC Bioinform. **9**(Suppl 11), S6 (2008)
59. Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., Roebuck, S., Tobin, R., Wang, X.: The ITI TXM corpora: tissue expressions and protein-protein interactions. In: Proceedings of LREC, vol. 8, Citeseer (2008)

60. Wang, X., Tsujii, J., Ananiadou, S.: Classifying relations for biomedical named entity disambiguation. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, vol. 3, pp. 1513–1522. Association for Computational Linguistics, Stroudsburg, PA, USA (2009)
61. Harmston, N., Filsell, W., Stumpf, M.P.H.: Which species is it? Species-driven gene name disambiguation using random walks over a mixture of adjacency matrices. Bioinformatics **28**(2), 254–260 (2012)
62. Sabol, V., Kow, W.O., Rauch, M., Ulbrich, E., Seifert, C., Granitzer, M., Lukose, D.: Visual ontology alignment system - an evaluation. In: Proceedings of SIGRAD (2012)