J Braz Comput Soc (2011) 17:103–116
DOI 10.1007/s13173-011-0031-9

ORIGINAL PAPER

A systematic review of named entity recognition in biomedical texts

Rodrigo Rafael Villarreal Goulart · Vera Lúcia Strube de Lima · Clarissa Castellã Xavier

Received: 5 August 2010 / Accepted: 3 February 2011 / Published online: 2 March 2011

Abstract Biomedical Named Entities (NEs) are phrases or combinations of phrases that denote specific objects or groups of objects in the biomedical literature. Research on Named Entity Recognition (NER) is one of the most disseminated activities in the automatic processing of biomedical scientific articles. We analyzed articles relevant to NER in biomedical texts, in the period from 2007 to 2009, through a systematic review. The results identify the main methods in the recognition of Biomedical NEs, features and methodologies for a NER system implementation. Aside from the tendencies identified, some gaps are detected that may constitute opportunities for new studies in the area.

Keywords Parsing · Named entity · Recognition · Biomedical

1 Introduction

A Biomedical Named Entity (NE) is a phrase or combination of phrases in a document that denotes a specific object or a group of objects in the biomedical literature [1]. Objects could be genes, proteins, drugs, etc. NF2, for example, refers to a human gene, according to Chen et al. [2]. The automatic processing of texts, including the indexing of scientific articles, paragraphs, sentences, and clauses, as well as the understanding of relations among parts of the text, make use of NE recognition. However, identifying and understanding the meaning of these terms is not a trivial task. In the previous example, just one capital letter distinguishes that entity from a rat gene (Nf2). But the term NF2 is also a gene, the protein that it produces, and the illness resulting from its mutation. These multiple meanings of terms (phrases) associated to polysemous genes come to 14.2% in the 23 species analyzed in [2].

Scientific research uses digital media to disseminate results. The strategic reading of articles has become a necessity, and an infrastructure for it has been developing over the past 20 years. The task of information retrieval is dependent on text processing and the problems that arise from it. This automatic processing undergoes NE recognition (NER—Named Entity Recognition) and then indexing of articles, paragraphs, sentences, clauses, etc.

Since NER is a subject that has received special attention in the area of Natural Language Processing (NLP) as an important preprocessing tool [3], this article presents the results of a systematic review to identify and analyze experimental proposals for NER. Biolchini et al. [4] define systematic review as "a specific methodology of research, developed in order to gather and evaluate the available evidence pertaining to a focused topic." A detailed presentation of the systematic review methodology, including its origins, can be found in [4] and [5].

The Named Entity Recognition in Biomedicine is the topic of this research. The research question of this study is: How is NER carried out in biomedical texts? In order to specify the research question, we had the following complementary questions:

R.R.V. Goulart (✉)
Feevale University, Novo Hamburgo, Brazil
e-mail: rodrigo@feevale.br

V.L. Strube de Lima · C.C. Xavier
PUCRS, Porto Alegre, Brazil

V.L. Strube de Lima
e-mail: vera.strube@pucrs.br

C.C. Xavier
e-mail: clarissa.xavier@pucrs.br

Question 1: What is the main method used in experimental NER of biomedical corpora?

Question 2: With respect to the main method, what are the features and methodologies employed in NER? What are the results?

Question 3: With respect to the main method, is there a corpus-independent methodology that can be freely employed?

In the next section, the research question and the basic information used in the systematic review are presented. Section 2 also establishes the criteria for source selection and article selection, for subsequent information extraction. Section 3 reports article searches and the information retrieved. Results and an analysis of this information are presented in Sect. 4. Tendencies (findings) and possible opportunities (gaps) are presented in Sect. 5. Finally, Sect. 6 presents the conclusions of this review.

2 Methodology and application

2.1 Methodology steps

To find scientific articles[1] related to the research questions, a topic sentence has been established to identify the main concepts. Based on "Named Entity Recognition and classification in Biomedicine texts," concepts, their synonyms, and some related terms were extracted. The result is a list of words presented in Table 1.

Articles and data the authors of this article had access before this review represent the control group. The authors had access to the corpus used in the bioentity recognition task from the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA) [6]. Despite not being a synonym for NE, the term "event" was added as a related term. The event recognition is an important topic in the 2009 edition of the JNLPBA, called BioNLP 2009, and points to the identification of relationships between biomedical NEs.

Table 1 Keywords, synonyms and related terms

Keywords	Synonyms and related terms
named entity	named entities, event
recognition	identification
classification	categorization
biomedicine	biomedical
text	corpus, corpora

[1] We are working only with articles in English. We consider that is the language that presents the most representative publication in the investigated subject.

For the source selection, two aspects are obligatory. Only the sources that provide access to the full article are selected, i.e., sources that offer access only to the abstracts are not considered. Sources that index third party articles are also excluded. The reason for this choice is to avoid sources with metasearches, e.g., Google Scholar, which includes results of an undetermined set of sources.

Moreover, the exclusive search in titles and abstracts is a desirable functional characteristic of the source. The objective here is to avoid the sources where research terms appear in the full text of the article, which can bring texts unassociated with the focus question or those employing terms with multiple meanings from multiple domains. This way, four sources were selected that fulfill these selection criteria, and present international recognition and circulation. They are:

- IEEE: IEEEXplore[2]
- ACM: ACM Digital Library[3]
- BMC: BioMed Central[4]
- OJ: Oxford Journals[5]

The search strings, presented in Table 2, are based on the research questions, keywords (Table 1) and their variations. The S1 string employs all of the terms related to the research question presented in Table 2. The S2, S3, and S4 strings are subsets of S1, where event, corpus, and biomedicine are terms respectively excluded in a cumulative manner.

The account of articles retrieved by a string is filtered by two exclusion criteria. They are:

- *C1*: excludes journal covers, tables of contents from annals, and any other document that is not a scientific article;

Table 2 Search Strings

Id	String
S1	(named entit* OR event) (classification OR recognition OR identification OR mention OR categorization) (biomedical OR biomedicine) (text OR texts OR corpora OR corpus)
S2	named entit* (classification OR recognition OR identification OR mention OR categorization) (biomedical OR biomedicine) (text OR texts OR corpora OR corpus)
S3	named entit* (classification OR recognition OR identification OR mention OR categorization) (biomedical OR biomedicine)
S4	named entit* (classification OR recognition OR identification OR mention OR categorization)

[2] http://ieeexplore.ieee.org/

[3] http://portal.acm.org/dl.cfm

[4] http://www.biomedcentral.com/

[5] http://www.oxfordjournals.org/

- *C2*: excludes articles not related to the context of the review (by reading of the title and, in the case of insufficient clarification, the abstract).

Using criteria C1 and C2, the articles retrieved were object of a preliminary reading, based on their titles and abstracts. During the reading, the articles were classified according to a hierarchical set of categories called *general categories*. The objective of this set of categories is to classify the articles according to those concepts related to the research question. The concepts are based on the authors previous knowledge on the topic and also on those concepts collected during the preliminary reading of the articles.

2.2 Application

The categories and the definitions for each concept are explained in the following paragraphs and summarized in Fig. 1.

```
1 Task
1.1 Anaphor resolution
1.2 Annotation
1.3 Disambiguation
1.4 Normalization
1.5 Recognition, identification or
    extraction
  1.5.1 NEs
  1.5.2 Events
  1.5.3 Relationships
  1.5.4 Facts and speculations
  1.5.5 Conflicts of interest
1.6 Classification or categorization
  1.6.1 NEs
  1.6.2 Events
  1.6.3 Texts
1.6.4 Relationships
2 Method
  2.1 Machine Learning Based
  2.2 Dictionary Based
  2.3 Rules Based
  2.4 Regular Expressions
  2.5 Other
3 Type of corpus
  3.1 Medical
  3.2 Biomedical
  3.3 No identification
4 Type of article
  4.1 Survey
  4.2 Overview
  4.3 Experimental
  4.4 System, framework, platform or
      corpus description
```

Fig. 1 General categories to classify the 130 articles selected

1. *Task*: the task refers to the type of problem discussed in the article.
   (a) *Anaphor resolution*: anaphora are references to objects in a text, e.g., "it" when referring to a certain gene. The process involving anaphora comprehension, with the identification of its two or more parts (e.g., "it" and the specific gene it refers to) and the representation for such a construction in a text, is called Anaphora Resolution.
   (b) *Annotation*: refers to corpora annotation as an objective as well as a procedure.
   (c) *Disambiguation*: ambiguous NEs in Biomedicine are understood as those which are similar to common words, such as English words for example, or those with multiple meanings in their domain [7]. These NEs can represent names of genes, molecules or chemical formula [8, 9]. This problem is treated as disambiguation.
   (d) *Normalization*: despite Hakenberg definition of normalization as a disambiguation procedure [10], in the present work, normalization is understood as the process of relating NEs mentioned in the texts to entries in databases structured with biological data, such as ontologies or genetic sequencing databases [11–13].
   (e) *Recognition, identification, or extraction*: the terms recognition, identification, and extraction are treated as synonyms in this taxonomy. However, they are subclassified as recognition, identification, or extraction of NEs, events, relationships, facts and speculations, or conflicts of interest. Events, specifically biological events, are biological processes such as gene expression, transcription, or regulation, among others [14]. Relationships, according to Shi et al. in [15], are functional biomedical relationships between specific concepts of the domain, for example, a bacterium, a protein, and a location (e.g., cytoplasm or membrane). However, for Liu et al. in [16], relationships can be characterized as interactions between proteins or genes, and this meaning is also given to the concept of event in this taxonomy. Facts or speculations, according to Cohen in [17], are the identification of the negation of events (e.g., '...the event *X* did not occur ...') or speculations about them (e.g., '...I affirm that event *X* should occur...'). Finally, conflicts of interest are situations in which opinions in a relationship between two parts can be conflicting. According to Aleman-Meza et al. [18], these situations can appear in personal relationships, in the elaboration of data structures, or texts of a domain.
   (f) *Classification or categorization*: the main objective of studies on classification or categorization is to relate phrases to a determined class (i.e., proteins,

genes, etc., in case they are classified as NEs) [14]. The classification or categorization task can be specified in the classification of NEs, events, texts, or relationships. The classification of events and relationships (entries 1.6.2 and 1.6.4 in Fig. 1) refers to the same idea of recognition, identification, or extraction as explained for item (e) above (entry 1.5 in Fig. 1).

2. *Methods*: computational method referred to in the article.
   (a) *Machine Learning Based*: according to Alpaydin [19] "Machine Learning (ML) is programming computers to optimize a performance criterion using example data or past experience." This method refers to the existing models of ML [20], e.g., Support Vector Machines (SVM), Hidden Markov Model (HMM) and Conditional Random Fields (CRF) presented in [21, 22], and [23], respectively.
   (b) *Dictionary Based*: the dictionary based method [20] refers to the use of word lists or databases containing NEs, with the purpose of comparison and identification.
   (c) *Rules Based*: rule-based systems [20] are used, for example, in the extraction of events based on the coocurrence of strings, prefixes, or syntactic tags.
   (d) *Regular expressions*: strings with particular prefixes, suffixes, or substrings can identify, for example, NEs or parts of certain events.
   (e) *Other*: a method is mentioned but not named in the abstract or title.
3. *Type of corpus*: determines the domain of the corpus created or used. A corpus is a set of texts organized with a specific purpose of study. In the majority of cases, a corpus is enriched with labels that allow studying it in detail, and the process of associating labels or tags with phrases, terms, or other parts of the corpus is named annotation.
   (a) *Medical*: studies using or building corpora in the medical domain.
   (b) *Biomedical*: studies using or building corpora in the biomedical domain.
4. *Type of article*: the type of article expresses the objective of the article.
   (a) *Survey*: articles that are presented as surveys. Survey is defined by Groves et al. [24] as "a systematic method for gathering information from (a sample of) individuals for the purposes of describing the attributes of the larger population of which the individuals are members."
   (b) *Overview*: articles that are presented as overviews. According to IEEE [25], "overview articles are intended to be of solid technical depth and lasting value and should provide advanced readers with a thorough overview of a field of interest."
   (c) *Experimental*: articles that report scientific experiments, with objectives, methodology, and results.
   (d) *System, framework, platform or corpus*: articles that describe NLP resources.

If it was not possible to identify a task, a method, a type of corpus, or a type of article, we labeled it "no classification." Articles can be classified in multiple categories, i.e., an article can be classified according to the method, such as Machine Learning based or dictionary based. Reading only the title and the abstract can lead to an erroneous classification of articles, which is the case for [10], which was classified as normalization. Upon a complete reading, it was understood that it should have been classified as disambiguation. Also, the use of dictionaries, for example, can be part of a more complex strategy where these data structures are employed in conjunction with other methods. In these cases, the identification of a method was done according to the emphasis of the paper which presented it.

For information extraction, the following conditions were established to select articles before reading the entire text:

- Articles should investigate NER, i.e., even if the main focus of the article is text classification, it should use and discuss the influence of NEs or NER on this task.
- The articles were subclassified in a new set of categories, presented in Fig. 2, called *specific categories*. The objective here is to highlight the information relevant to the context of the research question. The main concepts in this classification were:
  - *Methodology*: the methodology highlights the experimental model, with respect to the implementation of a system for NER. The methodologies are:
    - Cascaded: use of a sequence of increasingly more complex classifiers in cascade, e.g., performing the tasks of segmentation and classification separately [26].
    - External Resources: NER is one of the steps/tasks of the experiment and not the main goal. The use of external databases, such as Flybase in [22], is a curated alternative to build annotated corpora or ML training sets.
    - Multiagent: multiagent negotiation framework for integrating several agents to cooperate effectively to solve a problem [27].
    - Unified: experiments that make recognition of tokens/words and NE classification in a single step are called Unified, as shown in [15] where NEs and BLPs are simultaneously extracted.
  - *Purpose*: the objective is to specify the main task exposed by the article, such as NER, text classification, etc.
  - *Corpus*: indicates the corpus or the corpora involved in the experiments. The four main corpus are:

```
1) Methodology
  1.1 Cascaded
  1.2 External resources
  1.3 Multiagent
  1.4 Unified
2) Purpose
  2.1 Corpora Annotation
  2.2 NER
  2.3 Relationship recognition
  2.4 Relationship extraction
  2.5 Text classification
3) Corpus
  3.1 GENIA
  3.2 Flybase
  3.3 Medline
  3.4 PubMed
  3.5 RSC
  3.6 TREC Genomics
4) NE type
  4.1 BPL (Bacteria/Protein/Location)
  4.2 Formula
  4.3 GENIA
  4.4 Gene
  4.5 Molecules
  4.6 Mutation
  4.7 Protein
5) Features
  5.1 POS
  5.2 Unigram
  5.3 Bigram
  5.4 Trigram
  5.5 Ngram(n>3)
  5.6 Prefixes/Suffixes
  5.7 Frequency
  5.8 Class (preprocessing)
  5.9 Brief Word Class (BWC)
  5.10 Bag of Words (BOW)
  5.11 Chunks
  5.12 Morphologic
  5.13 Orthographic
```

Fig. 2 Specific Categories to classify the 13 selected articles

- GENIA: GENIA[6] corpus aims to provide reference materials to let NLP techniques work for bio-text mining [28]. The 3.0 version of the corpus consist of 1,999 MEDLINE[7] abstracts, more than 400,000 words and almost 100,000 annotations for biological terms. The corpus has been annotated semantically for part-of-speech, syntactic tree, and biomedical terms.
- PubMed/MEDLINE: PubMed[8] is an online database, comprising more than 20 million citations for biomedical literature from MEDLINE, life science journals, and online books. Many PubMed citations contain links to full text articles which are freely available. MEDLINE is a bibliographic database of biomedical scientific articles at National Library of Medicine (NLM).
- RSC: The Royal Society of Chemistry[9] (RSC) is an organization that aims the advance of the chemical sciences. It publishes journals, books, and databases, as well as host conferences, seminars, and workshops. Some of its texts have been annotated and used as corpus with information in the chemistry and biomedical domains. Two examples: the ART Corpus[10] consists of 225 papers manually annotated from papers from journals of the RSC Publishing; Corbett, Batchelor, and Teufel [29] have annotated a corpus of 42 chemistry papers provided from the RSC.
- Flybase: Flybase[11] is an online database of Drosophila genes and genomes. It also includes a bibliography of research on Drosophila genetics that can be used as a non-annotated corpus.

– *NE Class*: highlights the NE classes which are the object of classification in the article, e.g., GENIA classes; Bacteria, Protein, Location (BPL), etc.
– *Features*: identifies the specific features used in the experiments for NER.

3 Execution

The searches for the articles were carried out from October 26, 2009, to November 5, 2009. All results were stored in the BibTex[12] format and maintained with the use of the Jabref[13] reference manager. Quantitative data with respect to the sources are presented in Table 3. Together with the names of the sources, the total number of articles indexed by each one is indicated. The S1 string was selected as the search string in order to include the most concepts of the research question, which is the reason why the authors consider it the best to select articles.

Using the S1 string, 244 references were retrieved. Utilizing the criteria C1 and C2, 114 references were excluded. Journal covers and tables of contents represent half of the references excluded. Items not related to the research question represent the other half of the references excluded due to the multiple meanings of the words in string S1. For example, the word event refers to biomedical events, but also to software events; the word recognition refers to NER, but

[6] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA
[7] http://www.nlm.nih.gov/
[8] http://www.ncbi.nlm.nih.gov/pubmed
[9] http://www.rsc.org/
[10] http://www.ukoln.ac.uk/projects/ART_Corpus/
[11] http://flybase.org/
[12] http://www.bibtex.org/
[13] Reference manager available at http://jabref.sourceforge.net/.

Table 3 Results of preliminar articles selection

		Source				
		IEEE	ACM	BMC	OJ	Total
		2,455,406	263,029	80,951	1,167,839	3,967,225
String	S1	6	224	12	2	244
	S2	6	45	12	1	65
	S3	8	45	12	4	70
	S4	81	260	13	10	375

also to image recognition. The list of references with the remaining 130 articles is available online.[14] Table 4 reports the quantity of articles selected by the S1 string for each entry of the general categories.[15]

To answer the research questions, it was necessary to identify the main method used in the articles that do NER experimentally in biomedical corpora. This way, the articles for each method were counted and classified. We selected the papers classified as "experimental" (type of article), from the "biomedical domain" (domain of corpus) and as "recognition," identification, or extraction of NEs (task).

The result indicates, Machine Learning (ML) as the more frequent method, with 14 articles. As for the rest, 5 articles with dictionaries, 3 with rules, 3 with regular expressions, and 4 by other methods.

In fact, besides being the more frequent one, after a deeper analysis of the papers, we consider ML the main method in this population. It is directly or indirectly used in the great majority of the articles. Even inside the other methods, such as those dictionary based, there are embedded learning approaches. And the same happens when dealing with other tasks such as annotation, disambiguation, or recognition. Also present in platforms and frameworks, machine learning is undoubtedly a very important part of the solutions for NER. Nevertheless, we should notice that no objective relevance measure was previously defined to be adopted here.

With the main method identified, the 14 articles of ML were selected for reading and later classification with the specific taxonomy. The origin of this group of 14 articles can be more clearly understood with the help of Fig. 3 that illustrate graphically how they were selected through the intersection of the set of articles working with Biomedical corpus, the set of articles that uses ML method, and the set of articles that have NER as a task. Among the 14 selected articles, two refer to the same content, but were published in different sources. They are [8] and [31]. This way, we use only article [31] for this study and we discard [8], as show in Fig. 4.

[14] https://sites.google.com/site/biomedicalnesreview

[15] Some articles are classified in more than one group.

Table 4 Amount of articles selected with S1 and classified with the general taxonomy

Categories and Subcategories			Amount of articles
Task	Anaphor resolution		1
	Annotation		13
	Disambiguation		6
	Normalization		3
	Recognition, identification or extraction	NEs	38
		Events	23
		Relationships	18
		Facts and speculations	4
		Conflicts of interest	1
	Classification or categorization	NEs	10
		Events	0
		Texts	1
		Relationships	5
	No classification		25
Method	Machine Learning		56
	Dictionary		8
	Rules		15
	Regular expressions		8
	Other		6
	No classification		60
Corpus type	Medical		13
	Biomedical		75
	No classification		42
Article type	Survey		3
	Overview		5
	Experimental		94
	System, framework, platform or corpus		43
	No classification		12

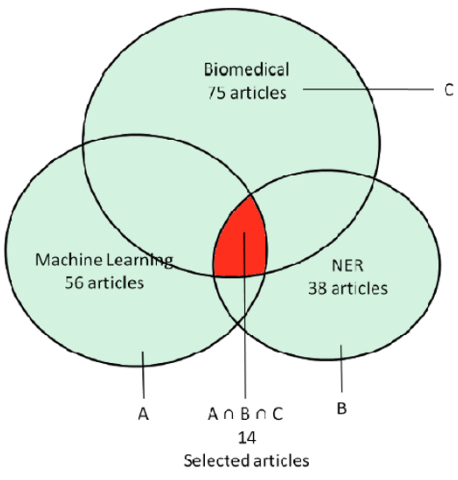

Fig. 3 The subset of 14 articles selected for a detailed analysis is obtained through the intersection of the set of articles working with Biomedical corpus, the set of articles that uses ML method and the set of articles that have NER as a task

Ref.	Source	Title	Methodology	Purpose	Corpus	NE type	Features	F	R	P
26	IEEEXplore	A Cascaded Approach to Biomedical Named Entity Recognition Using a Unified Model	Cascaded	NER	GENIA	GENIA	5.1, 5.2, 5.3, 5.4, 5.6, 5.8, 5.9, 5.13	71.95	74.63	69.45
30	Oxford Journals	Cascaded classifiers for confidence-based chemical named entity recognition	Cascaded	NER	RSC	Chemicals	5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, 5.13	80.80	82.90	78.70
					Pubmed			83.20	81.60	85.00
36	IEEEXplore	Recognizing Biomedical Named Entities in the Absence of Human Annotated Corpora	External resources	Corpora Annotation	GENIA	GENIA	5.1, 5.7, 5.12, 5.13	-	-	-
33	IEEEXplore	Two Approaches for Biomedical Text Classification	NER as part of the methodology	Text classification	TREC Genomics	-	5.10	-	-	-
32	ACM Digital Library	Extraction of biomedical events using case-based reasoning	-	Relationship extraction	-	-	-	-	-	-
14	ACM Digital Library	Biomedical event detection using rules, conditional random fields and parse tree distances	-	Relationship extraction	-	-	-	-	-	-
31	ACM Digital Library	How to make the most of NE dictionaries in statistical	Cascaded	NER	GENIA	Protein	5.1, 5.2, 5.3, 5.8, 5.13	73.78	79.85	68.58
15	ACM Digital Library	Simultaneous identification of biomedical named-entity and functional relations using statistical parsing techniques	Unified	Relationship recognition	Medline	BPL	-	-	-	-
9	ACM Digital Library	Mining, indexing, and searching for textual chemical molecule information on the web	Cascaded	NER	RSC	Molecules	5.1, 5.2, 5.3, 5.8, 5.11, 5.13	80.32	76.15	84.98
						Formulae	5.1, 5.2, 5.3, 5.11, 5.13	93.48	93.09	93.88
22	ACM Digital Library	Evaluating and combining biomedical named entity recognition systems	External resources	-	Flybase	Genes	5.1, 5.6, 5,13	81.86	75.68	89.14
27	IEEEXplore	Multi-Agent Classifiers Fusion Strategy for Biomedical Named Entity Recognition	Multiagent	NER	GENIA	GENIA	5.1, 5.2, 5.3, 5.4, 5.6, 5.7, 5.8, 5.11, 5.12, 5.13	77.88	77.41	78.36
28	ACM Digital Library	Extraction of named entities from tables in gene mutation literature	Unified	NER	Pubmed	Mutation	5.2, 5.10, 5.11, 5.13	83.90	84.10	83.90
35	ACM Digital Library	Reranking for biomedical named-entity recognition	Cascaded	NER	GENIA	GENIA	5.1, 5.2, 5.5, 5.6, 5.8, 5.11, 5.12, 5.13	72.65	-	-

Fig. 4 Selected articles with information, ordered by author

4 Preliminary analysis and results

With the execution of searches for articles and the classification of those selected by the search string and by the characteristics associated to the research questions, results analysis are presented in Sects. 4.1 and 4.2.

4.1 Numerical analysis of the results

The first analysis presented here considers publication sources and the number of articles indexed by each of them. Figure 5(a), shows the number of articles indexed by each source during the period in which the searches were done.

The IEEExplore source holds more than half of the articles indexed by the four sources (61.89%) and has twice as many articles as Oxford Journals (29.43%), which is the second largest one. However, the ACM Digital Library is the source of the broad majority of references selected by the S1 string. Two hundred and forty-four references were retrieved using string S1 (Fig. 5(b)). Using criteria C1 and C2, 114 references were excluded, representing 46% of the results.

As expected of the characteristics of string S1, Fig. 6 highlights a large number of articles related to the recognition, identification, or extraction of NEs (38 articles). Moreover, if we include the articles related to the identification of events and relationships, we have a greater number of articles in the same period (41 articles). Events are considered a problem of recognition, identification, or extraction, and not only of classification (23 articles), e.g., [32] and [14], NEs are also considered a problem of classification (8 articles), e.g., [26] and [33]. However, it is important to emphasize that for 25 articles (19.23% of 130 articles) the type of the

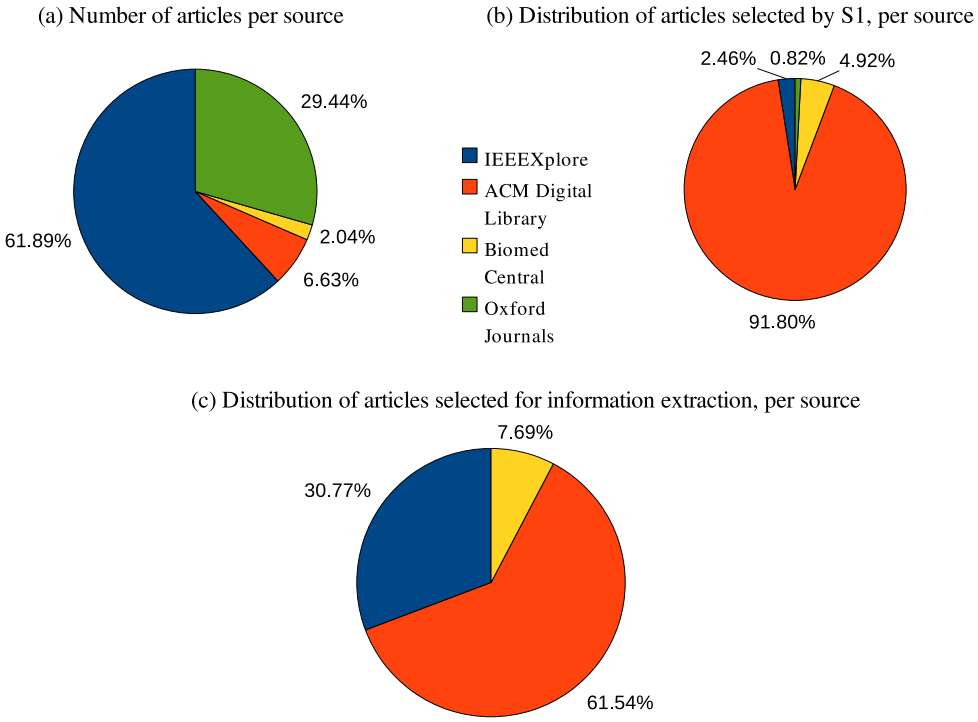

Fig. 5 Distribution of articles, per source

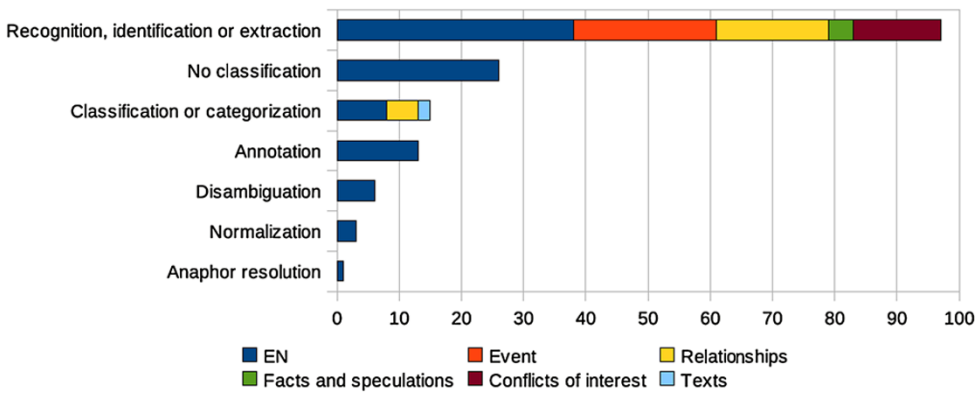

Fig. 6 Relationship among the amount of articles and types of tasks

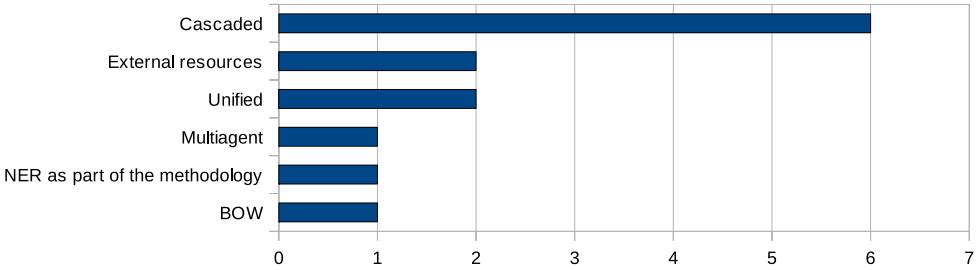

Fig. 7 Articles per methodology

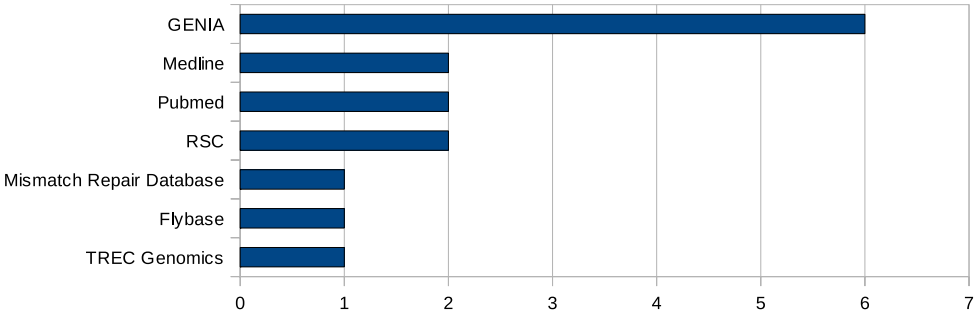

Fig. 8 Articles per corpus

article was not clearly identified by means of the title or the abstract.

The analysis of the methods indicates that the ML approach is the most cited in the abstracts or titles of articles, with 56 articles (43%). This frequency of occurrence is sometimes due to the fact that ML is the basis for several of the proposals appearing in the articles. For example, the use of the dictionaries (method = dictionary) is often associated with a ML approach. However, 60 articles (46.15%) did not explicitly identify the method used, which may point to a combination of methods or an underspecification of this information in the abstract. This question can be evaluated based on works that describe and highlight the annotation of corpora, problems, and results, or even the performance of a system. Such narratives do not exactly aid in the classification process of articles executed in this work.

The type of corpus, as expected, is dominated by biomedical texts, with 75 articles (57.69%). Despite this, 42 articles (32.30%) do not indicate in their abstract the corpus used. A large number of articles (94 articles, 72.30%) describe experiments. In a smaller proportion, 12 articles (9.23%) do not clearly mention their purpose in the abstract. Finally, 43 articles (33.07%) are presented as a description or proposal of a system, framework, platform, or corpus (so mainly describing resources or tools).

Despite being classified as a work related to NER, the article by Neves et al. [32] refers to event extraction using Case Based Reasoning (CBR). It mentions events in the abstract, calling them entities. The identification and classification of the tokens that constitute an event is done by the GENIA Tagger, in training as well as in testing. However, contribution to NER was not the subject here.

Of the 13 articles selected for information extraction by means of full text reading, some observations can be brought on methodologies, NE classes, and features. The cascaded methodology is the most common approach, representing 28.46% of the total (6 articles, Fig. 7). The cascaded methodology firstly appears in Physics: a method of attaining successively lower temperatures by utilizing the cooling effect of the expansion of one gas in condensing another less easily liquefiable, and so on.[16] Its use is more recent in NER, but the principle of successive levels of treatment is the same here: the use of a sequence of increasingly more complex classifiers in cascade [34].

The NE classes found in the articles reflect the fact that the GENIA corpus is the most referenced (6 articles, Fig. 8). Approximately 38% (5 articles, Fig. 9) use the GENIA treebank annotation schema. Word types morphosyntactic Part-of-speech (POS) tagging, orthographical rules, term frequency (TF), and Bag of Words (BOW) were other alternatives found.

4.2 Detailing Fig. 4

In Fig. 4, we find the 13 papers selected for a detailed analysis. Even if we included published results for F-score (for 10 of the selected articles) and Recall and Precision (for 9 of them), all these results cannot be directly compared. We should group articles with the same characteristics—purpose, corpus, type of the named entity they work with—before comparing results. This means that only those results by [27, 35], and [26] are comparable and, considering the

[16] Webster's Revised Unabridged Dictionary, published 1913 by C. & G. Merriam Co.

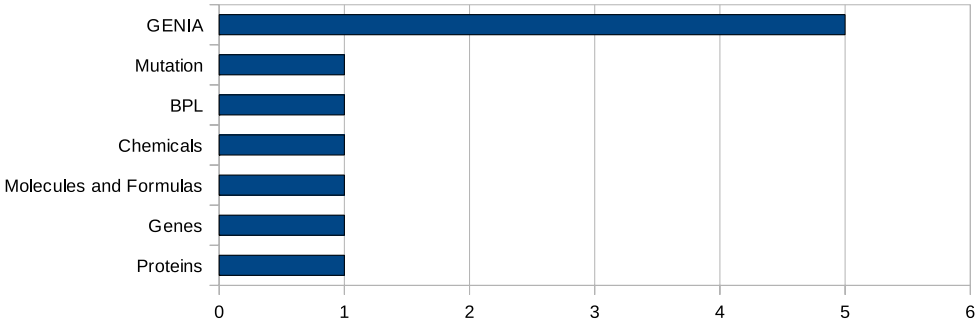

Fig. 9 Articles per NE class

F-score they report, the best results were obtained by [26]. In fact, this article brings a particular proposal, using a multiagent system to solve the NER problem. As a new paradigm for software development, multiagent systems did not get to achieve a top position if compared other alternatives like object oriented programming. However, the idea of distributed intelligence they carry may conduct to a well-organized architecture, where specialized agents (that can be simple programs working and communicating) interact in order to solve a problem. This is the case for [26].

Article [14], published in 2009, describes an application that aims at characterizing event types and their participating entities, considering a set of 9 event types (e.g., phosphorylation, protein catabolism, binding, or regulation). Depending on the event, it is also necessary to identify one or more participating proteins. This article reports a system which took part of the BioNLP'09, one of the contests in the area. This article brings important contribution to our study even if it does not provide solutions to NER. The NER methods already in use were remodeled for event detection, a task that appears as the next defy in the area. The results in the area are not yet comparable to those with NER, which points to future work to be done. Articles regarding NER for chemical entities in texts such as [9] were also studied, because of the applicability of these solutions when dealing with biomedical entities.

Paper [15] uses a statistical parsing-based method to identify NER and extract relations in form of BPL (bacterium, protein, location) triples from MEDLINE articles. It makes sintatic and semantic annotation. The authors do experiments using supervised and semisupervised methods, getting better results with the second one. We emphasize that the use of a semantic parser to identify relationships emerges as a trend in this research field.

The use of the cascaded methodology is emphasized by article [26]. The authors split the NER problem in two tasks: Segmentation (SEG) and Classification (CLASS). SEG groups all entity types of interest into one super-type. CLASS classifies each entity candidate into one of the entity types. A unified model called "maximum-entropy margin-based" (MEMB) is used in both tasks. This model considers the error between a correct and an incorrect output during training.

5 Findings, gaps, and opportunities

In this section, tendencies and opportunities identified along the review are reported. Aside from extracting information, the objective of these findings is to answer the research questions presented in Sect. 1: (1) What is the main method used in experimental NER of biomedical corpora? (2) With respect to the main method, what are the features and methodologies employed in NER? What are the results? (3) With respect to the main method, is there a corpus-independent methodology that can be freely employed?

Figure 10(a) presents the distribution of the methods with respect to the number of articles which use NER experimentally in the biomedical domain. There are 14 articles making use of ML, 5 articles using dictionary based techniques (together with ML), 3 articles using rule based techniques, and 3 using regular expressions. The remaining 4 articles represent other methods not considered in the general categories. Figure 10(a) helps in driving us to the same conclusion as Gu, Dahl, and Popowich in [36], already pointed out by Jurafsky and Martin in [22] placing Machine Learning as the most used method in recent years. As stated in [36] when studying Biomedical NER, "in recent years, supervised learning techniques have become dominant, with better performance and adaptability." The ML method appears as an undoubtable choice for those who start working in the area of NER. It is perfected with the Conditional Random Fields (CRF) alternative, which is already in use in many of the solutions studied. However, tuning the features is still an evolving area. We understand that this answers research question 1 for the limits of this review.

With the objective of providing information to answer research question 2, Fig. 4 reports methodologies, features, and F-score identified in the experiments presented in the 13 articles selected.

The number of articles selected for a review is an important point to be considered when making conclusions. We understand that the number of articles found and selected for this systematic review—130 articles analyzed, control done as a distributed task among 3 readers, 13 articles selected, 15 articles discarded—was adequate for this primary analysis on the subject and for finding tendencies and

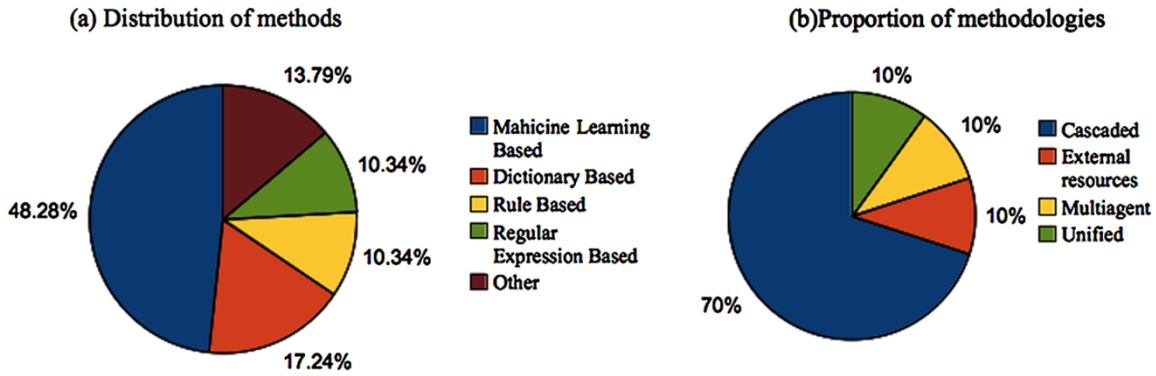

Fig. 10 Distribution of methods and methodologies

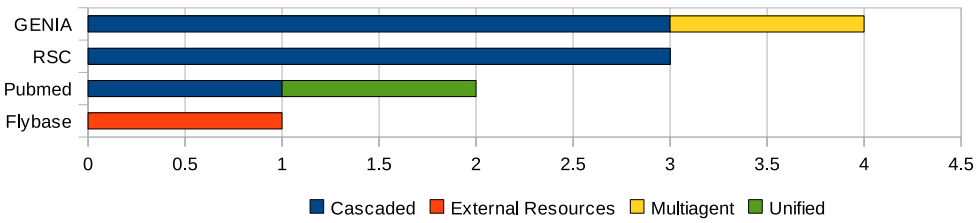

Fig. 11 Corpus vs. amount of articles, grouped by methodology

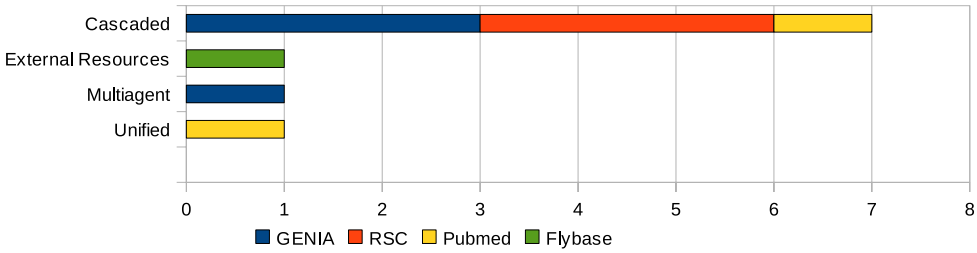

Fig. 12 Methodology vs. amount of articles, grouped by corpus

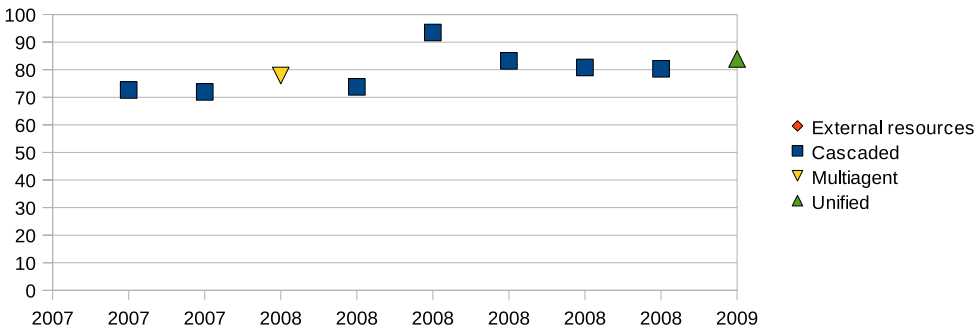

Fig. 13 F-score time vs. methodology

gaps. However, practical decisions regarding the research questions might claim for a confidence measure to be applied.

Figure 10(b) presents the proportion of each NER methodology in biomedical texts, considering the main experimental method, Machine Learning. Among these selected articles, there are 7 articles that employ the cascaded method, and 1 article for each one of the other methodologies, so with a strong presence of this first method.

Figure 11 relates the four corpora identified, and the amount of articles found that make use of these corpora in the experiments they report, grouped by methodology.

Figure 12 presents, on the other hand, the amount of articles in each methodology, grouped by corpus. From these two figures, we can observe that:

– The most used corpus in these articles is GENIA. In 2005, Cohen et al. [37] already indicated the importance of GENIA corpus, presenting it as the most used in the biomed-

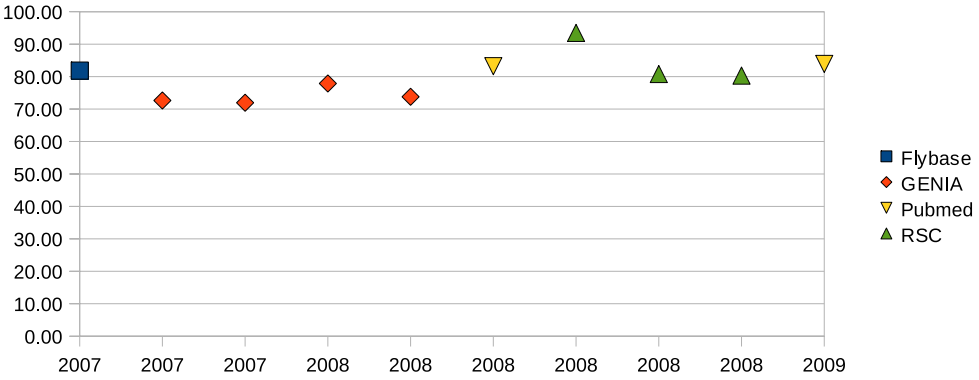

Fig. 14 F-score vs. time vs. corpus

ical field. Even if NER was not the primary focus of that paper, the authors explained GENIA's predominance in terms of structural and linguistic annotation, which keeps being one of the important factors for expressiveness and popularity of GENIA corpus.

- Once the source of the texts of the GENIA corpus is PubMed, the relevance of GENIA corpus for studies with NER is highlighted. We also remark that this strong presence is the result of a high level of preparation and an increasing number of applications and experiments that lie on GENIA corpus. Other more recent corpora still lack curation, also a problem for dictionaries and other external resources.
- The cascaded methodology is the most frequently used in the articles that report these experiments (3/4 of the total).
- Among the articles that report experiments with corpora, the fact that the RSC Corpus is only used in a cascaded methodology is not necessarily significant. In fact, the right observation could be that, considering that the RSC Corpus is not exactly a collection of biomedical texts, the use of this corpus as a testbed is achieved by means of a sequence of steps organized to use the best of the available technologies.
- The use of external resources with Flybase and PubMed should be connected to the fact that those corpora are linked with specialized ontologies. These ontologies are important resources that permit navigation among broader or narrower terms as well as a detailed description of their classes.
- The fact that a cascaded methodology may produce more readable results may influence the preference for this methodology.
- The importance of GENIA Corpus is also put in evidence by its use in the contests available in the area and, of course, regarding the possibility to compare results with those from other experiments. But would it be mandatory to participate and test systems in these contests?

Figure 13 relates, for the NER experimentation articles, the obtained F-score, the year of publication and the methodology used. Even if the F-score cannot be compared among all these experiments, of this relation, it is possible to identify that:

- During the years from 2007 to 2009, the cascaded method appears as the most frequent one.

A multiagent methodology, in fact, may claim for special infrastructure and implementation, so that the development of this solution is time consuming and demands more, in terms of computer science knowledge. Could this be a reason for its scarce presence?

Figure 14 presents the relation between the obtained F-score, in the period from 2007 to 2009 and the corpora identified in the articles. Based on the figure, it is possible to observe that:

- The experiments with the GENIA corpus and NER had their results published until 2008 (F-score between 71.95 and 77.88).
- Experiments using the RSC corpus start appearing in the year 2008. Note that the results with the RSC corpus present an F-score above 80% (between 80.32 and 93.48).

Aside from these observations, the following opportunities are identified:

- GENIA corpus has been extensively in use, what makes it a reference to the experimental studies in the area. So, the studies and experiments on this corpus can easily be compared.
- How do the differences and similarities between GENIA corpus and RSC corpus impact on the performance of NER?
- PubMed is the source of texts for GENIA corpus, so PubMed corpus can also benefit from working with GENIA.

In order to answer research question 3, considering these observations, we may argue that a relevant methodology for NER is cascaded. However, it is not possible to confirm it is an ideal one. Despite the majority of articles reporting experiments making use of a cascaded methodology (Fig. 11), this does not guarantee the best results (Fig. 13). The relation

between performance and corpus used in the experiments is very strict (Fig. 12), which leads one to believe that the features of the corpus together with the choice of the method can influence the performance in the experiments.

6 Conclusions

Most of the scientific studies in biomedical NER (92.1% among the references studied) are being developed by means of experimental work based on corpora and Machine Learning techniques. Other approaches represent only 7.89% of these studies in the period 2007–2009. Apparently, this is not different from previous studies reported on NER [1], and reflects a tendency already pointed out by Jurafsky and Martin in [38].

The experimental project for NER frequently involves the use of unified or cascaded approaches. Unified approaches—those that identify and classify NEs in a single stage—are common but the most usual practice is the cascaded processing, in two stages, that is, the NE identification followed by NE classification. Executing different stages in NER has been leading to interesting results. Moreover, the different stages make it possible to have a better comprehension of false-positives and false-negatives in each stage, as well as to understand the degrees of influence of these results in NE processing.

The authors usually use well-known corpora, as is the case with GENIA, but they also build new ones based on important databases, as with RSC. This may be a way of putting light on specific problems not found in GENIA texts but of interest in current research in the area, as well as discarding those specificities not present in the application-based corpora (e.g., RSC).

The GENIA corpus seems to be a solid reference for NER and is understood not to be neglected for purposes of comparison, including for other tasks. GENIA has been continuously complemented with new annotation with the purpose of using it in other tasks beyond NER, with the extraction of events, for example.

The number of articles found which are related to the identification of events (23 articles) and relationships (18 articles) leads to the conclusion that this topic is of great interest, being a task dependent on NER and maybe a next step for the research in the area, even if they are not classified as NER. Otherwise, facts, speculations and conflicts of interest were more rarely explored in the examined texts. The difficulty to present solutions for these tasks in the early stage they are, could be some reasons for that. Speculations were recently included in BioNLP shared task (2009) which may point to the publication of compared algorithms and results in a near future. Here, there is clearly a task dependence on NE, relationship and event recognition. In the long term, efforts to solve these more complex problems should be extensively considered.

Broad solutions, planned for treating a large spectrum of NER cases in different domains, seem to be seldomly effective here. Sasaki et al. [31] propose a more general methodology. However, it is important to note that this is limited to the classification of proteins. On the other hand, none of the studies found using GENIA corpus were able to surpass the results of the 2004 JNLPBA Workshop in terms of F-score, for the complete set of proposed classes.

Acknowledgements We thank Professor Marcelo Blois Ribeiro for his availability in reading this article, as well as in bringing suggestions and tips with respect to the text and the systematic review methodology.

References

1. Ananiadou S, McNaught J (2005) Text mining for biology and biomedicine. Artech House, Norwood
2. Chen L, Liu H, Friedman C (2005) Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics 21(2):248–256
3. Kozareva Z, Ferrandez O, Montoyo A, Munoz R, Suarez A (2007) Combining data-driven systems for improving named entity recognition. Data Knowl Eng 61(3):449–466
4. Biolchini J, Mian PG, Natali ACC, Travassos GH (2005) Systematic review in software engineering. System Engineering and Computer Science Department COPPE/UFRJ, Technical report ES, 679(05)
5. Kitchenham B (2004) Procedures for performing systematic reviews. Technical report, Keele University and NICTA
6. Kim J-D, Ohta T, Tsuruoka Y, Tateisi Y, Collier N (2004) Introduction to the bio-entity recognition task at JNLPBA. In: JNLPBA'04: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications, Morristown, NJ, USA. Association for Computational Linguistics, Menlo Park, pp 70–75
7. Tsai RT-H, Wu S-H, Chou W-C, Lin Y-C, He D, Hsiang J, Sungand T-Y, Hsu W-L (2006) Various criteria in the evaluation of biomedical named entity recognition. BMC Bioinform 7:92
8. Sasaki Y, Tsuruoka Y, McNaught J, Ananiadou S (2008) How to make the most of NE dictionaries in statistical NER. BMC Bioinform 9(Suppl 11):S5
9. Sun B, Mitra P, Giles CL (2008) Mining, indexing, and searching for textual chemical molecule information on the web. In: WWW '08: Proceedings of the 17th international conference on World Wide Web. ACM, New York, pp 735–744
10. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G (2008) Inter-species normalization of gene mentions with GNAT. Bioinformatics 24(16):126–132
11. Tan H, Lambrix P (2009) Selecting an ontology for biomedical text mining. In: BioNLP '09: Proceedings of the workshop on BioNLP, Morristown, NJ, USA. Association for Computational Linguistics, Menlo Park, pp 55–62
12. Vlachos A (2007) Evaluating and combining biomedical named entity recognition systems. In: BioNLP '07: Proceedings of the workshop on BioNLP 2007, Morristown, NJ, USA. Association for Computational Linguistics, Menlo Park, pp 199–206
13. Jijkoun V, Khalid MA, Marx M, de Rijke M (2008) Named entity normalization in user generated content. In: AND '08: proceedings of the second workshop on analytics for noisy unstructured text data. ACM, New York, pp 23–30

14. Sarafraz F, Eales J, Mohammadi R, Dickerson J, Robertson D, Nenadic G (2009) Biomedical event detection using rules, conditional random fields and parse tree distances. In: BioNLP '09: proceedings of the workshop on BioNLP, Morristown, NJ, USA. Association for Computational Linguistics, Menlo Park, pp 115–118
15. Shi Z, Sarkar A, Popowich F (2007) Simultaneous identification of biomedical named-entity and functional relations using statistical parsing techniques. In: NAACL '07: human language technologies 2007: the conference of the North American; Companion volume, Short papers on XX, Morristown, NJ, USA. Association for Computational Linguistics, Menlo Park, pp 161–164
16. Liu H, Blouin C, Keselj V (2009) Identifying interaction sentences from biological literature using automatically extracted patterns. In: BioNLP '09: proceedings of the workshop on BioNLP, Morristown, NJ, USA. Association for Computational Linguistics, Menlo Park, pp 133–141
17. Cohen KB, Verspoor K, Johnson HL, Roeder C, Ogren PV, Baumgartner WA Jr, White E, Tipney H, Hunter L (2009) High-precision biological event extraction with a concept recognizer. In: BioNLP '09: proceedings of the workshop on BioNLP, Morristown, NJ, USA. Association for Computational Linguistics, Menlo Park, pp 50–58
18. Aleman-Meza B, Nagarajan M, Ding L, Sheth A, Arpinar IB, Joshi A, Finin T (2008) Scalable semantic analytics on social networks for addressing the problem of conflict of interest detection. ACM Trans Web 2(1):1–29
19. Alpaydin E (2004) Introduction to machine learning. MIT Press, Cambridge
20. Jurafsky D, Martin JH (2009) Speech and language processing, 2nd edn. Prentice-Hall, New York
21. Joachims T (1999) Advances in kernel methods: support vector learning. In: Making large-scale support vector machine learning practical. MIT Press, Cambridge, pp 169–184
22. Vlachos A (2007) Evaluating and combining biomedical named entity recognition systems. In: BioNLP '07: proceedings of the workshop on BioNLP 2007, Morristown, NJ, USA. Association for Computational Linguistics, Menlo Park, pp 199–206
23. Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the eighteenth international conference on machine learning, pp 282–289
24. Groves RM, Fowler FJ Jr, Couper MP, Lepkowski JM, Singer E, Tourangeau R (2009) Survey methodology, 2nd edn. Wiley-Blackwell, New York
25. Overview articles. http://www.signalprocessingsociety.org/publications/overview-articles/
26. Chan S-K, Lam W, Yu X (2007) A cascaded approach to biomedical named entity recognition using a unified model. In: Data mining. ICDM 2007. Seventh IEEE international conference on, Oct 2007, pp 93–102
27. Wang H, Zhao T, Liu J (2008) Multi-agent classifiers fusion strategy for biomedical named entity recognition. In: BioMedical engineering and informatics. BMEI 2008. International conference on, May 2008, vol 1, pp 311–315
28. Kim J-D, Ohta T, Teteisi Y, Tsujii J (2003) GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics 19(Suppl 1):180–182
29. Corbett P, Batchelor C, Teufel S (2007) Annotation of chemical named entities BioNLP 2007: biological, translational, and clinical language processing, Prague, Czech Republic, pp 57–64
30. Corbett P, Copestake A (2008) Cascaded classifiers for confidence-based chemical named entity recognition. BMC Bioinformatics 9(Suppl 11):S4
31. Sasaki Y, Tsuruoka Y, McNaught J, Ananiadou S (2008) How to make the most of ne dictionaries in statistical ner. In: BioNLP '08: proceedings of the workshop on current trends in biomedical natural language processing, Morristown, NJ, USA. Association for Computational Linguistics, Menlo Park, pp 63–70
32. Neves ML, Carazo JM, Pascual-Montano A (2009) Extraction of biomedical events using case-based reasoning. In: BioNLP '09: proceedings of the workshop on BioNLP, Morristown, NJ, USA. Association for Computational Linguistics, Menlo Park, pp 68–76
33. Li Y, Lin H, Yang Z (2007) Two approaches for biomedical text classification. In: Bioinformatics and biomedical engineering. ICBBE 2007. The 1st international conference on, July 2007, pp 310–313
34. Viola P, Jones M (2001) Fast multi-view face detection. In: Proc of CVPR
35. Yoshida K, Tsujii J (2007) Reranking for biomedical named-entity recognition. In: BioNLP '07: proceedings of the workshop on BioNLP 2007, Morristown, NJ, USA. Association for Computational Linguistics, Menlo Park, pp 209–216
36. Gu B, Dahl V, Popowich F (2007) Recognizing biomedical named entities in the absence of human annotated corpora. In: Natural language processing and knowledge engineering. NLP-KE 2007. International conference on, August 30 2007–Sept 1, pp 74–81
37. Cohen KB, Fox L, Ogren PV, Hunter L (2005) Empirical data on corpus design and usage in biomedical natural language processing. In: AMIA symposium, pp 156–160
38. Jurafsky D, Martin JH (2000) Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall PTR, New York