OXFORD

Briefings in Bioinformatics, 17(1), 2016, 132–144

doi: 10.1093/bib/bbv024
Advance Access Publication Date: 1 May 2015
Paper

Community challenges in biomedical text mining over 10 years: success, failure and the future

Chung-Chi Huang and Zhiyong Lu

Corresponding author. Zhiyong Lu, National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), 8600 Rockville Pike, Bethesda, Maryland 20894, USA. Tel.: 301-594-7089; Fax: 301-480-2288; E-mail: Zhiyong.lu@nih.gov

Abstract

One effective way to improve the state of the art is through competitions. Following the success of the Critical Assessment of protein Structure Prediction (CASP) in bioinformatics research, a number of challenge evaluations have been organized by the text-mining research community to assess and advance natural language processing (NLP) research for biomedicine. In this article, we review the different community challenge evaluations held from 2002 to 2014 and their respective tasks. Furthermore, we examine these challenge tasks through their targeted problems in NLP research and biomedical applications, respectively. Next, we describe the general workflow of organizing a Biomedical NLP (BioNLP) challenge and involved stakeholders (task organizers, task data producers, task participants and end users). Finally, we summarize the impact and contributions by taking into account different BioNLP challenges as a whole, followed by a discussion of their limitations and difficulties. We conclude with future trends in BioNLP challenge evaluations.

Key words: biomedical natural language processing (BioNLP); BioNLP challenges; BioNLP shared tasks; critical assessment; text mining

Introduction

As most biomedical discoveries are communicated in scholarly publications, finding and reading relevant information in literature is essential for any researcher in life sciences [1–4]. However, the large size of the biomedical literature and its rapid growth in recent years (>3000 articles are published in biomedical journals every day) make literature search and information access a demanding task [1, 5, 6]. Health-care professionals in the clinical domain face the similar problem of information explosion and overload [7] when dealing with the increasingly available electronic medical/health records [8].

Because scholarly publications and clinical narratives are primarily written in text, natural language processing (NLP) becomes increasingly important in biomedical research, as it can greatly facilitate research productivity [9] by extracting key information from free text and converting it into structured knowledge for human comprehension. Since late 1990s, the interdisciplinary collaboration between the NLP and biomedical communities has become more common, forming a new research area known as biomedical natural language processing (**BioNLP**) or **text mining** with the goal of developing NLP methods for various kinds of biomedical applications. As illustrated in Figure 1, text-mining developers first use information retrieval (IE) techniques such as document classification and document/passage retrieval to select relevant documents [10–16]. This article selection process greatly narrows down the search space from the entire document collection to the ones of interest (associated BioNLP topics include biomedical literature retrieval, chemical patent retrieval, medical case retrieval and cohort identification). The selection process is known as article triaging [10, 17]. BioNLP developers then incorporate information extraction (IE) technologies (e.g. event extraction or entity-relation extraction) to identify the text segment that may represent a targeted information focus. The focus may be an entity–entity interaction (e.g. drug–drug interaction, or DDI [18], and protein–protein interaction, or PPI [19]), an entity–entity relation (e.g. protein–residue association [20], gene relation

Chung-Chi Huang is currently a visiting fellow at NCBI, NLM/NIH. His research interest includes machine translation, computer-assisted language learning, reader interest analysis, and text mining.
Zhiyong Lu is Earl Stadtman investigator at NCBI, NLM/NIH where he leads the text mining research group. His research focuses on text data mining and its applications for accelerating knowledge discovery. Dr. Lu is an organizer of the BioCreative challenge.

Submitted: 29 January 2015; **Received (in revised form):** 31 March 2015

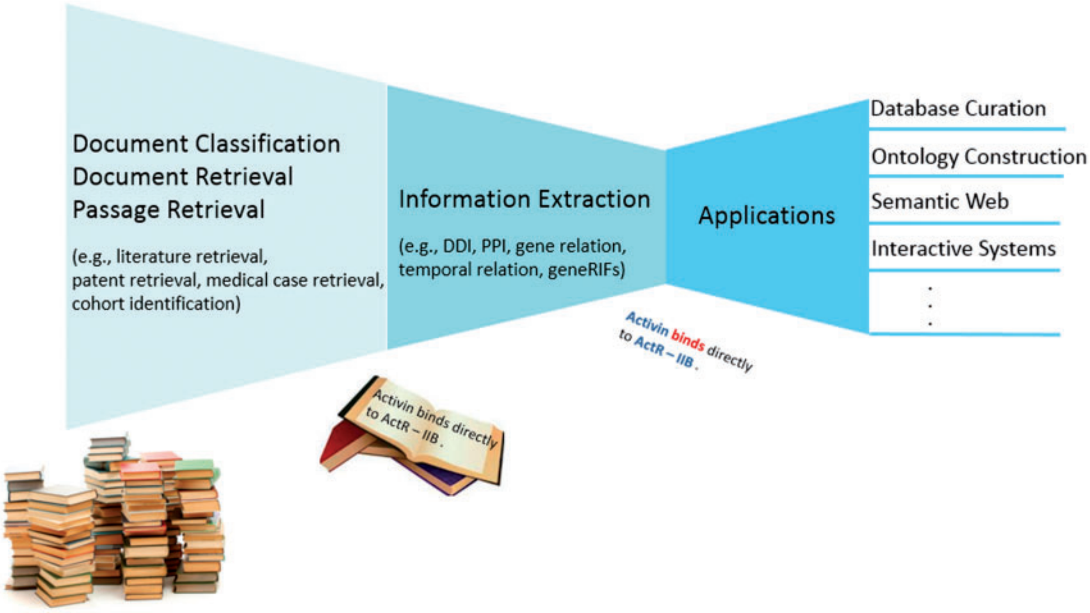

Document Classification
Document Retrieval
Passage Retrieval
(e.g., literature retrieval, patent retrieval, medical case retrieval, cohort identification)
Information Extraction
(e.g., DDI, PPI, gene relation, temporal relation, geneRIFs)
Applications
Database Curation
Ontology Construction
Semantic Web
Interactive Systems
⋮
Activin binds directly to ActR – IIB .
Activin binds directly to ActR – IIB .

Figure 1. Data channeling based on NLP technology and the application of the channeled data. NLP technology (e.g. *Document Retrieval* or *Information Extraction*) helps alleviate scientists in biology and life science from significant efforts of manual searching/researching for text snippets of interest by narrowing down the search space. The topics of BioNLP challenge tasks with the focus of the NLP technology are exemplified. For instance, topics associated with *Information Extraction* in BioNLP include, but not limit to, finding drug–drug interactions, protein–protein interactions, gene relations, clinical temporal relations and references into gene functions. The channeled/text-mined data, on the other hand, can be further used to curate databases, construct ontologies, build semantic networks or interactive systems and so on.

[21] or temporal relation in clinical record [22]), reference statements or experimental methods regarding bio-entities' functions or relations (e.g. references into gene function, or geneRIFs [23]), biological processes (e.g. phosphorylation) along with participating bio-entities (e.g. gene event extraction [24]), etc. Overall, the automatic NLP-enabled data channeling exposes users in life sciences to specific text snippets of interest without the significant effort of manual searching and researching. Subsequently, the extracted or text-mined information from biomedical literature and clinical records alike has a wide range of real-world applications. For instance, it can be exploited to assist database curation [25–29], construct ontologies [30], facilitate semantic Web search [31, 32] and help the development of interactive systems (e.g. computer-assisted curation tools [33, 34]).

To promote such interdisciplinary collaboration and to assess and advance the NLP techniques in biomedicine, a myriad of BioNLP shared tasks/challenges have been organized over the years following the success of *CASP* in 1994 [35, 36] on protein structure prediction. Some of the challenges focus on processing information in biomedical literature while others in medical records. Complementary to the surveys of general progress in biomedical text mining, such as [37, 38], this review serves as a brief introduction to the BioNLP shared tasks successfully held from 2002 to 2014. We specifically consider both the perspective of NLP research and the underlying biological problem, and provide a systematic summary and comparison of these tasks and their subtasks. We also outline the steps in organizing BioNLP challenges. Finally, we summarize the impact and contributions of these challenge tasks as a whole, followed by a discussion of their limitations and difficulties. Note that task participating systems (or implementations) are not in the scope of this review. We ask the reader to refer to the individual task overview papers for participants' methodology, system descriptions and system performance.

Task overview

Figure 2 chronologically lists the BioNLP challenge evaluations from 2002 to 2014. The challenges/competitions are shown in bold white font, whereas their specific task focus is shown in italic black font leveraging the short-hand track notations in Table 1, which details the challenge tasks. In Figure 2, the BioNLP challenge evaluations are primarily grouped by the text genre: biological tasks focus on scholarly publications while clinical tasks on clinical records. Challenge tasks were first introduced to the BioNLP community by the *ACM KDD Cup* in 2002, followed by *TREC Genomics* and many others in recent years.

KDD Cup, TREC Genomics/Chemical and CoNLL

Early challenges, such as *KDD Cup* and *TREC Genomics*, mostly focused on the document retrieval [23] or document classification [17] tasks. For example, the fly genetics task in *KDD Cup* 2002 [39] required participants to determine whether an article meets the curation criteria of fly gene expression. *TREC ad hoc* retrieval tasks [23] asked participants to perform document retrieval for the curation of gene functions (i.e. to select articles that discuss gene functions) in 2003 and to retrieve documents containing specific topics related to genes or other bio-entities in 2004 [10] and 2005 [40]. Passage/statement retrieval was also attempted by *TREC* [23] where participating systems were asked to extract texts that are references into gene functions (GeneRIFs) [41, 42]. Later in 2006 [43] and 2007 [44], *TREC Genomics* further formulated search topics into natural language questions (bio-entity-based questions in 2007) addressing biologists' quests in the paradigm of question answering (QA) via NLP. Continuing its previous efforts on document retrieval, *TREC* organized a chemical track from 2009 to 2011 [45–47] addressing the needs of document retrieval in chemical

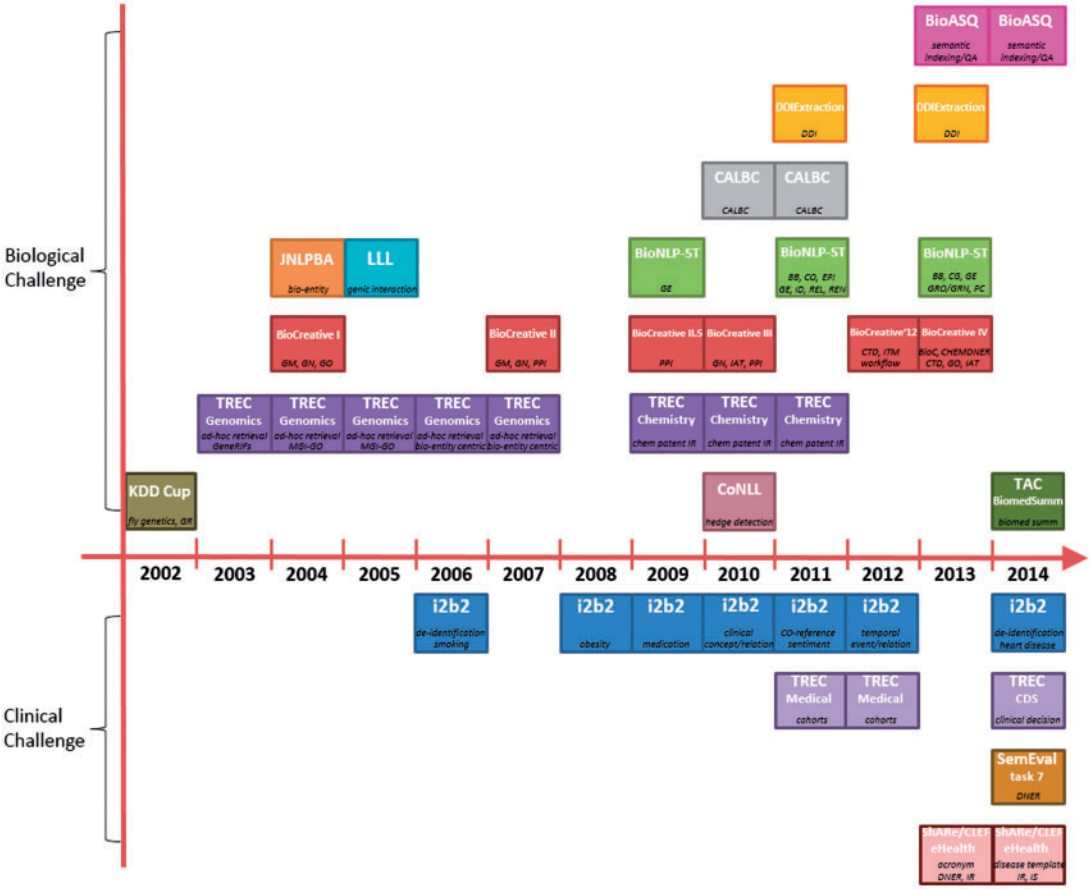

Biological Challenge
Clinical Challenge
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
BioASQ semantic indexing/QA
BioASQ semantic indexing/QA
DDIExtraction DDI
DDIExtraction DDI
CALBC CALBC
CALBC CALBC
JNLPBA bio-entity
LLL genic interaction
BioNLP-ST GE
BioNLP-ST BB, CO, EPI GE, ID, REL, REN
BioNLP-ST BB, CG, GE GRO/GRN, PC
BioCreative I GM, GN, GO
BioCreative II GM, GN, PPI
BioCreative II.5 PPI
BioCreative III GN, IAT, PPI
BioCreative'12 CTD, ITM workflow
BioCreative IV BioC, CHEMDNER CTD, GO, IAT
TREC Genomics ad-hoc retrieval GeneRIFs
TREC Genomics ad-hoc retrieval MGI-GO
TREC Genomics ad-hoc retrieval MGI-GO
TREC Genomics ad-hoc retrieval bio-entity centric
TREC Genomics ad-hoc retrieval bio-entity centric
TREC Chemistry chem patent IR
TREC Chemistry chem patent IR
TREC Chemistry chem patent IR
KDD Cup fly genetics, GR
CoNLL hedge detection
TAC BiomedSumm biomed summ
i2b2 de-identification smoking
i2b2 obesity
i2b2 medication
i2b2 clinical concept/relation
i2b2 CO-reference sentiment
i2b2 temporal event/relation
i2b2 de-identification heart disease
TREC Medical cohorts
TREC Medical cohorts
TREC CDS clinical decision
SemEval task 7 DNER
ShARe/CLEF eHealth acronym DNER, IR
ShARe/CLEF eHealth disease template IR, IS

Figure 2. BioNLP challenges in chronological order. Challenges are shown in bold white font, whereas their specific task focus is shown in italic black font following the task/track short-hands in Table 1.

patents. For text classification, *CoNLL* dedicated its 2010 shared task to the identification of uncertain sentences in biomedical literature and locating in-sentence hedging cues [48], as negations and speculations common in biomedical publications can have direct influence on the text-mined results.

BioCreative, JNLPBA and CALBC

In 2004, *BioCreative* and *JNLPBA* started to address the needs to automatically detect bio-entities in free texts. Specifically, the Gene Mention (GM) task [49] in *BioCreative* I [50] aimed for gene name detection, whereas the bio-entity task in *JNLPBA* [51] involved multiple entity types such as DNA, RNA and cell type. Biological named entity recognition (NER) is essential, as it is the building block for many higher-level NLP tasks such as protein–protein interaction or gene regulation (GR) extraction [51]. Following the GM tasks, the gene normalization (GN) task in *BioCreative* I [52], *BioCreative* II [53] and *BioCreative* III [54] was introduced where gene names located automatically are further mapped to unique identifiers in some standard lexicons/databases (e.g. EntrezGene). In addition to genes or proteins, automatic detection of other key biological entities such as chemicals and diseases is also examined in recent *BioCreative* tasks [55–57]. *CALBC* is another NER-oriented challenge with the goal of generating a large, shared corpus with annotated bio-entities [58, 59].

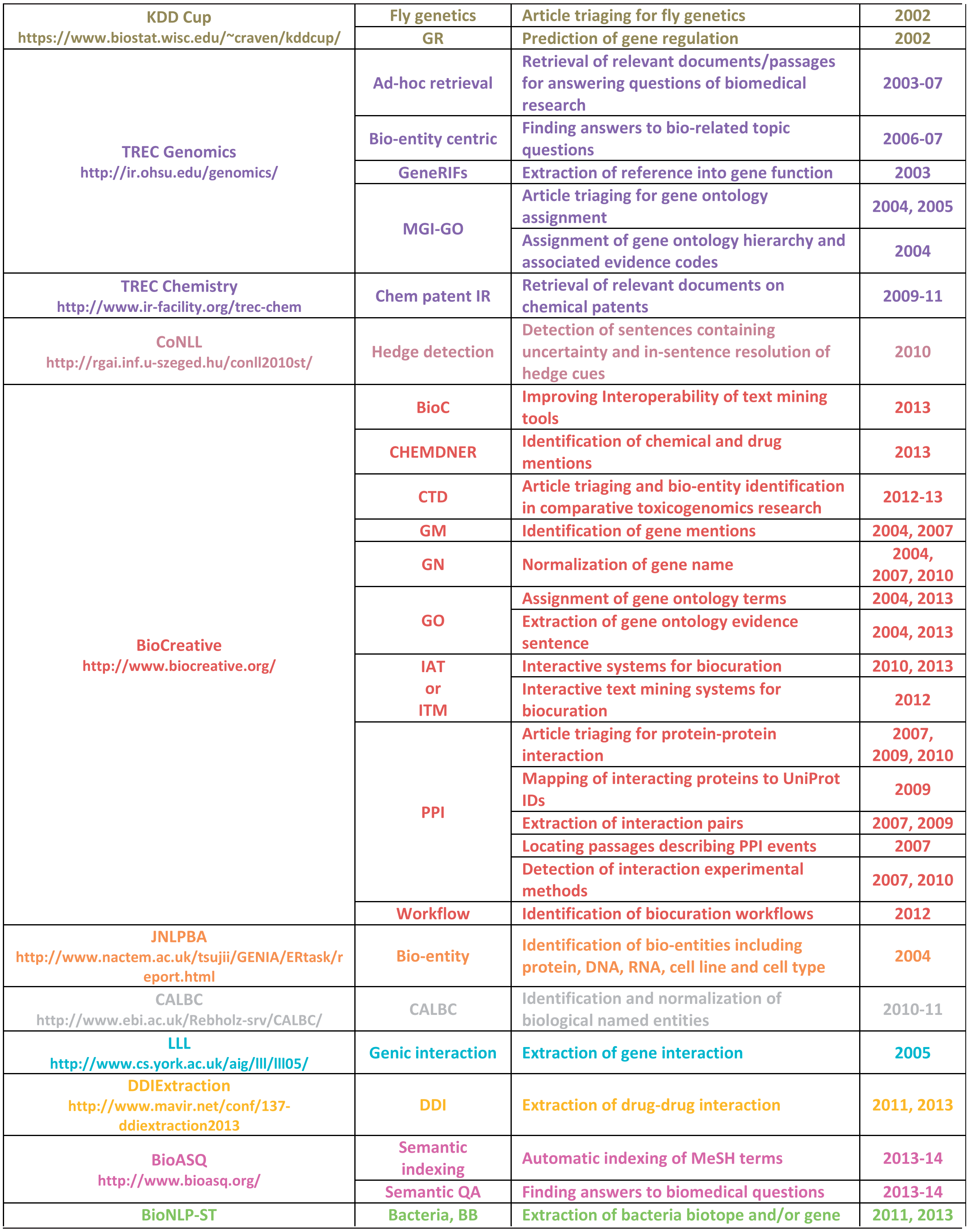
KDD Cup, LLL, BioCreative, DDIExtraction and BioASQ

Meanwhile, IE techniques have been tested in various bio-related topics such as the GR prediction track in *KDD Cup* 2002 and the genic interaction extraction task in *LLL* 2005 [21]. In *BioCreative*, two major IE tasks were introduced: the automatic assignment of gene ontology terms (i.e. the GO task in *BioCreative* I [60] and IV [61, 62]) and the extraction of protein–protein interaction (i.e. the PPI task in *BioCreative* II [19], II.5 [63] and III [12]). In 2011, the DDI task was first introduced in *DDIExtraction* [18] and then repeated in 2013 [64]. Facing similar challenge as the GO task in *BioCreative*, automatic assignment of MeSH terms to biomedical articles itself [65] is far more challenging owing to the size of the MeSH lexicon and rich expression of MeSH concepts in text (word sense disambiguation may be involved). Thus, it becomes one of the two tasks in the recent *BioASQ* challenge [66, 67] while the other task focused on obtaining precise and comprehensive answers to questions from real-life biomedical research.

BioNLP-ST

Compared with the *BioCreative* tasks and other IE-oriented events, *BioNLP-ST* has its unique semantics as to how to represent biological events/processes along with participating entities and devotes itself to event/relation extraction. For instance, the *BioNLP-ST* GENIA (GE) task in 2009 [24], 2011 [68] and 2013 [69] asked task participants to extract gene-related events such

Table 1. Brief task description



Challenge (Web site)	Abbreviated track names	Brief task description	Years
KDD Cup https://www.biostat.wisc.edu/~craven/kddcup/	Fly genetics	Article triaging for fly genetics	2002
	GR	Prediction of gene regulation	2002
TREC Genomics http://ir.ohsu.edu/genomics/	Ad-hoc retrieval	Retrieval of relevant documents/passages for answering questions of biomedical research	2003-07
	Bio-entity centric	Finding answers to bio-related topic questions	2006-07
	GeneRIFs	Extraction of reference into gene function	2003
	MGI-GO	Article triaging for gene ontology assignment	2004, 2005
		Assignment of gene ontology hierarchy and associated evidence codes	2004
TREC Chemistry http://www.ir-facility.org/trec-chem	Chem patent IR	Retrieval of relevant documents on chemical patents	2009-11
CoNLL http://rgai.inf.u-szeged.hu/conll2010st/	Hedge detection	Detection of sentences containing uncertainty and in-sentence resolution of hedge cues	2010
BioCreative http://www.biocreative.org/	BioC	Improving Interoperability of text mining tools	2013
	CHEMDNER	Identification of chemical and drug mentions	2013
	CTD	Article triaging and bio-entity identification in comparative toxicogenomics research	2012-13
	GM	Identification of gene mentions	2004, 2007
	GN	Normalization of gene name	2004, 2007, 2010
	GO	Assignment of gene ontology terms	2004, 2013
		Extraction of gene ontology evidence sentence	2004, 2013
	IAT or ITM	Interactive systems for biocuration	2010, 2013
		Interactive text mining systems for biocuration	2012
	PPI	Article triaging for protein-protein interaction	2007, 2009, 2010
		Mapping of interacting proteins to UniProt IDs	2009
		Extraction of interaction pairs	2007, 2009
		Locating passages describing PPI events	2007
		Detection of interaction experimental methods	2007, 2010
	Workflow	Identification of biocuration workflows	2012
JNLPBA http://www.nactem.ac.uk/tsujii/GENIA/ERtask/report.html	Bio-entity	Identification of bio-entities including protein, DNA, RNA, cell line and cell type	2004
CALBC http://www.ebi.ac.uk/Rebholz-srv/CALBC/	CALBC	Identification and normalization of biological named entities	2010-11
LLL http://www.cs.york.ac.uk/aig/lll/lll05/	Genic interaction	Extraction of gene interaction	2005
DDIExtraction http://www.mavir.net/conf/137-ddiextraction2013	DDI	Extraction of drug-drug interaction	2011, 2013
BioASQ http://www.bioasq.org/	Semantic indexing	Automatic indexing of MeSH terms	2013-14
	Semantic QA	Finding answers to biomedical questions	2013-14
BioNLP-ST	Bacteria, BB	Extraction of bacteria biotope and/or gene	2011, 2013

(Continued)

Table 1. Brief task description

Challenge (Web site)	Abbreviated track names	Brief task description	Years
http://www.nactem.ac.uk/tsujii/GENIA/SharedTask/ http://2011.bionlp-st.org http://2013.bionlp-st.org		interaction	
	CG	Extraction of biological process related to cancer	2013
	CO-reference	Resolution of co-reference	2011
	EPI	Extraction of epigenetic change	2011
	GE	Extraction of gene-related events like expression	2009, 2011, 2013
	GRO, GRN	Extraction of gene regulation	2013
	ID	Extraction of mechanism of infectious disease	2011
	PC	Extraction of pathway model	2013
	REL	Extraction of gene relations	2011
	REN	Extraction of gene renaming	2011
TAC BiomedSumm http://www.nist.gov/tac/2014/BiomedSumm/	Biomed summ	Identification of text spans in referenced papers reflecting citances	2014
		Classification of cited text spans	
		Summarization of the referenced papers	
i2b2 https://www.i2b2.org/NLP/	Clinical concepts	Extraction of clinical concepts	2010
	Clinical relations	Identification of clinical relations	2010
	CO-reference	Resolution of co-reference	2011
	De-identification	De-identification of clinical discharge summaries	2006, 2014
	Heart disease	Identification of heart disease risks	2014
	Medication	Extraction of medication-related information	2009
	Novel data use	Reuse of released clinical data sets for answering new clinical questions	2014
	Obesity	Prediction of obesity and its co-morbidity	2008
	Sentiment	Classification of sentiment at sentence levels	2011
	Smoking	Prediction of smoking status	2006
	Software usability	Evaluation of clinical NLP software usability	2014
	Temporal events	Identification of temporal event/expression	2012
	Temporal relations	Extraction of temporal relation	2012
TREC Medical http://trec.nist.gov/data/medical.html	Cohorts	Extraction of cohorts matching inclusion criteria	2011-12
TREC CDS http://www.trec-cds.org/	Clinical decision	Retrieval of relevant documents for clinical decision making/support	2014
ShARe/CLEF eHealth http://sites.google.com/site/shareclefehealth/ http://clefehealth2014.dcu.ie/	Acronym	Normalization of acronyms/abbreviations	2013
	Disease template	Disease template/attribute filling	2014
	DNER	Recognition and normalization of diseases	2013
	IR	Retrieval of relevant documents	2013-14
	IS	Interactive search system for eHealth data	2014
	Multilingual IR	Retrieval of relevant documents for query in different languages	2014
SemEval task7 http://alt.qcri.org/semeval2014/task7/	DNER	Recognition and normalization of diseases	2014

Challenges are ordered according to their appearance in the article, while their tasks are alphabetically ordered according to abbreviated names. The texts are color-coded and consistent with Figure 2. A colour version of this table is available at BIB online: http://bib.oxfordjournals.org.

as regulation, expression and transcription and to associate them with their corresponding event participants, localization or sites. *BioNLP-ST* 2011 Bacteria task [70] and 2013 Bacteria Biotope task [71] aimed to detect the habitats for bacteria, whereas the *BioNLP-ST* 2011 Infectious Diseases task [72] and 2013 Cancer Genetics (CG) task [73] targeted biomolecular mechanisms of infectious diseases and cancer genetics, respectively. *BioNLP-ST* also covered topics as high level as pathway curation (i.e. the Pathway Curation task in *BioNLP-ST* 2013 [74]) and as fundamental as co-reference resolution [i.e. the co-reference resolution task (CO-reference) in *BioNLP-ST* 2011 [75]] and name alias (i.e. the REN task in *BioNLP-ST* 2011 [76]). Though fundamental, co-reference and alias issues clearly impose an upper bound on the performance of event extracting systems. Gene regulation (e.g. GRO and GRN tasks in 2013 [77]) and gene interaction (e.g. Bacteria task in 2011 [70]) were addressed too.

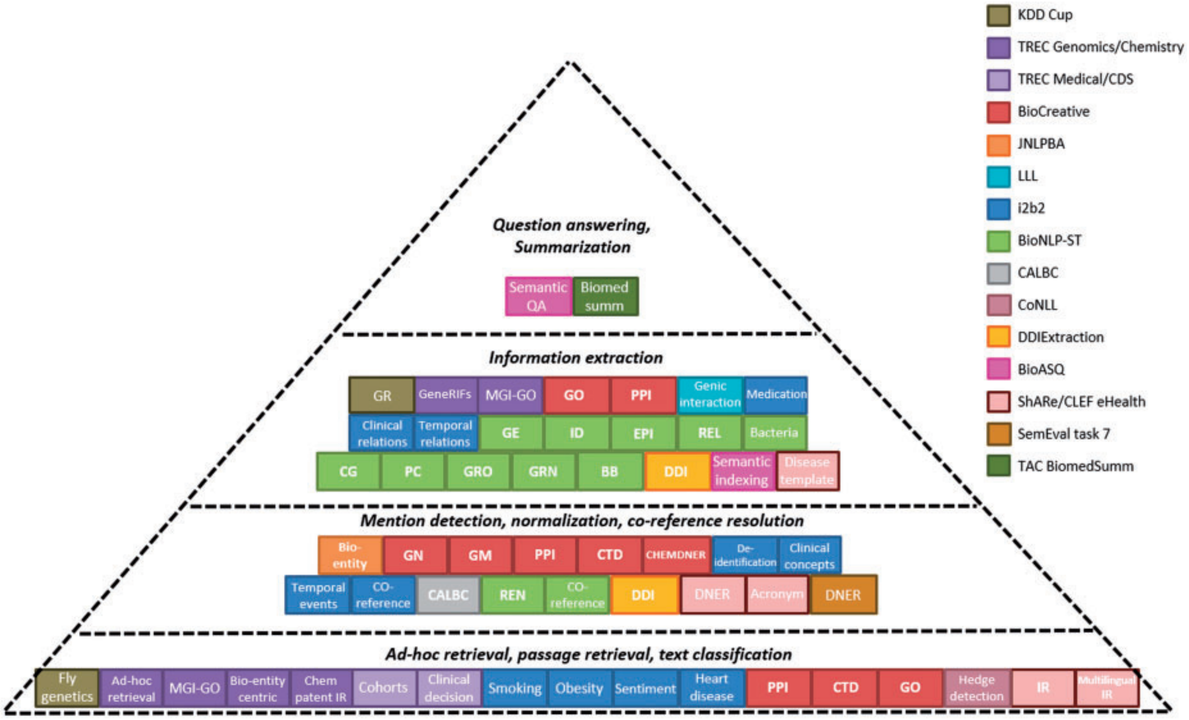


Figure 3. Challenges' subtasks/tracks organized based on NLP perspectives. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

TAC BiomedSumm

In 2014, the *TAC BiomedSumm* track challenged participants to leverage the set of citation sentences that reference a specific paper ('citances') for summarization, an important problem in BioNLP research [42, 78]. Specifically, the track included identifying the text spans in the referenced papers reflecting the citances, classifying those spans into paper facets and then generating summary for the referenced papers based on the community discussion of their citances.

i2b2, TREC Medical/CDS

The first clinically oriented challenge task was introduced by Informatics for Integrating Biology and the Bedside (*i2b2*) in 2006. An early focus was de-identification [79], a task similar in spirit to NER, as the sensitive private health information in the medical/clinical records needs to be removed before distribution. *i2b2* hosted text classification tasks on determining smoking status [80] at document level (i.e. clinical records) in the same year, on predicting obesity and its co-morbidities [81] at document level (i.e. clinical records) in 2008, on determining sentiments at sentence level (from suicide notes) in 2011 [82] and on predicting heart disease risks at document level (i.e. clinical records) in 2014. *i2b2* was also interested in mention detection and concept recognition but not for bio-entities. Instead they tackled with clinical concepts such as medical problems, tests, treatments, medication and dosage in clinical narratives in 2009 [83] and 2010 [84] and, in addition to the concepts, temporal expressions in 2012 [22]. Recognized entities were then analyzed with assertion information (e.g. the presence/absence of a medical problem) in 2010 [84] or were later linked with temporal relations (e.g. the dosage *before* the surgery) in 2012 [22]. *TREC*, on the other hand, shifted its focus from biomedical literature to clinical records in recent years. *TREC Medical* track was introduced in 2011 [85] and 2012 [86] in view of identifying cohorts matching specific 'inclusion criteria' (e.g. gender, age-group, treatment and disease present) for clinical research, clinical trials or epidemiological studies. In 2014, the *TREC CDS* track investigated NLP technologies for medical case retrieval for clinical decision support.

ShARe/CLEF eHealth and SemEval

In addition to *i2b2* and *TREC Medical/CDS*, a new evaluation event called *ShARe/CLEF eHealth* was piloted in 2013 [87]. It addressed three separate tasks (a) traditional NER on disease names in clinical notes and normalization, (b) mapping acronyms and abbreviations in clinical documents to UMLS CUIs and (c) retrieving relevant documents to address questions patients may have when reading discharge summaries. In 2014, task 7 of *SemEval* [88] repeated the disease NER and normalization task of *ShARe/CLEF eHealth* 2013 while *ShARe/CLEF eHealth* 2014 launched a different set of tasks [89]: (a) interactive search system for eHealth data, (b) disease template/attribute filling [90] and (c) *ad hoc* medical record retrieval [91] where task (c) is the first attempt to deal with multilingualism.

In Figure 3, we categorize the challenge tracks in Table 1 by the targeted problems in NLP research: from IR (*ad hoc* retrieval, passage retrieval and text classification), to NER (mention detection, normalization and co-reference resolution), to IE and to QA and summarization. For example, the cohort track of *TREC Medical* is in the category *text classification* while the *BioCreative*'s CHEMDNER track [92] is in *mention detection* (chemical, drug and disease detection to be exact) and so are its GM and GN tracks. In general, NLP tasks closer to the top of the pyramid are more difficult. We can see from Figure 3 that, among the shared tasks, *TREC* pays much of its attention on IR, whereas others (e.g. *BioCreative*, *LLL*, *BioNLP-ST* and *DDI*) focus on NER and IE.

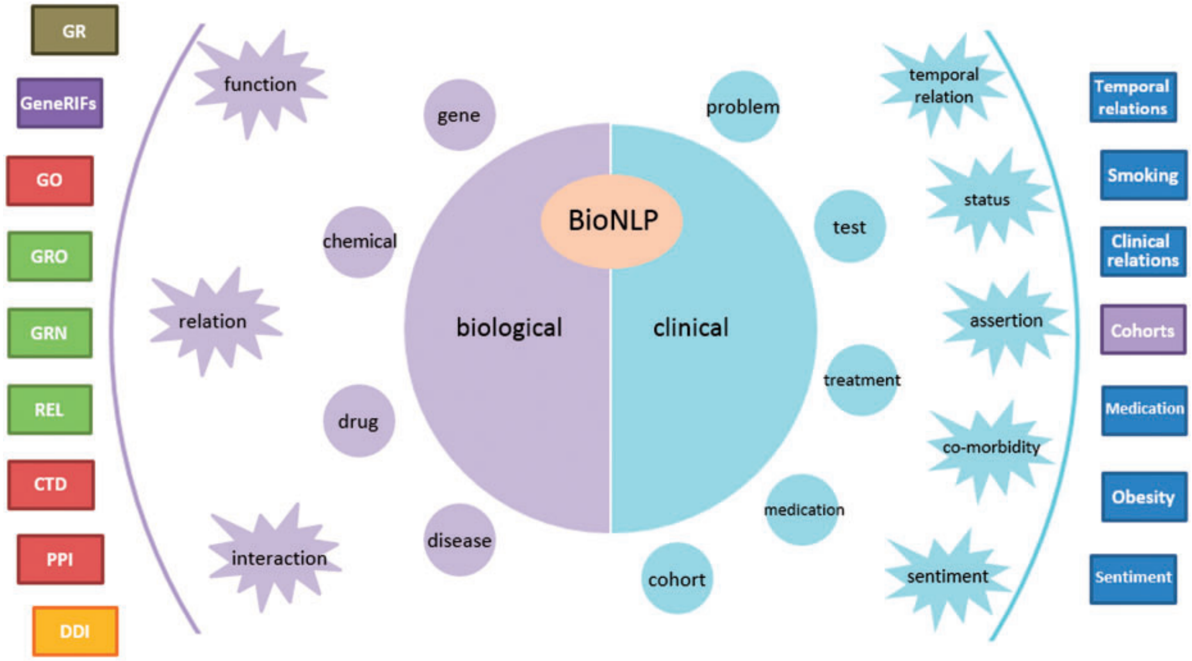


Figure 4. Different biological and clinical problems targeted by BioNLP challenges. Challenge subtasks are coded in the same colors as in Figure 2 (e.g. *BioNLP-ST* tasks are green marked). A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

From a different perspective, Figure 4 examines the BioNLP challenges based on targeted problems in biology and medicine. As can be seen, these two domains have their own interested entities and relations, respectively illustrated by circles and explosions in Figure 4. On the biological side, the key bio-entities include genes/proteins, chemicals, drugs and diseases [93–96], whereas on the clinical side, medical problems, tests, treatments, medication relevant information and cohorts are mostly concerned. As for important relations, biological tasks mainly aim for bio-entities' functions (e.g. gene functions in the GO task of *BioCreative*), relational events (e.g. gene-related events in the GE of *BioNLP-ST*) and interactions (e.g. drug–drug interaction in the DDI task of *DDIExtraction* and protein–protein interaction in the PPI task of *BioCreative*). Clinical tasks, on the other hand, deal with completely different sets of relations between clinical entities such as temporal relations, status, assertion/risk, co-morbidity and sentiment analysis. Figure 4 aligns individual tasks (in boxes) close to their associated entities and relations whenever possible. For example, the *i2b2* obesity task [81] is aligned with the relation of co-morbidity, as it aims at predicting co-morbidities of the medical problem of obesity whereas the *i2b2* temporal task [22] is aligned with the temporal relation, given its target of determining temporal relations between medical concepts (e.g. treatment and problem).

The organization of challenge tasks

Having discussed the roles of NLP community in life sciences and the topics in the past challenge events, we now turn our attention to how challenge tasks are typically organized.

Figure 5 depicts the typical workflow of organizing a challenge event, which generally involves four stakeholders: task organizers, task data producer/domain experts, task participants and end users. In the beginning, task organizers discuss and identify a domain-specific topic/task that is both important in biomedicine and difficult in NLP. As exemplified in the following scenarios, many tasks were designed to meet real-world needs in biomedical research: a researcher with an information need searches bibliographic databases to find relevant articles (e.g. *TREC Genomics/Chemistry* task) or a biocurator needs to identify protein–protein interactions in text (e.g. *BioCreative* PPI task) or evidence sentences for a Gene Ontology annotation or GO code (e.g. *BioCreative* GO task). It is an important and common practice for the task organizers to include end users in this planning stage. Challenge tasks are typically designed to be responsive to the need for critical mass in biomedical or NLP research. For that reason, selected task topics need to address new problems but must also relate to earlier studies to ensure adequate interest from the community ([97–100]). For instance, the document retrieval and gene NER tasks of *TREC Genomics* and *BioCreative* can be traced to the earlier pioneering studies such as [98] and [100], respectively.

Once a domain topic is determined, the task organizers examine existing resources to collect appropriate materials for preparing task data to be used in system development and evaluation. Typical text collections in BioNLP challenges include either scholarly publications (e.g. PubMed) or clinical records, which are distributed by health-care or medical centers with personal health information removed according to HIPAA rules. During this process, license and data privacy issues have to be examined and addressed.

Next, task organizers usually recruit domain experts to manually annotate the relevant documents for preparing task data or corpus. For example, in the *BioCreative* IV GO task, annotators were invited to mark up relevant gene/proteins, GO codes and associated evidence in scholarly publications. The human annotations then served as gold standards, against which the automated results from participating systems are compared. Typically, annotators need to reference external knowledge sources or databases when producing gold standards. For example, annotators may consult the gene identifiers in EntrezGene [101, 102], protein identifiers in UniProt [103] and

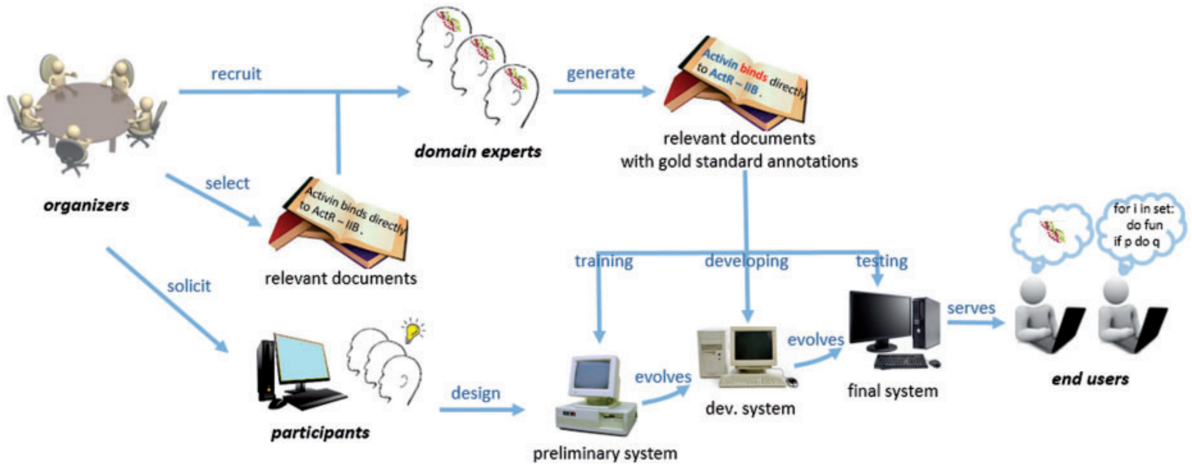


Figure 5. The typical workflow of organizing a shared task.

gene ontology [104] for tasks of GN and GO code assignment. To ease participants' burden and to focus their attention on the main task, sometimes part of the gold standards may be provided. For instance, most of *BioNLP-ST* tasks' bio-entities are provided so that task participants can focus solely on extracting events or biological processes. During data annotation, automatic tools may be used to speed up the process of manual annotation [105]. For example, OSCAR4 [106] was used for the corpus development in *BioNLP-ST* 2013 CG task. In such cases, domain experts may be prompted with automated pre-annotations and need only to correct erroneous annotations or to add missing ones. In addition to task data preparation/annotation, the organizers also provide the evaluation infrastructure for the challenge.

Meanwhile, task organizers will announce the competition and solicit participants through an open call. Participants typically have a few months to build their preliminary systems based on the training data distributed by task organizers and annotated by domain experts. And additional development data set may be released later to prevent over-fitting the training data. In the end, participants are given a few days to submit test results in different runs (e.g. different system parameters) and participating systems are evaluated using task-specific evaluation metrics (e.g. precision, recall, f-measure, reciprocal ranking) against the human-annotated gold standards.

After the competition, task organizers usually release training and testing data and task participants release their mature systems, to promote future improvement in the domain topic. In some events, an emphasis is also placed on pushing leading methods into real-world applications by engaging end users and system developers at the end of the challenge.

The impact and contributions of challenge tasks

Community-run BioNLP challenges not only assess the state of the art but also help advance the field in many aspects (see Figure 6):

Develop and release shared test collections

Manually created gold-standard annotations are critical for the development and evaluation of algorithms and systems in the BioNLP research. However, manual corpus annotation is also time-consuming and highly expensive, thus posing difficulties to individual research groups [107, 108]. Through community-run challenge evaluations, such costs are shared [57]. As a result, a large number of text corpora were made freely available to the BioNLP researchers including those de-identified health records that are otherwise difficult to obtain due to various access restrictions. These test data are widely (re-)used both during and after the challenges.

Because task data preparation and its sharing is critical for challenge evaluations, it has resulted in advancements in many related issues, such as annotation guideline standardization, alternatives to expert annotation and annotation tool development. The *i2b2* challenge evaluations have led to the standardization of the guidelines in annotating clinical narratives by domain experts. In several challenges, task organizers also successfully experimented with alternatives to expert annotation. For example, *i2b2* 2010 involved the research community (i.e. task participants) [84], whereas *BioCreative* III GN task used automatically inferred silver-standard [54]. Finally, the needs in corpus annotation also advanced the development of annotation tools such as BRAT [109] and PubTator [33].

Novel algorithms and improved results of difficult problems

Many BioNLP challenge evaluations led to the efforts of new NLP algorithm/tool development. And as a result, task performance sometimes increases significantly. For instance, performance was doubled for the *ad hoc* document retrieval tasks in *TREC Genomics* [10, 40, 43, 44]. Another example is the MeSH indexing task (i.e. automatically generate relevant MeSH headings for new PubMed articles) of the recent *BioASQ* challenge, where the best results (in F1-measure) moved from 0.538 to 0.591 at the end of its first year challenge in 2013 and further increased to 0.632 in 2014, improving the state of the art by 17% in 2 years [66, 67, 110].

Additional advances are seen in evaluation measures. To meaningfully evaluate results of different tasks, various evaluation measures have been adopted in addition to those traditional ones. For instance, TAP-k measures were introduced in the GN Task of *BioCreative* III because the traditional F-measure was unable to capture the ranking aspect of the predicted results [54].

Finally, challenge tasks help identify bottlenecks and emerging topics in BioNLP research. For instance, to improve interoperability and encourage combining efforts into more

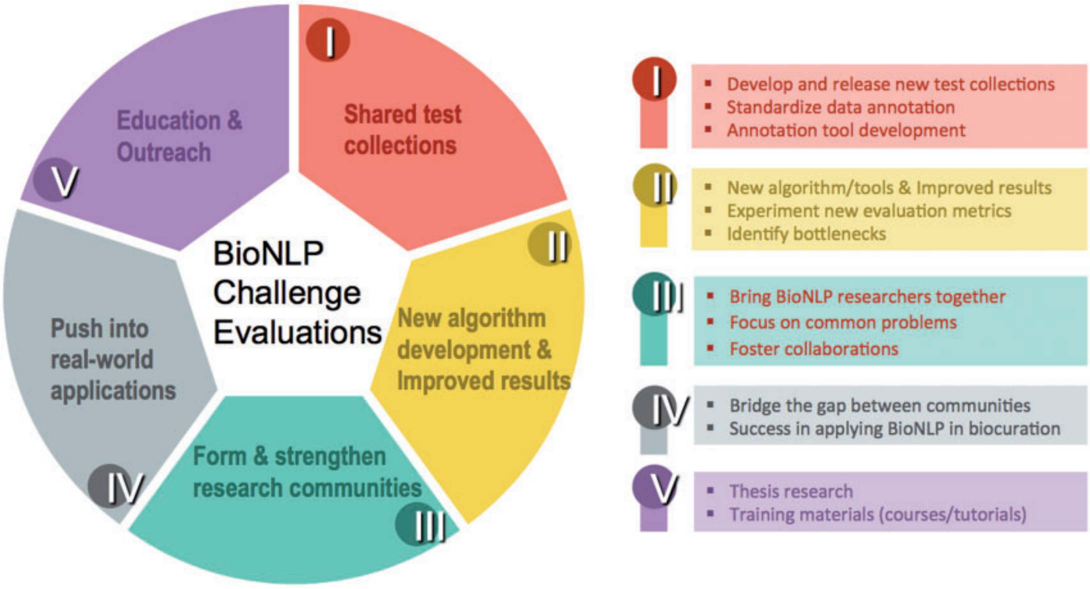

I
Shared test collections
II
New algorithm development & Improved results
III
Form & strengthen research communities
IV
Push into real-world applications
V
Education & Outreach
BioNLP Challenge Evaluations
I
Develop and release new test collections
Standardize data annotation
Annotation tool development
II
New algorithm/tools & Improved results
Experiment new evaluation metrics
Identify bottlenecks
III
Bring BioNLP researchers together
Focus on common problems
Foster collaborations
IV
Bridge the gap between communities
Success in applying BioNLP in biocuration
V
Thesis research
Training materials (courses/tutorials)

Figure 6. Impact/contributions from BioNLP challenges.

powerful and capable systems, recently there have been proposals such as BioC [111, 112] for new standards to share NLP tools and corpora. Challenges events provide unique opportunities for the community to develop, test and compare different proposals.

Form and strengthen research communities

Challenge evaluations help strengthen the community by focusing on important and challenging problems in BioNLP research (e.g. concept recognition, relation extraction and de-identification) and also by bringing in together researchers in associated workshops. Challenge participants benefit from in-depth discussion with their peers on common problems and publications in workshops and/or special issues. In addition, such events foster potential opportunities for collaborations.

Push BioNLP research into real-world applications

Some challenge evaluations also go beyond a single community and endeavor to bridge the gap between BioNLP research and new application domains. For example, since 2010, *BioCreative* has organized workshops in annual meetings of the biocuration society with a focus on better understanding biocuration workflows [113] and promoting the development and deployment of BioNLP tools into production curation pipelines. To date, there are already several reported successes (e.g. [33, 114]).

Education and outreach

Challenge tasks and associated data are often used by graduate students toward their thesis research (e.g. [115] and [116]), as well as have been used in developing training materials such as courses/tutorials (e.g. https://www.i2b2.org/NLP/DataSets/Courses.php). Students and trainees also greatly benefit from attending challenges workshops and participating in discussion.

To conclude, we find that these community-run challenge evaluations contribute significantly to the BioNLP research. And, indeed, they have become an integral part of the text-mining research, as many BioNLP studies experiment with the task data and use results during the challenge task as reference when developing their new techniques. Nonetheless, it is important to note that these formal evaluations are not without their own set of (internal) issues and challenges ranging from a focus on a small set of unrealistic tasks using small-size test collections, to insufficient participation and method homogenization, and to poor performance and limited benefits to practical applications.

The limitations of formal challenge evaluations

First, challenge tasks are always simplified or abstracted from the real-world problems due to the nature of open evaluations (tasks have to be well defined and with a 'proper' level of difficulty). For instance, owing to the restricted access and difficulty in processing full text, a common simplification step shared by many BioNLP challenge tasks is to use abstracts instead, despite the fact that individual researchers, data indexers and curators routinely read the full text [61, 65]. Other examples include using a modest number of artificial and well-structured questions of limited types for QA or IR tasks, whereas in realty, information seekers typically ask complex and open-ended natural language questions that are often ill-formed and ungrammatical [3].

Second, there may not be sufficient participation or innovation in method development. Insufficient participation can happen because of the task itself (too difficult or unattractive) or owing to the other competing tasks in the same time. As can be seen in Figure 2, in recent years there have always been more concurrent tasks than an individual group could address. As a result, some tasks had low participation (less than five). Also, we notice that when a task is broken down into multiple subtasks, fewer teams afford to choose to complete it end-to-end. For instance, in *BioNLP-ST* 2009 [24], there were 24 teams participating in GE subtask-1, although only two teams completed all three subtasks. In terms of technical advancement, challenge tasks are designed to stimulate the research community to progress by developing new and different methodologies. However, when an existing method is found to be effective and competitive, we often see a lack of diversity in team methods. For example, of the top 12 systems in the Smoking Status detection task at the 2006 *i2b2* challenge, 9 used support vector machines with no statistically significant difference in their micro-averaged F-measures [80].

Finally, there exists a gap between challenge tasks and real-world use for several reasons. First, for some tasks, even the best automated results are significantly lower than what is desired in practical settings. Furthermore, as previously discussed, challenge tasks are always an abstraction of the real problems so that even a high-performing system will be plagued by many other factors (e.g. different text input, system scalability or interoperability) in real-world applications. For instance, previous *BioCreative* Gene Normalization challenge showed that the task performance dropped significantly when tested on full texts [54] instead of abstracts [53]. Finally, a tradition for many participants in such challenges is to achieve competitive performance and publish results in a reputable journal. Not until recently, there was almost no incentive for teams to adapt and push their methods into real-world uses.

Future trends

With regards to future trends, we believe that challenge evaluations will continue playing critical roles in BioNLP given their success so far. Although some fundamental tasks (e.g. NER, IE) may continue, we also expect to see new challenge tasks in the near future to address different user needs in biomedical research and health care. Continued push of turning successful techniques/methods into real-world applications is also expected [117, 118]. Toward such a goal, the assessment of system scalability and interoperability should become important factors in future challenge evaluations in addition to the traditional evaluation of system accuracy [119, 120]. Another trend may be that given limited resources, in lieu of separate challenge evaluations, a more collaborative competition framework can be coordinated in a more efficient and cost-effective manner, which would benefit both task organizers and participants. Finally, formal challenge evaluations have been shown to be useful up to a certain point but collaborative competition might be the means to collectively solve those real-world problems that are too complex for a single participating team.

Key Points

- We review BioNLP challenges in biological and clinical domains held from 2002 to 2014.
- We summarize the challenge subtasks based on NLP and entity-relation perspectives.
- We describe the steps of organizing and running community challenge evaluations.
- We discuss the impact and contributions of BioNLP challenges in multiple aspects, as well as their difficulties and limitations.

Acknowledgement

The authors would like to thank Dr. Robert Leaman for his help of proofreading this article and the anonymous reviewers for their helpful comments and suggestions. We also thank all the task organizers and participants for their efforts in the community challenges.

Funding

This work was supported by the National Institutes of Health intramural research program, National Library of Medicine.

References

1. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database (Oxford)* 2011;**2011**: baq036.
2. Khare R, Leaman R, Lu Z. Accessing biomedical literature in the current information landscape. *Methods Mol Biol* 2014; **1159**:11–31.
3. Islamaj Dogan R, Murray GC, Neveol A, *et al*. Understanding PubMed user search behavior through log analysis. *Database (Oxford)* 2009;**2009**:bap018.
4. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* 2006;**7**:119–29.
5. Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: an overview. *J Comput Biol* 2003;**10**(6):821–55.
6. Hunter L, Cohen KB. Biomedical language processing: what's beyond PubMed? *Mol Cell* 2006;**21**(5):589–94.
7. Berner ES, Moss J. Informatics challenges for the impending patient information explosion. *J Am Med Inform Assoc* 2005; **12**(6):614–17.
8. Linder JA, Ma J, Bates DW, *et al*. Electronic health record use and the quality of ambulatory care in the United States. *Arch Intern Med* 2007;**167**:1400–5.
9. Cohen KB, Hunter L. Natural language processing and systems biology. *Artif Intell Syst Biol* 2004:147–75.
10. Hersh W, Bhuptiraju RT, Ross L, *et al*. TREC 2004 genomics track overview. In: Proceedings of the 13th Text Retrieval Conference, 2004
11. Chen D, Muller HM, Sternberg PW. Automatic document classification of biological literature. *BMC Bioinformatics* 2006;**7**:370.
12. Krallinger M, Vazquez M, Leitner F, *et al*. The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinformatics* 2011;**12** (Suppl 8):S3.
13. Wang P, Morgan AA, Zhang Q, *et al*. Automating document classification for the Immune Epitope Database. *BMC Bioinformatics* 2007;**8**:269.
14. Gobeill J, Teodoro D, Patsche E, *et al*. Report on the TREC 2009 experiments: chemical IR track. In: Proceedings of the 18th Text Retrieval Conference, 2009.
15. Lu Z, Kim W, Wilbur WJ. Evaluation of query expansion using MeSH in PubMed. *Inf Retr Boston* 2009;**12**(1):69–80.
16. Lu Z, Kim W, Wilbur WJ. Evaluating relevance ranking strategies for MEDLINE retrieval. *J Am Med Inform Assoc* 2009;**16**(1):32–6.
17. Cohen AM, Hersh WR. The TREC 2004 genomics track categorization task: classifying full text biomedical documents. *J Biomed Discov Collab* 2006;**1**:4.
18. Segura-Bedmar I, Martínez P, Sánchez-Cisneros D. The 1st DDIExtraction-2011 challenge task: extraction of drug–drug interactions from biomedical texts. In: Proceedings of the 1st Challenge Task on Drug–drug Interaction Extraction, 2011, 1–9.
19. Krallinger M, Leitner F, Rodriguez-Penagos C, *et al*. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biol* 2008;**9** (Suppl 2):S4.

20. Ravikumar K, Liu H, Cohn JD, *et al*. Literature mining of protein-residue associations with graph rules learned through distant supervision. *J Biomed Semantics* 2012;**3** (Suppl 3):S2.
21. Nédellec C. Learning language in logic - genic interaction extraction challenge. In: Proceedings of the Learning Language in Logic 2005 Workshop, 2005, 31–7.
22. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc* 2013;**20**(5):806–13.
23. Hersh W, Bhupatiraju RT. TREC genomics track overview. In: Proceedings of the 12th Text Retrieval Conference, 2003.
24. Kim JD, Ohta T, Pyysalo S, *et al*. Overview of BioNLP'09 shared task on event extraction. In: Proceedings of the Workshop on BioNLP: Shared Task, 2009.
25. Yeh AS, Hirschman L, Morgan AA. Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics* 2003;**19** (Suppl 1):i331–9.
26. Wei CH, Harris BR, Li D, *et al*. Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database (Oxford)* 2012;**2012**:bas041.
27. Burger JD, Doughty E, Khare R, *et al*. Hybrid curation of gene-mutation relations combining automated extraction and crowdsourcing. *Database (Oxford)* 2014;**2014**:bau-094.
28. Khare R, Li J, Lu Z. LabeledIn: cataloging labeled indications for human drugs. *J Biomed Inform* 2014;**52**:448–56.
29. Khare R, Burger J, Aberdeen J, *et al*. Scaling drug indication curation through crowdsourcing. *Database (Oxford)* 2015; **2015**:bav016.
30. Spasic I, Ananiadou S, McNaught J, *et al*. Text mining and ontologies in biomedicine: making sense of raw text. *Brief Bioinform* 2005;**6**(3):239–51.
31. Ananiadou S, Pyysalo S, Tsujii J, *et al*. Event extraction for systems biology by text mining the literature. *Trends Biotechnol* 2010;**28**(7):381–90.
32. Ananiadou S. *Advances of Biomedical Text Mining for Semantic Search. Web Science in the Medical Domain* 2011;5.
33. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res* 2013;**41**: W518–22.
34. Salgado D, Krallinger M, Depaule M, *et al*. MyMiner: a web application for computer-assisted biocuration and text annotation. *Bioinformatics* 2012;**28**(17):2285–7.
35. Moult J, Pedersen JT, Judson R, *et al*. A large-scale experiment to assess protein structure prediction methods. *Proteins* 1995;**23**(3):ii-v.
36. Moult J, Fidelis K, Kryshtafovych A, *et al*. Critical assessment of methods of protein structure prediction (CASP)–round x. *Proteins* 2014;**82** (Suppl 2):1–6.
37. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. *Brief Bioinform* 2005;**6**(1):57–71.
38. Simpson MS, Demner-Fushman D. Biomedical text mining: a survey of recent progress. In: Biomedical Text Mining: A Survey of Recent Progress, 2012, 465–517.
39. Yeh A, Hirschman L, Morgan A. Background and Overview for KDD Cup 2002 Task 1: information extraction from biomedical articles. *ACM SIGKDD Explor Newsl* 2002;**4**(2):87–9.
40. Hersh W, Cohen A, Yang J, *et al*. TREC 2005 Genomics Track Overview. In: Proceedings of the 14th Text Retrieval Conference, 2005.
41. Lu Z, Cohen KB, Hunter L. Finding GeneRIFs via gene ontology annotations. *Pac Symp Biocomput* 2006:52–63.
42. Lu Z, Cohen KB, Hunter L. GeneRIF quality assurance as summary revision. *Pac Symp Biocomput* 2007:269–80.
43. Hersh W, Cohen AM, Roberts P, *et al*. TREC 2006 Genomics Track Overview. In: Proceedings of the 15th Text Retrieval Conference, 2006.
44. Hersh W, Cohen A, Ruslen L, *et al*. TREC 2007 Genomics Track Overview. In: Proceedings of the 16th Text Retrieval Conference, 2007.
45. Lupu M, Piroi F, Huang X, *et al*. Overview of the TREC 2009 chemical IR track. In: Proceedings of the 18th Text Retrieval Conference, 2009.
46. Lupu M, Tait J, Huang J, *et al*. TREC-CHEM 2010:Notebook report. In: Proceedings of the 19th Text Retrieval Conference, 2010.
47. Lupu MH, Gurulingappa I. Filippov, *et al*. Overview of the TREC 2011 Chemical IR Track. In: Proceedings of the 20th Text Retrieval Conference, 2011.
48. Farkas R, Vincze V, Móra G, *et al*. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In: Proceedings of the 14th Conference on Computational Natural Language Learning: Shared Task, 2010.
49. Yeh A, Morgan A, Colosimo M, *et al*. BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics* 2005;**6** (Suppl 1):S2.
50. Hirschman L, Yeh A, Blaschke C, *et al*. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* 2005;**6** (Suppl 1):S1.
51. Kim JD, Ohta T, Tsuruoka Y, *et al*. Introduction to the Bio-Entity Recognition Task at JNLPBA. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, 2004, 28–29.
52. Hirschman L, Colosimo M, Morgan A, *et al*. Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics* 2005;**6** (Suppl 1):S11.
53. Morgan AA, Lu Z, Wang X, *et al*. Overview of BioCreative II gene normalization. *Genome Biol* 2008;**9** (Suppl 2):S3.
54. Lu Z, Kao HY, Wei CH, *et al*. The gene normalization task in BioCreative III. *BMC Bioinformatics* 2011;**12** (Suppl 8):S2.
55. Wiegers TC, Davis AP, Mattingly CJ. Collaborative biocuration–text-mining development task for document prioritization for curation. *Database (Oxford)* 2012;**2012**: bas037.
56. Krallinger M, Leitner F, Rabal O, *et al*. Overview of the chemical compound and drug name recognition (CHEMDNER) task. In: Proceedings of the Fourth BioCreative Challenge Evaluation Workshop, 2013, 2–33.
57. Krallinger M, Rabal O, Leitner F, *et al*. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J Cheminform* 2015;**7**:S2.
58. Rebholz-Schuhmann D, Jimeno Yepes A, Li C, *et al*. Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. *J Biomed Semantics* 2011;**2** (Suppl 5):S11.
59. Rebholz-Schuhmann D, Jimeno Yepes AJ, Van Mulligen EM, *et al*. CALBC silver standard corpus. *J Bioinform Comput Biol* 2010;**8**(1):163–79.
60. Blaschke C, Leon EA, Krallinger M, *et al*. Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics* 2005;**6** (Suppl 1):S16.
61. Mao Y, Van Auken K, Li D, *et al*. Overview of the gene ontology task at BioCreative IV. *Database (Oxford)* 2014;**2014**: bau086.

62. Van Auken K, Schaeffer ML, McQuilton P, *et al*. BC4GO: a full-text corpus for the BioCreative IV GO task. *Database (Oxford)* 2014;**2014**:bau074.
63. Leitner F, Krallinger M, Cesareni G, *et al*. The FEBS Letters SDA corpus: a collection of protein interaction articles with high quality annotations for the BioCreative II.5 online challenge and the text mining community. *FEBS Lett* 2010;**584**(19):4129–30.
64. Segura-Bedmar I, Martínez P, Herrero-Zazo M. SemEval-2013 Task 9:Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In: Proceedings of the Seventh International Workshop on Semantic Evaluation, 2013, 341–50.
65. Huang M, Neveol A, Lu Z. Recommending MeSH terms for annotating biomedical articles. *J Am Med Inform Assoc* 2011; **18**(5):660–7.
66. Partalas I, Gaussier E, Ngomo AN. Results of the First BioASQ Workshop. In: Proceedings of the first Workshop on BioASQ, 2013.
67. Balikas G, Partalas I, Ngomo ACN, *et al*. Results of the BioASQ Track of the Question Answering Lab at CLEF 2014. In: Results of the BioASQ Track of the Question Answering Lab at CLEF 2014. 2014. 1181–93.
68. Kim JD, Wang Y, Takagi T, *et al*. Overview of Genia Event Task in BioNLP Shared Task 2011. In: Proceedings of the BioNLP Shared Task 2011 Workshop, 2011, 7–15.
69. Kim JD, Wang Y, Yasunori Y. The Genia Event Extraction Shared Task, 2013 Edition - Overview. In: Proceedings of the BioNLP Shared Task 2013 Workshop, 2013, 8–15.
70. Bossy R, Jourde J, Manine AP, *et al*. BioNLP shared task—The bacteria track. *BMC Bioinformatics* 2012;**13** (Suppl 11): S3.
71. Bossy R, Golik W, Ratkovic Z, *et al*. BioNLP shared Task 2013 – An Overview of the Bacteria Biotope Task. In: Proceedings of the BioNLP Shared Task 2013 Workshop, 2013, 74–82.
72. Pyysalo S, Ohta T, Rak R, *et al*. Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011. *BMC Bioinformatics* 2012;**13** (Suppl 11):S2.
73. Pyysalo S, Ohta T, Ananiadou S. Overview of the Cancer Genetics (CG) task of BioNLP Shared Task 2013. In: Proceedings of the BioNLP Shared Task 2013 Workshop, 2013, 58–66.
74. Ohta T, Pyysalo S, Rak R, *et al*. Overview of the Pathway Curation (PC) task of BioNLP Shared Task 2013. In: Proceedings of the BioNLP Shared Task 2013 Workshop, 2013, 67–75.
75. Kim JD, Nguyen N, Wang Y, *et al*. The Genia Event and Protein Coreference tasks of the BioNLP shared task 2011. *BMC Bioinformatics* 2012;**13** (Suppl 11):S1.
76. Jourde J, Manine AP, Veber P, *et al*. BioNLP Shared Task 2011 – Bacteria Gene Interactions and Renaming. In: Proceedings of BioNLP Shared Task 2011 Workshop, 2011.
77. Nédellec C, Bossy R, Kim JD, *et al*. Overview of BioNLP Shared Task 2013. In: Proceedings of the BioNLP Shared Task 2013 Workshop, 2013, 1–7.
78. Jin F, Huang M, Lu Z, *et al*. Towards automatic generation of gene summary. In: Proceedings of BioNLP workshop, 2009.
79. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc* 2007;**14**(5):550–63.
80. Uzuner O, Goldstein I, Luo Y, *et al*. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc* 2008;**15**(1):14–24.
81. Uzuner O. Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc* 2009;**16**(4):61–70.
82. Xu Y, Wang Y, Liu J, *et al*. Suicide note sentiment classification: a supervised approach augmented by web data. *Biomed Inform Insights* 2012;**5**(Suppl. 1):31–41.
83. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;**17**(5): 514–8.
84. Uzuner O, South BR, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;**18**(5):552–6.
85. Voorhees EM, Tong RM. Overview of the TREC 2011 medical records track. In: Proceedings of the Text Retrieval Conference, 2011.
86. Voorhees EM, Hersh W. Overview of the TREC 2012 medical records track. In: Proceedings of the Text Retrieval Conference, 2012.
87. Suominen H, Salantera S, Velupillai S, *et al*. Overview of the ShARe/CLEF eHealth evaluation lab 2013. *Lect Notes Comput Sci* 2013;**8138**:212–31.
88. Pradhan S, Elhadad N, Chapman W, *et al*. SemEval-2014 task 7: analysis of clinical text. In: Proceedings of the 8th workshop on semantic evaluation, 2014, 54–62.
89. Kelly L, Goeuriot L, Suominen H, *et al*. Overview of the ShARe/CLEF eHealth evaluation lab 2014. In: Proceedings of the CLEF, 2014.
90. Mowery DL, Velupillai S, South BR, *et al*. Task 2: ShARe/CLEF eHealth evaluation lab 2014. In: Proceedings of the CLEF, 2014.
91. Goeuriot L, Kelly L, Li W, *et al*. ShARe/CLEF eHealth evaluation lab 2014, Task 3: User-centred health information retrieval. In: Proceedings of the CLEF, 2014.
92. Arighi CN, Wu CH, Cohen KB, *et al*. BioCreative-IV virtual issue. *Database (Oxford)* 2014;**2014**:bau039.
93. Van Landeghem S, Bjorne J, Wei CH, *et al*. Large-scale event extraction from literature with multi-level gene normalization. *PLoS One* 2013;**8**(4):e55814.
94. Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics* 2013;**29**(22):2909–17.
95. Wei CH, Kao HY, Lu Z. SR4GN: a species recognition software tool for gene normalization. *PLoS One* 2012;**7**(6): e38460.
96. Leaman R, Wei CH, Lu Z. tmChem: a high performance tool for chemical named entity recognition and normalization. *J Cheminform* 2015;**7**:S3.
97. Hirschman L, Grishman R, Sager N. From text to structured information: automatic processing of medical reports. In: Proceedings of the National Computer Conference and Exposition, 1976.
98. Hersh W, Buckley C, Leone TJ, *et al*. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: The 17th ACM SIGIR Conference on Research and Development in Information Retrieval, 1994, 192–201.
99. Ruch P, Baud RH, Rassinoux AM, *et al*. Medical document anonymization with a semantic lexicon. In: Proc AMIA Symp 2000, 729–33.
100. Franzen K, Eriksson G, Olsson F, *et al*. Protein names and how to find them. *Int J Med Inform* 2002;**67**(1–3):49–61.
101. Maglott D, Ostell J, Pruitt KD, *et al*. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2005;**33**: D54–8.

102. Coordinators NR. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2013;**41**:D8-D20.
103. Magrane M, Consortium UniProt. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011; **2011**:bar009.
104. Reference Genome Group of the Gene Ontology Consortium. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput Biol* 2009;**5**(7):e1000431.
105. Neveol A, Islamaj Dogan R, Lu Z. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. *J Biomed Inform* 2011;**44**(2):310–8.
106. Jessop DM, Adams SE, Willighagen EL, *et al*. OSCAR4: a flexible architecture for chemical text-mining. *J Cheminform* 2011;**3**(1):41.
107. Dogan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. *J Biomed Inform* 2014;**47**:1–10.
108. Doğan RI, Lu Z. An improved corpus of disease mentions in PubMed citations. In: Proceedings of BioNLP workshop, Association for Computational Linguistics, 2012.
109. Stenetorp A, Pyysalo S, Topić G, *et al*. brat: a Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012.
110. Mao Y, Wei CH, Lu Z. NCBI at the 2014 BioASQ challenge task: large-scale biomedical semantic indexing and question answering. In: Proceedings of the 2014 Question Answering Lab at CLEF, 2014.
111. Comeau DC, Islamaj Dogan R, Ciccarese P, *et al*. BioC: a minimalist approach to interoperability for biomedical text processing. *Database (Oxford)* 2013;**2013**:bat064.
112. Comeau DC, Batista-Navarro RT, Dai HJ, *et al*. BioC interoperability track overview. *Database (Oxford)* 2014; **2014**.
113. Lu Z, Hirschman L. Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. *Database (Oxford)* 2012;**2012**:bas043.
114. Cejuela JM, McQuilton P, Ponting L, *et al*. tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database (Oxford)* 2014; **2014**:bau033.
115. Lu Z. *Text Mining in GeneRIFs*. PhD thesis, University of Colorado Denver, 2007.
116. Wei CH. *The Recognition and Normalization of Biomedical and Biological Concepts*. PhD thesis, National Cheng Kung University, 2012.
117. Arighi CN, Roberts PM, Agarwal S, *et al*. BioCreative III interactive task: an overview. *BMC Bioinformatics* 2011;**12** (Suppl 8):S4.
118. Arighi CN, Carterette B, Cohen KB, *et al*. An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. *Database (Oxford)* 2013;**2013**: bas056.
119. Wiegers TC, Davis AP, Mattingly CJ. Web services-based text-mining demonstrates broad impacts for interoperability and process simplification. *Database (Oxford)* 2014;**2014**: bau050.
120. Khare R, Wei CH, MaoY, *et al*. tmBioC: improving interoperability of text-mining tools with BioC. *Database (Oxford)* 2014;**2014**.