OXFORD

Briefings in Bioinformatics, 19(6), 2018, 1400–1414

doi: 10.1093/bib/bbx057
Advance Access Publication Date: 13 June 2017
Paper

Biomedical text mining for research rigor and integrity: tasks, challenges, directions

Halil Kilicoglu

Corresponding author: Halil Kilicoglu, Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD 20894, USA. Tel.: +1(301)827-5014; Fax: +1(301)496-0673; E-mail: kilicogluh@mail.nih.gov

Abstract

An estimated quarter of a trillion US dollars is invested in the biomedical research enterprise annually. There is growing alarm that a significant portion of this investment is wasted because of problems in reproducibility of research findings and in the rigor and integrity of research conduct and reporting. Recent years have seen a flurry of activities focusing on standardization and guideline development to enhance the reproducibility and rigor of biomedical research. Research activity is primarily communicated via textual artifacts, ranging from grant applications to journal publications. These artifacts can be both the source and the manifestation of practices leading to research waste. For example, an article may describe a poorly designed experiment, or the authors may reach conclusions not supported by the evidence presented. In this article, we pose the question of whether biomedical text mining techniques can assist the stakeholders in the biomedical research enterprise in doing their part toward enhancing research integrity and rigor. In particular, we identify four key areas in which text mining techniques can make a significant contribution: plagiarism/fraud detection, ensuring adherence to reporting guidelines, managing information overload and accurate citation/enhanced bibliometrics. We review the existing methods and tools for specific tasks, if they exist, or discuss relevant research that can provide guidance for future work. With the exponential increase in biomedical research output and the ability of text mining approaches to perform automatic tasks at large scale, we propose that such approaches can support tools that promote responsible research practices, providing significant benefits for the biomedical research enterprise.

Key words: biomedical research waste; biomedical text mining; natural language processing; research rigor; research integrity; reproducibility

Introduction

Lack of reproducibility and rigor in published research, a phenomenon sometimes referred to as the ‘reproducibility crisis’, is a growing concern in science. In a recent *Nature* survey, 90% of the responding scientists agreed that there was a crisis in science [1]. It has become routine in recent years for scientific journals as well as for news media to publish articles discussing various aspects of this crisis as well as proposals and initiatives to address them. The reproducibility problem is perhaps most acutely felt in biomedical research, where the stakes are high because of the size of research investment and impact on public health. In 2010, the global spending on research in life sciences (including biomedical) was US$240 billion [2]. The problems in reproducibility and rigor of published research mean that a portion of this expenditure is wasted. Chalmers and Glasziou [3] estimate that avoidable waste accounts for ∼85% of the research investment.

A variety of factors, occurring at various stages of research and attributable to different stakeholders, can lead to reproducibility issues and, ultimately, waste. For example, at the conception, the scientist, unaware of the published literature, may propose to address a research question that can already be answered with existing evidence, or may fail to select the appropriate experimental design and methods [3, 4]. As the research is being conducted, the investigator, overwhelmed with administrative tasks, may not be able to provide adequate training/supervision to laboratory staff [5], who do not validate their experiments sufficiently. Only a subset of data that yields

Halil Kilicoglu is a Staff Scientist at the Lister Hill National Center for Biomedical Communications, US National Library of Medicine. His research focuses on biomedical language processing and text mining, literature-based discovery and question answering.

Submitted: 7 February 2017; **Received (in revised form):** 10 April 2017

statistical significant results may be reported, while negative results may be discarded completely (p-hacking, selective reporting or publication bias) [6–8]. The authors may neglect to identify the model organisms, antibodies and reagents necessary for other researchers to replicate the experiments [9]. Journal editors, valuing novelty over reproducibility, may be reluctant to publish negative results or replication studies [4]. A peer reviewer may miss methodological problems with the manuscript. An institutional review board (IRB) may fail to follow up on biased underreporting of the research that they approve [10]. Funding agencies may put too much emphasis on number of publications, citation counts and research published in journals with high impact factors for rewarding research grants [4, 11]. The so-called 'publish or perish' culture at academic institutions can create pressure to maximize research quantity with diminishing quality [4].

While research rigor and reproducibility in biomedical research is not a recent problem, discussions of the 'reproducibility crisis' are largely because of several recent high-profile studies. In one of the pioneering studies, Ioannidis [12] demonstrated how reliance on hypothesis testing in biomedical research frequently results in false-positive results, which he attributed to a variety of factors, such as effect size, flexibility in study design and financial interest and prejudice. More recently, Begley and Ellis [13] were unable to reproduce the findings reported in 47 of 53 landmark hematology and oncology studies. Studies with similar findings were conducted in other fields, as well [14–16]. Lack of reproducibility and rigor can mostly be attributed to questionable research practices (honest errors, methodological problems). At the extreme end of the reproducibility spectrum, fraudulent science and retractions constitute a small but growing percentage of the published literature. The percentage of retracted articles in PubMed has increased about 10-fold since 1975 and 67.4% are attributable to scientific misconduct: fraud, duplicate publication and plagiarism [17]. Owing to their pervasiveness, however, questionable research practices can be much more detrimental to science [18]. Biomedical research outcomes, estimated by life expectancy and novel therapeutics, have remained constant despite rising investment and scientific knowledge in the past five decades, partly attributed to non-reproducibility [11]. Such evidence establishes the lack of reproducibility and rigor as a major problem that can potentially undermine trust in biomedical research enterprise. All stakeholders involved in the biomedical research enterprise have a responsibility to ensure the accuracy, verifiability and honesty of research conduct and reporting to reduce waste and increase value.

Toward increased rigor and reproducibility, initiatives focusing on various aspects of reproducible science have been formed and they have been publishing standards, guidelines and principles. These include International Committee of Medical Journal Editors (ICMJE) trial registration requirement [19] and recommendations for the conduct and publication of scholarly work in medical journals [20], National Institutes of Health (NIH) efforts in enhancing research reproducibility and transparency [4] and data discovery [21], in addition to reporting guidelines (e.g. Consolidated Standards of Reporting Trials (CONSORT) statement [22] and Transparency and Openness Promotion (TOP) guidelines [23]), data-sharing principles [24], conferences (e.g. World Conference on Research Integrity), journals (e.g. Research Integrity and Peer Review) and centers (e.g. Center for Open Science, METRIC) dedicated to these topics.

We have used several terms (reproducibility, replication, rigor and integrity) somewhat interchangeably to describe related phenomena that differ in some fundamental aspects. The semantics of these terms are still somewhat muddled, leading to confusion and potentially hampering efforts to fix the problems [25]. To better focus the remainder of this article, we use the definitions below, provided in Bollen *et al.* [26].

- **Reproducibility:** The ability to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator.
- **Replicability:** The ability to duplicate the results of a prior study if the same procedures are followed but new data are collected.
- **Generalizability:** Whether the results of a study apply in other contexts or populations that differ from the original one (also referred to as translatability).

Results that are reproducible, replicable and generalizable are referred to as being robust.

The notions of rigor and integrity are also invoked to discuss related phenomena. The definitions below are taken from NIH's Grants and Funding resources:

- **Rigor:** Strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results (https://grants.nih.gov/reproducibility/index.htm).
- **Integrity:** The use of honest and verifiable methods in proposing, performing and evaluating research, reporting results with attention to adherence to rules, regulations, guidelines and following commonly accepted professional codes and norms (https://grants.nih.gov/grants/research_integrity/whatis.htm).

While reproducibility is often used as an umbrella term to cover all these related issues, in the remainder of this article, we use it in the limited sense given above. Instead, we focus on the notions of research rigor and integrity because (i) these notions emphasize reporting over duplication of prior experiments, and (ii) we are primarily interested in whether/how mining of textual research artifacts can contribute to open, transparent and rigorous biomedical science.

It is safe to assume that most biomedical scientists are well-intentioned and are doing their best to conduct rigorous research and report it accurately. Some of the reproducibility/rigor issues are natural consequences of difficulty and inherent uncertainty of empirical research. However, failures that cannot be attributed to these characteristics also occur, somewhat frequently, and our goal should be to minimize them. We envision that automatic text mining approaches play a dual role toward this goal by providing support for (i) scrutinizing reports of already conducted research (corrective role), and (ii) managing published literature to improve the quality of proposed/ongoing research (preventive role).

Mining of textual biomedical research artifacts is in the purview of biomedical natural language processing (referred to as bioNLP, henceforth), a field at the intersection of natural language processing (NLP) and biomedical informatics. In this article, we assume basic knowledge of bioNLP; for introductions and recent surveys, see [27–29]. In the next section, we turn to our original question: Can bioNLP provide tools that can help stakeholders in enhancing rigor and integrity of biomedical research?

BioNLP for research rigor and integrity

Text mining is largely concerned with unstructured text, the primary means of communication for biomedical research.

Unstructured biomedical text comes in various forms of textual artifacts, including:

- *Proposals* (grant applications and protocols), authored by scientists and assessed by funding agencies, reviewers and IRBs
- *Manuscript submissions*, authored by scientists; evaluated by journal editors, program chairs and peer reviewers; and edited by journal staff
- *Publications*, authored by scientists, read and assessed by other scientists, systematic reviewers, database curators, funding agencies, IRBs, academic institutions and policymakers

Clark *et al.* [30] conceptualize the ecosystem of biomedical communication as a cycle of nine activities, with inputs and outputs (the output of the last activity feeding back into the first):

1. Authoring
2. Reviewing for Publication
3. Editing and Publishing
4. Database and Knowledge Base Curation
5. Searching for Information
6. Reading
7. Discussion
8. Evaluating and Integrating Information
9. Experiment Design and Execution

Textual artifacts are primary inputs and outputs of some of these activities. For example, inputs for authoring include other relevant publications in the research space, as well as experimental data and observations, and the output is a manuscript for submission. By providing the ability to automatically process such artifacts at a large scale and extract relevant information for subsequent activities, bioNLP methods have the potential to assist scientists with the entire life cycle of biomedical communication. Though not explicit in Clark *et al.*'s conceptualization, the same capabilities can also benefit other stakeholders (journal editors, reviewers, funding agencies, etc.), who need to evaluate such artifacts based on their scientific merit.

What kinds of text mining tools can be envisioned? What kinds of benefits can they provide? We briefly outline several categories of tools and their potential benefits below.

1. *Plagiarism/fraud detection:* Although plagiarism and outright fraud are relatively rare (though seemingly growing [17]) in scientific literature, tools that can detect plagiarism/fraud can be helpful to journal editors in catching research misconduct before publishing an article and avoiding later retractions, which may reflect badly on the journal.
2. *Adherence to reporting guidelines:* Tools that can assess a manuscript against the relevant reporting guidelines [e.g. the CONSORT statement [22] for randomized clinical trials (RCTs)] and flag the issues would be useful for journal editors, who can then require the authors to fix the problems for publication.
3. *Managing information overload:* Text mining tools can help in managing information overload by summarizing and aggregating salient knowledge (e.g. hypotheses, claims and supporting evidence) in textual artifacts, a capability that can benefit all stakeholders. Efficient knowledge management can help research rigor and reduce research waste by ensuring that, for example, scientists are aware of all relevant studies before embarking on a research project [31] or that funding agencies are not awarding funds to redundant or unjustified proposals [32, 33].
4. *Accurate citation and enhanced bibliometrics:* Tools that can verify whether the authors cite relevant literature (or omit) accurately would be beneficial in reducing citation distortion, which has been shown to lead to unfounded authority [34]. Advanced citation analysis tools that can recognize the function of a citation and its significance for the publication can help funding agencies and academic institutions move beyond simple citation counts and make more informed decisions about the impact of a particular study.

Among these categories, the first two can be viewed as having corrective roles, while information overload management is mainly preventive, and citation-related tools can play both roles in ensuring rigor and integrity.

Although text mining has been used to address a variety of tasks that can be subsumed by the categories outlined above, there is little research on using text mining for broadly addressing research integrity and rigor issues in biomedical science. One nascent effort is a collaboration between the academic publisher Elsevier and Humboldt University (http://headt.eu), which aims to use text/data mining for early detection of integrity issues, focusing mainly on plagiarism/fraud, image manipulation and data fabrication, although no published results were available at the time of this writing.

In the remainder of this section, we review the existing NLP/bioNLP research on the four categories of tasks that we outlined above.

Plagiarism/fraud detection

Plagiarism is 'the appropriation of another person's ideas, processes, results, or words without giving appropriate credit' [35]. A serious problem in academia, especially with regard to student writing, plagiarism also occurs in medical publications [35]. Plagiarism comes in several forms: at one end of the spectrum is copy–paste plagiarism, which is relatively easy to detect, and on the other end is paraphrased or translated plagiarism, which can be challenging. Plagiarism detection is a well-studied information retrieval task, and dedicated tools have been developed [e.g. TurnItIn (http://www.turnitin.com) and Plagramme (http://www.plagramme.com)]. CrossRef Similarity Check (http://www.crossref.org/crosscheck), a TurnItIn-based tool used by publishers, specifically targets plagiarism in scholarly communication. It generates an overall similarity score between a manuscript and the articles in a large database of publications and flags the manuscript if its similarity score is over a publisher-determined threshold.

Generally, two plagiarism detection tasks are distinguished: extrinsic and intrinsic plagiarism detection. In extrinsic plagiarism detection, a document is compared with other candidate documents in a reference collection. Approaches to this task differ with respect to how documents are represented: document fingerprints based on substring hashing [36] or vectors [37]. Vector representations can be based on characters, words, word sequences (*n*-grams) or stopwords [38]. A high number of candidate documents can pose challenges in extrinsic plagiarism detection [37]; therefore, efficient document representations and candidate document selection can be critical. In intrinsic plagiarism detection, the goal is to recognize shifting writing styles within a document to spot plagiarism [39]. Methods for this task rely on stylometric features, such as word class usage and average word frequency, which indicate the author's vocabulary size and style complexity. Plagiarism detection has been the topic of PAN shared task challenges (http://pan.webis.de), the last edition of which took place in 2016 [40]. Performance for extrinsic plagiarism detection in these competitions has reached an F_1 score of 0.88 [41], while the state-of-the-art performance for intrinsic plagiarism detection is much lower at an F_1 score of 0.22 [42]. Plagiarism in the PAN corpora was simulated, whereas Nawab *et al.* [43] used a corpus of PubMed abstracts that were deemed to be suspiciously similar to other abstracts and used a query expansion approach based on Unified Medical Language System (UMLS) Metathesaurus [44] to detect plagiarism. Plagiarism detection tools are most beneficial to journal editors and peer reviewers, though scientists can also benefit from using such tools to prevent inadvertent plagiarism or self-plagiarism.

Plagiarism accounts for a relatively small fraction of retractions in biomedical research articles (9.8%), while the fraud accounts for 43.4% [17]. Such cases often involve data fabrication or falsification [45], types of misconduct that would typically be difficult to flag with text analysis alone. Focusing on text only, Markowitz and Hancock [46] investigated whether scientists write differently when reporting fraudulent research. They compared the linguistic style of publications retracted for fraudulent data with that of unretracted articles and articles retracted for reasons other than fraud. They calculated a linguistic obfuscation score based on stylistic and psycholinguistic characteristics of a document, including ease of reading, rate of jargon, causal and abstract words. They found that retracted articles were written with significantly higher levels of linguistic obfuscation and that obfuscation was positively correlated with the number of references. However, their score-based method had a high false-positive rate, and they suggested that NLP techniques could achieve higher classification performance. A task similar to fraud detection, considered in open-domain NLP, is deception detection, generally cast as a binary classification task (deceptive versus true) [47]. Supervised machine learning techniques [support vector machines (SVMs), Naive Bayes] using n-gram and psycholinguistic features have been applied to this task [47, 48], the latter achieving F_1 score of 0.90 [48]. Interestingly, inter-annotator agreement (Fleiss' $\kappa = 0.11$) and human judgements (0.60 F_1) were found to be lower than machine performance. The classification approach used for deception detection is likely to be beneficial in detecting fraudulent articles. Similarly to plagiarism detection, fraud detection tools would be most useful to journal editors.

We note that, in general, decisions regarding fraud or plagiarism should ultimately only be made by humans, as such accusations can be damaging to a scientist's career. However, text mining approaches can, to some extent, flag suspicious manuscripts, which can then be given extra scrutiny.

Adherence to reporting standards and guidelines

Reporting guidelines have been proposed for transparent and accurate reporting of biomedical research toward improving rigor and reproducibility. For various types of biomedical studies, reporting guidelines have been developed under the auspices of the EQUATOR Network [49]. These include CONSORT for RCTs [22] and ARRIVE for preclinical animal studies [50], among others. The CONSORT Statement consists of a 25-item checklist and a flow diagram. The CONSORT checklist for 'Methods' sections is provided in section S1 of the Supplementary Material as an example.

Although a large number of journals have adopted or supported such guidelines, adherence to them remains inadequate [51]. As a solution, structured reporting of key methods and findings has been proposed [52]. Until such proposals gain currency, however, most methodological information is likely to remain buried in narrative text or, in the worst case, be completely absent from the publication. Text mining tools can help journal editors enforce adherence to reporting guidelines by locating key statements corresponding to specific guideline items or giving alerts in their absence. For example, per CONSORT, the method used to generate the random allocation sequence as well as statistical details critical for reproducibility can be identified. Recognition of description of limitations and sources of possible bias can be beneficial for broader rigor and generalizability. Additionally, medical journals require or encourage inclusion of certain types of meta-information, such as funding, conflicts of interest, trial registration and data access statements. Identifying such meta-statements and locating statements corresponding to guideline items can both be categorized as information extraction tasks, and similar techniques (text/sentence classification, sequence labeling) can be applied to them. The difficulty of extracting these items varies widely: locating trial registration information seems relatively easy, as the trial registration numbers have a standard format, whereas extracting statements that indicate that interpretation is 'consistent with results, balancing benefits and harms, and considering other relevant evidence' (a CONSORT checklist item) seems challenging, as some subjectivity may be involved and a deeper text understanding may be needed. Commercial software has been developed to address adherence issues to some extent [e.g. Penelope Research (http://www.peneloperesearch.com/) and StatReviewer (http://www.statreviewer.com/)]; however, they currently have limited capabilities, and information about the underlying technologies is sparse. Tools that can address the guideline adherence would be useful not only for journals and reviewers but also to authors of systematic reviews, who aim to identify well-designed, rigorous studies, and to clinicians looking for the best available clinical evidence.

We are not aware of any published bioNLP research that aims to determine whether a manuscript fully complies with the relevant set of reporting guidelines. However, some research attempts to identify one or more guideline items as well as other meta-information, often for the purpose of automating systematic review process [53]. We discuss such research below; see O'Mara-Eves *et al.* [54] for a general discussion of using text mining for systematic reviews.

In the simplest case, some statistical details, such as P-values and confidence intervals, can be identified with simple regular expressions [55]. Some meta-information, such as funding, conflict of interest or trial registration statements, often appear in dedicated sections and are expressed using a limited vocabulary; hence, simple keyword-based techniques could be sufficient. For example, Kafkas *et al.* [56] mined data accession numbers in full-text articles using regular expressions.

Other items require more sophisticated techniques, often involving machine learning. For example, Kiritchenko *et al.* [57] extracted 21 key elements from full-text RCT articles (e.g. eligibility criteria, sample size, primary outcomes and registration number), some of which are included in CONSORT. Their system, trained on an annotated corpus of 132 articles, used a two-stage pipeline: first, given a particular element, a classifier predicted whether a sentence is likely to contain information about that element. Second, regular expressions were used to extract the exact mention of the element. Some meta-information (DOI, author name and publication date) was simply extracted from PubMed records. For each remaining element, an SVM model with *n*-gram features was trained for sentence classification. A post hoc evaluation on a test corpus of 50 articles yielded a precision of 0.66 for these elements (0.94 when partial matches were considered correct).

PICO frame elements (Problem, Population, Intervention, Comparison and Outcome) are recommended for formulating clinical queries in evidence-based medicine [58] and are widely used in the systematic review process to assess methodological rigor of a study. They also often appear in reporting guideline checklists (e.g. participants in CONSORT versus population in PICO). Some research focused on PICO or its variants. Demner-Fushman and Lin [59] identified PICO elements in PubMed abstracts for clinical question answering. Outcomes were extracted as full sentences and other elements as short noun phrases. The results from an ensemble of classifiers (rule-based, *n*-gram-based, position-based and semantic group-based), trained on an annotated corpus of 275 abstracts, were combined to recognize outcomes. Other elements were extracted using rules based on the output of MetaMap [60], a system that maps free text to UMLS Metathesaurus concepts [44]. Recently, Wallace *et al.* [61], noting that PICO elements may not appear in abstracts, attempted to extract PICO sentences from full-text RCT articles. They generated sentence-level annotations automatically from free-text summaries of PICO elements in the Cochrane Database of Systematic Reviews (CDSR), using a novel technique called supervised distant supervision. A small number of sentences in the original articles that were most similar to CDSR summary sentences were identified and a manually annotated subset was leveraged to align unlabeled instances with the structured data in CDSR. Separate models were learned for each PICO element with bag-of-words and positional features as well as features encoding the fraction of numerical tokens, whether the sentence contains a drug name, among others. Their technique outperformed models that used direct supervision or distant supervision only. A PICO variant, called PIBOSO (B: Background, S: Study design, O: Other), was also studied [62]. A corpus of 1000 PubMed abstracts was annotated with these elements (PIBOSO-NICTA corpus), and two classifiers were trained on this corpus: one identified PIBOSO sentences and the other assigned PIBOSO labels to these sentences. A conditional random field (CRF), a sequence labeling model, was trained using bag-of-words, *n*-gram, part of speech, section heading, position and sequence features as well as domain information from MetaMap. With the availability of the PIBOSO-NICTA corpus, several studies have explored similar machine learning-based approaches [63, 64] and state-of-the-art results, without using any external knowledge, were reported by Hassanzadeh *et al.* [64] (0.91 and 0.87 F_1 scores on structured and unstructured abstracts, respectively).

Marshall *et al.* [65] developed a tool called RobotReviewer to identify risk of bias (RoB) statements in clinical trials, which are included in CONSORT reporting guidelines. They used seven risk categories specified in the Cochrane RoB tool (e.g. random sequence generation, allocation concealment, blinding of participants and personnel and selective outcome reporting) and labeled articles as high or low risk with respect to a particular category. Similar to their approach for extracting PICO statements, they semiautomatically generated positive instances for training by leveraging CDSR, where systematic reviewers copy/paste a fragment from the article text to support their RoB judgements. An SVM classifier based on multitask learning mapped articles to RoB assessments and simultaneously extracted supporting sentences with an accuracy of 0.71 (compared with 0.78 human accuracy).

Availability of data and experimental protocols is important for replication of a study [66]. Extracting data deposition statements (i.e. where data used in the study can be retrieved from) from publications is, from a methodological perspective, a task similar to extracting PICO elements. Névéol *et al.* [67] focused on extracting deposition statements of biological data from full-text articles. They semiautomatically constructed a gold standard corpus. Their approach consisted of two machine learning models: one recognized data deposition components (data, action, general location and specific location) using a CRF model. The main model, a binary classifier, predicted whether a sentence contains a data deposition statement. This classifier, trained with Naive Bayes and SVM algorithms, used token, part of speech and positional features as well as whether the sentence included components identified by the CRF model. An article was considered positive for data deposition if the top-scored sentence was classified as positive. Their system yielded an F_1 score of 0.81.

Considering research rigor more broadly, Kilicoglu *et al.* [68] developed machine learning models to recognize methodologically rigorous, clinically relevant publications to serve evidence-based medicine. Several binary classifiers (Naive Bayes, SVM and boosting) as well as ensemble methods (stacking) were trained on a large set of PubMed abstracts previously annotated to develop PubMed Clinical Queries filter [69]. The base features used included token, PubMed metadata and semantic features, as extracted by MetaMap and SemRep [70], a biomedical relation extraction tool. Best results (F_1 score of 0.67) were achieved with a stacking classifier that used base models trained with various feature–classifier combinations (e.g. SVM with token features only).

Managing information overload

With the considerable size and the rapid growth of the biomedical literature, management of information overload becomes a critical part of research planning, conduct and assessment. By effectively managing literature, scientists and other stakeholders can make better decisions with respect to, for example, designing experiments, evaluating evidence and assessing research proposals, improving research quality and value. Thus, effective information management can serve a preventive role in supporting rigorous research. Unlike the tasks discussed so far, which did not require much deep natural language understanding, managing information overload is a broad problem that requires deeper semantic understanding because we need to be able to represent and extract meaning of a wide variety of biomedical text in a unified and scalable manner. In the NLP community, there are various proposals but little consensus regarding how to best represent meaning; therefore, the discussion in this section is somewhat more speculative than that in the previous sections.

A strategy for efficient management of the biomedical literature should support extraction of the hypotheses and the key arguments made in a research article (referred to as knowledge claims [71], henceforth) as well as their contextualization (e.g. identifying the evidence provided to support these claims, the level of certainty with which the claims are expressed and whether they are new knowledge). It should also allow aggregating such knowledge over the entire biomedical literature. A deeper text understanding is required for such capabilities, and we argue that the key to them is normalization of claims and the supporting evidence into computable semantic representations that can account for lexical variability and ambiguity. Such representations make the knowledge expressed in natural language amenable to automated inference and reasoning [72]. Furthermore, they can form the building blocks for advanced information seeking and knowledge management tools, such as semantic search engines, which can help us navigate the relevant literature more efficiently. For example, formal representations of knowledge claims can underpin tools that enable searching, verifying and tracking claims at a large scale, and summarizing research on a given biomedical topic, thus reducing the time spent locating/retrieving information and increasing the time spent interpreting it. Such tools can also address siloization of research [73, 74], putting research questions in a larger biomedical context and potentially uncovering previously unknown links from areas that the researcher does not typically interact with. Literature-scale knowledge extraction and aggregation on a continuous basis can also facilitate ongoing literature surveillance, with tools that alert the user when a new knowledge claim related to a topic of interest is made, when a claim of interest to the user is discredited or contradicted, increasing research efficiency (a service similar to Crossref's CrossMark, which indicates updates on a given publication, can be envisioned). Advanced knowledge management tools would be beneficial to all parties involved in biomedical research: (i) to researchers in keeping abreast of the literature, generating novel hypotheses and authoring papers; (ii) to funding agencies, IRBs and policymakers in better understanding the state of the art in specific research areas, creating research agendas/policies, verifying claims and evidence presented in proposals and assessing whether the proposed research is justified; (iii) to journal editors, peer reviewers, systematic reviewers, and database curators in locating, verifying and tracking claims and judging evidence presented in manuscripts and publications.

What do we mean by normalization of knowledge claims and evidence? With normalization, we refer to recognition of biomedical entities, their properties and the relationships between them expressed in text and mapping them to entries in a relevant ontology or knowledge base. As the basis of such formalization, we distinguish three levels of semantic information to be extracted: conceptual, relational and contextual. Roughly, the conceptual level is concerned with biomedical entities (e.g. diseases and drugs), relational level with biomedical relationships (e.g. gene–disease associations) and the contextual level with how these relationships are contextualized and related for argumentation. A knowledge claim, in the simplest form, can be viewed as a relation. We illustrate these levels on a PubMed abstract in Section S2 of the Supplementary Material.

Conceptual level is in the purview of the named entity recognition and normalization (NER/NEN) task, while relation extraction focuses on the relational level. These tasks are well studied in bioNLP. We provide a brief overview in Sections S3 and S4, respectively, of the Supplementary Material; see recent surveys [29, 75] for more comprehensive discussion. In the remainder of this subsection, we first briefly discuss tools that address information overload using concepts and relations extracted from the literature and then turn to research focusing on the contextual level.

Literature-scale relation extraction

Literature-scale relation extraction has been proposed as a method for managing information overload [76]. SemMedDB [77] is a database of semantic relations extracted with SemRep [70] from the entire PubMed. In its latest release (as of 31 December 2016), it contains about 89 million relations extracted from >26 million abstracts. It has been used for a variety of tasks, such as clinical decision support [78], uncovering potential drug interactions in clinical data [79], supporting gene regulatory network construction [80] and medical question answering [81]. It also forms the back end for the Semantic MEDLINE application [76], which integrates semantic relations with automatic abstractive summarization [82], and visualization, to enable the user navigate biomedical literature through concepts and their relations. Semantic MEDLINE, coupled with a literature-based discovery extension called 'discovery browsing', was used to propose a mechanistic link between age-related hormonal changes and sleep quality [83] and to elucidate the paradox that obesity is beneficial in critical care despite contributing to disease generally ('the obesity paradox') [84]. Another database, EVEX [85], is based on the Turku Event Extraction System (TEES) [86] and includes relations extracted from abstracts in PubMed and full-text articles in the PubMed Central Open Access subset(PMC-OA). It consists of ~40 million biomolecular events (e.g. gene expression and binding). A CytoScape plugin, called CyEVEX, is made available for integration of literature analysis with network analysis. EVEX has been exploited for gene regulatory network construction [87]. Other databases, such as PharmGKB [88] and DisGeNET [89], integrate relationships extracted with text mining with those from curated resources.

Contextualizing biomedical relations

Contextualizing relations (or claims) focuses on how they are presented and how they behave in the larger discourse. Two distinct approaches can be distinguished.

The first approach, which can be considered 'bottom-up', focuses on classifying scientific statements or relations along one or more meta-dimensions aiming to capture their contextual properties, for example whether they are expressed as speculation. One early task adopting this approach was distinguishing speculative statements from facts (hedge classification). For this task, weakly supervised learning techniques [90] as well rule-based methods using lexical and syntactic templates [91, 92] have been explored, yielding similar performance (0.85 F_1 score). Interesting from a research integrity/transparency perspective, the system developed in [91] was used to compare the language used in reporting industry-sponsored research and non-industry-reported research, which found that the former was on average less speculative [93].

Semantically more fine-grained, speculation/negation detection task has focused on recognizing speculation and negation cues in text (e.g. 'suggest', 'likely' and 'failure') and their linguistic scope, often formalized as a relation [94] or a text segment [95]. Speculation/negation detection has been studied in the context of BioNLP Shared Tasks on event extraction [96, 97] and the CoNLL'10 Shared Task on Hedge Detection [98]. Supervised machine learning methods [86, 99] as well as rule-based methods with lexico-syntactic patterns [100] have been applied to this task. The interaction of speculation and negation has been studied under the notion of factuality, and factuality values (Fact, Probable, Possible, Doubtful and Counterfact) of biological events were computed using a rule-based, syntactic composition approach [101].

Focusing on a more comprehensive characterization of scientific statements, Wilbur *et al.* [102] categorized sentence segments along five dimensions: focus (whether the segment describes a finding, a method or general knowledge), polarity (positive/negative), certainty (the degree of speculativeness expressed toward the segment on a scale of 0–3), evidence (four levels, from no stated evidence to explicit experimental evidence in text) and direction (whether segment describes an increase or decrease in the finding). A similar categorization ('meta-knowledge') was proposed by Thompson *et al.* [103], who applied it to events, rather than arbitrary text segments. They also proposed two hyper-dimensions that are inferred from their five categories: one indicates whether the event in question is New Knowledge and the other whether it is a Hypothesis. Studies that focused on predicting these meta-dimensions have been trained on the annotated corpora and used supervised machine learning techniques [104, 105]. The Claim Framework [106] proposed a categorization of scientific claims according to the specificity of evidence, somewhat similar to focus dimension in the schema of Wilbur *et al.* [102]. Five categories were distinguished (explicit claim, implicit claim, observation, correlation and comparison). A small corpus of full-text articles was annotated with these categories, and an approach based on lexico-syntactic patterns was used to recognize explicit claims.

The second approach ('top-down') focuses on classifying larger units (sentences or a sequence of sentences) according to the function they serve in the larger argumentative structure. Proposed by Teufel *et al.* [107, 108] for scientific literature on computational linguistics and chemistry, argumentative zoning assigns sentences to domain-independent zone categories based on the rhetorical moves of global argumentation and the connections between the current work and the cited research. The proposed categories include, for example, Aim (statement of specific research goal or hypothesis), Nov_Adv (novelty/advantage of the approach), Own_Mthd (description of methods used), among others. Mizuta *et al.* [109] adapted this classification to biology articles and presented an annotated corpus. Guo *et al.* [110] adopted a simplified version of argumentative zoning with seven classes (e.g. Background, Method, Result and Future Work). They used weakly supervised SVM and CRF models to classify sentences in abstracts discussing cancer risk assessment, which yielded an accuracy of 0.81. The CoreSC schema [111] is an extension of the argumentative zoning approach, in which sentences are classified along two layers according to their role in scientific investigation. The first layers consist of 11 categories (e.g. Background, Motivation, Experiment, Model, Result and Conclusion), and the second layer indicates whether the information is new or old. A corpus of chemistry articles annotated with these layers was presented. SVM and CRF classifiers that recognize the first layer categories were developed [112], achieving best results with Experiment, Background and Model classes (0.76, 0.62 and 0.53 F_1 scores, respectively). *N*-gram, syntactic dependency and document structure features (section headings) were found to be predictive. Such top-down classifications are similar to but more fine-grained than IMRaD rhetorical categories (Introduction, Methods, Results and Discussion) that underlie the structure of most scientific articles. As the sentences may not conform to the characteristics of the section that they appear in, some research considered classifying sentences into IMRaD categories. For example, Agarwal and Yu [113] compared several rule-based and supervised learning methods to classify sentences from full-text biomedical articles into these categories. The best results reported (0.92 accuracy and F_1 score) were obtained with a Naive Bayes classifier with *n*-gram, tense and citation features, and feature selection. Other similar categorizations have also been proposed [114]. Note that the methods applied in these approaches are largely similar to those discussed earlier for identification of specific statements, such as PICO or data deposition statements. Finally, a comprehensive, multilevel model of scientific argumentation, called Knowledge Claim Discourse Model (KCDM), has been proposed by Teufel [115]. Five levels varying in their degree of abstraction have been distinguished. At the most abstract level, rhetorical goals are formalized into four categories, often not explicit in text ('Knowledge claim is significant', 'Knowledge claim is novel', 'Authors are knowledgeable' and 'Research is methodologically sound'). Next level, rhetorical moves, addresses the properties of the research space (e.g. 'No solution to new problem exists') and the new knowledge claim (e.g. 'New solution solves problem'). The third level, knowledge claim attribution, is concerned with whether a knowledge claim is attributed to the author or others. At the fourth level are hinge moves, which categorize the connections between the new knowledge claim and other claims (e.g. 'New claim contrasts with existing claim'). The bottom and the most concrete layer, linearization and presentation, deals with how these rhetorical elements are realized within the structure of the article. Teufel reported the results of several annotation studies focusing on argumentative zoning and knowledge claim attribution (κ values of 0.71–0.78), and her argumentative zone detection system, based on supervised learning with verb features, word lists, positional information and attribution features, achieved a κ value 0.48, with respect to the annotated corpus.

Similarly, taking a top-down approach but focusing on the relations between individual discourse segments (similar to KCDM hinge moves) are models of discourse coherence. Such relations include elaboration, comparison, contrast and precedence, and are often indicated with discourse connectives (e.g. 'furthermore', 'in contrast'). Linguistic theories and treebanks have been proposed to address these relations, including Rhetorical Structure Theory (RST) [116] and the Penn Discourse TreeBank (PDTB) [117], each assuming a somewhat different discourse structure and relation inventory and differing in their level of formalization. In the biomedical domain, discourse relations remain understudied, with the notable exception of the Biomedical Discourse Relation Bank corpus [118], in which a subset of PDTB relation types was used to annotate abstracts in the GENIA corpus. Detection of discourse connectives was explored on this corpus, and an F_1 score of 0.76 was achieved with a supervised learning approach and domain adaptation techniques [119].

Some research considered combining bottom-up and top-down approaches for a fuller understanding of scientific discourse or contextual meaning. For example, a three-way characterization, based on meta-knowledge dimensions [103], CoreSC [111] and discourse segment classification [114], was attempted, and these components were shown to be complementary [120]. The Embedding Framework [121] is a unified, domain-independent semantic model for contextual meaning, consolidating the meta-dimensions and discourse coherence relations. A fine-grained categorization of contextual predicates was presented, with four top-level categories (Modal, Valence_Shifter, Relational and Propositional), where the Modal and Valence_Shifter categories overlap with meta-dimensions, and the Relational category overlaps with discourse relations. A dictionary of terms, classified according to this fine-grained categorization, was constructed and a rule-based interpretation methodology based on the dictionary and syntactic dependency composition was proposed. The framework is designed to complement existing relation extraction systems. While no specific corpus annotation was performed, the methodology has been applied to relevant tasks, such as speculation/negation detection [100], factuality assessment [101] and attribution detection [121], yielding good performance.

Although not a text mining approach, an effort that deserves discussion here is micropublications [30], a semantic model of scientific claims, evidence and arguments. Built on top of Semantic Web technologies, micropublications are intended for use in the research life cycle, where scientists create, publish, expand and comment on micropublications for scientific communication. They have been proposed as a potential solution to improve research reproducibility and robustness. At a minimum, a micropublication is conceived as a claim with its attribution, and in its full form, as a claim with a complete directed-acyclic support graph, consisting of relevant evidence, interpretations and discussion that supports/refutes the claim, either within the publication or in a network of publications discussing the claim. It has been designed to be compatible with claim-based models formalizing relationships (e.g. nanopublications [122]), as well as with claims in natural language text. The model can accommodate artifacts such as figures, tables, images and data sets, which text mining approaches generally do not consider. While it has been used for manual annotation [123], to our knowledge, the micropublications model has not been used as a target for text mining. An example of micropublication is presented in Section S5 of the Supplementary Material.

Accurate citation and enhanced bibliometrics

Citations are important for several reasons in ensuring research integrity/rigor. First, the performance of a scientist is often measured by the number of citations they receive and the number of articles they publish in high impact factor journals. Count-based measures, such as the h-index [124], are often criticized because they treat all citations as equal and do not distinguish between the various ways and reasons a paper can be cited. For example, a paper can be appraised in a positive light or criticized; it can be cited as the basis of the current study or more peripherally. Such differences should be accounted for enhanced bibliometric measures. More sophisticated measures have been proposed in response to such criticism [125]. Second, from an integrity perspective, it is important to ensure that all references in a manuscript (or any other scientific textual artifact) are accurately cited. Two kinds of reference accuracy problems are distinguished [126]: citation accuracy refers to the accuracy of details, such as authors' names, date of publication and volume number, whereas quotation accuracy refers to whether the statements from the cited papers are accurately reflected in the citing paper. Citation accuracy studies were found to report a median error rate of 39% and quotation accuracy studies a median error rate of 20% [126]. Greenberg [34] highlighted some types of citation distortions (i.e. quotation accuracy problems) that lead to unfounded authority. For example, citation transmutation refers to 'the conversion of hypothesis into fact through act of citation alone', and dead-end citation to 'citation to papers that do not contain content addressing the claim'. Another rigor issue is the continued citation of retracted papers, which may lead to spreading of misinformation. A study of retracted paper citations found that 94% of the citing papers did not acknowledge the retraction [127]. Automated citation analysis tools and accuracy checkers would be beneficial for journal editors and staff in their workflows, as well as for scientists in authoring manuscripts and academic institutions and funding agencies in considering quality of impact rather than quantity and improving decision-making.

Most text mining research on citations has focused on the computational linguistics literature, an area in which a corpus of full-text articles is available (ACL Anthology Corpus [128]). Citation analysis has been proposed for enhancing bibliometrics as well as for extractive summarization [129]. Several aspects of citations have been studied. Research on citation context detection aims to identify the precise span of the discussion of the reference paper in the citing paper. For example, to detect the surrounding sentences that discuss a reference paper, Qazvinian and Radev [130] proposed a method based on Markov Random Fields using sentence similarity and lexical features from sentences. Abu-Jbara and Radev [131] focused on reference scopes that are shorter than the full sentence. They explored several methods for this task: word classification with SVM and logistic regression, CRF-based sequence labeling and segment classification, which uses rules based on CRF results, achieving best performance with segment classification (F_1 score of 0.87). Other studies explored citation significance. Athar [132] presented a text classification approach to determine whether a citation is significant for the citing paper and achieved 0.55 F_1 score with a Naive Bayes classifier that used as features, number of sentences with acronyms, with formal citation to the paper and to the author's name, as well as average similarity of the sentence with the title. Similar text classification techniques were used to identify key references [133] and meaningful citations [134]; the number of times a paper is cited was identified as the most predictive feature. Citation sentiment (whether the authors cite a paper positively, negatively or neutrally) has also been proposed to enhance bibliometrics. Athar [135] annotated the ACL Anthology Corpus for citation sentiment and used an SVM classifier with *n*-gram and dependency features extracted from the citation sentence for sentiment classification, achieving a macro-F_1 score of 0.76. In the biomedical domain, Xu *et al.* [136] annotated the discussion sections of 285 RCT articles with citation sentiment. Using an SVM classifier with *n*-gram and various lexicon-based features (e.g. lexicons of positive/negative sentiment, contrast expressions), they reached a macro-F_1 score of 0.72. A more fine-grained citation classification concerns citation function, for which many classifications have been proposed. For example, Teufel *et al.* [137] presented a scheme, which contained 12 categories [e.g. Weak (weakness of the cited approach), PBas (cited work as the starting point) and CoCo (unfavorable comparison/contrast)] and measured inter-annotator agreement ($\kappa = 0.72$). Later, Teufel *et al.* [138] used a memory-based learning algorithm to recognize these categories. They used features based on cue phrases in the citation sentence, position of the citation and self-citation, which yielded a κ of 0.57. In the biomedical domain, Agarwal *et al.* [139] presented a corpus of 43 biomedical articles annotated with eight citation roles (e.g. Background/Perfunctory, Contemporary, Contrast/Conflict, Evaluation, Modality and Similarity/Consistency), achieving moderate inter-annotator agreement ($\kappa = 0.63$), though it seems difficult to think of some of their categories (Contemporary and Modality) as citation roles in a traditional sense. Using *n*-gram features with SVM and Naive Bayes classifiers, they obtained a macro-F_1 score of 0.75.

The first type of reference accuracy, referred to as citation accuracy above, is studied under the rubric of citation matching. We do not discuss this task here, as NLP has little relevance to it; see [140] for a comparison of several citation matching algorithms. Some tasks, such as author name disambiguation [141], focus on specific elements of citations, rather than on full citation matching. Ensuring quotation accuracy, on the other hand, can be viewed as a text mining task, in which the goal is to identify the segments of the reference paper that are discussed in the citing paper. Inability to find such a segment would indicate a dead-end citation, while finding inconsistencies between how a claim is presented in the reference paper versus the citing paper with respect to its factuality might indicate a citation transmutation [34]. However, identifying reference paper segments precisely can be challenging, as the citing paper usually does not simply quote the reference paper verbatim, but rather paraphrases its contents, and commonly, refers to its contents in an abstract manner. In the Text Analysis Conference 2014 Biomedical Summarization shared task (http://www.nist.gov/tac/2014/BiomedSumm), one subtask involved finding the spans of text in reference papers that most accurately reflect the citation sentence and identifying what facet of the reference paper it belongs to (e.g. Hypothesis, Method, Results and Implication). The task focused on a corpus of 20 biology articles, each with 10 corresponding reference articles. The inter-annotator agreement was found to be low. The results of this shared task were not available at the time of this writing; however, one of the reported systems [142] relied on calculating text similarity between the citation sentence and the sentences in the reference paper using tf.idf, as well as various methods to expand the citation context and the reference paper context for similarity calculation. The best results (F_1 score of 0.32) were obtained when using 50 sentences surrounding the citation sentence and all sentences from the articles that cite the reference paper for context. The same task has also been adapted to the computational linguistics literature [143], even though the results have been poorer, with the top-ranking system obtaining 0.1 F_1 score [144].

Challenges and directions

We examined four areas of concern for biomedical research integrity and rigor and discussed existing text mining research that has the potential to address them. We discuss below several general challenges facing bioNLP research focusing on these areas and highlight some promising avenues for future research.

The first challenge is concerned with availability of artifacts that can be used to train text mining methods. While most text mining research has focused on PubMed abstracts because of their availability, most biomedical knowledge relevant to the tasks discussed, including study details, knowledge claims and citations, can only be located in full text. Blake [106] found that only 8% of the explicit claims were expressed in abstracts. Furthermore, biomedical abstracts differ from full text in terms of structure and content [145]. The PMC-OA subset is amenable to automated approaches without much additional preprocessing effort; however, it contains only about a million full-text articles (4% of all PubMed abstracts). Owing to availability and access difficulties, researchers often use nonstandard PDF-to-text conversion tools to extract full text from PDF files [65, 146]. Considering that the progress in bioNLP is partly attributed to public availability of biomedical abstracts, a similar mode of access can further stimulate research in mining of full-text articles. We are not aware of research focusing on other textual artifacts discussed, though abstracts of NIH grant applications and the resulting publications are available via NIH RePORT (https://report.nih.gov/), and some journals (e.g. *British Medical Journal*) publish prepublication manuscripts and reviewer reports for transparency.

Collecting bibliographic data at a large scale also remains challenging. Two sources of scholarly citation considered most authoritative, Web of Science and Scopus, are neither complete nor fully accurate [147] and require high subscription fees. Others, like Google Scholar, have license restrictions. Open Citations Corpus (OCC) has been proposed as an open-access repository of citation data to improve citation access [148]. It relies on the SPAR ontologies [149], which define characteristics of the publishing domain. Citation information in PMC-OA has been made available in OCC. Although this is a small subset of the biomedical literature, the movement toward open-access citation data is encouraging for research.

Even when the text sources are plentiful, restrictions may apply to text mining of their contents. Publishers often adopt a license-based approach, allowing researchers from subscribing institutions to register for an API key to text-mine for research purposes. Negotiating a separate license with each publisher is not only impractical for both researchers and publishers but also ineffective, as some tasks (e.g. plagiarism detection, managing information overload and citation analysis) presuppose text mining at the literature scale with no publisher restrictions. The Crossref Metadata Application Program Interface (API) initiative [150] aims to solve this problem by providing direct links to full text on the publisher's site and a common mechanism for recording license information in Crossref metadata. Several publishers (e.g. HighWire Press, Elsevier and Wiley) as well as professional societies (e.g. American Psychological Association) have been involved in this initiative.

The next set of challenges are concerned with the text mining approaches themselves. Most approaches depend on annotated corpora and sizable corpora based on full-text articles or other text sources we discussed are lacking. The largest full-text corpus, CRAFT [151], contains 67 articles, and the annotation focuses mostly on low-level semantic information, such as named entities and concepts. Some tasks we discussed require higher-level annotation, such as annotation of argumentation, discourse structure, citation function and quotation, and are much more challenging, as they are less well defined and some subjectivity is involved in annotating them. Collaborative, cross-institution efforts would be beneficial for consolidating existing research in these areas and proposing more comprehensive characterizations. Ontology development research should also be taken into account, as some existing ontologies focus on scholarly discourse (e.g. SWAN [152]), and annotation efforts would benefit from the insights of such research. Another promising avenue is crowdsourcing of annotation, where the 'crowd' (a mix of lay people, enthusiasts and experts), recruited through an open call, provides their services for a given task. In the biomedical domain, crowdsourcing has been successfully applied to relatively low-level tasks such as named entity annotation, while it has been considered less suitable for complex, knowledge-rich tasks [153]. However, the design of crowdsourcing experiments plays a significant role in their success, and creative crowdsourcing interfaces could make collection of complex data (e.g. argumentation graphs) more feasible. It is also worth noting that frameworks like nanopublications [122] and micropublications [30] advocate the use of semantic models of scientific statements and argumentation, respectively, in the workflows of scientists as a means of knowledge generation and exchange. If such models are adopted more widely (not only among scientists but also publishers and other stakeholders), the knowledge generated would also be invaluable as gold standard data. The Resource Identification Initiative [154] promotes such a model for research resources (e.g. reagents and materials) and can be informative in this regard.

Representativeness and balance of a corpus is important for the generalizability of tools that are trained on it. Though corpus linguistics literature addresses the construction of balanced and representative corpora [155], in practice, most biomedical text corpora focus on a restricted domain of interest. For example, CRAFT [151] contains biology and genetics articles, while GENIA [156] contains abstracts about biological reactions involving transcription factors in human blood cells. Lippincott *et al.* [157] showed that subdomains in biomedical literature vary along many linguistic dimensions, concluding that a text mining system performing well on one subdomain is not guaranteed to perform well on another. Construction of wide-coverage, representative biomedical full-text article corpora, while clearly challenging, would be of immense value to text mining research in general. Also, note that a subfield of machine learning, domain adaptation, specifically focuses on model generalizability. Various methods (some requiring data from the new domain and some not) have been proposed [158], and such methods have been applied to biomedical text mining tasks [159]. Independently, some systems provided machine learning models that can be retrained on new annotated corpora [86, 160], while others attempted to generalize by appealing to linguistic principles [70, 100].

Important information in biomedical articles may only appear in tables, graphs, figures or even supplementary files. There is relatively little research in incorporating data from such artifacts into text mining approaches, even though some semantic models, such as micropublications [30], support them. Figure retrieval has been considered, mainly focusing on using text from figure captions [161], text within figures [162] and text from paragraphs discussing the figures as well as NER [163]. Research on information extraction from tables is rare [164–166], though this may change with recent availability of corpora [167]. Jimeno-Yepes and Verspoor [146] showed that most literature-curated mutation and genetic variant information existed only as supplementary material and used open-source PDF conversion tools to extract text from supplementary files for text mining.

The accuracy of text mining approaches varies widely depending on the task. In some classification tasks (e.g. identifying PICO categories), state-of-the-art performance is over 0.9 accuracy, whereas in recognition of citation quotation, the state-of-the-art performance is just over 0.3. Although text mining tools have shown benefits in curation and annotation [168, 169], it is critical to educate the users about the role of such tools in their workflows and their value/limitations, and not alienate them by setting their expectations impossibly high. It is also worth pointing out that human agreement on some tasks is not high; therefore, it may be unrealistic to expect that automated tools do well (e.g. Fleiss' κ of 0.11 for deceptive text annotation [48]). Depending on the task, a user may prefer not the setting that yields the highest F_1 score, generally considered the primary performance metric, but rather high recall or high precision. Providing the ability to tune a system for high recall or precision is likely to be advantageous for its adoption. Most machine learning systems are essentially black boxes, and the ability of systems to provide human-interpretable explanations for their predictions may also affect their adoption. Curation cycles, in which experts or the crowd manually 'correct' text mining results, providing feedback that is automatically incorporated into machine learning models, can also be effective in incrementally improving performance of such models. The majority of NLP publications focus on basic research rather than engineering, and generally report little with respect to certain characteristics of the tools developed, such as their computational complexity or usability [170], which can be critical for practical use. Finally, it is worth noting that reproducibility of results generated by NLP tools, the focus of some recent research [170–172], is likely to be increasingly important in their adoption.

Conclusion

Toward enhancing rigor and integrity of biomedical research, we proposed text mining as complementary to efforts focusing on standardization and guideline development. We identified four main areas (plagiarism/fraud detection, compliance with reporting guidelines, management of information overload and accurate citation), where text mining techniques can play preventive and corrective roles, and we surveyed the state of the art for these tasks. We believe that the tools that perform the following tasks can have the biggest and most immediate impact toward addressing some of the problems that lead to research waste:

1. Checking for adherence to all elements of the relevant reporting guideline:
    - Given a manuscript, such a tool can ensure that all methodological details needed for replication are provided and limitations are clearly specified.
2. Generating document-level and literature-level argumentation graphs:
    - With a document-level graph of a manuscript, we can check whether the conclusions are consistent with the results and limitations of the study.
    - A literature-level graph for a proposed research question can aid in determining whether it can already be answered with existing evidence.
    - For an unfamiliar research topic, a literature-level graph can help us quickly identify the main knowledge claims, their provenance and track their evolution.
3. Constructing citation quotation networks:
    - Given a manuscript, such a tool can ensure that existing research is accurately cited.
    - We can more accurately assess the contribution of a scientist to their field, based on the impact of their citations, rather than the citation count.

For some tasks, current state-of-the-art text mining techniques can be considered mature (e.g. extrinsic plagiarism detection, extracting PICO sentences), while for other tasks, substantial research progress is needed for practical tools (e.g. construction of argumentation graphs, identifying citation quotations). We argued that the main advantage of text mining comes in its ability to facilitate performing tasks at a large scale. By shortening the time it takes to perform tasks needed to ensure rigor and integrity, text mining technologies can promote better research practices, ultimately reducing waste and increasing value.

Key Points

- Lack of reproducibility and rigor in published research is a growing concern in science, and all stakeholders (scientists, journals, peer reviewers, funding agencies, etc.) have a responsibility to ensure the accuracy, verifiability and honesty of research conduct and reporting to reduce research waste and increase value.

- By providing the ability to automatically process textual artifacts of biomedical communication (proposals, manuscripts and publications) at a large scale and extract relevant information, biomedical text mining methods can assist all stakeholders in their research activities and complement efforts focusing on standards and guidelines.
- Biomedical text mining methods can support automated tools in four key areas: (a) plagiarism/fraud detection, (b) adherence to reporting guidelines, (c) managing information overload and (d) accurate citation/enhanced bibliometrics, providing preventive and corrective functions.
- Specific tasks that can have the biggest and most immediate impact include (a) checking for adherence to relevant reporting guidelines, (b) generating document-level and literature-level argumentation graphs and (c) constructing citation quotation networks.
- Challenges in using biomedical text mining methods in these areas include (a) potential publisher restrictions on text mining, (b) availability, representativeness and balance of text corpora and bibliographic data to use for training, (c) lack of consensus in how to best annotate relevant linguistic phenomena, (d) need to incorporate information from non-textual artifacts, such as tables, figures and graphs and (e) accuracy, complexity and black-box nature of most text mining techniques.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgements

The author thanks Jodi Schneider, Gerben ter Riet, Dina Demner-Fushman, Catherine Blake, Thomas C. Rindflesch, Olivier Bodenreider and Caroline Zeiss for their comments on earlier drafts of this article.

Funding

The intramural research program at the US National Library of Medicine, National Institutes of Health.

References

1. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016a;**533**:452–4.
2. Røttingen J-A, Regmi S, Eide M, *et al*. Mapping of available health research and development data: what's there, what's missing, and what role is there for a global observatory? *Lancet* 2013;**382**(9900):1286–307.
3. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *Lancet* 2009;**374**(9683): 86–9.
4. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. *Nature* 2014;**505**(7485):612–3.
5. Barham BL, Foltz JD, Prager DL. Making time for science. *Res Policy* 2014;**43**(1):21–31.
6. Head ML, Holman L, Lanfear R, *et al*. The extent and consequences of p-hacking in science. *PLoS Biol* 2015;**13**(3): e1002106.
7. Dwan K, Altman DG, Clarke M, *et al*. Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. *PLoS Med* 2014;**11**(6):e1001666.
8. Chan A, Hróbjartsson A, Haahr M, *et al*. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;**291**(20):2457–65.
9. Vasilevsky NA, Brush MH, Paddock H, *et al*. On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ* 2013;**1**:e148.
10. Chalmers I. Lessons for research ethics committees. *Lancet* 2002;**359**(9301):174.
11. Bowen A, Casadevall A. Increasing disparities between resource inputs and outcomes, as measured by certain health deliverables, in biomedical research. *Proc Nat Acad Sci USA* 2015;**112**(36):11335–40.
12. Ioannidis JPA. Why most published research findings are false. *PLoS Med* 2005;**2**(8):e124.
13. Begley CG, Ellis LM. Raise standards for preclinical cancer research. *Nature* 2012;**483**(29):531–3.
14. Kyzas PA, Denaxa-Kyza D, Ioannidis JPA. Almost all articles on cancer prognostic markers report statistically significant results. *Eur J Cancer* 2007;**43**(17):2559–79.
15. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* 2011;**10**(9):712.
16. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 2015;**349**(6251):aac4716.
17. Fang FC, Steen RG, Casadevall A. Misconduct accounts for the majority of retracted scientific publications. *Proc Natl Acad Sci USA* 2012;**109**(42):17028–33.
18. Bouter LM, Tijdink J, Axelsen N, *et al*. Ranking major and minor research misbehaviors: results from a survey among participants of four World Conferences on Research Integrity. *Res Integr Peer Rev* 2016;**1**(1):17.
19. De Angelis C, Drazen J, Frizelle F, *et al*. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N Engl J Med* 2004;**351**(12):1250–1.
20. International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. 2016. http://www.icmje.org/icmje-recommendations.pdf.
21. Ohno-Machado L, Alter G, Fore I, *et al*. bioCADDIE white paper - Data Discovery Index. Technical report, Figshare, 2015. http://dx.doi.org/10.6084/m9.figshare.1362572.
22. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;**340**:c332.
23. Nosek BA, Alter G, Banks GC, *et al*. Promoting an open research culture. *Science* 2015;**348**(6242):1422–5.
24. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al*. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018.
25. Baker M. Muddled meanings hamper efforts to fix reproducibility crisis. *Nature News* 2016.
26. Bollen K, Cacioppo JT, Kaplan RM, *et al*. Social, behavioral, and economic sciences perspectives on robust and reliable science. Technical Report, National Science Foundation, 2015.
27. Ananiadou S, McNaught J, *Text mining for biology and biomedicine*. Boston, MA: Artech House, 2006.

28. Cohen KB, Demner-Fushman D, *Biomedical Natural Language Processing*. Amsterdam: John Benjamins, 2014.
29. Gonzalez G, Tahsin T, Goodale BC, *et al*. Recent advances and emerging applications in text and data mining for biomedical discovery. *Brief Bioinform* 2016;**17**(1):33–42.
30. Clark T, Ciccarese P, Goble C. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. *J Biomed Semant* 2014;**5**(1):28.
31. Lund H, Brunnhuber K, Juhl C, *et al*. Towards evidence based research. *BMJ* 2016;**355**:i5440.
32. Robinson K, Goodman S. A systematic examination of the citation of prior research in reports of randomized, controlled trials. *Ann Intern Med* 2011;**154**(1):50–5.
33. Habre C, Tramèr MR, Pöpping DM, *et al*. Ability of a meta-analysis to prevent redundant research: systematic review of studies on pain from propofol injection. *BMJ* 2014;**349**:g5219.
34. Greenberg SA. How citation distortions create unfounded authority: analysis of a citation network. *BMJ* 2009;**339**: b2680.
35. Habibzadeh F, Shashok K. Rules of the game of scientific writing: fair play and plagiarism. *Croat Med J* 2011;**52**(4): 576–7.
36. Hoad TC, Zobel J. Methods for identifying versioned and plagiarised documents. *J Am Soc Inf Sci Technol* 2003;**54**: 203–15.
37. Stein B, Meyer zu Eissen S. Near similarity search and plagiarism analysis. In: *From Data and Information Analysis to Knowledge Engineering*, Springer Berlin Heidelberg, 2006, 430–7.
38. Stamatatos E. Plagiarism detection using stopword n-grams. *J Assoc Inf Sci Technol* 2011;**62**(12):2512–27.
39. Meyer zu Eissen S, Stein B. Intrinsic plagiarism detection. In: Lalmas M, MacFarlane A, Rüger S *et al*. (eds), 28th European Conference on IR Research (ECIR 06), volume 3936 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2006, 565–9.
40. Rosso P, Rangel F, Potthast M, *et al*. Overview of PAN'16—new challenges for authorship analysis: cross-genre profiling, clustering, diarization, and obfuscation. In: Fuhr N, Quaresma P, Larsen B *et al*. (eds), *7th International Conference of the CLEF Initiative (CLEF 16)*, Springer Berlin Heidelberg New York, 2016.
41. Sanchez-Perez M, Sidorov G, Gelbukh A. A winning approach to text alignment for text reuse detection at PAN 2014. In: L Cappellato, N Ferro, M Halvey *et al*. (eds), *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*, CEUR-WS.org, 2014.
42. Kuznetsov M, Motrenko A, Kuznetsova R, *et al*. Methods for intrinsic plagiarism detection and author diarization. In: K Balog, L Cappellato, N Ferro *et al*., (eds), *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, CEUR-WS.org, 2016.
43. Nawab RMA, Stevenson M, Clough P. An IR-based approach utilising query expansion for plagiarism detection in MEDLINE. *IEEE/ACM Trans Comput Biol Bioinform* 2016;**99**:1.
44. Lindberg DAB, Humphreys BL, McCray AT. The unified medical language system. *Methods Inf Med* 1993;**32**:281–91.
45. Fanelli D. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PLoS One* 2009;**4**(5):1–11.
46. Markowitz DM, Hancock JT. Linguistic obfuscation in fraudulent science. *J Lang Soc Psychol* 2015;**35**(4):435–45.
47. Mihalcea R, Strapparava C. The lie detector: explorations in the automatic recognition of deceptive language. In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Association for Computational Linguistics, Suntec, Singapore, 2009, 309–12.
48. Ott M, Choi Y, Cardie C, *et al*. Finding deceptive opinion spam by any stretch of the imagination. In: *Proceedings of the ACL:HLT 2011*, Association for Computational Linguistics, Portland, OR, USA, 2011, 309–19.
49. Simera I, Moher D, Hirst A, *et al*. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC Med* 2010;**8**(1):24.
50. Kilkenny C, Browne WJ, Cuthill IC, *et al*. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* 2010;**8**(6):e1000412.
51. Turner L, Shamseer L, Altman DG, *et al*. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Syst Rev* 2012;**1**(1):60.
52. Altman DG. Making research articles fit for purpose: structured reporting of key methods and findings. *Trials* 2015;**16**(1):53.
53. Tsafnat G, Glasziou P, Choong MK, *et al*. Systematic review automation technologies. *Syst Rev* 2014;**3**(1):74.
54. O'Mara-Eves A, Thomas J, McNaught J, *et al*. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst Rev* 2015;**4**(1):5.
55. Chavalarias D, Wallach JD, Li AH, *et al*. Evolution of reporting *P* values in the biomedical literature, 1990-2015. *JAMA* 2016;**315**(11):1141–8.
56. Kafkas Ş, Kim J-H, McEntyre JR. Database citation in full text biomedical articles. *PLoS One* 2013;**8**(5):e63184.
57. Kiritchenko S, de Bruijn B, Carini S, *et al*. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inform Decis Mak* 2010;**10**(1):56.
58. Sackett DL, Rosenberg WMC, Gray JAM, *et al*. Evidence based medicine: what it is and what it isn't. *BMJ* 1996;**312**(7023): 71–2.
59. Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. *Comput Linguist* 2007;**33**(1):63–103.
60. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. *JAMIA* 2010;**17**(3):229–36.
61. Wallace BC, Kuiper J, Sharma A, *et al*. Extracting PICO sentences from clinical trial reports using supervised distant supervision. *J Mach Learn Res* 2016;**17**(132):1–25.
62. Kim SN, Martínez D, Cavedon L, *et al*. Automatic classification of sentences to support Evidence Based Medicine. *BMC Bioinformatics* 2011;**12**(S-2):S5.
63. Verbeke M, Asch VV, Morante R, *et al*. A statistical relational learning approach to identifying evidence based medicine categories. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, Association for Computational Linguistics, Jeju Island, Korea, 2012, 579–89.
64. Hassanzadeh H, Groza T, Hunter J. Identifying scientific artefacts in biomedical literature: the evidence based medicine use case. *J Biomed Inform* 2014;**49**:159–70.
65. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Informat Assoc* 2015;193–201.

66. Huang Y, Gottardo R. Comparability and reproducibility of biomedical data. *Brief Bioinform* 2013;**14**(4):391.
67. Névéol A, Wilbur WJ, Lu Z. Extraction of data deposition statements from the literature. *Bioinformatics* 2011a;**27**(23): 3306–12.
68. Kilicoglu H, Demner-Fushman D, Rindflesch TC, *et al*. Towards automatic recognition of scientifically rigorous clinical research evidence. *J Am Med Informat Assoc* 2009;**16**(1):25–31.
69. Wilczynski NL, Morgan D, Haynes RB, *et al*. An overview of the design and methods for retrieving high-quality studies for clinical care. *BMC Med Inform Decis Mak* 2005;**5**:20.
70. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform* 2003;**36**(6):462–77.
71. Myers G. 'In this paper we report . . .': speech acts and scientific facts. *J Pragm* 1992;**17**(4):295–313.
72. Blackburn P, Bos J. *Representation and Inference for Natural Language. A First Course in Computational Semantics*. CSLI, Stanford, CA, USA, 2005.
73. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med* 1986;**30**(1):7–18.
74. Editorial: so long to the silos. *Nat Biotechnol* 2016;**34**:357.
75. Luo Y, Uzuner Ö, Szolovits P. Bridging semantics and syntax with graph algorithms state-of-the-art of extracting biomedical relations. *Brief Bioinform* 2016;**18**(1):160–78.
76. Kilicoglu H, Fiszman M, Rodriguez A, *et al*. Semantic MEDLINE: a web application to manage the results of PubMed searches. In: Salakoski T, Schuhmann DR, Pyysalo S, (eds), Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008), Turku Centre for Computer Science (TUCS), Turku, Finland, 2008, 69–76.
77. Kilicoglu H, Shin D, Fiszman M, *et al*. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics* 2012;**28**(23):3158–60.
78. Jonnalagadda S, Fiol GD, Medlin R, *et al*. Automatically extracting sentences from Medline citations to support clinicians' information needs. *JAMIA* 2013;**20**(5):995–1000.
79. Zhang R, Cairelli MJ, Fiszman M, *et al*. Using semantic predications to uncover drug-drug interactions in clinical data. *J Biomed Inform* 2014;**49**:134–47.
80. Chen G, Cairelli MJ, Kilicoglu H, *et al*. Augmenting microarray data with literature-based knowledge to enhance gene regulatory network inference. *PLoS Comput Biol* 2014;**10**(6):1–16.
81. Hristovski D, Dinevski D, Kastrin A, *et al*. Biomedical question answering using semantic relations. *BMC Bioinformatics* 2015;**16**(1):6.
82. Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. In: Proceedings of the HLT-NAACL Workshop on Compu tational Lexical Semantics, Association for Computational Linguistics, Boston, MA, USA, 2004, 76–83.
83. Miller CM, Rindflesch TC, Fiszman M, *et al*. A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men. *Sleep* 2012;**35**(2):279–85.
84. Cairelli MJ, Miller CM, Fiszman M, *et al*. Semantic MEDLINE for discovery browsing: using semantic predications and the literature-based discovery paradigm to elucidate a mechanism for the obesity paradox. In: *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, Washington, DC, USA, 2013, 164–73.
85. Van Landeghem S, Björne J, Wei C-H, *et al*. Large-scale event extraction from literature with multi-level gene normalization. *PLoS One* 2013;**8**(4):e55814.
86. Björne J, Salakoski T. Generalizing biomedical event extraction. In: Proceedings of BioNLP Shared Task 2011 Workshop. Association for Computational Linguistics, Portland, OR, USA, 2011, 183–191.
87. Hakala K, Van Landeghem S, Salakoski T, *et al*. Application of the EVEX resource to event extraction and network construction: shared task entry and result analysis. *BMC Bioinformatics* 2015;**16(Suppl 16)**:S3.
88. Hewett M, Oliver DE, Rubin DL, *et al*. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res* 2002;**30**(1):163–5.
89. Piñero J, Queralt-Rosinach N, Bravo A, *et al*. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* 2015;**2015**:bav028.
90. Szarvas G. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In: Proceedings of the 46th Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Columbus, OH, USA, 2008, 281–9.
91. Kilicoglu H, Bergler S. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics* 2008;**9(Suppl 11)**:s10.
92. Malhotra A, Younesi E, Gurulingappa H, *et al*. 'Hypothesis finder:' a strategy for the detection of speculative statements in scientific text. *PLoS Comput Biol* 2013;**9**(7):1–10.
93. ter Riet G, Chesley P, Gross AG, *et al*. All that glitters isn't gold: a survey on acknowledgment of limitations in biomedical studies. *PLoS One* 2013;**8**(11):e73623.
94. Kim J-D, Ohta T, Tsujii J. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* 2008;**9**:10.
95. Vincze V, Szarvas G, Farkas R, *et al*. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics* 2008;**9(Suppl 11)**:S9.
96. Kim J-D, Ohta T, Pyysalo S, *et al*. Overview of BioNLP'09 shared task on event extraction. In: Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop, Association for Computational Linguistics, Boulder, CO, USA, 2009, 1–9.
97. Kim J-D, 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.
98. Farkas R, Vincze V, Mora G, *et al*. The CoNLL 2010 shared task: learning to detect hedges and their scope in natural language text. In: *Proceedings of the CoNLL2010 Shared Task*, 2010, Association for Computational Linguistics, Uppsala, Sweden.
99. Morante R, van Asch V, Daelemans W. Memory-based resolution of in-sentence scopes of hedge cues. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, Uppsala, Sweden, 2010, 40–7.
100. Kilicoglu H, Bergler S. Biological event composition. *BMC Bioinformatics* 2012;**13(Suppl 11)**:S7.
101. Kilicoglu H, Rosemblat G, Cairelli M, *et al*. A compositional interpretation of biomedical event factuality. In: Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015), Association for Computational Linguistics, Denver, CO, USA, 2015, 22–31.

102. Wilbur WJ, Rzhetsky A, Shatkay H. New directions in biomedical text annotations: definitions, guidelines and corpus construction. *BMC Bioinformatics* 2006;**7**:356.
103. Thompson P, Nawaz R, McNaught J, *et al*. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics* 2011;**12**:393.
104. Shatkay H, Pan F, Rzhetsky A, *et al*. Multi-dimensional classification of biomedical text. *Bioinformatics* 2008;**24**(18):2086–93.
105. Miwa M, Thompson P, McNaught J, *et al*. Extracting semantically enriched events from biomedical literature. *BMC Bioinformatics* 2012a;**13**:108.
106. Blake C. Beyond genes, proteins and abstracts: Identifying scientific claims from full-text biomedical articles. *J Biomed Inform* 2009;**43**:173–89.
107. Teufel S, Carletta J, Moens M. An annotation scheme for discourse-level argumentation in research articles. In: *Proceedings of EACL*, Association for Computational Linguistics, Bergen, Norway, 1999, 110–17.
108. Teufel S, Siddharthan A, Batchelor CR. Towards domain-independent argumentative zoning: evidence from chemistry and computational linguistics. In: *Proceedings of EMNLP*, Association for Computational Linguistics, Singapore, 2009, 1493–502.
109. Mizuta Y, Korhonen A, Mullen T, *et al*. Zone analysis in biology articles as a basis for information extraction. *Int J Med Inform* 2006;**75**(6):468–87.
110. Guo Y, Korhonen A, Silins I, *et al*. Weakly supervised learning of information structure of scientific abstracts–is it accurate enough to benefit real-world tasks in biomedicine? *Bioinformatics* 2011;**27**(22):3179–85.
111. Liakata M, Teufel S, Siddhartan A, *et al*. Corpora for conceptualisation and zoning of scientific papers. In: *Proceedings of LREC* 2010, European Language Resources Association (ELRA), Valletta, Malta, 2010, 2054–61.
112. Liakata M, Saha S, Dobnik S, *et al*. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* 2012a;**28**(7):991–1000.
113. Agarwal S, Yu H. Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. *Bioinformatics* 2009;**25**(23):3174–80.
114. de Waard A, Buitelaar P, Eigner T. Identifying the epistemic value of discourse segments in biology texts. In: Proceedings of the 8th International Conference on Computational Semantics, 2009, 351–4.
115. Teufel S, *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. Center for the Study of Language and Information (CSLI), Stanford, CA, USA, 2010.
116. Mann WC, Thompson SA. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 1988;**8**(3):243–81.
117. Miltsakaki E, Prasad R, Joshi A, *et al*. The Penn discourse TreeBank. In: *Proceedings of Language Resources and Evaluation Conference*, 2004, European Language Resources Association (ELRA), Lisbon, Portugal.
118. Prasad R, McRoy S, Frid N, *et al*. The biomedical discourse relation bank. *BMC Bioinformatics* 2011;**12**:188.
119. Ramesh BP, Prasad R, Miller T, *et al*. Automatic discourse connective detection in biomedical text. *J Am Med Inform Assoc* 2012;**19**(5):800–8.
120. Liakata M, Thompson P, de Waard A, *et al*. A three-way perspective on scientific discourse annotation for knowledge extraction. In: Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, Association for Computational Linguistics, Jeju Island, Korea, 2012, 37–46.
121. Kilicoglu HH. Embedding Predications. PhD thesis, Concordia University, 2012.
122. Mons B, Velterop J. Nano-publication in the e-science era. In: Clark T, Luciano JS, Marshall MS, *et al*. (eds), *Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, CEUR-WS.org, 2009.
123. Schneider J, Ciccarese P, Clark T, *et al*. Using the micropublications ontology and the open annotation data model to represent evidence within a drug-drug interaction knowledge base. In: *Proceedings of the 4th Workshop on Linked Science 2014 - Making Sense Out of Data (LISC2014)*, CEUR-WS.org, 2014, 60–70.
124. Hirsch JE. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci USA* 2005;**102**(46): 16569–72.
125. Hutchins BI, Yuan X, Anderson JM, *et al*. Relative Citation Ratio (RCR): a new metric that uses citation rates to measure influence at the article level. *PLoS Biol* 2016;**14**(9):1–25.
126. Wager E, Middleton P. Technical editing of research reports in biomedical journals. *Cochr Database Syst Rev* 2008;**4**:mr00002.
127. Budd JM, Coble ZC, Anderson KM, Retracted publications in biomedicine: cause for concern. In: *Association of College and Research Libraries National Conference Proceedings*, Assoc. of College and Research Libraries, Philadelphia, PA, USA, 2011, 390–5.
128. Radev DR, Muthukrishnan P, Qazvinian V. The ACL anthology network corpus. In: *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, NLPIR4DL '09*, Association for Computational Linguistics, Suntec City, Singapore, 2009, 54–61.
129. Qazvinian V, Radev DR. Scientific paper summarization using citation summary networks. In: *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1, COLING '08*, Coling 2008 Organizing Committee, Manchester, UK, 2008, 689–96.
130. Qazvinian V, Radev DR. Identifying non-explicit citing sentences for citation-based summarization. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Uppsala, Sweden, 2010, 555–64.
131. Abu-Jbara A, Radev D. Reference scope identification in citing sentences. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT '12*, Association for Computational Linguistics, Montréal, Canada, 2012, 80–90.
132. Athar A. Sentiment analysis of scientific citations. Technical Report UCAM-CL-TR-856, University of Cambridge, Computer Laboratory, Cambridge, UK, 2014.
133. Zhu X, Turney PD, Lemire D, *et al*. Measuring academic influence: Not all citations are equal. *J Assoc Inf Sci Technol* 2015;**66**:408–27. abs/1501.06587.
134. Valenzuela M, Ha V, Etzioni O. Identifying meaningful citations. In: *Scholarly Big Data: AI Perspectives, Challenges, and Ideas, Papers from the 2015 AAAI Workshop*, AAAI Press, Palo Alto, California, 2015, 21–6.
135. Athar A. Sentiment analysis of citations using sentence structure-based features. In: *Proceedings of the ACL 2011 Student Session*, Association for Computational Linguistics, Portland, OR, USA, 2011, 81–7.
136. Xu J, Zhang Y, Wu Y, *et al*. Citation sentiment analysis in clinical trial papers. In: *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, San Francisco, CA, USA, 2015, 1334–41.

137. Teufel S, Siddharthan A, Tidhar D. An annotation scheme for citation function. In: *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, SigDIAL '06*, Association for Computational Linguistics, Sydney, Australia, 2006, 80–7.

138. Teufel S, Siddharthan A, Tidhar D. Automatic classification of citation function. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06*, Association for Computational Linguistics, Sydney, Australia, 2006, 103–10.

139. Agarwal S, Choubey L, Yu H. Automatically classifying the role of citations in biomedical articles. In: *AMIA Annual Symposium Proceedings*, Vol. 2010, American Medical Informatics Association, Washington, DC, USA, 2010, 11–15.

140. Olensky M, Schmidt M, van Eck NJ. Evaluation of the citation matching algorithms of CWTS and iFQ in comparison to the Web of science. *J Assoc Inf Sci Technol* 2016;**67**(10):2550–64.

141. Torvik VI, Smalheiser NR. Author name disambiguation in MEDLINE. *ACM Trans Knowl Discov Data* 2009;**3**(3):11.

142. Molla D, Jones C, Sarker A. Impact of citing papers for summarisation of clinical documents. In: *Proceedings of the Australasian Language Technology Association Workshop 2014*, Association for Computational Linguistics, Melbourne, Australia, 2014, 79–87.

143. Jaidka K, Chandrasekaran MK, Rustagi S, *et al*. Overview of the CL-SciSumm 2016 shared task. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, Vol. 1610, Association for Computing Machinery, New York, NY, USA, 2016, 93–102.

144. Cao Z, Li W, Wu D. PolyU at CL-SciSumm 2016. In: *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, Association for Computing Machinery, New York, NY, USA, 2016, 132–8.

145. Cohen KB, Johnson HL, Verspoor K, *et al*. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics* 2010;**11**:492.

146. Jimeno-Yepes A, Verspoor K. Literature mining of genetic variants for curation: quantifying the importance of supplementary material. *Database* 2014;**2014**:bau003.

147. Franceschini F, Maisano D, Mastrogiacomo L. Empirical analysis and classification of database errors in Scopus and Web of Science. *J Informetr* 2016;**10**(4):933–53.

148. Peroni S, Dutton A, Gray T, *et al*. Setting our bibliographic references free: Towards open citation data. *J Doc* 2015;**71**(2):253–77.

149. Peroni S, The semantic publishing and referencing ontologies. In: *Semantic Web Technologies and Legal Scholarly Publishing*. Springer, Cham, 2014, 121–193.

150. Lammey R. Using the Crossref Metadata API to explore publisher content. *Sci Editing* 2016;**3**(2):109–11.

151. Bada M, Eckert M, Evans D, *et al*. Concept annotation in the CRAFT corpus. *BMC Bioinformatics* 2012;**13**(1):1.

152. Ciccarese P, Wu E, Wong G, *et al*. The SWAN biomedical discourse ontology. *J Biomed Inform* 2008;**41**(5):739–51.

153. Khare R, Good BM, Leaman R, *et al*. Crowdsourcing in biomedicine: challenges and opportunities. *Brief Bioinform* 2016;**17**:23–32.

154. Bandrowski A, Brush M, Grethe JS, *et al*. The Resource Identification Initiative: a cultural shift in publishing. *J Comp Neurol* 2016;**524**(1):8–22.

155. Biber D. Representativeness in corpus design. *Lit Linguist Comput* 1993;**8**(4):243–57.

156. Kim J-D, Ohta T, Tateisi Y, *et al*. GENIA corpus - semantically annotated corpus for bio-text mining. *Bioinformatics* 2003;**19(Suppl 1)**:i180–2.

157. Lippincott T, Séaghdha DÓ, Korhonen A. Exploring subdomain variation in biomedical language. *BMC Bioinformatics* 2011;**12**:212.

158. Daumé H, III. Frustratingly easy domain adaptation. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Association for Computational Linguistics, Prague, Czech Republic, 2007, 256–63.

159. Miwa M, Thompson P, Ananiadou S. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics* 2012b;**28**(13):1759–65.

160. Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics* 2016;**32**(18):2839–46.

161. Hearst MA, Divoli A, Guturu H, *et al*. BioText Search Engine: beyond abstract search. *Bioinformatics* 2007;**23**(16):2196–7.

162. Rodriguez-Esteban R, Iossifov I. Figure mining for biomedical research. *Bioinformatics* 2009;**25**(16):2082.

163. Demner-Fushman D, Antani SK, Simpson MS, *et al*. Design and development of a multimodal biomedical information retrieval system. *J Comput Sci Eng* 2012;**6**(2):168–77.

164. Wong W, Martinez D, Cavedon L. Extraction of named entities from tables in gene mutation literature. In: *Proceedings of the BioNLP 2009 Workshop*, Association for Computational Linguistics, Boulder, CO, USA, 2009, 46–54.

165. Peng J, Shi X, Sun Y, *et al*. QTLMiner: QTL database curation by mining tables in literature. *Bioinformatics* 2015;**31**(10):1689.

166. Milosevic N, Gregson C, Hernandez R, *et al*. Disentangling the structure of tables in scientific literature. In: *21st International Conference on Applications of Natural Language to Information Systems (NLDB 2016) Proceedings*, Springer, Cham, 2016, 162–74.

167. Shmanina T, Zukerman I, Cheam AL, *et al*. A corpus of tables in full-text biomedical research publications. In*: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, 2016, 70–9.

168. Alex B, Grover C, Haddow B, *et al*. Assisted curation: Does text mining really help? In: *Proceedings of Pacific Symposium on Biocomputing*, World Scientific, Kohala Coast, HI, USA, 2008, 556–67.

169. Névéol A, Islamaj Doğan R, Lu Z. Semi-automatic Semantic Annotation of PubMed Queries. *J Biomed Inform* 2011b;**44**(2): 310–8.

170. Cohen KB, Xia J, Roeder C, *et al*. Reproducibility in natural language processing: a case study of two R libraries for mining PubMed/MEDLINE. In: LREC 4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language, European Language Resources Association (ELRA), Portorož, Slovenia, 2016, 6–10.

171. Fokkens A, van Erp M, Postma M, *et al*. Offspring from reproduction problems: what replication failure teaches us. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, 1691–701.

172. Névéol A, Cohen K, Grouin C, *et al*. Replicability of research in biomedical natural language processing: a pilot evaluation for a coding task. In Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, 2016, 78–84.