Biomedical Text Mining for Research Rigor and Integrity: Tasks, Challenges, Directions

Halil Kilicoglu
Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine
Bethesda, MD, 20894, USA
kilicogluh@mail.nih.gov

February 14, 2017

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, due to 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 end result 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 towards 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 add checks and balances that promote responsible research practices and can provide significant benefits for the biomedical research enterprise.
Supplementary information: Supplementary material is available at *BioRxiv*.

1 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 (Baker, 2016a). 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 due to 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 (Røttingen *et al.*, 2013). The problems in reproducibility and rigor of published research means that a portion of this expenditure is wasted. Chalmers and Glasziou (2009) estimate that avoidable waste accounts for approximately 85% of 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 (Chalmers and Glasziou, 2009; Collins and Tabak, 2014). As the research is being conducted, the investigator, overwhelmed with administrative tasks, may not be able to provide adequate training/supervision to lab staff (Barham *et al.*, 2014), who do not validate their experiments sufficiently. Only a subset of data that yields statistical significant results may be reported, while negative results may be discarded completely (p-hacking, *selective reporting*, or *publication bias*) (Head *et al.*, 2015; Dwan *et al.*, 2014; Chan *et al.*, 2004). The authors may fail to identify the model organisms, antibodies, reagents necessary for other researchers to replicate the experiments (Vasilevsky *et al.*, 2013). Journal editors, valuing novelty over reproducibility, may be reluctant to publish negative results or replication studies (Collins and Tabak, 2014). 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 (Chalmers, 2002). 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 (Collins and Tabak, 2014; Bowen and Casadevall, 2015). The so-called "publish or perish" culture at academic institutions can create pressure to maximize research quantity with diminishing quality (Collins and Tabak, 2014).

While research rigor and reproducibility in biomedical research is not a recent problem, discussions of the "reproducibility crisis" are largely due to several recent high-profile studies. In one of the pioneering studies, Ioannidis (2005) 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 (2012) 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 (Kyzas *et al.*, 2007; Prinz *et al.*, 2011; Open Science Collaboration, 2015). 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 (Fang *et al.*, 2012). Due to their pervasiveness, however, questionable research practices can be much more detrimental to science (Bouter *et al.*, 2016). Biomedical research outcomes, estimated by life expectancy and novel therapeutics, have remained constant despite rising investment and scientific knowledge in the last five decades, partly attributed to non-reproducibility (Bowen and Casadevall, 2015). 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.

Towards 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 ICMJE trial registration requirement (De Angelis *et al.*, 2004), NIH efforts in enhancing research reproducibility and transparency (Collins and Tabak, 2014) and data discovery (Ohno-Machado *et al.*, 2015), in addition to reporting guidelines (Simera *et al.*, 2010; Nosek *et al.*, 2015), data sharing principles (Wilkinson *et al.*, 2016), 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, 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 (Baker, 2016b). To better focus the remainder of this paper, we use the definitions below, provided in Bollen *et al.* (2015).

- **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[1].
- **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[2].

While reproducibility is often used as an umbrella term to cover all these related issues, in the remainder of this paper, 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 mainly interested in whether/how mining of textual research artifacts can contribute to openness and transparency in biomedical science.

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 Ananiadou and McNaught (2006); Cohen and Demner-Fushman (2014); Gonzalez *et al.* (2016). 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?

[1] `https://grants.nih.gov/reproducibility/index.htm`
[2] `https://grants.nih.gov/grants/research_integrity/whatis.htm`

2 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, protocols) authored by scientists and assessed by funding agencies, reviewers, and IRBs
- *Manuscript submissions*, authored by scientists and evaluated by journal editors, program chairs, and peer reviewers, 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.* (2014) 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 lifecycle 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 (Fang *et al.*, 2012)) 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 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, 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 (Lund *et al.*, 2016) or that funding agencies are not awarding funds to redundant or unjustified proposals (Robinson and Goodman, 2011; Habre *et al.*, 2014).
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 (Greenberg, 2009). 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.

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[3], 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.

[3] `http://headt.eu`

2.1 Plagiarism/Fraud Detection

Plagiarism is "the appropriation of another person's ideas, processes, results, or words without giving appropriate credit" (Habibzadeh and Shashok, 2011). A serious problem in academia, especially with regards to student writing, plagiarism also occurs in medical publications (Habibzadeh and Shashok, 2011). 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 very challenging. Plagiarism detection is a well-studied information retrieval task and dedicated tools have been developed (e.g., TurnItIn[4], Plagramme[5]). CrossRef Similarity Check[6], 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 to 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 (Hoad and Zobel, 2003) or vectors (Stein and Meyer zu Eissen, 2006). Vector representations can be based on characters, words, word sequences (n-grams), or stopwords (Stamatatos, 2011). High number of candidate documents can pose challenges in extrinsic plagiarism detection (Stein and Meyer zu Eissen, 2006); 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 (Meyer zu Eissen and Stein, 2006). 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[7], the last edition of which took place in 2016 (Rosso *et al.*, 2016). Performance for extrinsic plagiarism detection in these competitions has reached an F_1 score of 0.88 (Sanchez-Perez *et al.*, 2014), while the state-of-the-art performance for intrinsic plagiarism detection is much lower at an F_1 score of 0.22 (Kuznetsov *et al.*, 2016). Plagiarism in the PAN corpora was simulated, whereas Nawab *et al.* (2016) used a corpus of PubMed abstracts that were deemed to be suspiciously similar to other abstracts and used a query expansion approach based on UMLS Metathesaurus (Lindberg *et al.*, 1993) 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% (Fang *et al.*, 2012). Such cases often involve data fabrication or falsification (Fanelli, 2009), types of misconduct that would typically be difficult to flag with text analysis alone. Focusing on text only, Markowitz and Hancock (2015) 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 vs. true) (Mihalcea and Strapparava, 2009). Supervised machine learning techniques (Support Vector Machines (SVMs), Naive Bayes) using n-gram and psycholinguistic features have been applied to this task (Mihalcea and Strapparava, 2009; Ott *et al.*, 2011), the latter achieving F_1 score of 0.90 (Ott *et al.*, 2011). Interestingly, inter-annotator agreement (Fleiss' κ=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, since 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.

2.2 Adherence to Reporting Standards and Guidelines

The EQUATOR Network (Simera *et al.*, 2010) has promoted transparent and accurate reporting, and guidelines for various study types have been developed under its umbrella (e.g., CONSORT for randomized clinical trials (RCTs) (Schulz *et al.*, 2010), ARRIVE for preclinical animal studies (Kilkenny *et al.*, 2010)). For example, the CONSORT Statement consists of a 25-item checklist and a flow diagram. The CONSORT checklist for Methods sections is provided in the Supplementary File as an example.

Although a large number of journals have adopted or support such guidelines, adherence to them remains inadequate (Turner *et al.*, 2012). As a solution, structured reporting of key methods and findings

[4] `http://www.turnitin.com`
[5] `http://www.plagramme.com`
[6] `http://www.crossref.org/crosscheck`
[7] `http://pan.webis.de`

has been proposed (Altman, 2015). 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, since 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, since 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[8], StatReviewer[9]); 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 some guideline items as well as other meta-information, often for the purpose of automating systematic review process (Tsafnat *et al.*, 2014). We discuss such research below; see O'Mara-Eves *et al.* (2015) 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 (Chavalarias *et al.*, 2016). 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.* (2013) 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.* (2010) extracted 21 key elements (e.g., eligibility criteria, sample size, primary outcomes, registration number) from full-text RCT articles, 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, 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).

Névéol *et al.* (2011a) focused on extracting deposition statements of biological data (e.g., gene sequences) from full-text articles. They semi-automatically constructed a gold standard corpus. Their approach consisted of two machine learning models: one recognized data deposition components (Data, Action, General Location, Specific Location) using the Conditional Random Fields (CRF) sequence labeling algorithm. The main model, a binary classifier, predicted whether a sentence contains a data deposition statement. This classifier, trained with Naive Bayes and SVM algorithms, employed 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.

PICO frame elements (*Problem*, *Population*, *Intervention*, *Comparison*, and *Outcome*) are recommended for formulating clinical queries in evidence-based medicine (Sackett *et al.*, 1996). They often appear in reporting guideline checklists (e.g., Participants in CONSORT vs. *Population* in PICO). Some research focused on PICO and its variants. Demner-Fushman and Lin (2007) 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 (Aronson and Lang, 2010), a system that maps free text to UMLS Metathesaurus concepts (Lindberg *et al.*, 1993). Recently, Wallace *et al.* (2016), 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 regarding the fraction of

[8] `http://www.peneloperesearch.com/`
[9] `http://www.statreviewer.com/`

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 (Kim *et al.*, 2011). A corpus of 1,000 PubMed abstracts were annotated with these elements (PIBOSO-NICTA) and two classifiers were trained on this corpus: one identified PIBOSO sentences and the other assigned PIBOSO labels to these sentences. A CRF 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 explored similar machine-learning based approaches (Verbeke *et al.*, 2012; Hassanzadeh *et al.*, 2014), state-of-the-art results, without using any external knowledge, were reported by Hassanzadeh *et al.* (2014) (0.91 and 0.87 F_1 scores on structured and unstructured abstracts, respectively).

Marshall *et al.* (2015) developed a tool called RobotReviewer to identify risk of bias (RoB) statements in clinical trials. 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 semi-automatically 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 multi-task learning mapped articles to RoB assessments and simultaneously extracted supporting sentences with an accuracy of 0.71 (compared to 0.78 human accuracy).

Considering research rigor broadly, Kilicoglu *et al.* (2009) 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 (Wilczynski *et al.*, 2005). The base features used included token, PubMed metadata as well as semantic features, as extracted by MetaMap and SemRep (Rindflesch and Fiszman, 2003), 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).

2.3 Managing Information Overload

Tasks we discussed so far did not require much deep natural language understanding; surface-level features, such as n-grams and part-of-speech tags, positional information and limited semantic knowledge, extracted with tools like MetaMap, were mostly sufficient for models that yielded reasonable performance. In this section, we turn to tasks aiming to address information overload caused by the considerable size and the rapid growth of the biomedical literature.

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* (Myers, 1992) 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 (Blackburn and Bos, 2005). 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 (Swanson, 1986; Editorial, 2016), 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[10], increasing research efficiency. 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, 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, drugs), relational level

[10] A service similar to Crossref's CrossMark, which indicates updates on a given publication, can be envisioned.

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 the Supplementary File.

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 the Supplementary File; see recent surveys (Gonzalez *et al.*, 2016; Luo *et al.*, 2016) 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.

2.3.1 Literature-scale relation extraction

Literature-scale relation extraction has been proposed as a method for managing information overload (Kilicoglu *et al.*, 2008). SemMedDB (Kilicoglu *et al.*, 2012) is a database of semantic relations extracted with SemRep (Rindflesch and Fiszman, 2003) from the entire PubMed. In its latest release (as of December 31th, 2016), it contains about 89 million relations extracted from more than 26 million abstracts. It has been used for a variety of tasks, such as clinical decision support (Jonnalagadda *et al.*, 2013), uncovering potential drug interactions in clinical data (Zhang *et al.*, 2014), supporting gene regulatory network construction (Chen *et al.*, 2014), and medical question answering (Hristovski *et al.*, 2015). It also forms the back-end for the Semantic MEDLINE application (Kilicoglu *et al.*, 2008), which integrates semantic relations with automatic abstractive summarization (Fiszman *et al.*, 2004), 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 (Miller *et al.*, 2012) and to elucidate the paradox that obesity is beneficial in critical care despite contributing to disease generally ("the obesity paradox") (Cairelli *et al.*, 2013). Another database, EVEX (Van Landeghem *et al.*, 2013), is based on the Turku Event Extraction System (TEES) (Björne and Salakoski, 2011) and includes relations extracted from the full-text articles in PMC-OA as well as PubMed abstracts. It consists of approximately 40 million bio-molecular events (e.g., gene expression, 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 (Hakala *et al.*, 2015). Other databases, such as PharmGKB (Hewett *et al.*, 2002) and DisGeNET (Piñero *et al.*, 2015), integrate relationships extracted with text mining with those from curated resources.

2.3.2 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 or not. One early task adopting this approach was distinguishing speculative statements from facts (*hedge classification*). For this task, weakly supervised learning techniques (Szarvas, 2008), as well rule-based methods using lexical and syntactic templates (Kilicoglu and Bergler, 2008; Malhotra *et al.*, 2013) have been explored, yielding similar performance (0.85 F_1 score)[11]. Semantically more fine-grained, speculation/negation detection task has focused on recognizing speculation and negation cues in text (e.g., *suggest*, *likely*, *failure*) and their linguistic scope, often formalized as a relation (Kim *et al.*, 2008) or a text segment (Vincze *et al.*, 2008). Speculation/negation detection has been studied in the context of BioNLP Shared Tasks on event extraction (Kim *et al.*, 2009, 2012) and the CoNLL'10 Shared Task on Hedge Detection (Farkas *et al.*, 2010). Supervised machine learning methods (Björne and Salakoski, 2011; Morante *et al.*, 2010) as well as rule-based methods with lexico-syntactic patterns (Kilicoglu and Bergler, 2012) 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`, `Counterfact`) of biological events were computed using a rule-based, syntactic composition approach (Kilicoglu *et al.*, 2015).

Focusing on a more comprehensive characterization of scientific statements, Wilbur *et al.* (2006) 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 towards 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.* (2011), 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 (Shatkay *et al.*, 2008; Miwa *et al.*, 2012b). The

[11] Interesting from a research integrity/transparency perspective, the system developed in (Kilicoglu and Bergler, 2008) 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 (ter Riet *et al.*, 2013).

Claim Framework (Blake, 2009) proposed a categorization of scientific claims according to the specificity of evidence, somewhat similar to Focus dimension in the schema of Wilbur *et al.* (2006). 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.* (1999, 2009) 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.* (2006) adapted this classification to biology articles and presented an annotated corpus. Guo *et al.* (2011) 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 (Liakata *et al.*, 2010) 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 consists of 11 categories (e.g., Background, Motivation, Experiment, Model, Result, 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 (Liakata *et al.*, 2012a), achieving best results with Experiment, Background, and Model classes (0.76, 0.62, 0.53 F_1 scores, respectively). N-gram, 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, Discussion) that underlie the structure of most scientific articles. Since 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 (2009) 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 (e.g., (de Waard *et al.*, 2009)). 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, multi-level model of scientific argumentation, called Knowledge Claim Discourse Model (KCDM), has been proposed by Teufel (2010). 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", "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 to 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) (Mann and Thompson, 1988) and the Penn Discourse TreeBank (PDTB) (Miltsakaki *et al.*, 2004), 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 (BioDRB) corpus (Prasad *et al.*, 2011), in which a subset of PDTB relation types were 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 (Ramesh *et al.*, 2012).

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 (Thompson *et al.*, 2011), CoreSC (Liakata *et al.*, 2010), and discourse segment classification (de Waard *et al.*, 2009), was attempted and these components were shown to be complementary (Liakata *et al.*, 2012b). The Embedding Framework (Kilicoglu, 2012) proposed 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 4 top-level categories (`Modal`, `Valence_Shifter`, `Relational`, `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 this 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 (Kilicoglu and Bergler, 2012), factuality assessment (Kilicoglu *et al.*, 2015), and attribution detection (Kilicoglu, 2012), yielding good performance.

Although not a text mining approach, an effort that deserves discussion here is Micropublications (Clark *et al.*, 2014), a semantic model of scientific claims, evidence, and arguments. Built on top of Semantic Web technologies, micropublications are intended for use in the research lifecycle, 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 (Mons and Velterop, 2009)), as well as with claims in natural language text. The model can accommodate artifacts such as figures, tables, images, and datasets, which text mining approaches generally do not consider. While it has been used for manual annotation (Schneider *et al.*, 2014), to our knowledge, the Micropublications model has not been used as a target for text mining. An example micropublication is presented in the Supplementary File.

2.4 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 (Hirsch, 2005), 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 (e.g., (Hutchins *et al.*, 2016)). Secondly, 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 (Wager and Middleton, 2008): *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. Reference accuracy studies were found to have a median error rate of 39% and quotation accuracy studies a median error rate of 20% (Wager and Middleton, 2008). Greenberg (2009) 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 (Budd *et al.*, 2011). 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 (Radev *et al.*, 2009)). Citation analysis has been proposed for enhancing bibliometrics as well as for extractive summarization (Qazvinian and Radev, 2008). 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 (2010) proposed a method based on Markov Random Fields using sentence similarity and lexical features from sentences. Abu-Jbara and Radev (2012) 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 (2014) 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 (Zhu *et al.*, 2015) and meaningful citations (Valenzuela *et al.*, 2015); 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 (2011) 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.* (2015) annotated the discussion sections of 285 RCT articles with citation sentiment. Using an SVM clas-

sifier 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.* (2006a) presented a scheme, which contained 12 categories (e.g., `Weak` (weakness of the cited approach), `PBas` (cited work as the starting point), `CoCo-` (unfavorable comparison/contrast)) and measured inter-annotator agreement (κ=0.72). Later, Teufel *et al.* (2006b) 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.* (2010) presented a corpus of 43 biomedical articles annotated with eight citation roles (e.g., Background/Perfunctory, Contemporary, Contrast/Conflict, Evaluation, Modality, Similarity/Consistency), achieving moderate inter-annotator agreement (κ=0.63), though it seems difficult to think of some of their categories (Contemporary, 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 Olensky *et al.* (2016) for a comparison of several citation matching algorithms. 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* (Greenberg, 2009). 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 (TAC) 2014 Biomedical Summarization shared task[12], 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, 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 (Molla *et al.*, 2014) 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 (Jaidka *et al.*, 2016); even though the results have been poorer, with the top-ranking system obtaining 0.1 F_1 score (Cao *et al.*, 2016).

3 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 due to 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 (2009) 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 (Cohen *et al.*, 2010). The PMC-OA subset is amenable to automated approaches without much additional pre-processing effort; however, it contains only about a million full-text articles (4% of all PubMed abstracts). Due to availability and access difficulties, researchers often use non-standard PDF-to-text conversion tools to extract full-text from PDF files (e.g., (Marshall *et al.*, 2015; Jimeno-Yepes and Verspoor, 2014)). Considering that the progress in bioNLP is partly attributed to public accessibility 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[13] and some journals (e.g., British Medical Journal) publish pre-publication 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 (Franceschini *et al.*, 2016) 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 (Peroni *et al.*, 2015). They rely on the SPAR ontologies (Peroni, 2014), 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 towards open-access citation data is encouraging for research.

[12] `http://www.nist.gov/tac/2014/BiomedSumm`
[13] `https://report.nih.gov/`

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, since some tasks (e.g., plagiarism detection, managing information overload, citation analysis) presuppose text mining at the literature scale with no publisher restrictions. The Crossref Metadata API initiative (Lammey, 2016) 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, 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 (Bada *et al.*, 2012), 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 since 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, since some existing ontologies focus on scholarly discourse (e.g., SWAN (Ciccarese *et al.*, 2008)), 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, experts), recruited through an open call, provide 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 (Khare *et al.*, 2015). 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 (Mons and Velterop, 2009) and Micropublications (Clark *et al.*, 2014) 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 (Bandrowski *et al.*, 2016) promotes such a model for research resources (e.g., reagents, 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 (e.g., (Biber, 1993)), in practice, most biomedical text corpora focus on a restricted domain of interest. For example, CRAFT (Bada *et al.*, 2012) contains biology and genetics articles, while GENIA (Kim *et al.*, 2003) contains abstracts about biological reactions involving transcription factors in human blood cells. Lippincott *et al.* (2011) 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 very 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 (e.g.,(Daumé, III, 2007)), and such methods have been applied to biomedical text mining tasks (e.g., (Miwa *et al.*, 2012a)). Independently, some systems provided machine learning models that can be retrained on new annotated corpora (e.g., (Björne and Salakoski, 2011; Leaman and Lu, 2016)), while others attempted to generalize by appealing to linguistic principles (e.g., (Rindflesch and Fiszman, 2003; Kilicoglu and Bergler, 2012)).

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 (Clark *et al.*, 2014), support them. Figure retrieval has been considered, mainly focusing on using text from figure captions (Hearst *et al.*, 2007), text within figures (Rodriguez-Esteban and Iossifov, 2009), as well as text from paragraphs discussing the figures and NER (Demner-Fushman *et al.*, 2012). Research on information extraction from tables is rare (Wong *et al.*, 2009; Peng *et al.*, 2015; Milosevic *et al.*, 2016), though this may change with recent availability of corpora (Shmanina *et al.*, 2016). Jimeno-Yepes and Verspoor (2014) showed that most literature-curated mutation and genetic variant 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 vary 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 (Alex *et al.*, 2008; Névéol *et al.*, 2011b), 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 (Ott *et al.*, 2011)). Depending on the task, a user may prefer not the setting which 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.

4 Conclusion

Towards 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 a significant role and surveyed the state-of-the-art for these tasks. Among the tasks we discussed, we believe that the following can have the biggest and most immediate impact: given a document (e.g., manuscript, publication), i) checking for adherence to *all* elements of the relevant reporting guideline, ii) generating document-level and literature-level argumentation graphs, iii) constructing citation quotation networks. 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.

Acknowledgements

I thank 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 paper.

Funding

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

References

Abu-Jbara, A. and Radev, D. (2012). 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, pages 80–90.

Agarwal, S. and Yu, H. (2009). Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. *Bioinformatics*, **25**(23), 3174–3180.

Agarwal, S., Choubey, L., and Yu, H. (2010). Automatically classifying the role of citations in biomedical articles. In *AMIA Annual Symposium proceedings*, volume 2010, pages 11–15.

Alex, B., Grover, C., Haddow, B., Kabadjor, M., Klein, E., Matthews, M., Roebuck, S., Tobin, R., and Wang, X. (2008). Assisted curation: Does text mining really help? In *Proceedings of Pacific Symposium on Biocomputing*, pages 556–567.

Altman, D. G. (2015). Making research articles fit for purpose: structured reporting of key methods and findings. *Trials*, **16**(1), 53.

Ananiadou, S. and McNaught, J. (2006). *Text mining for biology and biomedicine*. Artech House, Boston, MA.

Aronson, A. R. and Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association (JAMIA)*, **17**(3), 229–236.

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, pages 81–87.

Athar, A. (2014). Sentiment analysis of scientific citations. Technical Report UCAM-CL-TR-856, University of Cambridge, Computer Laboratory.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Cohen, K. B., Verspoor, K., Blake, J. A., *et al.* (2012). Concept annotation in the CRAFT corpus. *BMC bioinformatics*, **13**(1), 1.

Baker, M. (2016a). 1,500 scientists lift the lid on reproducibility. *Nature*, **533**, 452–454.

Baker, M. (2016b). Muddled meanings hamper efforts to fix reproducibility crisis. *Nature*.

Bandrowski, A., Brush, M., Grethe, J. S., Haendel, M. A., Kennedy, D. N., Hill, S., Hof, P. R., Martone, M. E., Pols, M., Tan, S. C., Washington, N., Zudilova-Seinstra, E., and Vasilevsky, N. (2016). The Resource Identification Initiative: A cultural shift in publishing. *Journal of Comparative Neurology*, **524**(1), 8–22.

Barham, B. L., Foltz, J. D., and Prager, D. L. (2014). Making time for science. *Research Policy*, **43**(1), 21–31.

Begley, C. G. and Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, **483**(29), 531–533.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, **8**(4), 243–257.

Björne, J. and Salakoski, T. (2011). Generalizing Biomedical Event Extraction. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 183–191. Association for Computational Linguistics.

Blackburn, P. and Bos, J. (2005). *Representation and Inference for Natural Language. A First Course in Computational Semantics*. CSLI.

Blake, C. (2009). Beyond genes, proteins and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, **43**, 173–189.

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., and Olds, J. L. (2015). Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. Technical report, National Science Foundation.

Bouter, L. M., Tijdink, J., Axelsen, N., Martinson, B. C., and ter Riet, G. (2016). Ranking major and minor research misbehaviors: results from a survey among participants of four World Conferences on Research Integrity. *Research Integrity and Peer Review*, **1**(1), 17.

Bowen, A. and Casadevall, A. (2015). Increasing disparities between resource inputs and outcomes, as measured by certain health deliverables, in biomedical research. *Proceedings of the National Academy of Sciences of the United States of America*, **112**(36), 11335–11340.

Budd, J. M., Coble, Z. C., and Anderson, K. M. (2011). Retracted publications in biomedicine: Cause for concern. In *Association of College and Research Libraries National Conference Proceedings*, pages 390–395.

Cairelli, M. J., Miller, C. M., Fiszman, M., Workman, T. E., and Rindflesch, T. C. (2013). 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*, pages 164–173.

Cao, Z., Li, W., and Wu, D. (2016). PolyU at CL-SciSumm 2016. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 132–138.

Chalmers, I. (2002). Lessons for research ethics committees. *The Lancet*, **359**(9301), 174.

Chalmers, I. and Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet*, **374**(9683), 86–89.

Chan, A., Hróbjartsson, A., Haahr, M., Gøtzsche, P., and Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA*, **291**(20), 2457–2465.

Chavalarias, D., Wallach, J. D., Li, A. H., and Ioannidis, J. P. A. (2016). Evolution of reporting P values in the biomedical literature, 1990-2015. *JAMA*, **315**(11), 1141–1148.

Chen, G., Cairelli, M. J., Kilicoglu, H., Shin, D., and Rindflesch, T. C. (2014). Augmenting microarray data with literature-based knowledge to enhance gene regulatory network inference. *PLOS Computational Biology*, **10**(6), 1–16.

Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., and Clark, T. (2008). The SWAN Biomedical Discourse Ontology. *Journal of Biomedical Informatics*, **41**(5), 739–751.

Clark, T., Ciccarese, P., and Goble, C. (2014). Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. *Journal of Biomedical Semantics*, **5**(1), 28.

Cohen, K. B. and Demner-Fushman, D. (2014). *Biomedical Natural Language Processing*. John Benjamins, Amsterdam.

Cohen, K. B., Johnson, H. L., Verspoor, K., Roeder, C., and Hunter, L. E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, **11**, 492.

Collins, F. S. and Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature*, **505**(7485), 612–613.

Daumé, III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

De Angelis, C., Drazen, J., Frizelle, F., Haug, C., Hoey, J., Horton, R., Kotzin, S., Laine, C., Marusic, A., Overbeke, A., Schroeder, T., Sox, H., Van Der Weyden, M., and International Committee of Medical Journal Editors (2004). Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine*, **351**(12), 1250–1251.

de Waard, A., Buitelaar, P., and Eigner, T. (2009). Identifying the Epistemic Value of Discourse Segments in Biology Texts. In *Proceedings of the 8th International Conference on Computational Semantics*, pages 351–354.

Demner-Fushman, D. and Lin, J. (2007). Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, **33**(1), 63–103.

Demner-Fushman, D., Antani, S. K., Simpson, M. S., and Thoma, G. R. (2012). Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering*, **6**(2), 168–177.

Dwan, K., Altman, D. G., Clarke, M., Gamble, C., Higgins, J. P., Sterne, J. A., Williamson, P. R., and Kirkham, J. J. (2014). Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. *PLoS Medicine*, **11**(6), e1001666.

Editorial (2016). So long to the silos. *Nature Biotechnology*, **34**(357).

Fanelli, D. (2009). How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PLOS ONE*, **4**(5), 1–11.

Fang, F. C., Steen, R. G., and Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(42), 17028–17033.

Farkas, R., Vincze, V., Mora, G., Csirik, J., and Szarvas, G. (2010). The CoNLL 2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the CoNLL2010 Shared Task*.

Fiszman, M., Rindflesch, T. C., and Kilicoglu, H. (2004). Abstraction summarization for managing the biomedical research literature. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, pages 76–83.

Franceschini, F., Maisano, D., and Mastrogiacomo, L. (2016). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of Informetrics*, **10**(4), 933–953.

Gonzalez, G., Tahsin, T., Goodale, B. C., Greene, A. C., and Greene, C. S. (2016). Recent advances and emerging applications in text and data mining for biomedical discovery. *Briefings in Bioinformatics*, **17**(1), 33–42.

Greenberg, S. A. (2009). How citation distortions create unfounded authority: analysis of a citation network. *BMJ*, **339**, b2680.

Guo, Y., Korhonen, A., Silins, I., and Stenius, U. (2011). Weakly supervised learning of information structure of scientific abstracts–is it accurate enough to benefit real-world tasks in biomedicine? *Bioinformatics*, **27**(22), 3179–3185.

Habibzadeh, F. and Shashok, K. (2011). Rules of the game of scientific writing: fair play and plagiarism. *Croatian Medical Journal*, **52**(4), 576–577.

Habre, C., Tramèr, M. R., Pöpping, D. M., and Elia, N. (2014). Ability of a meta-analysis to prevent redundant research: systematic review of studies on pain from propofol injection. *BMJ*, **349**, g5219.

Hakala, K., Van Landeghem, S., Salakoski, T., Van de Peer, Y., and Ginter, F. (2015). Application of the EVEX resource to event extraction and network construction: Shared Task entry and result analysis. *BMC Bioinformatics*, **16**(Suppl 16), S3.

Hassanzadeh, H., Groza, T., and Hunter, J. (2014). Identifying scientific artefacts in biomedical literature: The Evidence Based Medicine use case. *Journal of Biomedical Informatics*, **49**, 159–170.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, **13**(3).

Hearst, M. A., Divoli, A., Guturu, H., Ksikes, A., Nakov, P., Wooldridge, M. A., and Ye, J. (2007). BioText Search Engine: beyond abstract search. *Bioinformatics*, **23**(16), 2196–2197.

Hewett, M., Oliver, D. E., Rubin, D. L., Easton, K. L., Stuart, J. M., Altman, R. B., and Klein, T. E. (2002). PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Research*, **30**(1), 163–165.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(46), 16569–16572.

Hoad, T. C. and Zobel, J. (2003). Methods for identifying versioned and plagiarised documents. *Journal of the American Society for Information Science and Technology*, **54**, 203–215.

Hristovski, D., Dinevski, D., Kastrin, A., and Rindflesch, T. C. (2015). Biomedical question answering using semantic relations. *BMC Bioinformatics*, **16**(1), 6+.

Hutchins, B. I., Yuan, X., Anderson, J. M., and Santangelo, G. M. (2016). Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. *PLOS Biology*, **14**(9), 1–25.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, **2**(8), e124.

Jaidka, K., Chandrasekaran, M. K., Rustagi, S., and Kan, M. (2016). 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)*, volume 1610, pages 93–102.

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

Jonnalagadda, S., Fiol, G. D., Medlin, R., Weir, C. R., Fiszman, M., Mostafa, J., and Liu, H. (2013). Automatically extracting sentences from Medline citations to support clinicians' information needs. *JAMIA*, **20**(5), 995–1000.

Kafkas, Ş., Kim, J.-H., and McEntyre, J. R. (2013). Database citation in full text biomedical articles. *PLOS ONE*, **8**(5), e63184.

Khare, R., Good, B. M., Leaman, R., Su, A. I., and Lu, Z. (2015). Crowdsourcing in biomedicine: challenges and opportunities. *Briefings in Bioinformatics*, pages 1–10.

Kilicoglu, H. and Bergler, S. (2008). Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, **9 Suppl 11**, s10.

Kilicoglu, H. and Bergler, S. (2012). Biological Event Composition. *BMC Bioinformatics*, **13 (Suppl 11)**, S7.

Kilicoglu, H., Fiszman, M., Rodriguez, A., Shin, D., Ripple, A., and Rindflesch, T. (2008). Semantic MEDLINE: A Web Application to Manage the Results of PubMed Searches. In T. Salakoski, D. R. Schuhmann, and S. Pyysalo, editors, *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, pages 69–76.

Kilicoglu, H., Demner-Fushman, D., Rindflesch, T. C., Wilczynski, N. L., and Haynes, R. B. (2009). Towards automatic recognition of scientifically rigorous clinical research evidence. *Journal of the American Medical Informatics Association*, **16**(1), 25–31.

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., and Rindflesch, T. C. (2012). SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, **28**(23), 3158–3160.

Kilicoglu, H., Rosemblat, G., Cairelli, M., and Rindflesch, T. (2015). A compositional interpretation of biomedical event factuality. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 22–31.

Kilicoglu, H. H. (2012). *Embedding Predications*. Ph.D. thesis, Concordia University.

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., and Altman, D. G. (2010). Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biology*, **8**(6), e1000412.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus - semantically annotated corpus for bio-text mining. *Bioinformatics*, **19 Suppl 1**.

Kim, J.-D., Ohta, T., and Tsujii, J. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, **9**, 10.

Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., and Tsujii, J. (2009). Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*, pages 1–9.

Kim, J.-D., Nguyen, N., Wang, Y., Tsujii, J., Takagi, T., and Yonezawa, A. (2012). The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011. *BMC Bioinformatics*, **13(Suppl 11)**, S1.

Kim, S. N., Martínez, D., Cavedon, L., and Yencken, L. (2011). Automatic classification of sentences to support Evidence Based Medicine. *BMC Bioinformatics*, **12**(S-2), S5.

Kiritchenko, S., de Bruijn, B., Carini, S., Martin, J., and Sim, I. (2010). ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Medical Informatics and Decision Making*, **10**(1), 56.

Kuznetsov, M., Motrenko, A., Kuznetsova, R., and Strijov, V. (2016). Methods for Intrinsic Plagiarism Detection and Author Diarization. In K. Balog, L. Cappellato, N. Ferro, and C. Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*.

Kyzas, P. A., Denaxa-Kyza, D., and Ioannidis, J. P. A. (2007). Almost all articles on cancer prognostic markers report statistically significant results. *European Journal of Cancer*, **43**(17), 2559–2579.

Lammey, R. (2016). Using the Crossref Metadata API to explore publisher content. *Science Editing*, **3**(2), 109–111.

Leaman, R. and Lu, Z. (2016). TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, **32**(18), 2839–2846.

Liakata, M., Teufel, S., Siddhartan, A., and Batchelor, C. (2010). Corpora for conceptualisation and zoning of scientific papers. In *Proceedings of LREC 2010*, pages 2054–2061.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., and Rebholz-Schuhmann, D. (2012a). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, **28**(7), 991–1000.

Liakata, M., Thompson, P., de Waard, A., Nawaz, R., Maat, H. P., and Ananiadou, S. (2012b). A three-way perspective on scientific discourse annotation for knowledge extraction. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 37–46.

Lindberg, D. A. B., Humphreys, B. L., and McCray, A. T. (1993). The Unified Medical Language System. *Methods of Information in Medicine*, **32**, 281–291.

Lippincott, T., Séaghdha, D. Ó., and Korhonen, A. (2011). Exploring subdomain variation in biomedical language. *BMC Bioinformatics*, **12**, 212.

Lund, H., Brunnhuber, K., Juhl, C., Robinson, K., Leenaars, M., Dorch, B. F., Jamtvedt, G., Nortvedt, M. W., Christensen, R., and Chalmers, I. (2016). Towards evidence based research. *BMJ*, **355**, i5440.

Luo, Y., Uzuner, Ö., and Szolovits, P. (2016). Bridging semantics and syntax with graph algorithms state-of-the-art of extracting biomedical relations. *Briefings in Bioinformatics*, **18**(1), 160–178.

Malhotra, A., Younesi, E., Gurulingappa, H., and Hofmann-Apitius, M. (2013). 'HypothesisFinder:' A Strategy for the Detection of Speculative Statements in Scientific Text. *PLOS Computational Biology*, **9**(7), 1–10.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, **8**(3), 243–281.

Markowitz, D. M. and Hancock, J. T. (2015). Linguistic obfuscation in fraudulent science. *Journal of Language and Social Psychology*, **35**(4), 435–445.

Marshall, I. J., Kuiper, J., and Wallace, B. C. (2015). RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, pages 193–201.

Meyer zu Eissen, S. and Stein, B. (2006). Intrinsic Plagiarism Detection. In M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, *28th European Conference on IR Research (ECIR 06)*, volume 3936 of *Lecture Notes in Computer Science*, pages 565–569.

Mihalcea, R. and Strapparava, C. (2009). The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312.

Miller, C. M., Rindflesch, T. C., Fiszman, M., Hristovski, D., Shin, D., Rosemblat, G., Zhang, H., and Strohl, K. P. (2012). A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men. *Sleep*, **35**(2), 279–285.

Milosevic, N., Gregson, C., Hernandez, R., and Nenadic, G. (2016). Disentangling the structure of tables in scientific literature. In *21st International Conference on Applications of Natural Language to Information Systems (NLDB 2016) Proceedings*, pages 162–174.

Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004). The Penn Discourse TreeBank. In *Proceedings of Language Resources and Evaluation Conference*.

Miwa, M., Thompson, P., and Ananiadou, S. (2012a). Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, **28**(13), 1759–1765.

Miwa, M., Thompson, P., McNaught, J., Kell, D. B., and Ananiadou, S. (2012b). Extracting semantically enriched events from biomedical literature. *BMC Bioinformatics*, **13**, 108.

Mizuta, Y., Korhonen, A., Mullen, T., and Collier, N. (2006). Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, **75**(6), 468–487.

Molla, D., Jones, C., and Sarker, A. (2014). Impact of citing papers for summarisation of clinical documents. In *Proceedings of the Australasian Language Technology Association Workshop 2014* , pages 79–87.

Mons, B. and Velterop, J. (2009). Nano-Publication in the e-Science era. In T. Clark, J. S. Luciano, M. S. Marshall, E. Prud'hommeaux, and S. Stephens, editors, *Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*.

Morante, R., van Asch, V., and Daelemans, W. (2010). Memory-based resolution of in-sentence scopes of hedge cues. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 40–47.

Myers, G. (1992). 'In this paper we report ...': Speech acts and scientific facts. *Journal of Pragmatics*, **17**(4), 295–313.

Nawab, R. M. A., Stevenson, M., and Clough, P. (2016). An IR-based Approach Utilising Query Expansion for Plagiarism Detection in MEDLINE. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **PP**(99), 1–1.

Névéol, A., Wilbur, W. J., and Lu, Z. (2011a). Extraction of data deposition statements from the literature. *Bioinformatics*, **27**(23), 3306–3312.

Névéol, A., Islamaj Doğan, R., and Lu, Z. (2011b). Semi-automatic Semantic Annotation of PubMed Queries. *Journal of Biomedical Informatics*, **44**(2), 310–318.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E. L., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R., and Yarkoni, T. (2015). Promoting an open research culture. *Science*, **348**(6242), 1422–1425.

Ohno-Machado, L., Alter, G., Fore, I., Martone, M., Sansone, S.-A., and Xu, H. (2015). bioCADDIE white paper - Data Discovery Index. Technical report, Figshare.

Olensky, M., Schmidt, M., and van Eck, N. J. (2016). Evaluation of the citation matching algorithms of CWTS and iFQ in comparison to the Web of science. *Journal of the Association for Information Science and Technology*, **67**(10), 2550–2564.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., and Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, **4**(1), 5.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, **349**(6251).

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the ACL:HLT 2011*, pages 309–319.

Peng, J., Shi, X., Sun, Y., Li, D., Liu, B., Kong, F., and Yuan, X. (2015). QTLMiner: QTL database curation by mining tables in literature. *Bioinformatics*, **31**(10), 1689.

Peroni, S. (2014). The semantic publishing and referencing ontologies. In *Semantic Web Technologies and Legal Scholarly Publishing*, pages 121–193. Springer.

Peroni, S., Dutton, A., Gray, T., and Shotton, D. (2015). Setting our bibliographic references free: Towards open citation data. *Journal of Documentation*, **71**(2), 253–277.

Piñero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., and Furlong, L. I. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford)*, **2015**, bav028.

Prasad, R., McRoy, S., Frid, N., Joshi, A., and Yu, H. (2011). The biomedical discourse relation bank. *BMC Bioinformatics*, **12**, 188.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, **10**(9), 712.

Qazvinian, V. and Radev, D. R. (2008). Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 689–696.

Qazvinian, V. and Radev, D. R. (2010). Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 555–564.

Radev, D. R., Muthukrishnan, P., and Qazvinian, V. (2009). The ACL Anthology Network Corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, NLPIR4DL '09, pages 54–61.

Ramesh, B. P., Prasad, R., Miller, T., Harrington, B., and Yu, H. (2012). Automatic discourse connective detection in biomedical text. *Journal of the American Medical Informatics Association*, **19**(5), 800–808.

Rindflesch, T. C. and Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, **36**(6), 462–477.

Robinson, K. and Goodman, S. (2011). A systematic examination of the citation of prior research in reports of randomized, controlled trials. *Annals of Internal Medicine*, **154**(1), 50–55.

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

Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., and Stein, B. (2016). Overview of PAN'16—New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation. In N. Fuhr, P. Quaresma, B. Larsen, T. Gonçalves, K. Balog, C. Macdonald, L. Cappellato, and N. Ferro, editors, *7th International Conference of the CLEF Initiative (CLEF 16)*.

Røttingen, J.-A., Regmi, S., Eide, M., Young, A. J., Viergever, R. F., Årdal, C., Guzman, J., Edwards, D., Matlin, S. A., and Terry, R. F. (2013). Mapping of available health research and development data: what's there, what's missing, and what role is there for a global observatory? *The Lancet*, **382**(9900), 1286–1307.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ*, **312**(7023), 71–72.

Sanchez-Perez, M., Sidorov, G., and Gelbukh, A. (2014). A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014. In L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*.

Schneider, J., Ciccarese, P., Clark, T., and Boyce, R. D. (2014). 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)*, pages 60–70.

Schulz, K. F., Altman, D. G., and Moher, D. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, **340**, c332.

Shatkay, H., Pan, F., Rzhetsky, A., and Wilbur, W. J. (2008). Multi-dimensional classification of biomedical text. *Bioinformatics*, **24**(18), 2086–2093.

Shmanina, T., Zukerman, I., Cheam, A. L., Bochynek, T., and Cavedon, L. (2016). 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)*, pages 70–79.

Simera, I., Moher, D., Hirst, A., Hoey, J., Schulz, K. F., and Altman, D. G. (2010). Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC Medicine*, **8**(1), 24.

Stamatatos, E. (2011). Plagiarism detection using stopword n-grams. *Journal of the Association for Information Science and Technology*, **62**(12), 2512–2527.

Stein, B. and Meyer zu Eissen, S. (2006). Near similarity search and plagiarism analysis. *From Data and Information Analysis to Knowledge Engineering*, pages 430–437.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, **30**(1), 7–18.

Szarvas, G. (2008). Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of the 46th Meeting of the Association for Computational Linguistics*, pages 281–289.

ter Riet, G., Chesley, P., Gross, A. G., Siebeling, L., Muggensturm, P., Heller, N., Umbehr, M., Vollenweider, D., Yu, T., Akl, E. A., Brewster, L., Dekkers, O. M., Mhlhauser, I., Richter, B., Singh, S., Goodman, S., and Puhan, M. A. (2013). All That Glitters Isn't Gold: A Survey on Acknowledgment of Limitations in Biomedical Studies. *PLOS ONE*, **8**(11), 1–6.

Teufel, S. (2010). *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. Center for the Study of Language and Information (CSLI).

Teufel, S., Carletta, J., and Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of EACL*, pages 110–117.

Teufel, S., Siddharthan, A., and Tidhar, D. (2006a). An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, SigDIAL '06, pages 80–87.

Teufel, S., Siddharthan, A., and Tidhar, D. (2006b). Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 103–110.

Teufel, S., Siddharthan, A., and Batchelor, C. R. (2009). Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In *Proceedings of EMNLP*, pages 1493–1502.

Thompson, P., Nawaz, R., McNaught, J., and Ananiadou, S. (2011). Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, **12**, 393.

Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., and Coiera, E. (2014). Systematic review automation technologies. *Systematic Reviews*, **3**(1), 74.

Turner, L., Shamseer, L., Altman, D. G., Schulz, K. F., and Moher, D. (2012). Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Systematic Reviews*, **1**(1), 60.

Valenzuela, M., Ha, V., and Etzioni, O. (2015). Identifying meaningful citations. In *Scholarly Big Data: AI Perspectives, Challenges, and Ideas, Papers from the 2015 AAAI Workshop*, pages 21–26.

Van Landeghem, S., Björne, J., Wei, C.-H., Hakala, K., Pyysalo, S., Ananiadou, S., Kao, H.-Y., Lu, Z., Salakoski, T., Van de Peer, Y., and Ginter, F. (2013). Large-scale event extraction from literature with multi-level gene normalization. *PLOS ONE*, **8**(4), e55814.

Vasilevsky, N. A., Brush, M. H., Paddock, H., Ponting, L., Tripathy, S. J., LaRocca, G. M., and Haendel, M. A. (2013). On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ*, **1**, e148.

Verbeke, M., Asch, V. V., Morante, R., Frasconi, P., Daelemans, W., and Raedt, L. D. (2012). 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*, pages 579–589.

Vincze, V., Szarvas, G., Farkas, R., Mora, G., and Csirik, J. (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, **9 Suppl 11**, S9.

Wager, E. and Middleton, P. (2008). Technical editing of research reports in biomedical journals. *Cochrane database of systematic reviews (Online)*, **4**(mr00002).

Wallace, B. C., Kuiper, J., Sharma, A., Zhu, M., and Marshall, I. J. (2016). Extracting PICO Sentences from Clinical Trial Reports Using Supervised Distant Supervision. *Journal of Machine Learning Research*, **17**(132), 1–25.

Wilbur, W. J., Rzhetsky, A., and Shatkay, H. (2006). New directions in biomedical text annotations: definitions, guidelines and corpus construction. *BMC Bioinformatics*, **7**, 356.

Wilczynski, N. L., Morgan, D., Haynes, R. B., and et al. (2005). An overview of the design and methods for retrieving high-quality studies for clinical care. *BMC Medical Informatics and Decision Making*, **5**, 20.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., â Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, **3**, 160018.

Wong, W., Martinez, D., and Cavedon, L. (2009). Extraction of named entities from tables in gene mutation literature. In *Proceedings of the BioNLP 2009 Workshop*, pages 46–54.

Xu, J., Zhang, Y., Wu, Y., Wang, J., Dong, X., and Xu, H. (2015). Citation sentiment analysis in clinical trial papers. In *AMIA Annual Symposium Proceedings*, pages 1334–1341.

Zhang, R., Cairelli, M. J., Fiszman, M., Rosemblat, G., Kilicoglu, H., Rindflesch, T. C., Pakhomov, S. V., and Melton, G. B. (2014). Using semantic predications to uncover drug-drug interactions in clinical data. *Journal of Biomedical Informatics*, **49**, 134 –147.

Zhu, X., Turney, P. D., Lemire, D., and Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *CoRR*, **abs/1501.06587**.