Database, 2016, 1–10
doi: 10.1093/database/baw161
Perspective

Perspective

Pressing needs of biomedical text mining in biocuration and beyond: opportunities and challenges

Ayush Singhal[1], Robert Leaman[1], Natalie Catlett[2], Thomas Lemberger[3], Johanna McEntyre[4], Shawn Polson[5], Ioannis Xenarios[6], Cecilia Arighi[1,5,*], and Zhiyong Lu[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, [2]Selventa, Cambridge, MA 02140, USA, [3]EMBO, Meyerhofstrasse 1, Heidelberg 69117, Germany, [4]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [5]Center for Bioinformatics and Computational Biology and Department of Computer and Information Sciences, Delaware Biotechnology Institute, University of Delaware, Newark, DE 19711, USA and [6]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

*Corresponding author Email: zhiyong.lu@nih.gov

Correspondence may also be addressed to Cecilia Arighi. Email: arighi@dbi.udel.edu

Citation details: Singhal,A., Leaman,R., Catlett,N., *et al*. Pressing needs of biomedical text mining in biocuration and beyond: opportunities and challenges. Database (2016) Vol. 2016: article ID baw161; doi:10.1093/database/baw161

Received 19 April 2016; Revised 10 November 2016; Accepted 11 November 2016

Abstract

Text mining in the biomedical sciences is rapidly transitioning from small-scale evaluation to large-scale application. In this article, we argue that text-mining technologies have become essential tools in real-world biomedical research. We describe four large scale applications of text mining, as showcased during a recent panel discussion at the BioCreative V Challenge Workshop. We draw on these applications as case studies to characterize common requirements for successfully applying text-mining techniques to practical biocuration needs. We note that system 'accuracy' remains a challenge and identify several additional common difficulties and potential research directions including (i) the 'scalability' issue due to the increasing need of mining information from millions of full-text articles, (ii) the 'interoperability' issue of integrating various text-mining systems into existing curation workflows and (iii) the 'reusability' issue on the difficulty of applying trained systems to text genres that are not seen previously during development. We then describe related efforts within the text-mining community, with a special focus on the BioCreative series of challenge workshops. We believe that focusing on the near-term challenges identified in this work will amplify the opportunities afforded by the continued adoption of text-mining tools. Finally, in order to sustain the curation ecosystem and have text-mining systems adopted for practical benefits, we call for increased collaboration between text-mining researchers and various stakeholders, including researchers, publishers and biocurators.

Introduction

The unprecedented advances in high-throughput technology and tools to support bioscience have led to a boom in biological and biomedical science research and an accompanying growth of the scientific literature. Access to the wealth of knowledge embedded in the literature is critical for enabling continued scientific advancements and breakthroughs. For this reason, several efforts over the last decade have focused on improving knowledge reusability through improved storage, representation and curation. These efforts include both public literature resources (e.g. PubMed and PubMed Central/Europe PMC) and biological knowledge bases (e.g. UniProt (1), NCBI Database Resources (2). Figure 1 illustrates the interconnection between literature services and biological databases, and their importance in biological research. As can be seen, researchers rely on literature services to keep up with the state of the art on topics of their interest, to generate novel hypotheses, and as a reference for developing research strategies. In addition, today's curated databases are critical in biomedical research by being a firsthand tool for researchers to investigate their hypothesis or research results (3).

Biological knowledge bases rely heavily on expert curation, however, and scaling to accommodate the growth of the scientific literature has been a continued challenge. Automatically annotating biological entities such as genes/protein and diseases (4, 5) and other scientific artifacts in biomedical literature, such as investigation techniques or the dataset used (6) is useful for improving the scalability of biocuration services. Surveys regarding the role of text mining for assisting literature curation were performed during the International Biocuration Conference and Workshop (Berlin, 2009) and the BioCreative 2012 Workshop (Washington, DC) (7, 8). The 2012 report indicates that more databases have adopted text mining into their curation workflows in some form than in 2009. A number of studies have indicated improved curation productivity with the assistance of text mining. In Table 1, we present a subset of studies benchmarking the quantitative significance of text-mining systems in database curation (9–13). We also refer the reader to the Interactive Annotation Task (IAT) at BioCreatives III–V (14–17), which investigated some aspects of usability and productivity of the text-mining systems for biocuration.

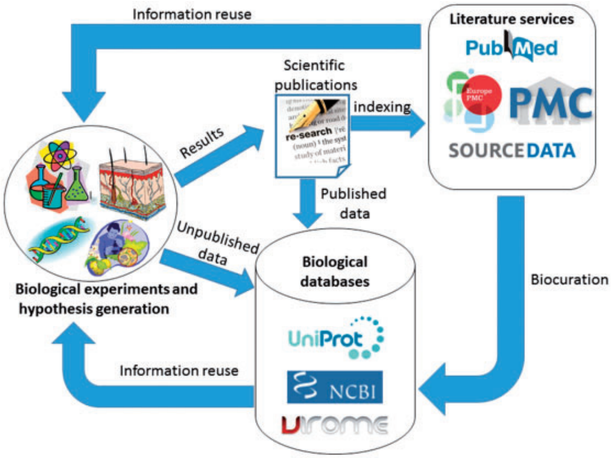

Figure 1. Interconnection between literature services and biological databases.

Given the earlier successes and increasing cost/limited resources in manual curation, we argue that computational approaches such as text mining are essential in the future to provide researchers and medical professionals efficient, comprehensive and up-to-date literature services to manage this growth according to customizable criteria such as clinical relevancy or specific genes or species. Since most of the discoveries and breakthroughs are first made available to the public through scholarly publications, the emphasis of this position article is with regard to text-mining applications in literature search and curation. Specifically, the four real-world applications discussed are (i) Literature search (Europe PMC), (ii) Data search (SourceData), (iii) BEL database curation and (iv) VIROME database curation.

In this article, we first discuss the two applications related to literature services, Europe PMC and SourceData, explaining both their value to the bioscience community and how text mining is essential for their continued progress. We next discuss two recent efforts supporting biological databases, BEL and VIROME, which curate information related to biological cause–effect relationships and microbiomes, respectively. Finally, we summarize the opportunities for text mining in such applications and the multiple challenges that hamper its immediate adoption in these applications. We also provide our understanding of a few strategies to facilitate an increased adoption of text mining in such applications.

Real world large scale applications

Europe PMC (Johanna McEntyre, EMBL-EBI)

Europe PMC (https://europepmc.org/) is a database of abstracts and full text articles (5). Partnering with PMC from the National Library of Medicine USA and PMC Canada

Table 1. A selection of studies demonstrating the benefit of text mining assistance for curation.

Citations	Database	Curation Task	TM System	Results
(9)	Wormbase	Cellular-component curation	Textpresso	8-fold increase in curation efficiency
(10)	dictyBase	Cellular-component curation	Textpresso	2.5-fold increase in curation efficiency
(10)	TAIR	Cellular-component curation	Textpresso	10-fold decrease in time for curation
(11)	TAIR	Genes	PubTator	45% increase in productivity
(12)	PIR	PPI involving protein phosphorylation	eFIP	2.5-fold increase curation efficiency
(13)	Flybase	Genes	Tagtog	2-fold decrease in curation time

as a PMC International node, it contains ~30 million abstracts (including PubMed) and over 3.6 million full text articles from the life sciences. In addition to serving general life science researchers who use *Europe PMC* to search the literature and access full text articles, *Europe PMC* also seeks to serve the specialized subset of users who are database curators. Database curators are professional literature readers, filterers, evaluators and extractors who work with the purpose of adding scientific value and context to public data resources.

Manual literature curation has resulted in many bioinformatics resources of excellent quality. It is clear, however, that some supportive computational approaches will be required in order for curation to scale to the accelerating pace of the biomedical literature while maintaining scientific quality. Since curators often require a wide variety of highly specific information, providing text-mining tools to fill each need may be a complex and never-ending task. However, text-mined outputs useful for curators are also likely to be useful for others in the broader scientific community; integrating text mining into *Europe PMC* therefore also opens the possibility for occasional users to contribute to community curation efforts and provide feedback on text-mining results.

Europe PMC is committed to enabling text mining. Currently, it provides features such as 'Highlight terms' to identify core biological entities such as genes/proteins and organisms within the article's abstract view; the entities are also linked to relevant databases. A similar feature is provided for full-text article views. *Europe PMC* is also being developed as a platform for third-party text-mining algorithms, allowing the output of these algorithms to be displayed in full text articles shown on the *Europe PMC* website.

In the future, outputs from the text-mining community could further semantically enrich *Europe PMC* content by including the annotation of additional entities, such as mutations, and/or relationships between various entities., such as genes/proteins and diseases. Search and browse features built on top of these annotations—e.g. references to other articles studying the same relationships (or, perhaps, contradictory relationships)—will help readers to better judge the article in light of related publications.

SourceData (Thomas Lemberger, EMBO and Ioannis Xenarios, SIB)

Hypothesis-driven research in molecular and cell biology primarily generates data from small-scale experiments. In scientific publications, such data are visually depicted in figures or tables. However the original data behind the figures—the 'source data'—are almost never available in structured format that would make them findable and reusable.

SourceData (http://sourcedata.embo.org/) (18) is building tools to allow researchers and publishers to generate machine-readable descriptions of data during the publication process and also to make this data searchable. To facilitate generating structured experimental descriptions, *SourceData* has developed an online tool for computer-assisted manual curation of figures and figure legends by data editors. The intention is to integrate a curation step into the publishing workflow to annotate figures of before article publication. Authors then verify and approve the curated information through a validation interface. The result is a machine-readable representation of the data (descriptive metadata) based on the information routinely provided by authors in the text of figure legends, thus respecting the traditional workflow adopted by scientists.

SourceData have also developed a search interface that allows users to search for specific experimental evidence and the articles where these data have been reported. This search function is incorporated into the '*SourceData SmartFigure*' viewer, which can easily be embedded in online publications. The *SmartFigure* application allows a specific figure panel to be linked with figures presenting similar data published elsewhere and therefore makes it possible for users to traverse the web of connected data by following these links across articles. Finally, programmatic access to the *SourceData* database is provided to the research community through a public API.

Integration of text mining with manual curation in the context of publishing seems to be a promising direction, as it will improve the efficiency and speed of the metadata extraction process and it will allow supervision of the automated results by both data editors and authors. In this context, text-mining methods will be useful for the automated semantic enrichment of figure legends or of the corresponding referring statements in the full text and also for identifying entity relationships that represent tested experimental hypotheses. Text mining is also envisioned to play a complementary role by linking curated figures with interpretative statements made in the article or with reagents listed in 'Materials and Methods' section. Finally, text-mining techniques developed for computer science publications (19, 20) might be useful to automatically prioritize a pool of candidate publications for further extraction of detailed experimental data and metadata.

OpenBEL: computable knowledge bases of cause–effect relationships (Natalie Catlett, Selventa)

Biological Expression Language (BEL) is a knowledge representation developed by Selventa to capture biological cause-and-effect relationships from the scientific literature in a format suitable for computation. BEL and its associated software platform are an open source project (www.openbel.org). BEL knowledge bases have been used to support inference from molecular profiling data (21–23) and to construct of network models representing specific biological processes (24). These approaches support precision medicine by illuminating the molecular mechanisms of disease, drug mechanisms of action, and supporting patient stratification.

BEL is designed to represent experimental observations in molecular biology, providing specific representations of various biological measurements including RNAs, proteins, post-translationally modified proteins, and protein activities, as well as biological processes and pathologies. This granular representation facilitates mapping of biological measurements to BEL networks to drive interpretation of molecular profiling data. BEL also represents the context for these experimental observations, such as the cell line or tissue used for the experiment, as well as a literature citation, allowing the creation of BEL networks that accurately represent the experiment and its context.

Over the last decade, Selventa has built a knowledge base comprised of >500 000 BEL statements primarily through manual curation. Many of these statements resulted from targeted curation efforts to support projects in various disease areas. This approach requires a significant effort from trained scientists to build a comprehensive knowledge base and keep it current.

Text mining promises to greatly improve the efficiency of building BEL knowledge bases. Accurate entity identification from the literature is critical to generating BEL knowledge bases useful for inference or building models. Another computational aspect important for automation is relation identification. Recently, Fluck and colleagues developed BELIEF, a text-mining work flow to improve the efficiency of BEL curation (25). BELIEF includes a UIMA-based text-mining workflow (with several state-of-the-art natural language processing, named entity recognition (NER) and relationship extraction tools) to facilitate a semi-automatic curation pipeline. Use of BELIEF was shown to significantly reduce human curation effort.

VIROME and building a knowledge base for microbiomes (Shawn Polson, University of Delaware)

Microbial communities and viral assemblages have been found to be both numerous and important drivers of biological processes globally. Recent research has linked microbiomes, microbes co-existing with a host, to many normal and pathological processes such as co-metabolism of food sources, exclusion of pathogens, fostering of host immune response, obesity, susceptibility to cancer and even mental disorders (26–31). Research aimed at unraveling the complex community-scale dynamics and functions of microbial communities, and the even more numerous viruses which play important roles in regulating and driving genetic diversity among them, are of paramount importance. Our ability to examine these systems was once limited by factors including the inability to cultivate the vast majority in a laboratory setting, but the advent of increasingly cost-effective platforms for deep sequencing of marker genes (e.g. 16S rRNA) and metagenomes in the past several years have finally opened the door for wide-spread research in this field.

These methods involve generation of raw sequence data elucidating the taxonomic or functional composition of the community at a specific geographic location, time, and environmental condition. Typically a study will include multiple samples varying across some spatial, temporal, or environmental variable allowing for testing of one or more specific hypotheses. The global nature of such data, however, means that its utility could extend far beyond the specific hypotheses it was collected to address. The results of such studies are typically published in peer-reviewed journals with deposition of only the raw data to public repositories such as the NCBI Sequence read archive. Other fields have seen the utility of publishing the analysed results of sequence-based studies (e.g. GEO for gene expression data). Some online tools such as VIROME (32) and MG-RAST (33) do provide a route for the analysis results themselves to be made public. Recent work by VIROME (http://virome.dbi.udel.edu/) and others are working to ensure that such results are accompanied by standardized metadata to make them useful when considered in alternate contexts, but the ability to look for trends across projects remains very limited. Leveraging microbial ecology results garnered from disparate projects could prove transformative for the field. Agreements are lacking, however, to populate centralized repositories with analysed data in a manner that would enable the creation of comprehensive microbial ecology resources, similar to what UniProt (1) and the Protein Information Resource (34), among others, provide for proteins. Development of such resources would enable large-scale observations and hypothesis testing, such as to assess the range of conditions under which a given microbe (or microbial protein) has been observed, thus providing key insights into its role, or assessing synergistic relationships by determining the consistency of co-occurrence for two or more microbes.

Text mining should play a key role in future microbiome studies by providing standalone tools to search for specific microbial relationships in the literature and populating databases designed to provide comprehensive views of such global data. Microbes almost always live in mixed communities, and thus cooperation and competition are key features; however, detecting such microbial dependencies is difficult and time-consuming. Similarly, defining the environmental parameters under which certain microbes or guilds of microbes exist can be very informative in understanding their roles. Single studies are rarely comprehensive enough to elucidate such trends, however. Text-mining tools may enable a comprehensive understanding of microbiomes by focusing on the NER of specific microbial entities, the extraction of biological conclusions (e.g. organism x can do y, but only in the presence of z), metadata extraction (description of time, place, and conditions of samples at collection), and methodological details of the original sample. The ENVIRONMENTS and EXTRACT tools presented at BioCreative V (35, 36), are examples of such tools, with emerging capability to extract environmental context and microbial taxonomy from published articles and map them to ontological standards such as Environmental Ontology (37).

Text-mining needs in large scale applications

The text-mining needs in aforementioned applications can be grouped into three primary tasks: NER, relation extraction (RE) and information visualization. NER involves automatically labeling bio-entities such as dataset name (SourceData), diseases, genes, proteins (BEL, Europe PMC) or microbial proteins (VIROME). Since NER is foundational to most text-mining applications, the availability of accurate application-specific NER tools is critical (38, 39). RE introduces the next higher level of knowledge discovery by automatically extracting relationships between the entities identified by NER. Such relationships may describe cause–effect relations (BEL) or microbe–environment relations (VIROME), and relationships may also involve metadata (such as spatio-temporal variables) to curate complex higher order relations. The final task is visualization of the text mined results. Some applications require visuals—summaries (or visual tags), links to other online databases (EuropePMC) and metadata highlight within text (SourceData)—to enhance knowledge representation. Text mining can help in selecting the most relevant outputs from large scale text-mined results, as not all text-mined outputs need be displayed even if they are correctly extracted. Although text-mining roles may be classified broadly into three tasks, the specific entities, relations and representation required for each application may be highly specific.

Challenges and opportunities in text mining

These domain applications above suggest several areas that remain challenging, namely 'accuracy', 'scalability', 'interoperability' and 'reusability'. These areas represent future opportunities for text mining to address the real world needs of large scale applications.

Accuracy

Although text-mining systems are rapidly transitioning to real world use, imperfect accuracy remains a limiting factor. Workflows incorporating text-mining systems must design processes that compensate for imperfect output. Although the importance of these considerations tapers as the output quality approaches that of human annotators, there are several limitations with the evaluations typically performed in the text-mining community. First, the evaluation most commonly performed is intrinsic, that is, it compares the output of the system to gold standard annotations performed by human annotators. Although such an evaluation provides several desirable properties, such as being quantifiable and providing a high degree of objectivity, it does miss some important considerations. Notably, it provides no feedback on whether the quality of the output is sufficient to support processes downstream in the workflow. Thus, while intrinsic evaluation of the system is important, the system must also be evaluated extrinsically, i.e. in place in the workflow.

Interoperability

Because system accuracy is critical and must be evaluated extrinsically in the workflow, each system evaluated must be fully integrated into the workflow. Thus, the difficulty in integrating the system must be kept to a minimum. Unfortunately, many factors reduce system interoperability, such as operating system dependencies and incompatibilities between input and output formats (40). Interoperability could be addressed in several different ways. For instance, UIMA (41) is a software architecture created by IBM in 2003 to provide uniform data formatting standards for different teams working on NLP projects. Although it uses a common analysis system (CAS), the ability to use different semantic tag sets creates an interoperability solution (42). Tools written in a system-independent language such as Java or Python do not require a specific operating system. Format incompatibilities can be addressed by creating a standard data format. The recent BioC project is such an example, which has created an interoperable data format that is both straightforward and sufficiently expressive to represent a wide variety of text-mining tasks (43, 44). Another solution may be web services, which hides all configuration and deployment details from the user by providing an API that can be accessed over the Internet, requiring no system installation or maintenance (45, 46). Despite these attempts, integrating text mining into mature database workflows remains difficult due to the complexities of curation workflow and existing infrastructure.

Scalability

A defining characteristic of large-scale text-mining applications is the requirement to scale to millions of documents. PubMed, e.g. contains over 25 million abstracts—at the relatively high rate of 100 abstracts per second, it therefore requires nearly 3 days of computational time to process; processing an equivalent amount of full text articles requires an order of magnitude longer. Text-mining implementations are therefore frequently paired with a database, allowing the text to be preprocessed and the results cached and indexed. Although this allows the text-mining results to be provided on demand for text available beforehand, this approach is insufficient for text that must be processed in real time. Moreover, this approach is also inconvenient for updates to the text-mining system, as all the cached results must be reprocessed. One approach to address scalability is the application of cluster computing: processing multiple documents in parallel on multiple hardware systems. Returning to our PubMed example, a cluster of 10 systems—each processing at the rate of 100 abstracts per second—is sufficient to reduce the processing time to under 7 h, a job which can be completed overnight.

Reusability

Text-mining systems are commonly applied to text somewhat different than the text used to train and evaluate them, making generalization—the ability to handle text previously unseen—very important. As an example, abstracts describing rare genetic diseases will contain significantly different information than those describing treatments for tropical infectious diseases, even though both will contain disease entities. A particular concern is the ability of the system to handle not only abstracts, but also full text documents (47–52). However, systems for dealing with many of the various nuances (such as figure captions, data in tables, information in supplementary materials, and various text cleaning issues) of full text are still not fully in place. Thus, a large improvement in the robustness of a system against shifts in the textual domain may be significantly more useful for real world applications than incremental improvements in system accuracy.

Future roles of researchers, publishers and curators

Bridging the gap between text-mining research and its application in real world databases requires a collaborative effort from the various stakeholders involved in advancing biomedical sciences. In this section, we provide a few perspectives which researchers, publishers and curators can use to advance biomedical sciences through text mining.

Research community

Community run challenges in biomedical text mining such as BioCreative can play a major role in realizing the potential of large scale text-mining applications, both by assessing the state of the art and also helping advance the field (53). The aim of conducting these challenges, in general, is to promote interdisciplinary collaboration, evaluate and advance the NLP techniques to facilitate biological research. Thus, these challenges are conducted as shared tasks where research teams from across the globe participate in fulfilling the goals of specified text-mining tasks. A myriad of such challenges have been organized over the years following the success of *CASP* in 1994 (54, 55) on protein structure prediction; Huang *et al.* (2016) (53) provides a comprehensive overview of several challenges conducted within the last decade.

In recent years, the community has introduced challenges that focus on bridging the gap between biomedical text-mining research and new application domains. For example, since 2010, BioCreative has organized workshops at the annual meetings of the International Society for Biocuration (http://biocuration.org/) with a focus on better understanding biocuration workflows (8) and promoting the development and deployment of biomedical text-mining tools into production curation pipelines. Several of these have been successfully integrated into existing curation workflows (e.g. 4, 13).

Nevertheless, there are several difficulties which must be resolved before community challenges can realize the potential of large scale text-mining applications. The foremost of these difficulties is that challenge tasks are often simplified or abstracted versions of the real-world problems. For example, although biocurators routinely use the full text of an article (56, 57), challenge tasks often only utilize the abstract due to difficulties in accessing full text articles and processing full text. A consequence of this simplification of the real-world problem is that even systems that perform well on challenge tasks yield significantly lower results when evaluated in practical real-world settings. For example, previous BioCreative Gene Normalization challenges have shown that the task performance dropped significantly when tested on full texts (58) instead of abstracts (59). These difficulties can be addressed by designing challenge tasks that focus on the unique problems presented by real world applications.

The BioCreative Collaborative Biocurator Assistant Task (BioC) and the BioCreative Interactive Text-Mining Task (IAT) serve as examples of such focused efforts. The BioC task centered on creating a text-mining system to support BioGRID curators by developing BioC-compatible text-mining modules complementing each other and integrated into one system. The IAT task involved biocurators in testing text-mining systems. In a similar vein, we describe below a few ideas that can be realized as challenge tasks in BioC workshops in the near term to help realize the opportunities of text-mining research in real-world applications more directly.

i. Creating a wide variety of manually curated benchmarks datasets for various text-mining problems. These benchmarks datasets are critical for text-mining researchers to train, test and compare their algorithms and also for organizations to determine the best fit for their large scale applications. These benchmarks should come from various sources including biomedical literature (both abstract and full text), clinical trials, clinical notes and Electronic Medical Records.

ii. Identifying metrics to measure critical system qualities in addition to accuracy. As application needs differ, so do their evaluation criteria for selecting text-mining tools. Identifying or creating metrics addressing performance aspects beyond accuracy, such as scalability, usability, and cost-of-adoption (such as database management and front-end design) will greatly help both researchers and application developers to identify text-mining tools that best fit their performance dimensions. In this direction, BioCreative-IAT task has included both performance and usability metrics in the evaluation of the text-mining systems by curators, which were also adopted in the BioC task. These metrics should be extended to include scalability and cost-of-adoption.

iii. Like BioC's focus on BioGRID, challenge tasks can be designed to focus on individual large scale applications such as SourceData, BEL and VIROME. Involving the data indexers and curators in the task design step will enrich the utility of the challenge task for real-world use. Parameters such as evaluation criteria can be designed specifically for the individual application. Moreover, the data bottleneck such as full text access and processing can be addressed with help of literature services such as Europe PMC.

Publishers' role

The SourceData project provides a good example of how publishers could actively encourage innovative knowledge curation and representation. As described in the SourceData section, the publishers collaborate with researchers to generate machine-readable descriptions of datasets during the publication process and also to make this data searchable. In addition to the role of text-mining expressed earlier, as the databases grow, text-mining systems can be employed in the future to provide automatic recommendations of machine-readable tags or descriptions for the datasets. Similar to SourceData project's initiative to enrich articles during in-publication or pre-publications phase, the publishers' role can be to enrich articles in pre-publication phase by employing text-mining systems.

In the future, the curation step may not wait until after publication, as is the current practice. A possibility is to move the curation step 'upstream' i.e. capturing knowledge at the time of peer review and prior to publication. Such an initiative would require development of very high quality and sustainable text-mining systems, and possibly require a greater involvement of the article authors in validating some of the text-mined results.

Curators' role

It is central to keep the human curators/experts in the loop in any newly proposed text-mining-based curation ecosystem. Curators are critical for defining text-mining requirements, providing annotation guidelines and standards, and providing training data for the initial system development and evaluation. Curators should be involved in evaluating the text-mined results and decide their fitness for curation. Curators should help system developers iteratively improve the text-mining algorithms and make any necessary system customizations for their specific database curation needs. This would be the ideal way to incorporate text mining into curation workflows.

Conclusions

In this work, we presented four large scale applications of text mining in the biological and life sciences, as showcased during a recent panel at BioCreative V. We used these applications as case studies in the challenges encountered in adopting text-mining solutions into realistic tasks and discussed several areas of opportunity for text mining to support real world services in the near term. Finally, we presented a few actionable steps that the BioCreative community can take to bridge the gap between text-mining research and real world biomedical services.

Funding

This work was supported by the National Institutes of Health [R13-GM109648-01A1 to CNA, P20-GM103446 to SWP, Intramural Research Program at National Library of Medicine to A.S., R.L., Z.L.], the US Department of Energy [DE-SC0010838 to C.N.A.], the US National Science Foundation [DBI-1356374 to S.W.P.] for the VIROME project and the Swiss Federal Government through the State Secretariat for Education, Research and Innovation (SERI); SyBIT project of the SystemsX.ch, the Swiss Initiative in Systems Biology (in part) (to IX). The Robert Bosch Foundation and EMBO are acknowledged for funding of the SourceData project. The Wellcome Grant, UK PubMed Central Phase 3 Developments 2012–2016 (grant number 098231/Z/12/Z) for the EuropePMC project.

Acknowledgement

We also acknowledge the BioCreative steering committee http://www.biocreative.org

Conflict of interest. None declared.

References

1. UniProt,C. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, 43, D204–D212.
2. NCBI Resource Coordinators (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 43, D6–D17.
3. Baxevanis,A.D., and Bateman,A. (2006) The importance of biological databases in biological discovery. *Curr. Protoc. Bioinformatics*, **50**, 1.1.1.–1.1.8.
4. Wei,C.H., Kao,H.Y., and Lu,Z. (2013) PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.*, 41, W518–W522.
5. The Europe PMC Consortium (2015) Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Res.*, 43, D1042–D1048.
6. Lemberger,T. (2014) Tools of discovery. *Mol. Syst. Biol.*, 10, 715.
7. Hirschman,L., Burns,G.A., Krallinger,M. *et al.* (2012) Text mining for the biocuration workflow. *Database (Oxford)*, 2012, bas020.
8. Lu,Z., and Hirschman,L. (2012) Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. *Database (Oxford)*, 2012, bas043.
9. Van Auken,K., Jaffery,J., Chan,J. *et al.* (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. *BMC Bioinformatics*, 10, 228.
10. Van Auken,K., Fey,P., Berardini,T.Z. *et al.* (2012) Text mining in the biocuration workflow: applications for literature curation at WormBase, dictyBase and TAIR. *Database (Oxford)*, 2012, bas040.
11. Wei,C.H., Harris,B.R., Li,D. *et al.* (2012) Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database (Oxford)*, 2012, bas041.
12. Tudor,C.O., Arighi,C.N., Wang,Q. *et al.* (2012) The eFIP system for text mining of protein interaction networks of phosphorylated proteins. *Database (Oxford)*, 2012, bas044.
13. Cejuela,J.M., McQuilton,P., Ponting,L. *et al.* (2014) tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database (Oxford)*, 2014, bau033.
14. Arighi,C.N., Roberts,P.M., Agarwal,S. *et al.* (2011) BioCreative III interactive task: an overview. *BMC Bioinformatics*, 12, S4.
15. Arighi,C.N., Carterette,B., Cohen,K.B. *et al.* (2013) An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. *Database (Oxford)*, 2013, bas056.
16. Matis-Mitchell,S., Roberts,P., Tudor, C.O., and Arighi,C.N. (2013) BioCreative IV interactive task. *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, vol 1. Bethesda, MD, pp. 190–203.
17. Wang,Q.H., Abdul,S.S., Almeida,L. *et al.* (2016) Overview of the interactive task in BioCreative V. *Database* (Oxford), 2016, baw119.
18. Liechti,R., George,N., El-Gebali,S. *et al.* (2016) SourceData - a semantic platform for curating and searching figures. *bioRxiv*, 058529.
19. Singhal,A., Kasturi,R., Sivakumar,V. *et al.* (2013) Leveraging web intelligence for finding interesting research datasets. *Proceedings of the International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). 2013 IEEE/WIC/ACM. IEEE*, vol. 1, pp. 321–328.

20. Singhal,A. (2014) *Leveraging Open Source Web Resources to Improve Retrieval of Low Text Content Items*. Department of Computer science. University of Minnesota, Minneapolis, MN, 144.
21. Martin,F., Thomson,T.M., Sewer,A. *et al.* (2012) Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks. *BMC Syst. Biol.*, 6, 54.
22. Thomson,T.M., Sewer,A., Martin,F. *et al.* (2013) Quantitative assessment of biological impact using transcriptomic data and mechanistic network models. *Toxicol. Appl. Pharmacol.*, 272, 863–878.
23. Catlett,N.L., Bargnesi,A.J., Ungerer,S. *et al.* (2013) Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC Bioinformatics*, 14, 340.
24. Boue,S., Talikka,M., Westra,J.W. *et al.* (2015) Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Database (Oxford)*, 2015, bav030.
25. Fluck,J., Madan,S., Ansari,S. *et al.* (2014) *BELIEF-a Semiautomatic Workflow for BEL Network Creation. Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (SMBM)*, University of Aveiro, Portugal, 109–113.
26. Gomez,A., Rothman,J.M., Petrzelkova,K. *et al.* (2016) Temporal variation selects for diet-microbe co-metabolic traits in the gut of Gorilla spp. *ISME J.*, 10, 532.
27. Sampson,T.R., and Mazmanian,S.K. (2015) Control of brain development, function, and behavior by the microbiome. *Cell Host Microbe*, 17, 565–576.
28. Neish,A.S. (2014) Mucosal immunity and the microbiome. *Ann. Am. Thoracic Soc.*, 11, S28–S32.
29. McDermott,A.J., and Huffnagle,G.B. (2014) The microbiome and regulation of mucosal immunity. *Immunology*, 142, 24–31.
30. Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, 486, 207–214.
31. Schulz,M.D., Atay,C., Heringer,J. *et al.* (2014) High-fat-diet-mediated dysbiosis promotes intestinal carcinogenesis independently of obesity. *Nature*, 514, 508–512.
32. Wommack,K.E., Bhavsar,J., Polson,S.W. *et al.* (2012) VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand. Genomic Sci.*, 6, 421–433.
33. Glass,E.M., Wilkening,J., Wilke,A. *et al.* (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb. Protoc.*, 2010, pdb. prot5368.
34. Wu,C.H., Huang,H., Arminski,L. *et al.* (2002) The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res.*, 30, 35–37.
35. Pafilis,E., Buttigieg,P.L., Ferrell,B. *et al.* (2016) EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation. *Database (Oxford)*, 2016.
36. Pafilis,E., Frankild,S.P., Schnetzer,J. *et al.* (2015) ENVIRONMENTS and EOL: identification of Environment Ontology terms in text and the annotation of the Encyclopedia of Life. *Bioinformatics*, 31, 1872–1874.
37. Buttigieg,P.L., Morrison,N., Smith,B. *et al.* (2013) The environment ontology: contextualising biological and biomedical entities. *J. Biomed. Semant.*, 4, 43.
38. Leaman,R., and Lu,Z. (2016) TaggerOne: joint named entity recognition and normalization with semi-Markov models. *Bioinformatics*, 32, 2839–2846.
39. Wei,C.H., Kao,H.Y., and Lu,Z. (2012) SR4GN: a species recognition software tool for gene normalization. *PloS ONE*, 7, e38460.
40. Wiegers,T.C., Davis,A.P., and Mattingly,C.J. (2014) Web services-based text-mining demonstrates broad impacts for interoperability and process simplification. *Database (Oxford)*, 2014, bau050.
41. Ferrucci,D., and Lally,A. (2004) UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10, 327–348.
42. Verspoor,K., Baumgartner, W., Jr, Roeder, C. *et al.* (2009) Abstracting the types away from a UIMA type system. In C. Chiarcos, R. Eckhart de Castilho, M. Stede (Eds.), Form to Meaning: Processing Texts Automatically, pp. 249–256.
43. Comeau,D.C., Islamaj Dogan,R., Ciccarese,P. *et al.* (2013) BioC: a minimalist approach to interoperability for biomedical text processing. *Database (Oxford)*, 2013, bat064.
44. Comeau,D.C., Batista-Navarro,R.T., Dai,H.J. *et al.* (2014) BioC interoperability track overview. *Database (Oxford)*, 2014, bau053.
45. Pafilis,E., O'donoghue,S.I., Jensen,L.J. *et al.* (2009) Reflect: augmented browsing for the life scientist. *Nat. Biotechnol.*, 27, 508–510.
46. Wei,C.H., Leaman,R., and Lu,Z. (2016) Beyond accuracy: creating interoperable and scalable text-mining web services. *Bioinformatics*, 32, 1907–1910.
47. Gerner,M., Nenadic,G., and Bergman,C.M. (2010) LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinformatics*, 11, 85.
48. Gerner,M., Sarafraz,F., Bergman,C.M. *et al.* (2012) BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. *Bioinformatics*, 28, 2154–2161.
49. Sun,Z., Errami,M., Long,T. *et al.* (2010) Systematic characterizations of text similarity in full text biomedical publications. *PloS ONE*, 5, e12704.
50. Thomas,P., Starlinger,J., Vowinkel,A. *et al.* (2012) GeneView: a comprehensive semantic search engine for PubMed. *Nucleic Acids Res.*, 40, W585–W591.
51. Caporaso,J.G., Deshpande,N., Fink,J.L. *et al.* (2008) Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. *Pac Symp Biocomput*, 640–651.
52. Cohen,K.B., Johnson,H.L., Verspoor,K. *et al.* (2010) The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11, 492.
53. Huang,C.C., and Lu,Z. (2016) Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief. Bioinformatics*, 17, 132–144.
54. Moult,J., Pedersen,J.T., Judson,R. *et al.* (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins*, 23, ii–iv.
55. Moult,J., Fidelis,K., Kryshtafovych,A. *et al.* (2014) Critical assessment of methods of protein structure prediction (CASP)–round x. *Proteins*, 82, 1–6.

56. Mao,Y., Van Auken,K., Li,D. *et al*. (2014) Overview of the gene ontology task at BioCreative IV. *Database (Oxford)*, 2014, bau086.
57. Huang,M., Neveol,A., and Lu,Z. (2011) Recommending MeSH terms for annotating biomedical articles. *J. Am. Med. Inform. Assoc*., 18, 660–667.
58. Lu,Z., Kao,H.Y., Wei,C.H. *et al*. (2011) The gene normalization task in BioCreative III. *BMC Bioinformatics*, 12, S2.
59. Morgan,A.A., Lu,Z., Wang,X. *et al*. (2008) Overview of BioCreative II gene normalization. *Genome Biol*., 9, S3.