Tagging Assistant for Scientific Articles

Zara Nasar, Syed Waqar Jaffry and Muhammad Kamran Malik

Artificial Intelligence and Multidisciplinary Research Lab, Punjab University College of Information Technology, University of the Punjab, Lahore-54000, Pakistan
zara.nasar@pucit.edu.pk

Abstract. With the advent of World Wide Web (WWW), world is being overloaded with huge data. This huge data carries potential information that once extracted, can be used for betterment of humanity. Information from this data can be extracted using manual and automatic analysis. Manual analysis is not scalable and efficient, whereas, the automatic analysis involves computing mechanisms that aid in automatic information extraction over huge amount of data. WWW has also affected overall growth in scientific literature that makes the process of literature review quite laborious, time consuming and cumbersome job for researchers. Hence a dire need is felt to automatically extract potential information out of immense set of scientific articles in order to automate the process of literature review. Such service would require machine learning models to train. Whereas, such model in turn require training dataset. To construct a quality dataset often involves employment of annotation tools. There exist wide variety of annotation tools, but none are tailored to assist annotation of scientific articles. Hence in this study, web-based annotation tool for scientific articles is developed using Python language. The developed assistant employs state of the art machine learning models to extract metadata from scientific articles as well as to process article's text. It provides various filters in order to assist annotators. An article is divided into various textual constructs including sections, paragraphs, sentences, tokens and lemmas. This division can help annotators by addressing their information need in an efficient manner. Hence, this annotation tool can significantly reduce time while preparing dataset for full-text scientific articles.

Keywords: Metadata, Key-insights, Information Extraction, Annotation Assistant, Tagging Tool, Research Articles, Scientific Literature.

1 Introduction

In last few decades, advent of computers and later World Wide Web (WWW) have changed human civilization dramatically. Now we live in the world which is being overloaded with the data and the information. This information overload is posing new challenges to human intellect and hence creating opportunities for innovation. The WWW has resulted into rapid growth of scientific literature. A research study presented in [1], concludes that amount of scientific articles tend to doubles every ten to fifteen years. There are other studies as well that have compiled the stats regarding

published scientific articles in 2016 only, and number goes around 2.2. million [2], [3].

This enormous increase in scientific content poses significant challenges for the researchers who want to determine state of art in their respective field of interest. As literature review involves literate acquisition, its pruning followed by reading of filtered articles and finally consolidation of findings. Hence, due to almost exponential growth of this data, the process of literature review becomes very time consuming, laborious and cumbersome. At the same time, this whole process of performing systematic literature review is of utmost importance for researchers to identify research gaps in existing literature. According to one of the systematic literature review guideline, time require to conduct a quality review can take up to one year [4]. Another study points that systematic literature review can take up to 186 weeks with single/multiple human resources [5].

To provide researchers with assistance during literature acquisition, many research organizations and scientific publishers such as ACM, IEEE and Springer etc. have provided digital research repositories. These libraries tend to offer search filters that provide ease to users while querying through millions of research articles. These digital research repositories employ metadata information from scientific articles in order to provide various searching facilities. Hence, metadata extraction from scientific articles eventually helps in saving researcher's time while performing literature acquisition. In order to perform literature review, next step is to read and consolidate findings from acquired literature. This step requires to go through bulk of scientific articles in order to determine the state-of-the-art in a specific domain of interest. From a researcher's point of view, this whole process is of utmost importance but time-consuming, laborious and cumbersome.

In the light of above points, it is evident that study of research papers by means of automated analysis will eventually aid researchers. Pertinent question in this regard is that how potential information from scientific articles can be automatically extracted. In order to address this and related problems, a whole domain named Information Extraction (IE) is dedicated for extraction of potential information nuggets from data. The IE is majorly focused on extraction of structured data from unstructured or semi-structured data. It is being widely used across multiple domains, for example, in the domain of medical sciences, IE is applied in order to extract information about patient's information, their previous medical history, causes and respective cures [6]. The domain of IE is comprised of concepts and techniques of Machine Learning, Natural Language Processing (NLP), Text Mining (TM) and Information Retrieval (IR).

As far as IE application on scientific articles is concerned, progress is limited. The main reason is unavailability of benchmark datasets. For any IE problem, dataset is critical. An article consists of various sections; its metadata or header, full-body text and references section. Metadata usually include title, authors, affiliations, venue, date and abstract of a scientific article. Full-text refers to the whole text part of scientific article from abstract till conclusion. References refer to bibliography section and it is either included in metadata or dealt separately in literature.

Each of these can be used to make IE from scientific articles more beneficial to community. Metadata Extraction is being widely studied in literature with pioneer

studies dating back to 1999 [7]. Reference parsing, also known as citation metadata extraction, is also studied in literature and work is going on after a comprehensive dataset is made publicly available [8]. Both these problems had their initial benchmark datasets created in early 2000s as part of CORA project [7], [9]. Full-text processing on other hand, is still in preliminary phases. All these previous advancements, in metadata extraction of reference processing, adopt the approach of Named Entity Recognition (NER) to extract phrases and assign rhetorical categories to them.

In case of full-length scientific articles, prior advancements are made by ART project [10], [11], but the project focused on sentence level classification in various rhetorical categories such as background, method, result, conclusion etc. First attempt to extract domain, techniques from scientific articles' abstracts was made in 2011 [12], where the technique relies on rules and bootstrapping approaches. Recently, several contributions are made that are focused on annotation of scientific articles' abstracts [13], [14]. Most of these researches rely on annotation tool in order to annotate the data. The annotation tool makes the task of annotation easier by providing automatic way of tag assignment that result in human error reduction and is time efficient.

2 Existing Tools

In the light of literature regarding annotation tools, BRAT [15] is the widely used open-source tool for annotation of IE-oriented problems. Many datasets have been prepared using this tool. It provides a great UI interface with many features including collaboration, comparison of annotations among annotators etc. Moreover, BRAT tends to convert the input into sentences, and later provide support to annotate phrases as sentences. On top of it, it further gives facility to annotate relationships between entities. In order to set it up, very minor configurations are required. Primary weakness of BRAT with respect to scientific articles annotation is its incapability to process PDF processing or complex text, as it requires plain text as input. Therefore, this solution cannot be used to annotate scientific articles. Another open-source annotation solution include Callisto [16] which provides great linguistic support but also supports plain text only and requires configurations that are cumbersome in comparison to other available solutions.

Adobe Acrobat itself offers primitive highlighting and notes as well as commenting support, but manually inputting the respective tagged information and later compiling this annotation information requires a lot of manual effort and is prone to human errors. Additionally, Acrobat is a proprietary solution and hence does not enable automatic extraction of highlighted snippets, comments, tags and notes. Hence, it also does not serve the purpose for annotating scientific articles. Recently OpenCalais [17], a project by Thomson Reuters has also started services for annotation. It offers demo version as well as API support. It does not offer features to extend annotation markers though. Currently, it is focused on extracting general entities from PDF documents as well as from plain text. Table 1 presents overall attributes of various annotation tools that are being used in literature.

In following table, type refers to the medium that is provided to consume services. Input refers to the input format that are being supported by the respective solutions. Export refers to the availability of export feature that will assist in exporting user annotations. Vis refers to the availability of visualization support while making annotations. Free refers that solution is available either free-of-cost or under a subscription fee. OS field refers that either respective tool is open-source. Ease refers to the ease of usage of respective solution i.e. how many configurations or prior knowledge are required to employ the respective tool. Lastly, Ext points towards the availability of the solution to provide support to extend annotation markers set. In other word, it means that if a solution provides means to customize annotation markers or not.

Table 1. Summary of Existing Tools

Tool Name	Type	Input	Export	Vis	Free	OS	Ease	Ext
BRAT	Web	Text	Yes	Yes	Yes	Yes	High	Yes
A. Acrobat	Desktop	PDF	No	Yes	No	No	Med	No
Mendely	Desktop	PDF	No	Yes	Yes*	No	Med	No
Callisto	Desktop	Text	Yes	Yes	Yes	No	Low	Yes
OpenClais	Web/API	PDF/Text	No	Yes	Yes*	No	High	No

* refers that upgraded solutions are available against subscription fees

Thus, it is evident that researchers tend to use annotation tools for quality data preparation and effective time-utilization. So far, there exist no open-source annotation tool for scientific articles that provides options to process PDF along with customized annotation markers. Hence, in this study a primitive annotation tool for scientific articles is developed that takes scientific article into PDF format and later convert PDF to text. After that, metadata extraction and citation metadata extraction is performed by means of state of the art solutions. Section 3 briefly explains the major use-cases developed in the study along with major use-cases. Section 4 concludes the study followed by discussion of future prospects and advantages of developed tool. Last section compiles the related bibliography.

3 Tool Development

This section briefly explains the major use cases provided for the annotation of scientific articles. It briefly explains the User Interface developed to assist the annotators while performing data annotation. Currently, the tool supports articles in PDF format only, as PDF is the most widely used format for scientific research dissemination. The current tool is developed as Web application in Python using Django server, JavaScript and JQuery for scripting purposes and SQLite DB for storage. It employs Python NLTK for text processing. The whole system is developed on Linux operating system.

3.1 Use Cases

The first version of developed tool provides facilities to upload articles. After the user uploads an article, next step is to convert the document into text. PDF to text conversion is rather tricky due to various styling and formatting variation across scientific articles Thus, various tools were employed to carry out the task including PDFBox, PDFToText, TextSharp, AbbyReader etc. These tools parsed single formatted article just fine, whereas in double column format, output was most of the time not usable due to text disorientation making sentence meaning incomplete. Hence, a comprehensive search for various scientific artilces processing tools was carried out. During this course, many tools were discovered including Parsict, Docear, GROBID and CERMINE [18]–[21]. Out of these, GROBID was selected due to its wide usage across various research platforms and on-going development. In addition to converting scientific articles to PDF, GROBID also extracts primitive metadata information and citations as well. GROBID tends to covert PDF document into XML using Text Encoding Initiative format.

By means of parsing this format, metadata and citation information is separated from the document. The remaining text carries broken passages. Hence, by means of various heuristics and language processing techniques, this text is compiled and further classified into sections, paragraphs, sentences and tokens using natural language toolkit. Furthermore, in the light of primitive survey conducted on a national graduate symposium, search filters are provided. These filters enable annotators to search through the various occurrences of terms across a document. All this textual processing is carried out using Python Natural Language Toolkit (NLTK).

After processing of textual content, next major assistance provided is regarding annotation markers. Annotators can add as many annotation markers as they can, with an option to provide distinct color against any marker. This color selection further aids annotators while performing data annotation, as it tends to highlight the annotated text with the color associated with respective annotation marker. This feature tends to provide visual assistance to the annotators regarding annotations made so far.

In addition to visually aiding the annotators, all the annotated texts are visible by means of a drop-down option. These annotations can be searched and deleted by the annotators to provide ease in case, that an annotation is made by mistake. Lastly, annotators can export the annotations into variety of formats including ann format that is being used by BRAT, IOB format: that is widely used format for various NER problems and XML format: that was adopted by pioneer search study regarding entity extraction from scientific articles [12]. Exporting options are available against a single article as well as against whole set of annotated articles. Major use-cases of the developed tool are listed in Table 2.

Table 2. Major Use-cases against developed tool

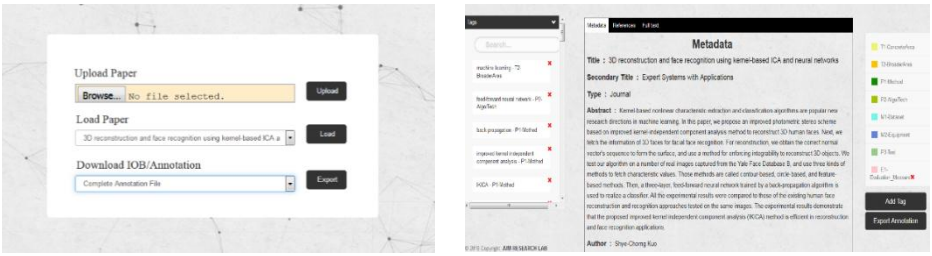
No.	Name	Description	Success
1	Upload PDF	In this module user will upload the research paper in the PDF format.	PDF is uploaded successfully and user is redirected to document details.

2	Load Existing Article	User can select any article that was processed before from the drop-down.	Document loaded successfully, and user is redirected to document details.
3	Download All Annotations	User can download all annotations which are made up till now.	Data downloaded in archive format.
4	View Meta Data	User can view meta data of uploaded PDF/ existing article.	Metadata of the research paper is displayed on the frontend.
5	View References	User can view references against uploaded PDF/ existing article.	References of the research paper are displayed on the frontend.
6	View Full Text	User can view the full-text of article by selecting respective option from drop-down menu.	All plain text is displayed on the frontend.
7	View Sections	User can view the sections of article by selecting respective option from drop-down menu.	All sections are displayed in left side-bar on the frontend.
8	View Sentences	User can view the sentences of article by selecting respective option from drop-down menu.	All sentences are displayed in left side-bar on the frontend.
9	View Lemma	User can view the lemmas of article by selecting respective option from drop-down menu.	All lemmas are displayed in left side-bar on the frontend.
10	Click Side-Bar Entry For Details	User can click any side-bar entry to view respective details e.g. if lemmas are selected using drop-down menu, on click upon any individual lemma entry will result in detailed view of its respective occurrence information in various sentences.	Respective information is loaded on front-end.
11	View Tags	User can view all tags previously stored are displayed on front-end by default.	Tags are displayed on the frontend.
12	Delete Tag	User can delete an annotation marker/tag, only if there is no annotation text associated with it.	Tag is deleted successfully.
13	Add Tag	User can add more tags as per the need.	Tag is added successfully.
14	Select Tag	User can select an annotation marker from the list of all available markers so far. It is the first step in making annotation.	Respective tag is selected and shown on the frontend.
15	Perform Tagging	User can select a text span in order to assign it the selected tag.	Respective text span is annotated and highlighted in the

			frontend. It is added in the list of annotations made so far as well.
16	View Annotations	User can view all annotations previously made.	All annotations are displayed on the frontend.
17	Delete Tagged Text	User can delete any previously made annotation by deleting it from the annotations list.	Annotation text and its respective highlighting information is removed.
18	Search Annotations	User can search through the annotation list with a phrase.	All the annotations carrying input phrase will be filtered and shown to the user.
19	Export Annotation	User can export all annotations with multiple facilities and options.	Export file is created and downloaded.

3.2 User Interface

This section briefly explains the user interface developed against the annotation tool for scientific articles. Fig. 1(a) presents the main screen of the developed web-tagging tool. It offers three major operations that cover the use-cases numbered 1, 2 and 3. After fulfilling either use case 1 or use case 2, the major working screen of annotation tool is shown in Fig. 1(b). This screen has three major structures: left side-bar, middle panel and right side-bar.



(a) (b)

Fig. 1. Major screens of developed tool (a) Start/Main Screen, (b) Working Screen

The left side-bar is used in order to provide various filters on text including full-text, sections, lemmas and annotations (tags). Left side-bar is used to cover the use-cases numbered 6-9 and 16 as shown in Fig. 2. Here if an entity from side-bar is clicked, respective information is shown in middle portion as shown in Fig. 2 (c). Similarly, by clicking the entry from section, respective section is loaded in the full-text tab explained in next passage.

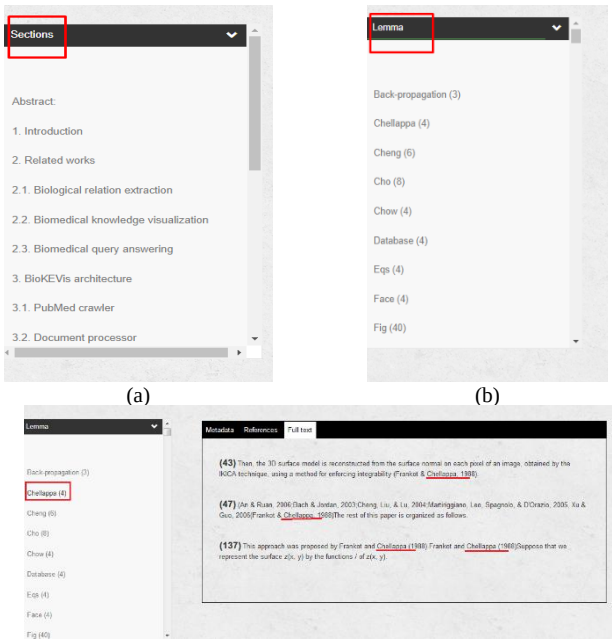


Fig. 2. Left side bar functions (a) Sections (b) Lemmas (c) Sentences carrying a lemma

Middle portion consist of a tabbing interface with three tabs where each tab is focused on one major aspect of a scientific article. First tab presents metadata information of the article that is loaded via main screen. Second tab presents the citation information contained in the loaded article while third tab carries the full-text of respective scientific article as shown in Fig. 3.

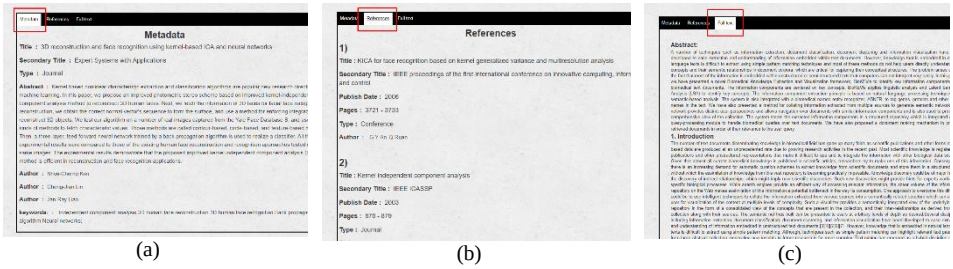
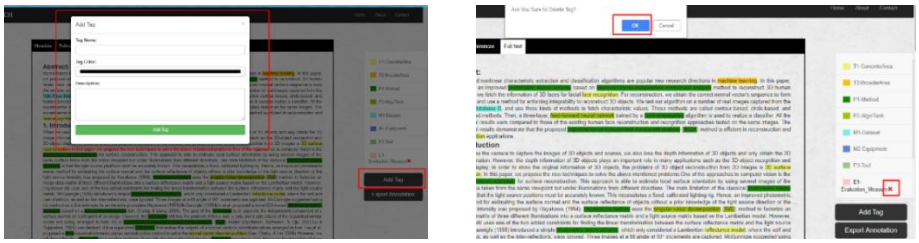


Fig. 3. View of tabbing interface (a) Metadata View (b) References view (c) Full-text view

Right-side bar consists of options related to annotation-markers management. By default, it shows all annotation markers that are used in data annotation previously. It includes addition of a new maker, deletion of maker and selection of a marker for annotation purpose thus covers the use-cases numbered 11-14. Insertion option requires name of annotation marker, its respective color that will be used for highlighting and its brief description. Respective screens are being presented in Fig. 4.

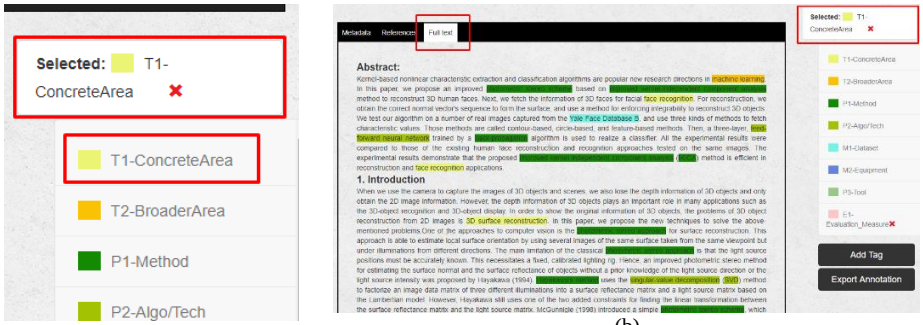


(a) (b)

Fig. 4. Function of Right-side (a) Addition of new annotation marker (b) Deletion of existing marker

Currently, the tool does not let annotator delete any annotation maker that is used in annotations made so far. This is done as a precautionary measure to avoid any data loss. Hence, deletion operation for an annotation marker requires that no previous data is annotated using respective maker.

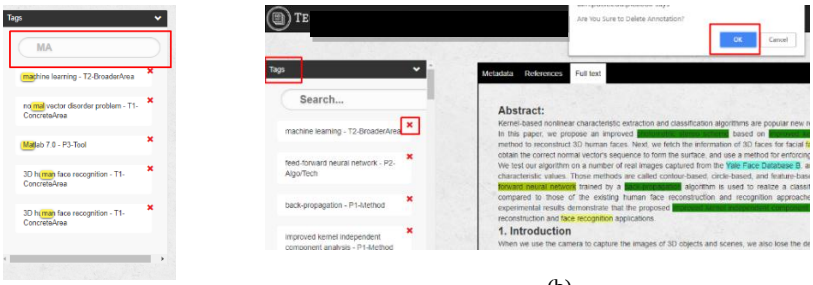
In order to perform annotation on text; two operations are required. First one is selection of annotation marker that is to be applied. After selecting the appropriate marker using right-tag screen, next step is to select the text span from middle screen that is going to be annotated. By selecting text, selected annotation marker is applied and respective background text of selected text span is also highlighted with respective annotation marker color. Operations against annotation making are presented in Fig. 5.



(a) (b)

Fig. 5. Performing annotation, (a) Select the annotation marker (b) Select the text

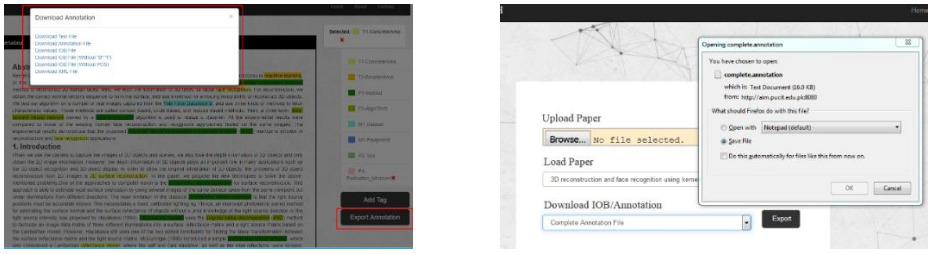
Whenever an annotation is made, it can be also viewed using drop-down menu available in left-side bar. Using this menu; two annotations related operations can be performed that include searching amongst annotation and deletion of an annotation as shown in Fig. 6. When annotation is done, annotator can simply download the file using export options by either downloading annotation against current loaded document from major working screen or by means of exporting bulk-annotations using main screen as presented in Fig. 7.



(a) (b)

Fig. 6. Annotation-related operations (a) Searching annotation using free-text (b) Deleting an annotation

Hence, the developed tool provides various features to reduce the time required in annotation of scientific articles. Various textual constructs based filters including sections, paragraphs, sentences and lemmas are provided to provide effective means during annotation. In addition, users can easily add annotation markers using UI without dealing with configurations files as required in other systems.



(a) (b)

Fig. 7. Exporting annotations (a) Exporting option from working screen (b) Exporting option from main screen

In view of developments so far, addition of various filters, support provided to export the annotations in widely used formats and ease-of-use regarding annotation markers management is amongst the most distinctive features of the developed tool. Lastly, accuracy of header and citation level metadata extraction is dependent on GROBID, that is currently giving state-of-the-art results [22] in the light of recent study followed by CERMINE and ParsCit. Therefore, the developed tool can be used to annotate scientific articles in a time-efficient manner.

4 Conclusion and Future Work

Due to rapid growth of scientific literature, there is a dire need for systems that can perform automatic information extraction from ever-growing scientific articles. Such system development would require quality annotated datasets. As scientific article annotation is rather a hefty and laborious task. Hence, in this study, a web-based an-

notation tool is presented which can provide assistance during scientific articles' annotation. In literature there exist different text annotation tools but, to the best of our knowledge, there is no easy-to-use web-based annotation tool for scientific articles' annotation. The current version of tool provides rapid and intuitive means to annotate scientific articles by offering various text filters and highlighting support. In addition, the tool also extracts an article's metadata information and citation information using state of the art text processing libraries. Thus, it has capability to help in generating comprehensive datasets for scientific articles including metadata, citation and full-text information.

The tool currently provides PDF processing facility and is primarily developed to annotate scientific articles. In future, this tool would be made generic to annotate general text datasets. Furthermore, collaboration support between annotators along with comparison of various annotations made by different annotators can also be incorporated. Integration of the tool with different BRAT visualization features is also among the possible future extensions.

References

1. D. de S. Price, *Science since babylon*. Yale University Press, 1961.
2. B. Mudrak, "Scholarly Publishing in 2016 | AJE | American Journal Experts," 2016. [Online]. Available: https://www.aje.com/en/arc/scholarly-publishing-trends-2016/. [Accessed: 02-Apr-2018].
3. NSF, "S&E Indicators 2018 | NSF - National Science Foundation," 2018. [Online]. Available: https://www.nsf.gov/statistics/2018/nsb20181/. [Accessed: 03-Apr-2018].
4. B. Morin, "LibGuides: Systematic Reviews: Intro," 2017. [Online]. Available: https://researchguides.library.tufts.edu/c.php?g=249130&p=1658802. [Accessed: 27-Mar-2018].
5. R. Borah, A. W. Brown, P. L. Capers, and K. A. Kaiser, "Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry," *BMJ Open*, vol. 7, no. 2, p. e012545, Feb. 2017.
6. H. Harkema, I. Roberts, R. Gaizauskas, and M. Hepple, "Information extraction from clinical records," in *Proceedings of the 4th UK e-Science All Hands Meeting*, 2005.
7. K. Seymore, A. Mccallum, and R. R. T, "Learning Hidden Markov Model Structure for Information Extraction," in *Proc. AAAI'99 Workshop Machine Learning for Information Extraction*, 1999, pp. 37–42.
8. S. Anzaroot and A. Mccallum, *A New Dataset for Fine-Grained Citation Field Extraction*. 2013.
9. A. McCallum, D. Freitag, and F. C. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation.," in *Icml*, 2000, vol. 17, pp. 591–598.
10. M. Liakata, "Aberystwyth University - ART," 2009. [Online]. Available: https://www.aber.ac.uk/en/cs/research/cb/projects/art/. [Accessed: 12-Feb-2018].
11. M. Liakata, "Zones of Conceptualisation in Scientific Papers: A Window to Negative and Speculative Statements," in *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, Stroudsburg, PA, USA, 2010, pp. 1–4.

12. S. Gupta and C. Manning, “Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers,” presented at the Proceedings of 5th International Joint Conference on Natural Language Processing, 2011, pp. 1–9.
13. Y. Tateisi, T. Ohta, S. Pyysalo, Y. Miyao, and A. Aizawa, “Typed Entity and Relation Annotation on Computer Science Papers.,” in *LREC*, 2016.
14. I. Augenstein, M. Das, S. Riedel, L. Vikraman, and A. McCallum, “SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications,” *ArXiv170402853 Cs Stat*, Apr. 2017.
15. P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, “BRAT: a web-based tool for NLP-assisted text annotation,” in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 102–107.
16. C. Mitre, “Callisto - Home Page,” 2013. [Online]. Available: https://mitre.github.io/callisto/index.html. [Accessed: 07-Jul-2018].
17. “Open Calais,” *Open Calais*, 2008. [Online]. Available: http://www.opencalais.com/. [Accessed: 06-Sep-2017].
18. J. Beel, S. Langer, M. Genzmehr, and A. Nürnberger, “Introducing Docear’s Research Paper Recommender System,” in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY, USA, 2013, pp. 459–460.
19. I. Councill, C. L. Giles, and M.-Y. Kan, “ParsCit: an Open-source CRF Reference String Parsing Package,” in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*, Marrakech, Morocco, 2008.
20. P. Lopez, “GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications,” in *International Conference on Theory and Practice of Digital Libraries*, 2009, pp. 473–474.
21. D. Tkaczyk, P. Szostek, M. Fedoryszak, P. Dendek, and Ł. Bolikowski, “CERMINE: automatic extraction of structured metadata from scientific literature,” *Int. J. Doc. Anal. Recognit. IJDAR*, vol. 18, no. 4, pp. 317–335, 2015.
22. D. Tkaczyk, A. Collins, P. Sheridan, and J. Beel, “Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers,” in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 2018, pp. 99–108.