Scientometrics (2018) 117:1931–1990
https://doi.org/10.1007/s11192-018-2921-5

Information extraction from scientific articles: a survey

Zara Nasar[1] · Syed Waqar Jaffry[1] · Muhammad Kamran Malik[1]

Received: 19 May 2018 / Published online: 29 September 2018

Abstract
In last few decades, 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 to automate the process of literature review. Therefore, in this study, aim is to present the overall progress concerning automatic information extraction from scientific articles. The information insights extracted from scientific articles are classified in two broad categories i.e. metadata and key-insights. As available benchmark datasets carry a significant role in overall development in this research domain, existing datasets against both categories are extensively reviewed. Later, research studies in literature that have applied various computational approaches applied on these datasets are consolidated. Major computational approaches in this regard include Rule-based approaches, Hidden Markov Models, Conditional Random Fields, Support Vector Machines, Naïve-Bayes classification and Deep Learning approaches. Currently, there are multiple projects going on that are focused towards the dataset construction tailored to specific information needs from scientific articles. Hence, in this study, state-of-the-art regarding information extraction from scientific articles is covered. This study also consolidates evolving datasets as well as various toolkits and code-bases that can be used for information extraction from scientific articles.

Keywords Metadata extraction · Key-insights extraction · Text mining · Information extraction · Machine learning · Research articles · Scientific literature

✉ Zara Nasar
zara.nasar@pucit.edu.pk

Syed Waqar Jaffry
swjaffry@pucit.edu.pk

Muhammad Kamran Malik
kamran.malik@pucit.edu.pk

[1] Punjab University College of Information Technology, University of the Punjab, Lahore, Pakistan

Introduction

In last few decades, advent of computers and later World Wide Web (WWW) has 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. WWW has affected the overall growth in scientific literature. According to a study carried out in Price (1961), amount of research data doubles every ten to 15 years. Additional resources (Mudrak 2016; NSF 2018) pointed that around 2.2 million new scientific-articles were published in 2016. Some of the major reasons regarding rapid growth in number of scientific-articles include increased number of publication venues, online digital libraries and ease of access in acquiring scientific literature; whereas these facilities were not available in pre-digital age. In the light of report issued by International Association of Scientific, Technical and Medical Publishers, there is an increase in publishing scientists by 4–5% annually. Additionally, as of 2014, there exist around 28,100 peer-reviewed scholar journals in English (Ware and Mabe 2015).

This increase in scientific content poses significant challenges for the researchers who want to determine state of art in their respective field of interest. To perform literature review, firstly literature is required from variety of relevant research repositories. Later, the acquired results are filtered by means of manual analysis. After acquiring the relevant literature, the findings from these scientific-articles are consolidated in order to determine state-of-the-art of desired field. This whole process of performing systematic literature review is of utmost importance for researchers as it helps in performing gap analysis and determining room for innovation. At the same time, this is very time consuming, cumbersome and laborious task. According to one of the systematic literature review guideline, amount of time that is required to conduct a quality review can take up to 1 year (Morin 2017). In the light of another study, systematic literature review can take up to 186 weeks with single/multiple human resources (Borah et al. 2017).

To provide researchers with basic filters, 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 (Harkema et al. 2005). The domain

of IE is comprised of concepts and techniques of Machine Learning, Natural Language Processing (NLP), Text Mining (TM) and Information Retrieval (IR). There exist various research studies that focus on describing state-of-the-art in the domain of IE (Simoes et al. 2009; Sirsat et al. 2014).

The survey study presented in Simoes et al. (2009) has its major focus on categorizing various tasks of IE reported in literature and respective techniques used to perform those tasks. This study categorized IE tasks into five major classes that include segmentation, classification, association, normalization and co-reference resolution. Segmentation refers to the task of segmenting the data into atomic segments like tokens. Classification task deals with assigning each segment to its suitable class called entity. According to Simoes et al. (2009) major techniques employed in literature to perform classification include Hidden Markov Models (HMM) and Maximum Entropy Markov Models (MEMM). Association task focuses on extraction of relations between related various entities. Major algorithms that are being used for association mining task include context free grammars, MEMM and Conditional Random Fields (CRF). As far as normalization and co-reference resolution tasks are concerned, these are less-generic as they require domain-specific information. Normalization refers to the task where different representations of a similar entity are transformed into single entity. This task is usually carried out via human-designed conversion rules and regular expressions. Co-reference resolution refers to the problem of identifying various senses of text fragments that point towards a same real-world entity.

Amongst the various tasks mentioned for IE in Simoes et al. (2009), classification task is usually regarded as Named Entity Recognition and Classification (NERC). The NERC refers to a sub-problem in domain of IE that deals with extraction of named entities (NEs) while keeping surrounding context under consideration. The NERC deals with problem of recognition of named entities followed by their classification in rhetorical categories. It holds utmost importance in other IE, NLP and TM oriented tasks including relation extraction, event detection, question answering systems and machine translation. Table 1 represents the NEs that can be extracted from the following short paragraph.

> Valencia is on her way to Wal-Mart super-store in Austin. She is asked to bring couple of coffee bags. Her nephews from Valencia are waiting for her arrival.

Here, in this example it could be observed that Valencia is a person name in opening sentence of paragraph, whereas in last sentence, it is a geographical location. Thus, NERC tends to recognize senses of entities based on the surrounding context.

There exist multiple survey studies that presents the current progress in domain of NERC (Kanya and Ravi 2012; Palshikar 2013; Patil et al. 2016; Sharnagat 2014). These surveys classify NER literature in terms of various factors. Some are focused on employed approaches that include rule-based and machine-learning oriented solutions. Whereas, some surveys perform primitive classification based on underlying resources' language. Most of this literature is focused on developments of NERC in various news datasets and well-formatted English language where primary task is to identify person names, location and organization. These annotated benchmark datasets are available in variety of languages including English, Spanish, Arabic and Chinese.

In addition to survey studies focused on conventional NERC problems, there exist surveys that present developments of NERC when applied on medical scientific articles. In the past years, many developments in the domain of medical sciences, genetics and other biology domains (Abdelmagid et al. 2014; Duck et al. 2016; Shickel et al. 2017) are being made. Major reason of rapid development in the domain of biology and related domains is

Table 1 A sample NERC/IE task

Named entity	Named entity value
Name	Valencia
Location	Austin, Valencia
Organization	Wal-Mart

availability of formal ontologies, extensive corpuses and lexicons. These language resources and sophisticated rule-based as well as machine-learning approaches are usually employed to extract various entities. In addition, these entities are often related to genomics, gene relations, various proteins and molecular information. These survey studies often include bio-specific entities and hence are not generic in nature.

As research literature is exponentially increasing across various disciplines. Hence, there is a need to consolidate findings that have been made so far regarding information extraction from this ever-growing scientific literature. Further, the emphasis of this paper is to consolidate findings from studies that can be applicable to wider range of domains. Therefore, developments against bio-specific entities' extraction are not included in current survey. However, if research study extracts generic insights from medical dataset, then such studies are part of current survey.

In order to present survey focused on generic IE from scientific literature, current survey presents ongoing advancements against two major information constituents of a scientific article as explained above. These constituents include its metadata and its body. To the best of our knowledge, there does not exist a comprehensive survey that is focused on presenting such insights from scholarly literature. Although, there exist comparative studies to evaluate performance of various information extractors from scientific articles, but these studies are focused on developed tools and more inclined towards practical aspects.

In the light of above points, it is evident that a survey study focused on presenting state of the art advancements along with open-areas carries huge importance. Therefore, the current work compiles and analyzes research work and applications of NERC task of IE, when applied on research papers with respect to metadata and article's body. This study covers major datasets for scientific articles, respective evaluation metrics on various datasets against studies along with employed approaches to perform IE from scientific articles. As survey is majorly focused on general insights extraction from articles, therefore, emphasis has not been given to describe various employed tools or techniques to preprocess the scientific articles' content. Hence, preprocessing techniques for scientific literature that are required to convert input into desired feature vectors are not part of current study.

This survey aims to assist researchers interested to learn about recent advancements and to have an overview regarding automatic IE from scientific articles. Current study further highlights the open research areas as well as future prospects in this domain. In addition to that, as metadata and insights from full-text can include many sub-fields. Therefore, study is focused towards providing detailed results to give a brief overview about on-going progress rather than reporting average results only. This is because, results against coarse-grained fields can provide better insights about current gaps in literature by letting the readers know about specific subset of fields that are currently performing lower than the rest. Hence, this study will be very helpful for researchers interested in mining of scientific literature.

Rest of this study is organized in following sections. "Methodology" section describes methodology to conduct this study. It briefly explains the primary classification of

literature followed by widely used evaluation metrics in current domain in "Evaluation metrics" section. "Metadata extraction" and "Key-insights extraction" sections describe state-of-the-art in metadata and key-insights extraction from scientific articles respectively. Finally, "Conclusion and future work" section provides overall conclusion with future prospects with bibliography presented in "References".

Methodology

In order to conduct this study, first of all literature review is performed to determine the state-of-the-art of current domain. For this purpose, two famous research repositories including ACM and IEEE were used to get the relevant domain papers till 2017. Most relevant seed words to scientific literature were firstly identified by means of exploring synonyms and related words. Later, both research repositories were queried with the identified seed words within titles of publications only. All queries were made via advanced search options whereas in case of ACM, ACM Guide to Computing Literature bibliographic database is used for wider coverage.

Querying mechanism enforced the presence of all words in acquired scientific article's titles i.e. AND operation is being performed among query strings. Further double quotes ensure that whole phrase appears together in title. Out of all seed words, "research article" seed word resulted into huge number of results. When the acquired results were analyzed, it was noticed that there was huge noise in form of conference proceedings that used to end with: "Research Articles". In addition, there were some publishers that were using similar words for proceedings name. Hence, in order to avoid such results, resultant literature was acquired by means of refining respective particular query and adding more fields to avoid noise. After filtering such records, around 200 results were acquired against "research article" query from ACM. Statistics regarding initial results acquired against each seed word are shared in Table 2.

These acquired results were later manually filtered based on their relevance and categorized in major classes. This categorization was made after reading the titles of scientific articles only. The tentative counts of articles against each category is also mentioned.

1. Information Extraction ($\sim$ 80)
2. Recommender Systems ($\sim$ 45)
3. Classification and Clustering ($\sim$ 20)
4. Summarization ($\sim$ 20)

Table 2 Stats against initial queries

Queries	ACM	IEEE	Total
Research article	2720	12	2732
Research literature	44	8	52
Research paper	226	29	255
Research publication	20	4	24
Scientific paper	164	16	180
Scientific article	70	6	76
Scientific literature	106	41	147
Scientific publication	82	7	89
	3432	123	3555

5. Citation Analysis ($\sim$ 50)
6. Structural studies ($\sim$ 40)

A brief overview of overall methodology that is followed in order to perform this study is presented in Fig. 1. After categorization of acquired literature; articles regarding information extraction were studied. These articles were further categorized into two types; metadata and key-insights. Later, state-of-the-art approaches and datasets against each category were determined. In the light of this whole process, research findings against this study are consolidated to present the current state in this domain.

Many researchers have contributed their researches in order to extract the information from scientific articles. A scientific article generally consists of two major constructs: metadata and full-body text. Therefore, existing research studies can be broadly classified in two categories:

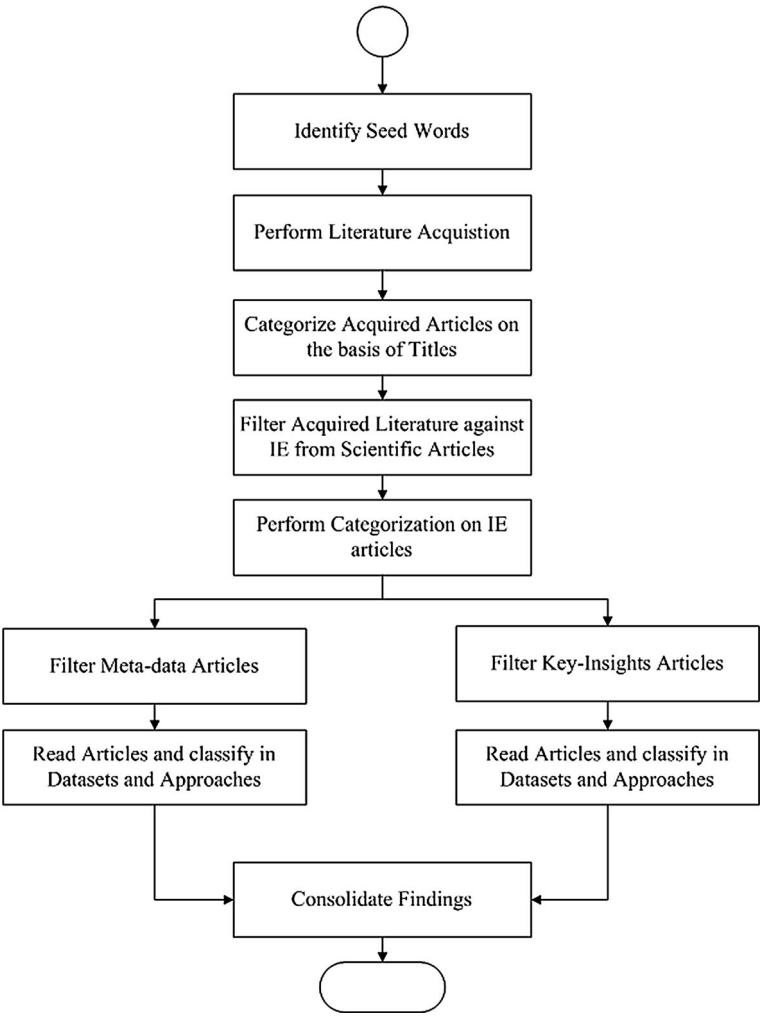

Fig. 1 Overall flow of study

1. Metadata Extraction
2. Key-insights Extraction

Metadata Extraction: In order to automatically extract metadata, semi-structured format of scientific articles can be exploited. This metadata information holds great importance in context of digital research repositories. This information includes title of a scientific article, respective authors, publication venue, date of publication and keywords etc. In addition, metadata information within citation also carries immense importance especially in the domain of Scientometrics. As, metadata information can be used to perform variety of other tasks including article recommendation and citation analysis etc., current study compiles and presents research progress in this domain.

Key-Insights Extraction: Apart from structured format, the text of a research paper has got its own importance. A researcher can have various research questions while reading a scientific article. Some of these research questions include:

1. Problem addressed in a scientific article
2. Domain of a research study
3. Methodology/Algorithms/Processes used in order to address the problem
4. Datasets that are used in order to conduct experiments
5. Tools used to perform the experiments
6. Evaluation measures to gauge performance
7. Results achieved in a research study
8. Limitation of a research study
9. Future extensions

Automatic extraction of such insights can provide substantial ease to researchers while performing literature review. In addition, if these insights are extracted from bulk of scientific data, literature gaps can be identified efficiently. Hence, this study covers on-going advancements towards automatic key-insights extraction form scientific articles.

Evaluation metrics

One of the very important aspect to measure the progress within any research area is its evaluation. Due to their avid importance, this section briefly describes the evaluation metrics that are being employed in reported literature for IE. Evaluation of an IE system is usually performed by means of comparing the extracted information with the respective gold standard data-set. These gold-standard datasets are mostly annotated by humans and serve as the ground truth in any problem. Hence, to compare and evaluate performance using gold-standard datasets, major evaluation metrics include Precision, Recall, F-measure and Accuracy. Precision focuses on evaluating how many of the extracted information is correct. Recall, on the other hand, is focused on evaluating that how much of the correct information is extracted. Usually, a confusion matrix is constructed to calculate various

Table 3 A confusion matrix for two class problem

	Predicted (positive)	Predicted (negative)
Actual (positive)	True positive (TP)	False negative (FN)
Actual (negative)	False positive (FP)	True negative (TN)

evaluation measures for a classification problem. Table 3 shows a confusion matrix for binary classification problem. This concept can be further extended into multiple classes.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

where FP is regarded as type-1 error and FN is regarded as type-2 error. Increase in FP tends to decrease precision whereas increase in FN tends to decrease recall. In order to take into account both of these measures, F-score is widely used that is harmonic mean between precision and recall.

$$\text{Fscore} = \left(1 + \beta^2\right) \frac{\text{Precision} * \text{Recall}}{\beta^2 * \text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{4}$$

$$\text{Error} - \text{rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{5}$$

Equation 3 facilitates researchers to weight precision and recall as per their information need. For $\beta = 1$, this equation gives similar weight to precision and recall and usually termed as F-measure or balanced F-score/F1-score. Accuracy measure represents the ratio between total number of correct results over total results generated via system as shown in Eq. 4. Error-rate, on the other hand, represents ratio of total no of incorrect results made by the respective algorithm, over the total results as presented in Eq. 5.

"Metadata extraction" and "Key-insights extraction" briefly explains current state-of-the-art regarding metadata insights and key-insights extraction from research articles. All evaluations results reported in this study are taken from respective research articles and all evaluation measures are being presented in percentages. F-score reported throughout the study is balanced F-score. In some studies, evaluation measures against token and field are presented. Token-level evaluation measures are based on the number of individual word tokens that are correctly classified in the respective label class. Field-level scores are based on the number of exact fields that are classified correctly as whole, whereas a field can contain multiple tokens. Thus, in case of field-level scores, there is no partial credit for subset of correct predictions at token level. In all the tables carrying evaluation measures; Precision, Recall, F-measure and Accuracy are represented as Prec., Rec., F1 and Acc. respectively.

Metadata extraction

Metadata is broadly classified into three types by NISO (2004) that includes descriptive, structural and administrative. Descriptive metadata is used for discovery and identification, which in turns helps in finding and searching tasks such as title, author, keywords etc. On the other hand, structural metadata helps in determining how a paper is organized. For example, an outline of a paper can give an insight about paper structure. Administrative metadata provides information regarding resource management such as file type, creation date etc.

In the context of research articles, metadata is usually of descriptive nature and it holds a great importance. It provides a brief overview about a scientific article by providing information such as title of an article, its authors and bibliography etc. Hence, researchers tend to decide paper relevance with their domain of interest based on metadata information such as title, abstract, references, authors, citing articles and affiliations. In addition to that, digital research repositories also make use of metadata in order to provide support regarding literature acquisition for research community. These libraries aid researchers by providing intelligent search tools that include search filtering based on keywords, authors, organizations, publication venues etc. that are part of metadata information. In addition to that, this information can also be used to recommend articles (Haruna et al. 2017; Knoth et al. 2017).

Further, by extracting citation level metadata; one can also provide statistical information regarding an article's citations count and popularity over time. Citation-level metadata extraction is also very useful in the domain of Scientometrics. (Alam et al. 2017; Insights 2013). Table 4 presents the NEs that can be extracted from following reference strings.

REF # 1	Ramadge, P., & Wonham, W. (1989). The control of discrete event systems. Proceedings of the IEEE, 77 (1), 81–98
REF # 2	W. H. Enright. Improving the efficiency of matrix operations in the numerical solution of stiff ordinary differential equations. ACM Trans. Math. Softw., 4(2), 127–136, June 1978

Table 4 A sample NERC task from references

Entities	Ref # 1	Ref # 2
Author names	Ramadge, P; Wonham, W.	W. H. Enright
Publication date	1989	June 1978
Publication title	The control of discrete event systems	Improving the efficiency of matrix operations in the numerical solution of stiff ordinary differential equations
Proceedings	Proceedings of the IEEE	
Journal		ACM Trans. Math. Softw
Volume	77	4
Issue	1	2
Pages	81–89	127–136


Journal Year Volume
Scientometrics (2018) 116 721–750 Pages
https://doi.org/10.1007/s11192-018-2768-9 URL

CrossMark

Three new bibliometric indicators/approaches derived from keyword analysis

Title

Mengyang Wang[1] · Lihe Chai[1]

Author Names

Received: 17 July 2017 / Published online: 11 May 2018

Abstract

Abstract Keyword analysis has been an important research theme in bibliometrics. The deduction of new valuable bibliometric indicators/approaches through keyword analysis is important for prompting the further development of this subject area. In this study, the following three bibliometric indicators/approaches were thus derived. Indicator *K* was derived using the ratio between the average unique keyword number and average keyword frequency of a discipline for quantitatively describing the discipline's development stages highlighted by scientific-philosopher Kuhn. Next, the correlation matrix analysis was used after *k*-core filtration to quantitatively expose the detailed correlations between topics for a large network. Thirdly, indicators *I* (node betweenness divided by node degree) and *C* (clustering coefficient) were collectively introduced to predict potential growth keywords. Diverse topical evolutions were categorized into a strategic diagram according to the tendencies of *I* and *C*. With sustainable development as a case study, we verified that the three new bibliometric indicators/approaches work well and can realize many new concepts beyond the scope of available indicators or approaches. In summary, the present paper makes a renewed effort to promote the development of bibliometrics. We hope our work could catalyze the further studies from the communities in the scientometric fields.

Keywords Keyword analysis · *K* indicator · *k-Core* · Correlation matrix · *I*–*C* indicator · Co-word network

Keywords

Introduction

For publications, authors are requested to add several terms or phrases to best describe the document topic. These terms or phrases are called keywords, and are always regarded as important and condensed contents of academic publications for any discipline and

✉ Mengyang Wang
vellio@163.com Author Email

[1] School of Environmental Science and Engineering, Tianjin University, Tianjin 300354, China
Author Affiliation

Springer

Fig. 2 Sample Header metadata extraction

Table 5 Information against CORA Header dataset

	Field name	Field description
1	Title	It refers to title of scientific article
2	Author	It refers to name(s) of the article's authors
3	Affiliation	It refers to affiliation of author(s) i.e. where they work at
4	Address	It refers to address of author(s)
5	Email	It refers to email addresses of authors
6	Date	It refers to publication date of a scientific article
7	Phone number	It refers to author's phone number
8	Web/URL	It refers to URL of author's webpage
9	Degree	It refers to mention of any degree i.e. if an article is a thesis submitted in partial fulfillment of M.S. or Ph.D.
10	Publication number/ISSN	It refers to code that is used to identify journals, books, magazines etc.
11	Note	It refers to acknowledgements, copyright, notices etc. that are made in an article
12	Abstract	Abstract and introduction together are regarded as summary in Shuxin et al. (2013)
13	Introduction	
14	Keyword	It refers to the main author-assigned keywords of an article
15	Page	It refers to the page end

Figure 2 on the other hand, presents sample task of header level metadata extraction from Wang and Chai (2018). Header level metadata extraction deals with identification and extraction of title, authors, affiliations, emails, publication venue, DOI, keywords, abstract and other related fields usually from the title page of a scientific article. In respective figure, title, authors and their respective affiliation is being recognized from the title page of a scientific article.

In the light of above points, it is evident that metadata extraction carries huge importance in many research oriented tasks. As there exist wide variety of reporting styles in forms of various journals, conferences, technical reports and wide variety of citation formats; the task of header-level metadata extraction as well as citation-level metadata extraction becomes quite challenging. In the remaining section, first of all major datasets for metadata extraction will be discussed followed by widely used approaches that are being used to solve this problem.

Datasets

There are three widely used datasets namely CORA, FLUX-CiM and UMASS that were developed in 1999–2000, 2007 and 2013 respectively. Amongst these, CORA dataset is split into two parts: one is focused towards document metadata whereas other one is focused on metadata extraction from citation strings. Other two datasets are also focused on metadata extraction from citations.

CORA[1] dataset consists of computer science articles' data. The widely used CORA-Header dataset for document header metadata extraction is presented in Seymore et al.

[1] https://people.cs.umass.edu/∼mccallum/data.html.

Table 6 Information against CORA reference dataset

	Field name	Field description
1	Author	It refers to the authors of a scientific article
2	Book-title	It refers to the name of conference proceedings and books, if respective content is a conference proceeding or a book chapter or book
3	Date	It refers to publication date of a scientific article
4	Editor	It refers to editor of journals and books
5	Institution	It refers to institution of author(s)
6	Journal	It refers to journal name, if article is published in journal
7	Location	It refers to location of conference
8	Note	It refers to any additional notes; such as "Submitted to EuroPar'97"
9	Pages	It refers to page-range where respective article resides within a book or proceeding
10	Publisher	It refers to publisher information
11	Tech	It refers to technical report
12	Title	It refers to title of scientific article
13	Volume	It refers to volume of book/journal along with issue number

(1999). This dataset has total fifteen (15) fields that are explained in Table 5 and comprises 935 records in total with 500 training records and 435 testing records. CORA-reference dataset (McCallum et al. 2000) contains 500 references in total, whereas 350 records are usually used for training and the remaining 150 for testing. This dataset contains total of thirteen (13) fields. Tables 5 and 6 compile the attributes that are part of CORA header and reference dataset respectively.

FLUX-CiM dataset consists of articles from varied domains including Computer Science (CS), Health Science (HS) and Social Sciences (SS) articles. CS article dataset carries total of 300 reference strings, where each reference is further segmented into ten fields. HS dataset contains total of 2000 reference strings and is developed using PubMed Central data and each reference string is further segmented into six fields (Cortez et al. 2007). While SS dataset also share same fields as that of HS dataset and is constructed using data from Scielo Digital Library. Mapping of entities from CORA and FLUX-CiM is presented in Table 7. FLUX-CiM dataset majorly differs from CORA in terms of variety as it includes citations from HS and SS as well. In addition, FLUX-CiM does not cover all the fields that are present in CORA.

UMASS[2] dataset consisting of bibliography information from 5000 research papers is presented in 2013 (Anzaroot and Mccallum 2013). It consists of citations from total of 5000 articles from Arxiv. These articles are evenly distributed in four major domains that include physics, mathematics, computer science and quantitative biology. Dataset comprises variety of formats and styles, including journal pre-prints, conference papers and technical reports. Each of these citation strings is labeled in a hierarchical manner, with both coarse-grain labeled segments, as well as fine-grain labeled segments that are presented in Tables 8 and 9 respectively.

[2] http://www.iesl.cs.umass.edu/data/data-umasscitationfield.

Table 7 Mapping of CORA fields against FLUX-CiM

	CORA	FLUX-CiM	CS	HS	SS
1	Author	Author	✔	✔	✔
2	Title	Title	✔	✔	✔
3	Journal	Journal	✔	–	–
4	Date	Date	✔	✔	✔
5	Pages	Pages	✔	✔	✔
6	Book-title	Conference	✔	–	–
7	Location	Place	✔	–	–
8	Publisher	Publisher	✔	–	–
9	Volume	Number	✔	–	–
10		Volume	✔	✔	✔

Table 8 UMass dataset coarse-grained entities

Coarse-grained entities	Description
Ref-marker	It refers to citation marker in the paper
Authors	It refers to the list of authors in a citation
Title	It refers to the primary title of a cited work
Date	It refers to the publication date of cited work
Venue	It refers to venue including its publication venue, volume etc.

Approaches

During the last decades, many researchers have contributed in domain of IE from research papers. Multiple machine learning and NLP techniques are being used to extract metadata from scientific literature. Some of the widely used techniques include Rule-based systems and machine learning systems. Amongst the machine learning techniques: markov models, conditional random fields and support vector machines are being frequently used. Following section describes developments for metadata extraction with respect to each technique.

Rule-based approaches

Rule-based systems refer to the systems that rely on a set of predefined instructions that specify how to extract desired information from data. In the context of metadata extraction, many researchers have used rule-based approaches based on text structure and layouts. A study reported in Klink et al. (2000) uses rules that rely on textual and geometrical features. It focuses on extraction of following entities from an article's metadata: abstract-body, abstract-heading, affiliation, biography, caption, drop-cap, highlight, keyword-body, keyword-heading, membership, page-number, pseudo code, publication-info, reader-service, synopsis, and text-body. They used rule-base that can be applied on multiple domains. Study claims to have reasonable results when rules are used along with fuzzy matching. Results are evaluated on 979 journal pages from University of Washington corpus.

Table 9 UMASS dataset fine-grained fields

Coarse-grained	Fine-grained	Description
Venue	Note	It refers to plain text note about the citation
	Web	It refers to any web address that is included in citation
	Status	It refers to the current status of respective article, e.g. in preparation, submitted, accepted, revised, available
	Language	It refers to the language information of the cited work
	Book-title	It refers to the name of a book or conference proceedings in which an article is published
	Date	It refers to the date on which the venue of cited work was published
	Address	It can refer to the location of a conference or of a publisher
	Pages	It refers to the pages on which the article appears in book or proceedings
	Organization	It refers to the sponsoring organization of a conference
	Volume	It refers to the volume of cited work
	Number	It refers to the issue number of article
	Publisher	It refers to the publisher of the journal, conference, book etc.
	Editor	It refers to the list of journal editors
	Tech	It refers to the words that describe the tech report or type of unpublished material with possible tech report number, e.g. eprints, preprints
	Institution	It refers to the organization that publishes the tech report
	Series	It refers to the series name in which a cited book is published
	Chapter	It refers to a book chapter that a citation is referencing
	Thesis	It refers to the part of the citation mentioning that the cited work is a thesis, e.g. Ph.D. Thesis
	School	It refers to the school of authors that have published the thesis
	Department	It refers to the department of authors that have published the thesis
Person names	Person-first	It refers to the first name or initial name of a person
	Person-middle	It refers to the middle name or initial of a person
	Person-last	It refers to the last name/surname of a person
	Person-affix	If refers to any affix of a person, e.g. Jr., Sr

Table 10 Accuracies against Post-Script (Giuffrida et al. 2000) format, OCR system (Mao et al. 2004) and Template Matching framework (Huang et al. 2006)

	Giuffrida et al. (2000)	Mao et al. (2004)	Huang et al. (2006)
Title	92.00	96.36	96.30
Author	87.00	89.09	80.20
Affiliations	75.00	92.73	71.90
Author-affiliation	71.00	–	–
Table of contents	76.00	–	–
Abstract	–	98.18	88.40
Keywords	–	–	84.70
Overall	–	94.09	–

Metadata extraction from research article in post-script format is reported in Giuffrida et al. (2000). This study employs knowledge base carrying rules against various metadata fields including title, author, affiliations, author-affiliation mapping and table of contents. The knowledge base makes use of visual and spatial knowledge in order to identify these metadata entities with fuzzy-logic. For example, rules such as "title is usually in big font and in start of the text" or "title should be above the abstract section" are used to extract metadata. In order to demonstrate the effectiveness of proposed approach, data set of hundred articles is used. This dataset has 70% conference articles and remaining consist of journal articles and technical reports. Respective accuracies against proposed approach are reported in Table 10.

Study reported in Mao et al. (2004) makes use of OCR in order to identify the respective metadata spans. Additionally, it presents a dynamic feature update system that tends to generate and improve features, whereas these features include geometrical as well as contextual features that include font size, font type and bounding box. The distribution of these features is calculated using data from OCR and saved against each journal's style. Feature generation algorithm later employs various string-matching algorithms to extract the feature vectors. Feature vectors learnt over previous journal issues and/or other journal issues are applied to extract information from current issues. These features are later used in a rule-based system to extract metadata information. In order to evaluate the proposed system, title pages of 309 medical research articles are used. These articles are scanned images from two medical journals and dataset include various types of articles including short papers, correspondences etc. Results are evaluated on 166 title pages of Indian Journal of Experimental Biology and 143 pages from Journal of Clinical Monitoring and Computing, where both these journals are scanned medical journals. Experiments results show that employment of multiple journal issues for feature learning yield better results than using one issue. Optimum labeling accuracies against this study are presented in Table 10.

The study proposed in Huang et al. (2006) makes use of template matching in order to extract header metadata information that includes title, author, authors' affiliations, abstract and keywords. By analyzing the four widely used publication styles that include Springer Lecture Notes in Computer Science (LNCS), Elsevier, ACM and IEEE JNS formats; authors proposed a template that can carry various fields from these publication styles. Later, a finite state automaton is presented that is being used to perform template matching. Results are evaluated on 400 sampled articles from ACM, IEEE, Springer LNCS and Elsevier. Title extraction accuracy is highest, while affiliation extraction accuracy is the lowest one amongst the rest in the light of acquired results as shown in Table 10.

A hierarchical template-based citation metadata extraction for scholarly publications is presented in Day et al. (2007). A hierarchical knowledge representation framework that extracts important concepts from natural language texts is used. In order to cover major domain-specific constructs, proposed framework named INFOMAP consists of domain-specific concepts along with related sub-concepts, relevant categories, attributes and actions. This information eventually helps in maintaining relationships between various concepts and ultimately transforms this knowledge base into taxonomy. Using this taxonomy, INFOMAP classifies citation strings in concepts as well as their related concepts. A powerful feature of the framework is its ability to represent and match complicated template structures. The proposed template extraction framework is evaluated on self-generated dataset covering six major citation styles including APA, IEEE, ACM, ISR, MISQ and JMIS from 160,000 citations. Results are mentioned in Table 11 with overall average accuracy of 92.39%.

Table 11 Accuracies against INFOMAP (Day et al. 2007)

		Fields							
		Author	Title	Journal	Volume	Issue	Year	Pages	Average
Journals	APA	92.32	71.80	94.33	97.39	84.92	96.48	95.09	90.33
	IEEE	94.17	89.05	92.07	95.45	84.49	97.18	89.81	91.75
	ACM	88.36	91.10	99.41	80.28	87.73	96.47	83.95	89.61
	ISR	91.93	78.33	95.32	95.28	87.00	96.34	90.61	90.69
	MISQ	97.73	97.92	100.0	99.99	99.98	99.94	99.64	99.31
	JMIS	76.55	72.57	99.99	99.98	99.97	99.93	99.69	92.67
	Average	90.18	83.46	96.85	94.73	90.68	97.72	93.13	92.39

A template based metadata extraction architecture has been presented in Flynn et al. (2007). This work is majorly focused on processing of various type of data including data from various agencies, laboratories, universities etc. PDF containing either scanned images or text is taken as input. Data from Defense Technical Information Center (DTIC) and National Aeronautics and Space Administration (NASA) reports is used in study. DTIC dataset usually contains Report Document Page (RDP) forms. Hence, major emphasis of the proposed architecture is processing of form and non-form based data. Due to the layout of RDP, templates are very suitable choice. For inputs containing no RDP forms, a non-form based process is executed that firstly converts respective input into XML format. Results against form-based inputs show higher precision and recall, whereas achieved accuracy against non-form based metadata extraction is 66% and 64% against DTIC and NASA reports.

An unsupervised system for metadata extraction named FLUX-CIM is proposed in Cortez et al. (2007, 2009). This approach differs from existing rule-based/knowledge-based systems as this study automatically creates knowledge base using existing metadata records. In order to validate this approach, two types of dataset are constructed. First dataset consists of Computer Science articles and carries total of 300 reference strings, where each reference is further segmented into ten classes including Author, Title, Journal, Date, Pages, Conference (Book-title), Place (Location), Publisher, Number and Volume. Second dataset consist of articles from medical sciences and contains total of 2000 reference string and each reference string is further segmented into six fields including Author, Title, Journal, Date, Pages and Volume (Cortez et al. 2007). Another dataset from Social Sciences articles is constructed. Both health sciences and social sciences datasets carry uniform format of citations, hence, these datasets are referred as organized, because they follow similar citation formats and thus are relatively simpler to deal with. The automatic construction of knowledge-base is handled using existing data; e.g. for CORA; corresponding bibTex entries against training set were parsed and included in knowledge-base. The filed level Precision/Recall/F-measure against the developed dataset using the proposed unsupervised approach is presented in Table 12. One interesting claim of the authors which is backed with respective experiment is that, if extracted entities are straight away added into knowledge base, it can also improve the results as knowledge base size affects the overall performance. For future directions, author suggest learning implicit

Table 12 Evaluation measures against FLUX-CIM various datasets

	CS (Cortez et al. 2007)			CORA (Cortez et al. 2009)			HS (Cortez et al. 2009)			SS (Cortez et al. 2009)		
Field	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Author	93.59	95.58	94.57	97.21	98.67	97.93	98.57	99.04	98.8	96.48	99.17	97.81
Title	93.00	93.00	93.00	93.01	96.67	94.8	84.88	85.14	85.01	91.2	96.67	93.86
Journal	95.70	97.80	96.75	93.45	91.5	92.46	97.23	89.35	93.12	97.99	93.68	95.79
Date	97.75	97.44	97.59	96.01	90.01	92.91	99.85	99.5	99.67	99.57	97.01	98.27
Pages	97.00	97.84	97.41	97.98	98.81	98.39	99.7	99.2	99.45	99.65	98.45	99.05
Book-title	97.47	95.45	96.45	93.16	91.73	92.44	–	–	–	–	–	–
Location	96.83	97.6	97.21	90.01	94.78	92.33	–	–	–	–	–	–
Publisher	100.0	100.0	100.0	90.00	96.56	93.16	–	–	–	–	–	–
Number	97.87	97.87	97.87	100.0	99.14	99.57	–	–	–	–	–	–
Volume	100	98.2	99.12				96.41	98.75	97.57	98.67	98.66	98.66
Tech	–	–	–	94.67	91.9	93.26	–	–	–	–	–	–
Note	–	–	–	91.00	88.81	89.89	–	–	–	–	–	–
Editor	–	–	–	90.76	91.66	91.21	–	–	–	–	–	–
Institution	–	–	–	92.15	91.11	91.63	–	–	–	–	–	–
Average	96.62	97.01	97	96.28	95.8	96.01	96.11	95.16	95.6	97.26	97.27	97.24

Table 13 F1-score against FLUX-CiM (Cortez et al. 2009), CRF and Template Extraction (TE) (Guo and Jin 2011)

	CORA		HS		SS		SGD[a]
Field	FLUX-CiM	CRF	FLUX-CiM	CRF	FLUX-CiM	CRF	TE
Author	94.20	99.40	96.62	95.48	99.54	94.31	94.7
Title	93.57	98.30	99.56	96.16	99.78	97.14	88.1
Journal	92.62	91.30	93.71	89.30	94.01	88.89	90.2
Date	95.66	98.90	99.87	96.57	99.84	96.19	91.2
Pages	95.67	98.60	97.83	96.47	93.18	90.67	83.8
Book-title	93.64	93.60	–	–	–	–	91.8
Location	93.15	87.20	–	–	–	–	
Publisher	92.50	76.10	–	–	–	–	
Others	94.08	89.40	–	–	–	–	
Volume	–	–	99.95	95.92	97.20	92.14	82.4
Issue	–	–	–	–	–	–	86.1
URL	–	–	–	–	–	–	93.3
Average	93.90	92.54	97.92	94.98	97.26	93.22	

[a]SGD denotes self-generated dataset. This dataset consisted of 97 computer science articles from IEEE and ACM

styling and improved matching functions to distinguish between similar entities such as author's name or editor's name.

In addition, experiments are conducted to compare the proposed approach with CRF that tend to provide state-of-the-art results in statistical modeling techniques. For the sake of comparison, CORA dataset was used as computer science dataset besides self-constructed dataset of social and health sciences. The respective F-scores against various datasets using proposed approach and CRF are presented in Table 13.

Text formatting information is also used in Groza et al. (2009) to extract Title, Author, Sections and references from research articles in PDF format. This study proceeds with firstly carrying out a pilot study to determine the habits, beliefs and opinions regarding metadata reporting in research articles. Later, in light of the learnt insights, heuristics and rules are prepared that exploit formatting and font styling features. There are two major modules of the proposed approach namely first-page content extraction and full-text content extraction. First-page extraction deals with extraction of Title, Abstract and Author Names and full-content extraction specifically refers to extraction of section information and references. Evaluation is performed on 1203 documents following ACM or Springer LNCS format. Results show F-measure greater than 90% against all entities. One thing to note in the evaluation set is that all selected articles were correctly parsed from PDF format. By individually analyzing performance against Springer and ACM, extraction on Springer LNCS outperforms ACM due to less variation. This study has proposed several feature oriented mathematical functions in order to extract metadata information from scientific articles published in PDF format. Authors have presented two major applications of proposed system that include metadata extraction web service and personal research assistant. Various evaluation metrics against this study are reported in Table 14.

The methodology used in Adefowoke Ojokoh et al. (2009) combines segmentation based on keywords and pattern matching techniques (regular expressions) to extract general metadata from documents such as Title, Table of Contents and Abstract etc. This study was tested on dataset consisting forty thesis using precision, recall, accuracy and F-measure scores, whereas results against these evaluation measures are presented in Table 14.

Another study in this regard is presented in Guo and Jin (2011) that employs knowledge base and template extraction. Initially, templates are constructed using formation of citations. Total of 576 templates are created covering various reference styles. In addition to that, a knowledge base carrying names of authors, venues and publisher is populated. This knowledge base is basically used to determine that in which class; a particular input element belongs to. After getting primitive idea about possible and most-likely classes of input elements in a citation, template matching is performed using most similar template in the light of extracted insights earlier. Once these elements are extracted, metadata knowledge base is again queried to check if it has records against input citations. If record exists, results from knowledge base are returned as they are more accurate. Thus incorporation of knowledge base helps in improving the overall results. The proposed approach is evaluated on 97 computer science articles from IEEE and ACM, where these articles can be journals or conference papers. Table 13 shows the accuracies against extracted fields using proposed approach. This approach is not robust enough to handle articles carrying complex structures.

Another template based approach is proposed in Chen et al. (2012). This approach treats citation string as text data carrying fields to be extracted along with delimiters. The study is focused to extract seven attributes out of a citation string including Author, Title, Venue, Volume, Issue, Page, and Date. Venue fields is later post processed to identify journal,

Table 14 Evaluation measures against Groza et al. (2009) and Adefowoke Ojokoh et al. (2009)

	Groza et al. (2009)				Adefowoke Ojokoh et al. (2009)			
	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.
Title	96.00	98.00	96.00	95.00	75.00	75.00	75.00	75.00
Authors	92.00	96.00	93.00	90.00	–	–	–	–
Table of contents	–	–	–	–	73.00	81.00	77.00	68.00
Preface	–	–	–	–	86.00	100.0	92.00	98.00
Abstract	99.00	96.00	97.00	96.00	77.00	92.00	84.00	78.00
Sections	97.00	93.00	94.00	92.00	–	–	–	–
Acknowledgment	–	–	–	–	64.00	90.00	75.00	70.00
Introduction	–	–	–	–	68.00	68.00	68.00	60.00
Conclusion	–	–	–	–	90.00	68.00	77.00	70.00
References	96.00	93.00	94.00	91.00	100.0	91.00	95.00	93.00

Table 15 Evaluation Measures against template-matching approach (Chen et al. 2012)

–	INFOMAP				CORA				FLUX-CIM (HS)			
	Token			Field	Token			Field	Token			Field
	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.
Author	99.37	99.33	99.35	98.13	96.01	97.47	96.72	89.55	95.20	99.71	97.40	93.20
Title	99.54	96.90	98.21	94.93	97.12	94.35	95.71	90.07	99.85	95.91	97.84	93.06
Venue	98.14	98.14	98.14	96.80	92.48	88.54	90.39	78.79	97.26	99.44	98.3	98.90
Volume	96.37	98.07	97.21	97.59	84.81	88.41	86.41	86.75	99.85	98.64	99.24	98.95
Issue	99.04	94.79	96.87	93.38	77.77	93.85	84.92	78.51	–	–	–	–
Page	99.65	98.85	99.25	98.10	97.28	96.71	96.99	95.08	100.0	97.07	98.51	96.99
Date	99.31	99.50	99.41	99.10	97.76	97.34	97.04	94.25	98.38	100.0	99.18	99.00
Journal	99.12	92.80	95.86	92.2	66.65	88.13	75.7	85.6	97.97	89.26	93.39	93.19
Book-title	–	–	–	–	97.39	82.38	89.14	65.86	–	–	–	–
Tech-report	–	–	–	–	86.93	78.07	81.72	58.53	–	–	–	–
Average	98.77	97.94	98.35	96.86	91.89	93.81	92.6	87.57	97.53	98.31	97.89	96.48

book-title and tech-report. Proposed approach has three major modules namely canonicalization algorithm, template database construction followed by query processing. In order to identify structural elements from a citation string, rule-base algorithm is employed that is being termed canonicalization algorithm. It employs various heuristics and makes use of patterns and reserved words in order to retain structural information in a contextual string. This information is later used in template-extraction module as well as query-processing module to define templates and search templates based on structured citation respectively. This algorithm is evaluated on three datasets including INFOMAP, CORA and FLUX-CiM. Respective evaluation metrics against each dataset are show in Table 15.

Rule-based systems tend to have very good performance due to the manual effort and human observations, but it certainly has obvious disadvantages. These systems are less adaptable than machine learning based systems, due to their dependence on text formatting, text location and graphical attributes of text. Rule formation in itself is a laborious and a time-consuming task. Complexity in the rules makes them powerful. But consequently, the processing of rules becomes time expensive with increase in time complexity. Hence, the overall time complexity of the system rapidly increases with the number of rules as concluded in Klink et al. (2000).

Machine-learning based approaches

Following section compiles the major approaches that employ machine learning concepts to perform metadata extraction from scientific articles.

Hidden Markov model Hidden Markov Model (HMM) has strong statistical grounds that are robust in nature and efficient to develop. Its major weakness is its reliance on training data. It is widely being used across many domains including Speech Recognition (Juang and Rabiner 1991) and machine learning related problems. In current domain of interest, HMM is used along with multiple state-merging options in Seymore et al. (1999). It makes use of distantly labeled data-set (bibTex) to improve the accuracy of HMM model. It primarily deals with extraction of CORA Header entities. It was tested on manually tagged

Table 16 Accuracy against CORA dataset in Seymore et al. (1999) and McCallum et al. (2000)

	Seymore et al. (1999)	McCallum et al. (2000)
Abstract	98.70	98.60
Address	95.10	95.00
Affiliation	90.70	90.60
Author	97.20	97.10
Date	97.20	97.00
Degree	73.20	73.20
Email	86.90	86.50
Keyword	98.90	98.80
Note	89.00	88.70
Phone	87.40	86.90
ISSN	60.60	60.30
Title	97.80	97.90
URL	41.70	41.70
Overall	92.90	92.70

data along with bibTex Collection with 92.9% accuracy against all headers classes including 97.8% for Title and 97.2% for Authors. Detailed results against each field are shared in Table 16.

HMMs are also used in development of CORA system proposed in McCallum et al. (2000). This system is used as Internet portal for computer science articles providing various features such as searching and identification of metadata entities from scientific articles. Proposed approach is generic enough to apply on various other Internet portals. In the development system, one HMM is used to identify the fields such as author, title, affiliation from paper's header. Second HMM is used to extract metadata information from references. With respect to HMMs for IE, primary focus of this study is to learn the parameters and transition structures using labeled and unlabeled text. Study shows that distant supervision tends to improve the results, whereas parameter estimation using forward-Baum–Welch (Baum 1972) degrades the performance. One primary reason can be that forward-Baum–Welch algorithm tends to get stuck in local maxima; therefore, it is sensitive towards initial parameter settings. Here, distant supervision refers to incorporation of data that is annotated for some other purpose such as bibTex that carries marked authors against an article, whereas it doesn't carry all the required fields. Error-rate and accuracy against various fields are presented in Table 16.

A research study carried out in Hetzner (2008) employs HMM by means of Viterbi algorithm and string manipulation methods. In order to improve the performance, separate set of cue-words are constructed that are good indicators of fields to be extracted. Results of this study are evaluated on CORA dataset. A quite similar approach is proposed in Ni and Xu (2009) that is also focused towards citation metadata extraction by means of HMM. It makes use of Baum–Welch (BW) algorithm in order to learn the weights during HMM transitions. It also forms multiple states against potential information to be extracted from citation. This HMM-BW based model has been comparatively evaluated using existing

Table 17 Evaluation measures against various HMM models

	Hetzner (2008)			Ni and Xu (2009)		
Field	Prec.	Rec.	F1	Prec.	Rec.	F1
Author	99.50	99.60	99.50	99.81	99.81	99.81
Book-title	90.40	93.80	92.10	99.52	89.19	94.07
Date	96.70	99.40	98.00	99.07	99.13	99.1
Editor	89.20	92.20	90.70	91.16	92.81	91.98
Institution	61.90	92.90	74.30	–	–	–
Journal-title	89.00	73.00	80.20	93.84	99.46	96.56
Location	79.20	76.00	77.60	67.36	93.49	78.3
Note	34.10	75.00	46.90	–	–	–
Number	95.70	100.0	97.80	–	–	–
Pages	98.80	98.20	98.50	99.95	99.13	99.54
Publisher	73.20	83.30	77.90	84.95	91.42	88.06
Tech-title	93.10	93.10	93.10	99.64	98.85	99.24
Title	99.20	96.60	97.90	99.75	98.31	99.02
Volume	96.10	81.70	88.30	99.81	99.81	99.81
All	94.30	94.40	94.40	–	–	–
Macro-average	85.40	89.60	86.60	94.08	96.49	95.04

Table 18 Evaluation measures against Cui and Chen (2010)

	Accuracy-way		Location-based-way	
	Prec.	Rec.	Prec.	Rec.
Title	97.52	92.55	100.0	100.0
Author	89.45	94.82	99.33	98.50
Affiliation	86.95	82.56	93.05	92.86
Address	73.87	85.86	94.32	86.91
Email	93.40	79.27	94.13	98.88
Abstract	96.15	99.21	99.08	99.47

Table 19 Evaluation Measures against Ojokoh et al. (2011) on CORA and FLUX-CiM datasets

	CORA				FLUX-CIM (HS)			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Title	98.32	96.36	97.14	96.73	99.80	99.61	99.94	99.78
Author	99.43	98.31	99.54	98.91	99.76	99.60	99.65	99.62
Date	99.70	97.93	98.39	98.16	99.98	100.0	99.58	99.58
Pages	99.66	96.33	98.10	97.20	99.95	99.50	99.76	99.63
Volume	99.71	97.04	94.24	95.53	99.94	99.88	98.74	99.30
Journal	98.83	90.31	85.66	87.76	99.88	99.98	98.75	99.36
Book-title	98.10	91.56	93.92	92.71	–	–	–	–
Publisher	99.60	91.04	88.87	89.82	–	–	–	–
Location	99.33	89.04	88.76	88.53	–	–	–	–
Tech	99.38	83.64	85.60	83.69	–	–	–	–
Institution	99.49	85.50	90.47	87.55	–	–	–	–
Note	99.28	82.11	54.88	60.33	–	–	–	–
Editor	99.41	92.78	85.36	88.61	–	–	–	–
Overall	99.25	95.06	95.18	95.12	99.88	99.66	99.65	99.65
Average	99.25	91.69	89.3	89.66	99.88	99.76	99.33	99.54

HMM model (Hetzner 2008) as well as CRF (Peng and McCallum 2006) in respective study as well. Table 17 represents various evaluation measures against aforementioned HMM models.

Another study proposed in Cui (2009) makes uses of HMM with text block as basis of Viterbi algorithm (Forney 1973) instead of words, along with some heuristic for email, phone numbers, keywords and web. The fields that are being extracted in this study include title, author, address, affiliation, email, web, phone, date, abstract and keyword. It is trained on 800 headers and tested with data of 135 headers with all fields' precision and recall greater than 85%. This study is further extended in Cui and Chen (2010) to improve Viterbi algorithm in HMM model. It makes use of the idea that transition probability between the same states in the same line is far greater than that in different lines. Further it employs location based information to further improve the results of Viterbi model. As existing dataset does not contain location information, a new dataset consisting of 458 articles was

Table 20 Evaluation measures against Ojokoh et al. (2011) and Yin et al. (2004) on ManCreat dataset

	Bi-gram HMM (Yin et al. 2004)			Tri-gram HMM (Ojokoh et al. 2011)			
	Prec.	Rec.	F1[a]	Acc.	Prec.	Rec.	F1
Title	89.59	92.12	90.84	97.01	94.17	94.11	94.14
Author	97.70	95.87	96.78	98.72	98.05	97.20	97.61
Date	92.39	94.47	93.42	99.07	98.05	89.73	93.63
Pages	95.52	96.30	95.91	99.32	95.82	91.60	93.60
Volume	84.05	89.97	86.91	99.60	86.08	84.12	84.44
Issue	88.60	89.95	89.27	99.61	84.43	88.94	86.47
Journal	90.52	80.80	85.38	96.48	90.72	91.24	90.98
Url	87.70	95.53	91.45	99.52	84.53	95.17	89.36
Publisher	64.36	83.30	72.62	99.69	78.69	78.37	77.89
Location	66.89	78.27	72.13	98.92	86.59	91.17	88.78
Other	42.88	61.80	50.63	98.84	93.49	93.24	93.37
Total	90.15	90.15	90.15	98.84	87.19	89.47	87.76

[a]Denotes self-computed values using balanced F-score formula

constructed from VLDB conferences to train the location-based model. Table 18 presents evaluation measure when location based heuristics are excluded and included.

Tri-gram HMMs are being employed in Ojokoh et al. (2011) to extract citation metadata. Total of twenty features are used as emission vocabulary to improve the model. These features include full-stop, comma, capital letter, all numbers etc. In order to further improve the results, shrinkage is employed. Shrinkage refers to technique that is usually used to handle the sparse data transitions while training HMMs. Results are evaluated on CORA and FLUX-CiM datasets. In addition, effect of data size on model is also experimented using FLUX-CiM dataset which shows that with increase in data, F-score and recall tend to decrease whereas precision increases. Moreover, one-third dataset was able to achieve 98% accuracy. Further data addition increments this overall gain minimally. The results are being shown in Table 19. Comparison is also made with existing bi-gram HMM study (Yin et al. 2004) that also employed similar idea of shrinkage but used bi-grams for network training. The study employing bi-grams for network training used self-created dataset for evaluation consisting of 713 citation strings obtained from 250 scientific articles. Tri-gram model (Ojokoh et al. 2011) is evaluated against self-annotated data of bi-gram model (Yin et al. 2004) as well that is being referred as ManCreat dataset. Evaluation metrics using ManCreat dataset against both bi-gram and tri-gram models are presented in Table 20.

HMM tends to compute a probability distribution over possible sequences of labels followed by selection of best label sequence. Parameters in HMM are trained to maximize the joint likelihood of training examples. This requires enumerating all possible observation sequences. Due to that, long-range dependencies and interacting features can't be represented into this model. These are the pioneer statistical models to be applied in order to solve sequence oriented problems. These models made the foundation of further improved models such as Maximum Entropy Markov models and Conditional Random Fields.

Table 21 Evaluation Measures against Peng and McCallum (2006) against CORA and self-annotated dataset

	CORA header		Self-annotated		CORA Reference		
Overall-accuracy	98.30		98.60		95.37		
Instance-accuracy	73.30		75		77.37		
Field	Acc.	F1	Acc.	F1	Field	Acc.	F1
Title	99.70	97.10	99.60	96.90	Title	98.90	98.30
Author	99.80	97.50	99.80	97.70	Author	99.90	99.40
Affiliation	99.70	97.00	99.70	98.10	Institution	99.70	94.00
Address	99.70	95.80	99.60	95.20	Book-title	97.70	93.70
Note	98.80	91.20	99.30	83.20	Editor	99.50	87.70
Email	99.90	95.30	99.80	95.50	Journal	99.10	91.30
Date	99.90	95.00	99.90	98.90	Date	99.80	98.90
Abstract	99.60	99.70	99.80	99.80	Note	99.70	80.80
Phone	99.90	97.90	99.90	97.80	Pages	99.90	98.60
Keyword	99.70	88.80	99.80	91.90	Publisher	99.40	76.10
Web	99.90	94.10	99.90	93.00	Tech	99.40	86.70
Degree	99.80	84.90	100.0	100.0	Location	99.30	87.20
Pubnum	99.90	86.60	99.90	62.50	Volume	99.90	97.80
Average-F1		93.9	97.7	93.1	Average-F1		91.5

Conditional random fields Conditional Random Field (CRF) is a statistical model that has the ability to incorporate effect of neighbors as well. CRFs are currently being used as an alternative to HMMs in Named Entity Recognition, Pattern Matching and other Machine Learning problems. Many researchers have applied concepts of CRF in domain of IE from research papers.

Research study reported in Peng and McCallum (2004) makes use of CRF with Gaussian priors, regularization and hyperbolic priors to extract metadata fields including: author, affiliation, address, note, email, date, abstract, introduction, phone, keywords, web, degree, publication number and page. In addition to that, CRF is also used to perform citation metadata extraction. This technique, when applied on a standard benchmark dataset, resulted in reduced error in average F-score and word error rate by 36% and 78% respectively, in comparison with the previous best SVM results of study (Han et al. 2003) with average F-score being 93.9% and overall accuracy being 98.3%. Study employs CORA header and reference dataset for evaluation. The extension of this study is presented in Peng and McCallum (2006) that provides mechanism to exploit co-reference citations using CiteSeerX 2007, which results in error rate reduction by 6-14% on self-annotated datasets that are tagged with co-reference information. Another dataset is developed in extended study that consists of 450 headers. This dataset contains font information as it is used as a feature to improve identification of field boundaries. For this dataset, scientific articles were randomly selected amongst 8000 articles that are crawled from internet from various sources. In order to train the model, 300 records were used while remaining 150

Table 22 Evaluation Results in Yu and Fan (2007)

Header					Reference				
	English		Chinese			English		Chinese	
	Prec.	Rec.	Prec.	Rec.	Tags	Prec.	Rec.	Prec.	Rec.
Title	97.40	92.00	96.20	95.70	Title	97.40	95.00	98.80	92.60
Author	98.30	90.60	99.30	95.00	Author	98.60	95.70	99.70	92.80
Affiliation	98.60	93.70	99.00	92.40	Journal	94.70	94.50	96.70	94.00
Address	93.80	91.60	96.30	90.50	Volume	96.00	89.60	90.80	85.60
Zip code	99.10	97.00	99.20	95.80	Year	98.60	97.50	96.80	95.30
Abstract	98.00	100.0	97.50	100.0	Pages	86.00	89.50	80.90	83.20

Table 23 Evaluation measures against Councill et al. (2008) using various datasets

Field	CORA Reference			CiteSeer			FLUX-CIM (CS)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Author	98.70	99.30	99.00	95.80	95.70	96.00	98.80	99.00	99.00
Book-title	92.70	94.20	93.00	72.50	92.90	81.00	95.70	99.30	97.00
Date	100.0	98.40	99.00	98.80	89.80	94.00	99.80	94.50	97.00
Editor	92.00	81	86.00	95.60	51.10	67.00	–	–	–
Institution	90.90	87.90	89.00	70.90	76.70	74.00	–	–	–
Journal	90.80	91.20	91.00	88.00	78.60	83.00	97.10	82.90	89.00
Location	95.60	90	93.00	91.90	78.40	85.00	96.90	88.40	89.00
Note	74.20	59	65.00	88.90	17.20	29.00	–	–	–
Pages	97.70	98.40	98.00	90.30	91.50	91.00	94.70	99.30	97.00
Publisher	95.20	88.70	92.00	88.70	74.80	81.00	98.80	75.90	85.00
Tech	94.00	79.60	86.00	76.10	70.00	73.00	–	–	–
Title	96.00	98.40	97.00	91.90	93.90	93.00	98.80	98.30	96.00
Volume	97.30	95.50	96.00	89.30	85.00	87.00	95.30	89.70	92.00
Average	95.70	95.70	95.00	–	–	–	97.40	97.40	94.00

were used for testing. Table 21 shows the results against CORA and self-annotated dataset carrying font information.

The study presented in Yu and Fan (2007) applies CRF in order to extract metadata from Chinese research papers. It uses three different types of features that include local features regarding character specifics, layout features that carry information regarding word occurrence and external features that carry information from external lexicons such as family names and location names etc. Comparison is made with HMMs as well and results show that CRF tends to perform better in both languages. For English dataset, CORA header and reference dataset is used for evaluation whereas for Chinese, dataset was constructed using data from China National Knowledge Infrastructure. Header dataset for Chinese consist of 600 headers whereas reference dataset consists of 1500 references.

Table 24 Base line Results against Anzaroot *and* Mccallum (2013) using CRF

	Field			Token		
Label	F1	Prec.	Rec.	F1	Prec.	Rec.
Authors	96.13	94.05	98.31	98.00	97.78	98.21
Person-affix	40.00	50.00	33.33	80.00	100.0	66.67
Person-first	95.05	92.31	97.96	97.67	96.32	99.05
Person-last	95.20	92.58	97.97	97.69	96.32	99.09
Person-middle	92.84	89.72	96.19	96.23	95.22	97.26
Date-year	90.91	87.72	94.34	92.61	92.16	93.07
Ref-marker	97.64	96.42	98.90	99.69	100.0	99.38
Reference-Id	87.10	87.10	87.10	96.10	100.0	92.50
Title	87.07	84.96	89.30	97.09	95.13	99.14
Venue	48.00	60.00	40.00	52.17	60.00	46.15
Address	85.71	94.29	78.57	92.11	98.13	86.78
Book title	41.86	42.86	40.91	55.56	48.70	64.66
Category	–	100.0	–	–	100.0	–
Chapter	–	100.0	–	–	100.0	–
Date-month	62.50	50.00	83.33	87.50	77.78	100.0
Date-year	92.82	91.38	94.31	96.17	95.02	97.35
Department	100.0	100.0	100.0	100.0	100.0	100.0
Edition	61.54	80.00	50.00	54.05	100.0	37.04
Editor	60.61	52.63	71.43	67.86	59.38	79.17
Person-first	69.57	72.73	66.67	75.00	81.82	69.23
Person-last	72.00	81.82	64.29	70.59	75.00	66.67
Person-middle	72.73	80.00	66.67	72.73	80.00	66.67
Institution	30.77	100.0	18.18	26.67	100.0	15.38
Journal	91.37	87.89	95.13	95.52	93.39	97.75
Language	–	–	–	20.00	100.0	11.11
Note	–	100.0	–	–	100.0	–
Number	74.07	75.00	73.17	88.89	91.95	86.02
Organization	66.67	50.00	100.0	44.44	28.57	100.0
Pages	94.51	91.81	97.36	98.45	97.34	99.58
Publisher	77.23	76.47	78.00	87.93	82.70	93.87
Reference-Id	72.73	80.00	66.67	81.08	88.24	75.00
School	–	–	100.0	–	–	100.0
Series	25.00	40.00	18.18	25.00	58.82	15.87
Status	36.36	50.00	28.57	57.14	72.73	47.06
Tech	57.14	72.73	47.06	72.00	90.00	60.00
Thesis	–	100.0	–	–	100.0	–
Volume	93.91	91.61	96.32	95.90	95.34	96.46
Web	–	100.0	–	–	100.0	–
Overall	91.16	90.17	92.16	94.79	94.08	95.50

Table 25 Results against Shuxin et al. (2013) using optimized Particle Swarm Optimization algorithm

CORA header			CORA reference		
	Acc.	F1		Acc.	F1
Title	99.70	97.10	Author	99.90	99.40
Author	99.80	97.50	Book-title	97.70	93.70
Unit	99.70	97.00	Date	99.80	98.90
Address	99.70	95.80	Editor	99.50	87.70
Summary	98.80	91.20	Institution	99.70	94.00
E-Mail	99.90	95.30	Journal	99.10	91.30
Date	99.90	95.00	Address	99.30	87.20
Abstract	99.60	99.70	Summary	99.70	80.80
Telephone-number	99.90	97.90	Page-number	99.90	98.60
Keyword	99.70	88.80	Press	99.40	76.10
URL	99.90	94.10	College	99.40	86.70
Degree	99.80	84.9	Title	98.90	98.30
ISSN	99.90	86.60	Volume	99.90	97.80
Average	98.30	93.90	Average	95.37	91.50

In both languages sets, similar six fields are selected for experiments. Results for header and reference dataset against both languages are presented in Table 22.

Another study employing CRF for the task of metadata extraction is presented in Councill et al. (2008). This study is a pioneer contribution in open-source domain and provides features for automatic reference string extraction followed by its segmentation into multiple classes. Additionally, study also focuses towards extraction of citation context. Citation context refers to those areas/sentences, which corresponds to a citing article. In order to develop this framework, CRF and heuristics are used. Heuristics are primarily used for extraction and identification of reference strings and citation contexts. CRF, on the other hand, is used in order to segment reference string into further categories. In order to evaluate the model's performance, various experiments are performed including CORA, CiteSeer and FLUX-CiM datasets. Here, CiteSeer dataset consist of randomly sampled 200 reference strings from millions of reference strings available in CiteSeer system. Respective results against various datasets are presented in Table 23. The proposed system is integrated into CiteSeer system.

Study presented in Anzaroot and Mccallum (2013) also use CRF in order to provide baseline results against developed UMASS dataset. Study argues about limitations of conventional CRFs in making predictions due to Markov's assumptions. Therefore, future work is directed towards development of improved CRF models. In addition, dataset presented is to be revised and extended with time, as increase in tagged dataset eventually helps in improving the accuracy of machine learning systems. Baseline results against fine-grained dataset are presented in Table 24 with field level as well as token level evaluation. Other studies that are focused on improvement of underlying CRF models to improve the global context coverage include (Anzaroot et al. 2014; Vilnis et al. 2015). These studies discuss citation extraction as an application of improved CRF models on UMASS dataset.

Another approach that is using CRF for Information Extraction makes use of Particle Swarm Optimization algorithm (Kennedy and Eberhart 1995) which is used to evaluate the

Table 26 F1-scores against Souza et al. (2014) using two-layer CRF model

	40-papers	100-papers
Title	100.0	100.0
Author-name	99.41	98.91
Email	100.0	100.0
Affiliation	99.83	99.64
Venue-name	85.20	85.94
Venue-year	100.0	100.0
Venue-date	100.0	98.89
Venue-publisher	100.0	100.0
Venue-location	93.86	96.60
ISBN	100.0	98.82
Average-(F1)	97.83	97.88

optimal value keeping evolution in context. This approach uses an optimized version of Particle Swarm Optimization algorithm in order to avoid local convergence by using iterative likelihood ratio as stop criterion (Shuxin et al. 2013). It improves results of existing CRF based studies of Peng and McCallum (2004, 2006) with average F-score of 93.9% and accuracy of 98.3%. Detailed results are presented in Table 25.

Another study employing CRF for the task of metadata extraction is presented in Souza et al. (2014). This study presents two-layer model of CRF. The study takes into account first page of research article as it carries the potential header metadata information. The first layer identifies larger components from article text that may contain metadata information. These components are header, title, author information, body, and footnote. The header usually holds important information about the conference/journal in which the paper has been published. The title class represents the title of the paper. Author information contains data about the authors, such as: name, affiliation, and email. As the body class does not include useful data for the task of metadata extraction, hence it is not further processed. On the other hand, as footnotes usually contain information about the publisher, conference, and additional information about the authors that may include authors 'email and affiliation. Hence, a second CRF layer was created for header, author information and footnote. This extra layer allows to extract the actual metadata and define section-specific features. Results are evaluated on 100 papers whereas dataset and respective corpus is freely available over github.[3] Out of these 40 papers belong to an existing study that is focused towards extraction of structural contents from paper presented in Kan et al. (2010) using single-layer CRF. F1-score results against initial 40 papers from existing extraction study and total 100 papers are presented in Table 26.

Another study (Cuong et al. 2015) has focused on improvement of conventional CRF results by introducing concepts of higher order semi-CRFs. These models have the capability to model the transition between variable length sequence segments, hence, giving them more power than traditional linear chain CRFs. Proposed approach is applied to variety of problems including author names, authors' affiliation extraction as well as citation metadata extraction from scientific articles. The experiments are conducted using ParsCit dataset[4] with linear-chain CRFs as baseline and first order, second order and third

[3] https://github.com/alansouzati/artic-poc.

[4] https://github.com/knmnyn/ParsCit/tree/master/crfpp/traindata.

Table 27 F1-scores against Cuong et al. (2015)

Fields	Base-line	1st order	2nd order	3rd order
Citation metadata extraction				
Author	99.00	98.97	99.02	98.78
Book-title	93.60	93.67	94.15	93.71
Date	93.61	92.98	93.26	93.11
Editor	75.33	71.54	75.60	75.60
Institution	79.17	79.17	79.17	96.43
Journal	89.31	89.60	90.12	88.22
Location	89.20	89.18	89.91	90.68
Note	57.14	57.14	57.53	60.00
Pages	95.91	95.59	94.56	95.49
Publisher	83.33	83.68	83.33	84.39
Tech	46.15	46.15	46.15	62.5
Title	94.53	94.74	95.35	95.22
Volume	91.28	92.20	87.74	90.00
Micro-average	94.01	94.01	94.35	94.26
Header metadata extraction				
Author	93.64	93.53	94.06	93.21
Affiliation	98.33	98.50	98.50	98.50

order semi-Markov CRFs. Results show that second-order CRFs tends to give better results than the rest as shown in Table 27.

CRF is currently giving state-of-the-art results in metadata extraction tasks. These models majorly deal with limitations of HMM. One potential drawback of CRF is that they are computationally very expensive. It is currently widely used statistical model for sequence labeling tasks.

Support vector machines Support Vector Machines (SVM) (Cortes and Vapnik 1995) is another technique that has been widely used in literature for automatic metadata extraction. It is primarily a supervised learning technique that is generally used for classification and regression.

A research study in Han et al. (2003) used SVM in order to extract structured metadata from scientific literature. It applies SVM classifiers for two major classifications. One is line classification that is being performed using word and line specific features including word position, line number, and capitalized words. It is used to extract main feature that in turn help in classification. This classified line set is later being passed to another SVM classifier that performs chunk classification that is applied only to multi-line data. It is required to classify multi-line data into their respective categories and makes use of boundary heuristic and punctuation marks. The evaluation is performed using CORA header dataset and respective results are presented in Table 28.

A research study proposed in Kovačević et al. (2011) makes use of SVM classifiers in order to extract eight fields of metadata that includes: title, authors, affiliation, address, email, abstract, keywords and publication note. This study employed SVM in variety of ways. It compared the results when a single classifier is used to classify all fields and when

Table 28 Results against Han et al. (2003) and CRIS system (Kovačević et al. 2011)

	Han et al. (2003)			Kovačević et al. (2011)		
Class	Acc.	Prec.	Rec.	Prec.	Rec.	F1
Title	98.90	94.10	99.10	99.38	98.17	98.77
Author	99.30	96.10	98.40	94.25	90.11	92.13
Affiliation	98.10	92.20	95.40	90.44	89.78	90.37
Address	99.10	94.90	94.50	87.50	88.11	87.80
Note	95.50	88.90	75.50	90.82	83.18	86.83
Email	99.60	90.80	92.70	98.73	98.10	98.41
Date	99.70	84.00	97.50	–	–	–
Abstract	97.50	91.10	96.60	91.83	91.22	91.52
Phone	99.90	93.80	91.00	–	–	–
Keyword	99.20	96.90	81.50	93.88	71.88	81.42
Web	99.90	79.50	96.90	–	–	–
Degree	99.50	80.50	62.20	–	–	–
ISSN	99.90	92.20	86.30	–	–	–

multiple classifiers are used to classify each category. It includes experiments on several classifiers including Decision Tree Classifier, K-Nearest Neighbor Classifier, Naïve Bayes Classifier and SVM. It concludes that best results are achieved when eight separate SVM classifiers are used, where each classifier is used to classify one category. It resulted in above 85% F-score measure for all categories except keywords. In addition to that, it is different from existing techniques as it takes into account the actual text of PDF files with font and styles. Other techniques proposed in Peng and McCallum (2004, 2006), Shuxin et al. (2013), Seymore et al. (1999) makes use of text only. Results are reported on self-annotated corpus of 100 computer science articles belonging to the domain of automatic term recognition and are shown in Table 28.

Others There are several studies that either use hybrid approach to perform metadata extraction or makes use of other techniques that cannot be classified in aforementioned sections. This sub-section compiles such studies to provide brief overview of other on-going advancements.

Study performed in Marinai (2009) makes use of multi-layer perceptron in order to extract metadata information from PDF scientific articles. This extraction is done by means of exploiting visual and layout features of text. Proposed approach performs low level image processing to extract graphical features from initial pages of PDF articles, as they tend to carry major metadata information. Furthermore, DBLP indexing engine is also incorporated to improve the author extraction results. Tool is developed using Greenstone package development library that is focused on extraction of document title, author and related information. In order to evaluate the developed tool, eighty (80) articles from two conference proceedings including ICDAR and GREC, having double and single column document format respectively, are selected. The developed tool was later incorporated with Greenstone packages and results show that substantial work is required to improve the overall extracted results.

A Markov-based model study presented in Kern et al. (2012) makes use of entropy Markov model to extract metadata information from PDF articles called TeamBeam

Table 29 Results against TeamBeam (Kern et al. 2012)

	E-prints		Mendeley		PubMed	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
First-name	53	36	70	69	90	84
Sur-name	86	68	84	82	92	87
Title	87	94	70	92	75	94
E-mail	–	–	–	–	96	98
Affiliation-start	–	–	–	–	68	48
Affiliation	–	–	–	–	65	55

algorithm. TeamBeam uses variety of features and heuristic to identify various metadata fields. The procedure consists of three steps; first step deals with text block classification to identify major blocks carrying various metadata fields. Next step deals with token level classification of text contained in blocks. Final step is to extract metadata using block level as well as token-level classification information. Study performs extensive experimentation with three types of datasets including Mendely, E-prints and PubMed. In addition, classification performance against various algorithms are presented. Lastly, study also performs variety of experiments to see the impact of increased training data on overall extraction performance. E-prints dataset carries 2542 entries while Mendely and PubMed carry 20,672 and 19,581 entries respectively. All three datasets differ each other in terms of layout and formatting styles; whereas this informatin is being exploited as primary feature set in proposed approach. Metadata extraction results against various fields using TeamBeam algorithm are presented in Table 29.

The study carried out in Tkaczyk et al. (2015) focuses on the task of automatic metadata extraction by means of various machine learning constructs whereas the system is named as CERMINE. It divides the task into multiple independent modules that include layout analysis, content extraction, metadata classification and bibliography extraction. Layout analysis deals with character reading, page segmentation and order preservation. Content extraction deals with feature extraction that can identify various zones i.e. a particular piece of text belongs to either metadata, body, bibliography or others class. Using these features, SVM classifier is trained to perform primary zone classification. Metadata extraction deals with further classification of classified zones into pre-determined classes such as authors, affiliations etc. by means of SVM. Further, rule-based approach is also employed to extract metadata. Finally, last phase deals with bibliography extraction that has two major sub-modules namely reference strings extraction and reference parsing. Reference strings extraction deals with separation of individual references which is carried out using K-means clustering. Reference parsing, on the other hand, deals with metadata information extraction from individual references using CRF. Various datasets are used in order to evaluate each individual module. Comparative analysis is also presented with other freely available metadata extractors that include ParsCit (Councill et al. 2008), GROBID (Lopez 2009) and PDFX (Constantin et al. 2013). Compiled results show that overall CERMINE tends to outperform existing solutions. The results reported in Table 30 are evaluated on dataset from selected articles of PubMed Central (PMC). This tool was the top-performing in Semantic publication 2015 challenge for contextual information extraction (SemPub2015 2015).

The study presented in An et al. (2017) makes use of deep neural network in order to extract citation metadata. It employs deep learning model along with CRF in order to

Table 30 Results against CERMINE system (Tkaczyk et al. 2015) and other existing solutions on PMC data

	CERMINE-SVM			PDFx-rule-based			GROBID-CRF			ParsCit-CRF		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Title	95.50	93.40	94.50	85.70	84.70	85.20	82.50	77.40	79.80	34.10	39.60	36.60
Authors	90.20	89.00	89.60	71.20	71.50	71.30	85.90	90.50	88.10	57.90	48.60	52.80
Affiliation	88.20	83.10	85.60	–	–	–	90.80	51.80	66.00	72.20	44.30	54.90
Email	51.70	42.60	46.70	53.00	73.60	61.60	26.90	7.80	12.10	28.80	36.20	32.10
Abstract	82.80	79.90	81.30	71.10	66.70	68.80	70.40	67.70	69.00	47.70	61.30	53.70
Keywords	89.90	63.50	74.40	–	–	–	94.20	44.20	60.20	15.60	03.00	05.10
Journal	80.30	73.20	76.60	–	–	–	–	–	–	–	–	–
Volume	93.30	83.00	87.80	–	–	–	–	–	–	–	–	–
Issue	53.70	28.40	37.10	–	–	–	–	–	–	–	–	–
Pages	87.00	80.40	83.50	–	–	–	–	–	–	–	–	–
Year	96.30	95.00	95.60	–	–	–	95.70	40.40	56.80	–	–	–
DOI	98.20	75.00	85.10	–	–	–	99.10	65.40	78.80	–	–	–
Reference	96.10	89.80	92.80	91.30	88.90	90.10	79.70	66.70	72.60	81.20	71.8.00	76.20

Table 31 Results using Bi-LSTM-CRF (An et al. 2017) framework against UMASS dataset

	Without fine-tuning			Trained model on UMASSS			UMASS baseline		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Address	82.69	65.65	73.19	86.49	97.71	91.76	98.13	86.78	92.11
Author	98.19	98.91	98.55	99.57	99.93	99.75	97.78	98.21	98.00
Booktitle	79.80	53.02	63.71	80.19	82.89	81.52	48.70	64.66	55.56
Chapter	–	–	–	–	–	–	100.0	–	–
Edition	81.82	52.94	64.29	70.59	70.59	70.59	100.0	37.04	54.05
Editor	97.37	75.51	85.06	100.0	91.84	95.74	59.38	79.17	67.86
Institution	06.90	13.33	09.09	45.45	33.33	38.46	100.0	15.38	26.67
Journal	83.33	90.41	86.73	97.18	94.34	95.74	93.39	97.75	95.52
Language	–	–	–	100.0	100.0	100.0	100.0	11.11	20.00
Month	72.73	80.00	76.19	83.33	100.0	90.91	77.78	100.0	87.50
Note	–	–	–	42.86	10.00	16.22	100.0	–	–
Number	68.64	85.26	76.06	86.87	90.53	88.66	91.95	86.02	88.89
Organization	–	–	–	–	–	–	28.57	100.0	44.44
Pages	96.40	89.94	93.06	96.75	99.79	98.25	97.34	99.58	98.45
Publisher	64.15	71.20	67.49	84.39	90.58	87.37	82.70	93.87	87.93
School	73.08	86.36	79.17	100.0	63.64	77.78	–	100.0	–
Series	62.07	22.50	33.03	73.75	73.75	73.75	58.82	15.87	25.00
Status	100.0	78.57	88.00	100.0	96.43	98.18	72.73	47.06	57.14
Tech	100.0	61.90	76.47	97.14	80.95	88.31	90.00	60.00	72.00
Title	92.42	96.64	94.48	98.95	99.45	99.2	95.13	99.14	97.09
Volume	84.12	89.70	86.82	97.21	95.65	96.42	95.34	96.46	95.90
Year	92.19	95.34	93.74	97.72	99.69	98.69	95.02	97.35	96.17
Web	17.65	25.00	20.69	58.33	58.33	58.33	100.0	–	–
Average	90.39	90.96	90.38	96.65	96.82	96.67	94.08	95.50	94.79

perform the extraction. This hybrid Neuro-CRF approach is currently giving state-of-the-art results in general Information extraction tasks as well (Huang et al. 2015; Ma and Hovy 2016; Strubell et al. 2017; Lee 2017). In this study, bi-directional LSTM (Britz 2015) model is used as deep learning framework with Glove's 100 dimensional word embeddings at input layer, which are later fine-tuned. As deep learning frameworks require huge dataset to train, the model is trained on self-generated dataset of 50,000 citations. These citations belong to various domains including computer science, physics, philosophy, etc. with total of twenty-four (24) fine-grained fields. These fields are very close to UMASS dataset. Table 31 shows performance of proposed algorithms on UMASS dataset without any fine-tuning in first column, followed by results achieved when developed deep learning framework is trained using UMASS dataset. Final column contains baseline results of UMASS study that only employs CRF.

In addition to these studies, there exist multiple tools that are dedicated to perform metadata extraction from research articles, citations or both (Beel et al. 2013; Councill et al. 2008; Lopez 2009; Zahedi and Haustein 2017). Some comparisons of these tools are reported in literature (Atdağ and Labatut 2013; Granitzer et al. 2012). Currently, CERMINE and GROBID both are being actively developed and provide good performance over

Table 32 Tools for metadata extraction

Name of tool	Approach used	Link
Docear's PDF inspector	Rule-based	http://docear.org
GROBID	CRF	https://github.com/kermitt2/GROBID
Mendeley Desktop	SVM, External Information using Web	http://www.mendeley.com/
ParsCit	CRF	http://aye.comp.nus.edu.sg/ParsCit/
PDFMeat	Query based system	http://code.google.com/p/pdfmeat/
PDFSSA4MET	Rule-based on XML	http://code.google.com/p/pdfssa4met/
SciPlore Xtract	Rule-based on XML	http://sciplore.org/
SVMHeaderParse	SVM	https://sourceforge.net/projects/citeseerx/
CERMINE	SVM, CRF, Rule-based, K-means clustering	http://cermine.ceon.pl/index.html
ScienceParse	CRF	https://github.com/allenai/science-parse

others such tools. List of tools along with their primary algorithm/mechanisms and available links are provided in Table 32. In the light of recent comparison study against various tools including GROBID, CERMINE, ParsCit, SciecneParse and PDFSSA4MET presented in Tkaczyk et al. (2018), GROBID tends to give better results followed by CERMINE and ParsCit.

Conclusion

In the light of literature reviewed regarding metadata extraction from scientific articles, a comprehensive summary is presented in Table 33. Reference field in table header represents respective research study. Type field represents that which type of information is being extracted i.e. either study performs header metadata extraction or citation metadata extraction. Format refers to the format of input required by the proposed methodology for further processing e.g. PDF, plain text etc. Approach refers to algorithm(s) applied to perform desired extraction from data. Features/Improvement refers to major distinctive contribution or features that are incorporated in study to improve the performance. Dataset refers to dataset name that is used for evaluation. No. field refers to total number of metadata entries that are being extracted in respective study. Lastly, Metric represents evaluation measure(s) applied to report results respectively. Here, in Metrics column: A, P, R, F, E represents Accuracy, Precision, Recall, F1-score and Error-rate respectively.

In the light of Table 32 and Table 33, it is evident that most studies are employing CRF to perform metadata extraction. Initially, linear-chain CRFs were being mainly used. But recent trends show the application of higher order CRFs to incorporate flexibility and to further improve results. Many other studies have opted various improvements to existing implementations. In case of CRF, highest order markov chains are being developed to capture probability of various segments having variable lengths. Similarly, performance gains over HMMs are also achieved by making use of higher order n-grams. Other

Table 33 Summary of Metadata Extraction articles

References	Type	Format	Approach	Features/improvement	Dataset	No.	Metric
Day et al. (2007)	Citation	HTML	Template	Word, punctuation	INFOMAP	7	A
Souza et al. (2014)	Citation	PDF	CRF	Line, word, two-layer model	Self	10	F
An et al. (2017)	Citation	Text	Bi-LSTM CRF	Dropout, early stopping	UMASS	23	P,R,F
Anzaroot and Mccallum (2013), Anzaroot et al. (2014) and Vilnis et al. (2015)	Citation	Text	CRF	Word, external, numeric, punctuation	UMASS	29	P,R,F
Councill et al. (2008)	Citation	Text	CRF + Heuristics	Word, external, numeric, punctuation	CORA, FLUX-CiM,CiteSeer	13	P,R,F
Yin et al. (2004)	Citation	Text	HMM	Bigram	ManCreat	11	P,R
Ojokoh, Zhang, and Tang (2011)	Citation	Text	HMM	Trigram	CORA, FLUX-CiM,ManCreat	13	P,R,F
Cortez et al. (2007, 2009)	Citation	Text	Knowledge-base	Blocking and neighborhood information	FLUX-CiM, CORA	13	P,R,F
Chen et al. (2012)	Citation	Text	Rule, Knowledge-Base	Reserved words, delimiter	FLUX-CiM, CORA, INFOMAP	10	P,R,F,A
Cui and Chen (2010)	Header	HTML	HMM	Spatial	Self	6	P,R,F
Kern et al. (2012)	Header	PDF	MEMM	Beam search, spatial, language model, formatting	E-prints, Mendely, PMC	7	P/R
Marinai (2009)	Header	PDF	MLP	Word, formatting, neighbor	Self	2	–
Groza et al. (2009)	Header	PDF	Rule	Formatting, spatial	Self	5	P,R,F,A
Adefowoke Ojokoh et al. (2009)	Header	PDF	Rule	Segmentation using keywords	Self	8	P,R,F,A
Guo and Jin (2011)	Header	PDF	Rule, knowledge-Base	Formatting	Self	9	A
Tkaczyk et al. (2015)	Header	PDF	SVM, CRF, Rule	Word, line	CORA, PMC, CiteSeer	13	P,R,F
Flynn et al. (2007)	Header	PDF	Template	RDP format exploitation	NASA, DTIC	–	P, R, A
Huang et al. (2006)	Header	PDF	Template	Formatting, spatial	Self	5	P,R,F,A

Table 33 continued

References	Type	Format	Approach	Features/improvement	Dataset	No.	Metric
Giuffrida et al. (2000)	Header	PostScript	Rule	Spatial, word	Self	5	A
Mao et al. (2004)	Header	Scanned	Rule	Formatting, spatial	Self	4	A
Klink et al. (2000)	Header	Scanned	Rule	Word, formatting	U-Wass Corpus	15	P,R
McCallum et al. (2000)	Header	Text	HMM	Reinforcement learning	CORA	13	E
Seymore et al. (1999)	Header	Text	HMM	Smoothing	CORA	13	A
Hetzner (2008)	Header	Text	HMM	Word, smoothing	CORA	13	P,R,F
Kovačević et al. (2011)	Header	Text	SVM	Formatting, word, NER, lexicon	Self	8	P,R,F
Han et al. (2003)	Header	Text	SVM	Word, Line	CORA	13	P,R,F,A
Cuong et al. (2015)	Header, Citation	Text	CRF	Higher-order semi-Markov	CORA, FLUX-CiM	15	F
Shuxin et al. (2013)	Header, Citation	Text	CRF	Improved optimization algorithm	CORA	13	P,R,F
Peng and McCallum (2004)	Header, Citation	Text	CRF	Spatial, word, external	CORA	13	F,A
Yu and Fan (2007)	Header, Citation	Text	CRF	Spatial, word, external	Self, CORA	6	P,R
Peng and McCallum (2006)	Header, Citation	Text with font information	CRF	Formatting, spatial, word, external	Self	13	F,A

improvements include smoothing techniques, improved error functions and optimization algorithms etc.

Other than algorithmic improvements, studies have also reported improved performance by means of employing various features of input data at hand. Regarding major features listed in Table 33, word features refer to any features that correspond to a word itself including its content, character length, casing features etc. Line features include number of words in line, total line length by characters and words. Spatial features refer to the location of any particular field in text. Formatting features include font stylings and font size information. External features refer to incorporation of external lexicons in the system. Neighbor features refer to incorporation of neighborhood information by means of contextual words or distance. Numeric features capture information about a word being a number or an alphanumeric sequence.

Amongst the primary challenge in metadata extraction is information loss that occur while converting input from one format to another. Many PDF to text conversion libraries result into errors during the phase of conversion. These errors tend to affect the performance of extraction task as described in Kern et al. (2012). Pre-processing techniques required to transform scientific literature in PDF to text format are not part of this study. But, it is a crucial part for all studies dealing with PDF format. On the other hand, studies employing OCR to identify blocks from visual format tends to perform really well and usually exploit layout and font styling information to improve results.

In addition to various tools and many scientific research studies, semantic publishing challenge (ceurws/lod 2014) has also been introduced that deals with various type of insights extraction from scholarly data. These insights include quality analysis, metadata extraction and interlinking of information among scholarly data. A recent study (Dimou et al. 2017) compares various challenges regarding semantic publishing and is focused on analyzing the current trends in various semantic challenges. This study further consolidates various insights that are analyzed and studied in the light of conducting various semantic challenges. Study aims to improve the quality of organized challenges and workshops by means of employing learnt insights from previous experiences that include feedback incorporation, dataset updates, evolution in tasks etc.

As the scientific community is contributing into this domain for past many years. Thus, now there exist variety of open-source platforms that assist in automatically extracting this information from scientific articles. These systems currently suffer with the issues of layout and formatting primarily due to format conversion. Recent comparison study (Tkaczyk et al. 2018) shows that amongst the various open source extractors, GROBID, CERMINE and ParsCit presents best results.

Key-insights extraction

In scope of current paper, key-insights refer to any valuable information enclosed within a research paper's text that can be beneficial for researches. Key-insights refer to potential information nuggets contained in a scientific article. In literature, there exist wide terminologies to refer to similar concepts. (Augenstein et al. 2017) regard this task of key-insights extraction as Information Extraction, (QasemiZadeh and Schumann 2016) names this similar concepts as term recognition and classification. There exist other names as well including typed entity recognition, entity recognition, entity extraction, core scientific-concepts and argumentative zoning (Liakata et al. 2010; Tateisi et al. 2016). Examples of key-insights include underlying methodology or technique used, evaluation criteria,

Table 34 Phrase-level key-insights

Key-insight tag	Key-insight value
Domain	Computational modeling
Problem	Trust-based SA
Process	ABM, PBM

results, future work and limitations. These insights, if automatically retrieved, provide a researcher with a clear and concise concept of a research paper. This can be very fruitful for researchers who have to go through a bundle of research papers in order to have an idea about what is going on in their respective research domain. Table 34 presents the extracted key-insights from following passage taken from (Nasar and Jaffry 2018). Sentence-level insights are color-coded within passage where red presents **Aim**, Green presents **Goal** and blue presents **Extension**.

> Decisions and beliefs of human beings about surroundings and their environments are affected by their trust on other agents they are communicating with. Hence, in this study, primary aim is to extend computational model of SA presented in [2] to trust-based SA using ABM and PBM techniques. Keeping this in view, key goal of current research is to analyze the proposed model with both computational modeling paradigms i.e. ABM and PBM, along with a comparative analysis on the basis of their dynamics. Rest of the paper describes related background, outlines methodology opted to build the system that is an extension to a previous model proposed in [2], briefly explains the conducted experiments and respective results, followed by conclusion and future directions.

If such information is automatically extracted from scientific articles, it would aid in variety of applications including automated literature review, trend analysis and personalized research assistance. Thus, reset of this section is focused on presenting progress in this area. Following section firstly highlights major datasets available followed by state-of-the-art approaches that are being employed to perform key-insight extraction from scientific articles.

Datasets

Datasets for key-insights extraction can be majorly classified into two major classes: sentence-level and phrase-level. There exist multiple datasets for sentence-level key-insights extraction, but majority work done belong to the domain of medical sciences. In addition, there are two types of potential insights that are being annotated. One insight is regarding the potential named entities i.e. concepts such as domain, results, technique etc. Other insights are related to relation between entities. For example, a technique or algorithm is applied to solve a particular task. So a relation of application between a TECHNIQUE and TASK can be established namely Apply(TECHNIQUE, TASK). Similarly, results achieved against various evaluation measures can also be expressed as relations e.g. Result(F-measure, 98). There exist very few studies that are focused towards relation extraction between entities from scientific articles though. As relations are usually expressed between core concepts, therefore, phrase-level datasets can be extended to further have relation information as well. Whereas, sentence level datasets cannot be used for this purpose as sentence itself is composed of multiple entities.

Table 35 Annotation tags against Sentence-level Datasets

Dataset	Entity	Description
Section names (Hirohata et al. 2008)	Objective	It refers to the background and the aim of the scientific article.
	Method	It refers to the way to achieve the goal desired in paper.
	Result	It refers to the principle findings that are reported in study.
	Conclusion	It refers to analysis, discussion and the main conclusions presented in study.
Argumentative Zoning (Teufel and Moens 2002)	Background	It refers to the circumstances pertaining to the current work, situation and history etc.
	Objective	It refers to a thing aimed at or sought, a target or goal.
	Method	It refers to a way of doing research, esp. according to a defined and regular plan.
	Result	It refers to the effect, consequence, issue or outcome of an experiment.
	Conclusion	It refers to a judgment or statement arrived at by any reasoning process.
	Related work	It refers to a comparison between the current work and the related work.
	Future work	It refers to future directions presented in work.
Core Scientific Concepts (Liakata et al. 2012)	Hypothesis	It refers to a statement that has not been yet confirmed.
	Motivation	It refers to the reason for carrying out the investigation.
	Background	It refers to the description of accepted background knowledge and previous work.
	Goal	It refers to the target state of the investigation where intended discoveries are made.
	Object	It refers to an entity which is a product or main theme of the investigation.
	Experiment	It refers to experiment details.
	Model	It refers to a statement about a theoretical model or framework.
	Method	It refers to the means by which the authors seek to achieve a goal of the investigation.
	Observation	It refers to the data/phenomena recorded within an investigation.
	Result	It refers to factual statements about the outputs of an investigation.
	Conclusion	It refers to statements inferred from observations and results, relating to research hypothesis.
MAZEA (Dayrell et al. 2012)	Background	It refers to the context of the study, including any reference to previous work on the topic, relevance of the topic and main motivations behind the study.
	Gap	It refers to any indication that the researched topic has not been explored, that little is known about it, or that previous attempts to overcome a given problem or issue have not been successful
	Purpose	It refers to the intended aims of the paper or hypotheses put forward
	Method	It refers to the methodological procedures adopted as well as the description of the data/materials used in the study.
	Result	It refers to main findings presented in the paper.
	Conclusion	It refers to general conclusion of the paper; subjective opinion about the results, suggestions and recommendations for future work.

Table 35 continued

Dataset	Entity	Description
Dr. Inventor (Ronzano and Saggion 2015)	Challenge	It refers to the current situation faced by the researcher: it will normally include a Problem Statement, the Motivation, a Hypothesis and/or a Goal.
	Background	It refers to information which is helpful for understanding the situation or problem that is the subject of the publication.
	Approach	It refers to approach carried out by author to carry out the investigation.
	Outcome	It refers to results, discussions, contributions and conclusions presented in a scientific article.
	Future work	It refers to future directions presented in work.

(Teufel and Moens 2002) has applied concept of argumentative zoning in order to summarize biomedical articles. This annotation scheme is further extended in Teufel et al. (2009) by means of improved granularity. All existing tags except TEXTUAL are further classified into multiple categories. Hence, in improved scheme, there exist total of fifteen rhetorical classes for sentence-level key-insights extraction across full-length articles. This scheme is used to annotate articles from chemistry and computational linguistics domain. Results show that this annotation scheme can be used for data annotation by non-experts as well. This was established by making an expert, a semi-expert and a non-expert person to annotate articles and later by calculating agreement between them.

Research work to extract sentence-level insights from full-text of the article is performed as part of ART project (Liakata 2009). This project formed the basis of semantic annotation project (Liakata 2010), that is focused towards semantic annotation of scientific articles and its various applications are being studied in the domain of life sciences and cancer research (Guo et al. 2011; Liakata et al. 2012). Another notable addition for full-length sentence-based key insight dataset is Dr. Inventor framework (Ronzano and Saggion 2015), that carries total of forty articles belonging to computer imaging domain only. In addition to full-length articles set, many of the studies related to sentence level key-insights extraction are focused on abstracts.

A recent and diverse study in this regard is Multi-label Argumentative Zoning for English Abstracts (MAZEA) (Dayrell et al. 2012). This study has used total of 645 abstracts from Physical Sciences and Engineering (PE) and 690 abstracts from Life and Health Sciences (LH). Existing datasets for sentence-level key-insights tends to classify a sentence into one category. Primary contribution of this study is that it allows to assign multiple labels to a sentence; hence, multiple labels can be applied to a single sentence. The respective dataset is publicly available.[5] Widely used sentence-level annotation schemes that are applied in both abstract-only and full-length articles are presented in Table 35.

As far as entity level datasets are concerned, progress has been made in this direction recently. Pioneer study (Gupta and Manning 2011) in this regard comprises 475 abstracts from ACL. Another project named Term Entity Recognition (QasemiZadeh and Schumann 2016) is intended to perform task and entity recognition from ACL anthology corpus. This dataset (Handschuh and QasemiZadeh 2014) consists of three hundred annotated abstracts

[5] http://www.nilc.icmc.usp.br/mazea-web/.

from ACL paper collection, where publication year of respective article ranges from 1965 to 2006.

Entity and Relation extraction project (Tateisi et al. 2016) is focused on phrase level entity extraction from Japanese as well as English scientific-articles. For English, it uses total 400 abstracts from scientific-articles where 250 belong to ACL anthology corpus and remaining 150 from ACM digital library. Out of 250 articles from ACL corpus, 100 are randomly selected from Gupta-Manning dataset (Gupta and Manning 2011). Entities used in this project are inspired from internet artifact ontology (IAO) (IAO 2015). This study further extends the dataset by annotating relation information as well. Total of twenty distinct relations are being annotated in the underlying dataset. A base study regarding this dataset was carried out in Tateisi et al. (2014) with three primitive entities. The base study was only focused on Japanese articles and it dealt with sixteen distinct relation types.

Science-IE project was organized as part of Semantic-Evaluation (Sem-Eval) in 2017, where Sem-Eval is ongoing series of evaluations related to computational semantic analysis of systems, that is usually held on yearly basis. Science-IE (Augenstein et al. 2017) project is collaboration effort among various universities. This project is focused on annotation of scientific-articles belonging to three major domains that include material sciences, physical sciences and computer sciences. The data consists of 500 passages that are selected from open-access scientific-publications available on the research repository of ScienceDirect. The annotated dataset includes three entities namely Task, Process and Material. This study also includes two primitive relations that are "synonym-of" and "hyponym-of".

"Synonym-of" relation is being used to deal with abbreviations. For example, take this sentence: "This study is related to Information Extraction (IE)". Here, if "Information Extraction" is assigned any class, a relation of "synonym-of" should be expressed between "Information Extraction" and "IE". This will help in determining various mentions to a similar concept. "Hyponym-of" relation is used to describe hierarchy of objects. For Example: In sentence; "Apple is a fruit", apple is a hyponym-of fruit. Similarly, in context of scientific article, if sentence appears saying, "NER is a sub-task of IE", NER would be hyponym-of IE.

An alternate on-going effort in the direction of phrase-level key insights is the project of Information Retrieval Group at Iowa State University (Projects | ISU Information Retrieval Group 2017), which is related to automatic extraction of information from scientific-articles with primary focus on animal studies. Some phrase-level datasets along with the entities they cover and respective description of these entities are presented in Table 36.

Most of these datasets are recently developed. Therefore, there exists no substantial progress regarding algorithm application against these datasets. One thing to note here is that, in the domain of biology, there exist multiple resources and databases that help in identifying genes, proteins, diseases etc. Variety of datasets exist that are focused towards annotation of bio-centered entities such as gene–gene interaction, protein identification etc. Thus, there exist multiple studies that are focused on biology oriented information extraction from scientific articles (Friedman et al. 2001; Hirschman et al. 2005; Li et al. 2015) exploiting available information. In current review study, focus is to extract general phrase-level insights that are applicable and useful in other domains as well such as Problem, Domain, Process and Result etc. Hence, studies focused on bio-specific information extraction are not included in this study.

Table 36 Annotation tags against Phrase-level Datasets

Dataset	Entity	Description
Science-IE (Augenstein et al. 2017)	Task	It refers to smaller concrete research tasks (e.g. 'powder processing', 'dependency parsing') and broader research areas (e.g. 'machine learning'). Generally, these are problems tackled in a paper.
	Process	It refers to methods/techniques/algorithms, physical equipment and tools. Generally, these are solutions proposed to solve problems in a paper.
	Material	It refers to physical materials, datasets and corpora. These are resources studied in the paper or resources used to solve problems in a paper.
ACL-RD-TEC (QasemiZadeh and Schumann 2016)	Technology, system, and method	It refers to methods, processes, and approaches that are employed to solve practical tasks.
	Tool or library	It refers to an actual implemented technology.
	Language resource	It refers to components of natural language processing (NLP) systems that contain linguistic knowledge, for example, lexical databases, corpora, and so on.
	Language resource product	It refers to actual language resources. For example, "Princeton WordNet" is a lexical database which can be obtained and used in a project.
	Model	It refers to method-specific knowledge resources.
	Measures and measurements	They refer to components of evaluation systems used for measuring and measurement processes.
	Other nominals	Any category other than listed above (e.g., theories, formalism, linguistic entities).
Typed Entity and Relation Extraction from Japanese's and English articles (Tateisi et al. 2016)	Thing-occurrent-process	It refers to "Processual Entity" such as running, computation.
	Thing-occurrent-time	It refers to temporal information such as before 2013, waiting time.
	Thing-continuant-artifact	It refers to any physical object that is created for a purpose such as mobile devices.
	Thing-continuant-data-item	It refers to Data Item and textual entities in IAO.
	Thing-continuant-location	It refers to a spatial region such as United States, Asia, Space.
	Thing-continuant-person	It refers to individual or group of people.
	Thing-continuant-plan	It refers to "Processual Entity" that realizes a plan.
	Thing-continuant-quality	It refers to quality concept as described in IAO.
	Quantity	It refers to numbers, with or without units.

Table 36 continued

Dataset	Entity	Description
	Modality	It refers to terms expressing modality such as can, can't.
	Reference	It refers to anaphoric expressions such as it, they.
	External-reference	It refers to external literature reference or citation such as [1].
	Language	It refers to languages employed for inter-human communication such as English.
	Domain	It refers to primary area of study such as NLP, Bio-medicine.
	Organization	It refers to group of people that is established for a purpose such as MIT.
	Formula	It refers to any mathematical formulaic expression such as F = 90.
	Plan-or-process	It refers to an expression such as "web search" that can denote a process, a function that realizes the process, or steps of instructions to achieve the function.
	Judging-process	It refers to an expression that describes a system's behavior and also the author's subjective judgment, e.g., outperform in "The current system outperforms the baseline".
	Intelligent-agent	It refers to an expression that can be interpreted as people or artifacts/programs that emulate human behavior, e.g., players (of video games).
Japanese articles (Tateisi et al. 2014)	Object	It refers to the name of concrete entities such as a program, a person, and a company.
	Measure	It refers to the value, measurement, necessity, obligation, expectation, and possibility.
	Term	Other
ACL (Gupta and Manning 2011)	Focus	It refers to an article's main contribution.
	Technique	It refers to a method or a tool used in an article.
	Domain	It refers to an article's application domain, such as speech recognition and classification.

Approaches

In the past years, many researchers have contributed in domain of information extraction from research papers. Multiple Machine Learning and NLP techniques are used to extract key-insights from scientific literature. For sentence-level key-insights extraction, many studies make use of rule-based approaches. In addition, many machine learning approaches are also applied including Bayesian classifiers, CRFs and SVM. Due to unavailability of benchmarked datasets for phrase-level insights during past years, there exist not much development in this regard. Majority approaches for phrase-level insights extraction makes use of rule-based and CRF on self-generated datasets.

Rule-based approaches

A research study carried out in Hanyurwimfura et al. (2012) takes into account abstract and conclusion text along with some assumptions regarding the orientation of sentences in these two. It majorly relies on rule-based approach. Some examples for the used heuristics in this study are: words such as 'results', 'experiments' and 'evaluation' are used to represent result in a research article and phrases such as 'this paper', 'our approach' are used to represent main-idea of paper. In addition, title of the study as well as its authors are

also being extracted using simple heuristics. In order to determine results, experiment was conducted on 200 papers in group of 40 papers, which resulted in 89.4% precision and 91.2% recall. Apart from that, a survey on 20 papers is also conducted and extracted information was later manually evaluated which resulted in 7.75 ranks by readers (20 readers were employed in conduction of survey study) with range [0-10], with 10 being the highest.

Another study in this regard extract focus and technique as well along with domain (Gupta and Manning 2011) from scientific articles. In this study, pattern matching and dependency trees of sentences are being used along with seed-rules to identify focus, technique and domain. Later more patterns are being identified using bootstrapping approach. After extraction of focus, technique and domain concepts, LDA clustering (Blei et al. 2001) is performed in order to find topics. ACL anthology data-set (Bird et al. 2008) is used for evaluation. Four hundred and seventy four abstracts were hand-labeled for testing, which resulted in high recall and low precision.

Research study proposed in Houngb and Mercer (2012) primarily focuses on technique extraction from Biology Journals. Initially phrases are extracted containing Method-Mention terms such as algorithms, technique, method etc. Rules are formulated in order to extract such sentences from text and identifying the respective techniques used. Machine Learning techniques are also employed which makes use of word, POS, Word-shape (capitalized, start with capital letter, all lower case, all capital case, mixed case), Word-position (start of sentence, end of sentence, not beginning of sentence, not end of sentence), Token prefixes, Token suffixes, and Bigrams as features for CRF. Results are evaluated on two self-generated datasets. First dataset clearly mentions the method and consist of 918 sentences (dataset 1); whereas second one consists of 211 sentences (dataset 2) and does not contain method keyword. Each dataset contains pairs of sentences against every entry: where first sentence carries the method while other carries its potential usage. Later these sentences are tokenized and converted into BIO data tagging format for phrase-level method mention extraction. Results show Precision/Recall/F-measure of (85.40 100 91.89) and (81.8 75.00 78.26) against rule based system and CRF-based Machine Learning system respectively. Where, rule-based systems are being evaluated on dataset 1 whereas CRF system is being evaluated on dataset 2.

Machine-learning based approaches

Following section compiles the major approaches that employ machine learning concepts to perform key-insights extraction from scientific articles.

Naïve Bayes Pioneer study to perform sentence based classification from abstracts is presented in Teufel and Moens (2002). It makes use of Naïve-Bayes classification in order to classify sentences in aim, contrast, basis and background. In order to evaluate the system, total eighty conference articles from computational linguistics domain are annotated. Two types of evaluation are being performed, one deals with rhetoric classification performed using Naive-Bayes. Other is relevance based evaluation that tells that according to humans, how much relevant results are being extracted. Tags, their descriptions and respective evaluation measures are presented in Table 37.

A sentence-level key-insights extraction study in medical sciences is being proposed in Ruch et al. (2007). It makes use of Naïve-Bayes classifier in order to classify sentences from abstract in four categories including purpose, methods, conclusion and results.

Table 37 Evaluation measure using NB in Teufel and Moens (2002)

Field	Description	Rhetoric			Relevance	
		Prec.	Rec.	F	Prec.	Rec.
Aim	It refers to specific goal of current research paper	44	65	52	96.2	69.8
Contrast	It refers to statement of contrast with other's work	34	20	26	70.1	23.8
Textual	It refers to sentences stating structure of section	57	66	61	–	–
Own	It refers to description of own work	84	88	86	–	–
Background	It refers to generally accepted Scientific background	40	50	45	38.4	88.2
Basis	It refers to statement of agreement with other's work	37	40	38	70.5	39.4
Other	It refers to description of work done by others	52	39	44	–	–

Table 38 Evaluation measures against in Lin et al. (2006)

Section	Acc.	Prec.	Rec.	F1
Introduction	95.70	93.00	84.00	88.50
Methods	92.10	81.00	87.50	84.30
Results	92.10	89.80	89.80	89.80
Conclusions	96.30	89.80	89.60	89.70

Results show F-score of 85. The dataset used for evaluation comprises 12,000 abstracts from MEDELINE that carries implicit tags against these four categories.

In order to extract domains from research articles, study presented in Lakhanpal et al. (2015) makes use of preposition disambiguation. It relies on rules that are based on prepositions in a sentence. By following rules, phrases are identified which are later classified using Naïve-Bayes classification. Results shows 90% precision and 91% recall when applied on ACM SIGKDD (1995) papers from 2010–2014.

Hidden Markov model The study carried out in Lin et al. (2006) use HMM in order to assign rhetoric categories to sentences. The study is focused on medical abstracts, which generally follow the pattern of Introduction, Method, Result and Conclusion. Latent Discriminative Analysis (LDA) is also employed in order to further improve the performance. Multiple experiments are performed where HMM with LDA performed best against abstracts selected from MEDELINE. Respective evaluation measure against best approach is presented in Table 38.

Another study that employs HMM (Wu et al. 2006) is used in order to extract Move structures. Move structures refer to the categories of functional roles. These structures include Background, Purpose, Method, Result and Conclusion. Total 709 sentences are tagged that belong to 106 abstracts from CiteSeer. Study tends to exploit Move-constructs and collocation information to improve HMM model. This approach results into best precision of 80.54.

Conditional random fields The work presented in Hirohata et al. (2008) is focused on extraction of section related information from article abstracts. It makes use of CRFs in order to identify sentences from abstract against major sections that include Objective, Method, Result and Conclusion. In order to develop the model, corpus of 51,000 abstracts is developed. The corpus consists of abstracts that have the exact four section labels. The proposed method achieved 95.5% per-sentence accuracy and 68.8% per-abstract accuracy.

Table 39 Evaluation measures against Lin et al. (2010)

Entity	Description	Prec.	Rec.	F1
Author	It refers to the author names of a scientific article.	89.00	85.30	87.1
E-mail	It refers to corresponding authors' email addresses of research article.	100.0	97.30	98.6
Institution	It refers to the author's affiliations.	91.30	78.00	84.1
Macro-average against metadata entries		93.40	86.60	89.9
Key study parameters				
Age group	It refers to the age range of the subjects of the study.	64.30	35.40	45.70
Data analysis name	It refers to the name of the method or software used in the analysis of data collected for the study.	79.30	37.20	50.60
Data collection method	It refers to the data collection methods for the study.	20.00	01.60	02.90
Database name	It refers to the name of any biomedical databases used or mentioned in the study.	42.50	10.50	16.80
Data type	It refers to the type of data involved in the study.	70.00	19.70	30.70
Geographical area	It refers to the names of the geographical area in which an experiment takes place or the subjects are from.	43.70	10.40	16.80
Intervention	It refers to the name of medical intervention used in the study.	40.00	02.70	05.10
Number of observations	It refers to the number of cases or subjects observed in the study.	43.40	10.70	17.10
Time period	It refers to the duration of an experiment or observation in the study.	82.70	69.40	75.50
Macro-average against key study parameters		48.5	19.7	26.1

A research study proposed in Kondo et al. (2009) analyze research paper's title in order to identify underlying Technique and Research Field of the respective research paper. In order to extract the desired fields, firstly Cue words are identified using Rule-based approaches. Later these words are searched in research paper's titles. This helps in identifying research paper Goal, underlying Methodology and major Topic or Research Field of paper. CRF are used with word POS, word being a Method, Goal or Head word as features in order to classify the identified words into their respective classes. Experiments were performed on Japanese and English literature, which resulted into 82.5% precision and 81.6% recall for the Japanese research papers, while for English literature it resulted in scores of 73.5% precision and 78% recall.

The study presented in Lin et al. (2010) also makes use of CRF in order to extract metadata information as well as key-insights information from medical articles. This metadata is regarded as formulaic author metadata in this study. It includes Author Name, Email and Institution. For key-insights, this study extracts entities that are part of full-text and depict information related to nature of study. In order to perform the training and later evaluation, gold set is prepared by means of annotating 185 open-accesses PubMed articles. This article set belongs to studies performed from 2008 to 2009 and strictly consist of research articles excluding any reviews, case-studies, editorials and perspectives. Annotators were provided with Rich Text Format (RTF), generated by means of processing respective HTML version of research article, along with primitive annotation guidelines. Results show that CRF is very effective in determining formulaic author metadata with

Table 40 Evaluation metrics against Kovačević et al. (2012)

	Prec.	Rec.	F1
Task	69.59	43.25	53.35
Implementation	86.11	69.33	76.82
Method	70.46	42.51	53.03
Resource	67.61	55.39	60.89

Table 41 F-measures against various annotation schemes in Guo et al. (2010)

Section names			AZ		CoreSC	
	NB	SVM	NB	SVM	NB	SVM
Accuracy	82.00	89.00	76.00	90.00	53.00	81.00
Background	–	–	79.00	94.00	–	62.00
Objective	85.00	90.00	25.00	88.00	–	85.00
Method	75.00	81.00	70.00	85.00	32.00	89.00
Result	85.00	90.00	83.00	91.00	59.00	86.00
Conclusion	71.00	90.00	66.00	88.00	62.00	87.00
Observation	–	–	–	100.0	61.00	82.00
Motivation	–	–	–	–	56.00	82.00
Goal	–	–	–	–	–	62.00
Experiment	–	–	–	–	30.00	70.00
Future-work	–	–	–	100.0	–	–

average F-score of 89.9%, whereas key-insights extraction shows relatively poor performance with 26.1% F-measure as shown in Table 39.

Another study that performs both sentence level and phrase level KIE from articles is carried out in (Kovacevic et al. 2012). This study makes use of various features to perform extraction. First of all, sentence level extraction is made using similar annotation scheme and categories as used in (Teufel and Moens 2002). After primary classification of sentences, the sentences of OWN category are further sub-divided into results, solution and else category. Further, solution category's sentences are later annotated to extract phrase level concepts including method, task, tools and resources. This classification is being performed by means of CRF. The evaluation metrics against these insights are being presented in Table 40. Study has rigorously experimented with various features. In the light of results, all entries except resources tend to perform optimally when all features are incorporated. These features include lexical, syntactic, citation and frequency features.

Support vector machines A relevant study that is focused towards sentence extraction from scientific articles is presented in (Guo et al. 2010). It performs comparative analysis between three various annotation schemes for sentence-level key-insights extraction. These schemes are based on section names (Hirohata et al. 2008), argumentative zones (AZ) and core-scientific concepts (CoreSC). Later two schemes are associated with ART project (Liakata 2009), a pioneer project to deal with sentence-level key-insights extraction from full-text scientific articles of medical sciences. This proposed approach makes use of Naïve-Bayes and SVM classifiers to perform IE from abstracts only. Results show that SVM presents better results than Naïve-Bayes classifier as shown in Table 41.

Table 42 Results against Tateisi et al. (2016)

	Entity			Relation		
Dataset	Prec.	Rec.	F1	Prec.	Rec.	F1
Random 250	62.90	62.80	62.90	54.30	45.20	49.30
Gupta-manning	68.00	70.60	69.30	41.60	52.30	46.30

An SVM based solution is presented in (Ronzano and Saggion 2015) to extract sentence-level key-insights. Linear kernel was used for training. As for the data, total 40 computer graphics paper from the Dr. Inventor Rhetorically Annotated Corpus (Fisas et al. 2015) containing total of 8877 sentences were used. Annotation categories for sentences are almost same as followed by the ART project. All the sentences of the Corpus have been manually characterized by three annotators with inter-annotator agreement of 65.67%. Proposed SVM model takes into account both lexical and syntactic features to model each sentence. Java based machine learning library of Weka2.0 is used to perform all the tasks related to rhetorical sentence classification. The model resulted into F1 score equal to 76.4 against a tenfold cross validation.

Others There are several studies that either use hybrid approach to perform metadata extraction or makes use of other techniques that cannot be classified in existing techniques as in stated above sections. This sub-section highlights such studies.

The baseline results against Typed Entity and Relation Extraction project (Tateisi et al. 2016) are calculated using joint modeling approach presented in (Miwa and Sasaki 2014). This approach uses tables in order to maintain history. Tables are filled using history based approach where every cell is mapped with labels. In order to map sequence to tables, tables are firstly transformed into one dimensional form using static ordering. Preceding assignments in cells are taken into account while adding labels in the cells in order to avoid any illegal assignment. A structured learning approach using margin is used in order to learn the weights and multiple training algorithms are employed including Perceptron, AdaGrad and SVM (Chang and Yih 2013; Collins 2002; Duchi et al. 2011; Mejer and Crammer 2010). These weights help in mapping entities and relations into a table. As this dataset contains total 400 articles from ACM and ACL, where 100 articles belong to Gupta-Manning dataset (Gupta and Manning 2011). Results against 10-cross validation for randomly selected 250 articles excluding Gupta-Manning as well as results against Gupta-Manning dataset only are reported in Table 42. Annotated dataset against Japanese and English scientific articles is publicly available.

ScienceIE project was conducted as Sem-Eval Task in 2017. Against the developed dataset in ScienceIE, a competition was held. This competition has total of three evaluation scenarios. First scenario was focused on information extraction when only plain text of scientific article's content is provided. Second scenario provides additional key-phrases along with plain-text. Third scenario provides partial information regarding key-phrases along with their rhetorical class i.e. Task, Process and Material. Various groups participated in this project to compete. Hybrid models of recurrent neural networks with CRF performed best, with maximum F-measure score of 43 against first evaluation scenarios. Lexical feature based SVM model provided maximum F-measure of 64 in second evaluation scenario. For third evaluation scenario, convolution neural network based approach performed better than the rest with F-measure of 64. Detail of overall evaluation and subtasks involved along with dataset is publicly available.

Table 43 Summary against key-insights extraction from scientific articles

References	Level	Origin	Approach	Domain	Size	Entities	Metrics
Houngb and Mercer (2012)	Phr	–	CRF, Rule	BL	918,211*	Meth	P, R, F
Gupta and Manning (2011)	Phr	Abs	Boot-Strapping, rule	CL	474	Dom, focus, tech	F
Tateisi et al. (2016)	Phr	Abs	Joint modeling	CS	400	Typed entity **	P, R, F
Hanyurwimfura et al. (2012)	Phr	Abs, Conc	Rule	–	200	Main content, Res	P, R, F, SE
Lin et al. (2010)	Phr	FA	CRF	HS	185	Metadata, Subj Props	P, R, F
ScienceIE (Sem-Eval 2017)	Phr	A Para	Hybrid Approaches	PS, MS, CS	500	Task, process, material	P, R, F
(Kondo et al. 2009)	Phr	Title	CRF	–	–	Tech	P, R
Lakhanpal et al. (2015)	Phr	Title, Abs, KW	NB	CS	272	Dom	P, R, A
Kovačević et al. (2012)	Phr, Sen	FA	CRF	CS	45	AZ**, Resource, Implementation, Meth, Task	P, R, F
Hirohata et al. (2008)	Sen	Abs	CRF	MeS	51,000	Obj, Meth, Res, Conc	A
Wu et al. (2006)	Sen	Abs	HMM	CS	106	Back, Purp, Meth, Res Conc	P
Lin et al. (2006)	Sen	Abs	HMM, LDA	MeS	–	Intro, Meth, Res, Conc	P, R, F, A
Ruch et al. (2007)	Sen	Abs	NB	BL	12,000	Purp, Meth, Res, Conc	F
Guo et al. (2010)	Sen	Abs	SVM	BM	1000	AZ, CoreSC**	F
Teufel and Moens (2002)	Sen	FA	NB	CL	80	Aim, Contrast, Basis, Back	P, R
Ronzano and Saggion (2015)	Sen	FA	SVM	CV	40	Chall, Back, App, Res, FW	F

Meth method, *Dom* domain, *Tech* technique, Res result, *Obj* objective, *Back* background, *Chall* challenge, *App* approach, *FW* futurework

*Represents no. of sentences

**Represents annotation schemes/datasets names applied here

Table 44 Available datasets for KIE

Name	Domain	Size	Links
Sonal-Gupta	Computer science	475	https://nlp.stanford.edu/pubs/FTDDataset_v1.txt
ACL-RD-TEC	Computational linguistics	–	https://github.com/languagerecipes/the-acl-rd-tec
ScienceIE	Physical, material and computer sciences	500	https://scienceie.github.io/resources.html
Typed Entity Recognition	Computer science	400	https://github.com/mynlp/ranis/tree/master/EN/data
Dr. Inventor	Computer vision	41	http://sempub.taln.upf.edu/dricorpus#download (on demand)
ART project	Physical and biochemistry	225	http://www.aber.ac.uk/en/cs/research/cb/projects/art/art-corpus/

Conclusion

In the light of literature reviewed regarding KIE from scientific articles, a comprehensive summary is presented in Table 43. Reference entry in table header represents respective research study. Level entry represents that which type of information is being extracted i.e. phrase-level (Phr) or sentence level (Sen). Origin refers to the data sections taken from an article including abstract (abs), conclusion (Con), keywords (KW), Full-article (FA) etc. Approach refers to algorithm(s) applied to perform desired extraction from data. Domain refers to the area of study that is selected for evaluation e.g. computer science, physical science, etc. Size refers to total number of articles/abstracts included in a study, whereas exceptions are marked with asterisk (*) and represents number of sentences. Entities represent type of key-insights that are being extracted in a research study. Lastly, metric represents evaluation measure(s) applied to report results respectively. Here, in metrics column: A, P, R, F, SE represents Accuracy, Precision, Recall, F1-score and Subjective Evaluation measures respectively. In domain header: BL, CL, CS, HS, PS, MS, MeS, BM and CV represent biology, computational linguistics, computer science, health sciences, physical sciences, material sciences, medical sciences, biomedicine and computer vision respectively.

In the light of Table 43, it is very much evident that majority of work has been reported on abstracts only. The primary issue of very limited studies on full-text article can be the complexity involved regarding annotation task. As entities grow, the time of annotating an article can exponentially explode while dealing with full-text scientific articles. Even, in case of abstract, fine-grained annotation can take lot of time as reported in (Augenstein et al. 2017) due to subjectivity of classes at hand. This time can be saved by using crisp annotation guidelines. As "Datasets" section points recent contributions regarding annotation guidelines and datasets for KIE, progress is yet to be made to perform phrase-level KIE on full-text scientific articles. Table 44 compiles all open source datasets along with their details.

Conclusion and future work

This study is focused towards determining state-of-the-art regarding potential information that can be extracted from scientific articles. As a scientific article follows a semi-structured format. Therefore, on the basis of its structure: information to be extracted from an article is broadly classified into two major categories namely metadata extraction (ME) and key-insights extraction (KIE). ME from scientific articles refers to identification and extraction of metadata elements such as Title, Author, Affiliations etc. In order to perform ME, there exist multiple datasets that vary on the basis of article's sources, publication venues, data size and granularity of fields. On these datasets, multiple approaches including Rule based approaches and machine learning approaches including HMM, CRF and SVM are applied. Amongst these, CRF tends to outperform other approaches with reported F-measure of more than 0.95. Currently, deep learning approaches are not being widely employed to perform ME. As, hybrid deep learning frameworks are performing really well and currently governing the state-of-the-art in general Information Extraction tasks. Thus, application of deep learning frameworks and their hybrid versions are an open-area in context of ME. Apart from various techniques and datasets, there are variety of open-source tools that aids in automatic extraction of meta-data entities from research articles' header as well as bibliography. One of the primary challenges in ME is to minimize information loss while converting scientific article from one format to another.

As far as KIE is concerned, there exist two broad classifications regarding insights to be extracted namely sentence-level key-insights and phrase-level key-insights. Sentence-level KIE processes are focused on classification of sentences in pre-defined categories based on insights they carry. Widely used approaches to perform sentence-level KIE include Rule-based approaches, Bayesian classification, SVM and CRFs. Majority of work regarding sentence-level KIE is based on medical studies. Although, there are a couple of studies that perform sentence level KIE on full-length articles, most developments in this area are based on articles' abstracts only.

Phrase-level KIE, on the other hand, is focused towards extraction of phrases carrying potential information. Mostly work done with phrase-level KIE; such as Problem, Domain, Technique, Results etc. is reported on self-created datasets that are not publicly available. In addition, the guidelines and inter-annotator agreements while developing these datasets are also not reported. Apart from that, various other limitations in existing studies were found which include lack of proper dissemination of achieved results; lack of expression regarding the methodology used to perform desired task; ambiguity in explanation of the corpus used for data evaluation and deficiency in performing cross-validation for various techniques while reporting results (Houngb and Mercer 2012; Kavila and Rani 2016).

Regarding available datasets for phrase-level key-insights; in past several years, researchers have been working to create benchmark datasets to extract phrase-level key-insights. There exist wide varieties of key-insights that are being annotated in currently available datasets. Some are specific to a domain such as computational linguistics; others are generic and cover a variety of disciplines. One of the major limitations of existing phrase-level annotated datasets is that they only consist of a single passage. Another challenge is the unavailability of crisp definitions for various key-insights. This gives rise to subjective notions across phrase-level key-insights that are being identified in various datasets. Therefore, in order to minimize the subjective individual biases regarding any key-insight, respective definitions should be crisp and clear. Hence, primary open research task with regard to phrase-level key-insights dataset is identification of specific concepts or

key-insights to be extracted from scientific articles. Once these are identified, next question would be to devise the criterion that helps in determining particular phrase as a key-insight. These criteria will eventually help in development of annotation guidelines. Once annotation guidelines are developed, next major contribution would be to prepare a dataset in light of these guidelines.

Additionally, as majority of phrase-level datasets are developed recently, therefore, a great deal of development is required in order to efficiently extract the potential information insights followed by relation extraction (RE) between extracted conceptual insights. In scientific articles, relation can express application of a technique to solve a problem, results generated against various evaluation measures etc. This information can serve multiple benefits such as ontology construction and question answering systems. Hence, datasets preparation and algorithm development for RE is an open research area as well. Other open research questions include analysis and application of various state-of-the-art IE approaches on various existing datasets. These analyses will further reveal the potential advantages and pitfalls of existing techniques. CRF is generally regarded as the state-of-the-art statistical technique for ME, but recently, after identification of its limitations in one of the dataset, several research studies were carried out to improve those limitations (Anzaroot et al. 2014; Vilnis et al. 2015). Similarly, by acquiring brief understanding after application of existing solutions on KIE, analysis of primary reasons for achieved results followed by ways to improve and mitigate the identified challenges remains an open research area.

Regarding primary limitations of current survey study, it only contains those articles that are focused on extracting generic insights from scientific articles. Thus, articles focused on key-insights extraction specific to any domain are not catered. Furthermore, pre-processing techniques that are applied to convert data from one format to another as well as to generate textual, layout, and formatting features are not part of study.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Abdelmagid, M., Himmat, M., & Ahmed, A. (2014). Survey on information extraction from chemical compound literatures: Techniques and challenges. *Journal of Theoretical and Applied Information Technology, 67*(2), 284–289.

Adefowoke Ojokoh, B., Sunday Adewale, O., & Oluwole Falaki, S. (2009). Automated document metadata extraction. *Journal of Information Science, 35*(5), 563–570. https://doi.org/10.1177/0165551509105195.

Alam, H., Kumar, A., Werner, T., & Vyas, M. (2017). Are cited references meaningful? Measuring semantic relatedness in citation analysis. In *BIRNDL@SIGIR (1)* (Vol. 1888, pp. 113–118). CEUR-WS.org.

An, D., Gao, L., Jiang, Z., Liu, R., & Tang, Z. (2017). Citation Metadata Extraction via Deep Neural Network-based Segment Sequence Labeling. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 1967–1970). New York, NY, USA: ACM. https://doi.org/10.1145/3132847.3133074.

Anzaroot, S., & Mccallum, A. (2013). A new dataset for fine-grained citation field extraction. In *ICML Workshop on Peer Reviewing and Publishing Models,* 2013.

Anzaroot, S., Passos, A., Belanger, D., & McCallum, A. (2014). Learning soft linear constraints with application to citation field extraction. arXiv:1403.1349 [Cs]. Retrieved from http://arxiv.org/abs/1403.1349.

Atdağ, S., & Labatut, V. (2013). A comparison of named entity recognition tools applied to biographical texts. In *2nd International conference on systems and computer science* (pp. 228–233). https://doi.org/10.1109/IcConSCS.2013.6632052.

Augenstein, I., Das, M., Riedel, S., Vikraman, L., & McCallum, A. (2017). SemEval 2017 task 10: ScienceIE—extracting keyphrases and relations from Scientific Publications. arXiv:1704.02853 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1704.02853.

Baum, L. E. (1972). an inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In O. Shisha (Ed.), *Inequalities III: Proceedings of the third symposium on inequalities* (pp. 1–8). Los Angeles: University of California.

Beel, J., Langer, S., Genzmehr, M., & Müller, C. (2013). Docear's PDF inspector: title extraction from PDF files. In *Proceedings of the 13th ACM/IEEE-CS joint conference on digital libraries* (pp. 443–444). ACM Press. https://doi.org/10.1145/2467696.2467789.

Bird, S., Dale, R., Dorr, B. J., Gibson, B., Joseph, M. T., Kan, M.-Y., & Tan, Y. F. (2008). The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Sixth International Conference On Language Resources And Evaluation (LREC'08)*, 2008, pp. 1755–1759.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2001). Latent dirichlet allocation. In *Advances in neural information processing systems* (pp. 601–608).

Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *British Medical Journal Open, 7*(2), e012545. https://doi.org/10.1136/bmjopen-2016-012545.

Britz, D. (2015). Recurrent neural network tutorial, part 4—implementing a GRU/LSTM RNN with python and theano. Retrieved August 16, 2017, from http://www.wildml.com/2015/10/recurrent-neural-network-tutorial-part-4-implementing-a-grulstm-rnn-with-python-and-theano/.

Ceurws/lod. (2014). Retrieved August 6, 2018, from https://github.com/ceurws/lod.

Chang, M.-W., & Yih, W. (2013). Dual coordinate descent algorithms for efficient large margin structured prediction. *Transactions of the Association for Computational Linguistics, 1,* 207–218.

Chen, C.-C., Yang, K.-H., Chen, C.-L., & Ho, J.-M. (2012). BibPro: A citation parser based on sequence alignment. *IEEE Transactions on Knowledge and Data Engineering, 24*(2), 236–250.

CiteSeerX. (2007). Retrieved January 20, 2018, from http://citeseerx.ist.psu.edu/index.

Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on empirical methods in natural language processing-volume 10* (pp. 1–8). Association for Computational Linguistics.

Constantin, A., Pettifer, S., & Voronkov, A. (2013). PDFX: Fully-automated PDF-to-XML conversion of scientific literature. In *Proceedings of the 2013 ACM symposium on document engineering* (pp. 177–180). New York, NY, USA: ACM. https://doi.org/10.1145/2494266.2494271.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018.

Cortez, E., da Silva, A. S., Gonçalves, M. A., Mesquita, F., & de Moura, E. S. (2007). FLUX-CIM: Flexible unsupervised extraction of citation metadata. In *Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries* (pp. 215–224). New York, NY, USA: ACM. https://doi.org/10.1145/1255175.1255219.

Cortez, E., da Silva, A. S., Gonçalves, M. A., Mesquita, F., & de Moura, E. S. (2009). A flexible approach for extracting metadata from bibliographic citations. *Journal of the American Society for Information Science and Technology, 60*(6), 1144–1158. https://doi.org/10.1002/asi.v60:6.

Councill, I., Giles, C. L., & Kan, M.-Y. (2008). 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: European Language Resources Association (ELRA). Retrieved August 29, 2016, from http://www.lrec-conf.org/proceedings/lrec2008/pdf/166_paper.pdf.

Cui, B. (2009). Scientific literature metadata extraction based on HMM. In Y. Luo (Ed.), *Cooperative design, visualization, and engineering* (Vol. 5738, pp. 64–68). Berlin: Springer. Retrieved December 4, 2017, from http://link.springer.com/10.1007/978-3-642-04265-2_9.

Cui, B.-G., & Chen, X. (2010). An improved hidden Markov model for literature metadata Extraction. In D.-S. Huang, Z. Zhao, V. Bevilacqua, & J. C. Figueroa (Eds.), *Advanced intelligent computing theories and applications* (Vol. 6215, pp. 205–212). Berlin: Springer. Retrieved December 26, 2017, from http://link.springer.com/10.1007/978-3-642-14922-1_26.

Cuong, N. V., Chandrasekaran, M. K., Kan, M.-Y., & Lee, W. S. (2015). Scholarly document information extraction using extensible features for efficient higher order semi-CRFs. In *Proceedings of the 15th ACM/IEEE-CS joint conference on digital libraries* (pp. 61–64). ACM.

Day, M.-Y., Tsai, R. T.-H., Sung, C.-L., Hsieh, C.-C., Lee, C.-W., Wu, S.-H., et al. (2007). Reference metadata extraction using a hierarchical knowledge representation framework. *Decision Support Systems, 43*(1), 152–167. https://doi.org/10.1016/j.dss.2006.08.006.

Dayrell, C., Candido, A., Lima, G., Machado, D., Copestake, A. A., Feltrim, V. D., & Aluísio, S. M. (2012). Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora. In *LREC*.

de Price, D. S. (1961). *Science since babylon*. New Haven: Yale University Press.

Dimou, A., Vahdati, S., Iorio, A. D., Lange, C., Verborgh, R., & Mannens, E. (2017). Challenges as enablers for high quality linked data: Insights from the semantic publishing challenge. *PeerJ Computer Science, 3,* e105. https://doi.org/10.7717/peerj-cs.105.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research, 12*(Jul), 2121–2159.

Duck, G., Nenadic, G., Filannino, M., Brass, A., Robertson, D. L., & Stevens, R. (2016). a survey of bioinformatics database and software usage through mining the literature. *PLoS ONE, 11*(6), e0157989. https://doi.org/10.1371/journal.pone.0157989.

Fisas, B., Saggion, H., & Ronzano, F. (2015). On the discoursive structure of computer graphics research papers. In *Proceedings of the 9th linguistic annotation workshop* (pp. 42–51).

Flynn, P., Zhou, L., Maly, K., Zeil, S., & Zubair, M. (2007). Automated template-based metadata extraction architecture. In *Proceedings of the 10th international conference on Asian digital libraries: Looking back 10 years and forging new frontiers* (pp. 327–336). Berlin: Springer. Retrieved December 26, 2017, from http://dl.acm.org/citation.cfm?id=1780653.1780708.

Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE, 61*(3), 268–278.

Friedman, C., Kra, P., Yu, H., Krauthammer, M., & Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. In *ISMB* (*supplement of bioinformatics*) (pp. 74–82).

Giuffrida, G., Shek, E. C., & Yang, J. (2000). Knowledge-based metadata extraction from PostScript files. In *Proceedings of the fifth ACM conference on digital libraries* (pp. 77–84). ACM.

Granitzer, M., Hristakeva, M., Knight, R., Jack, K., & Kern, R. (2012). A comparison of layout based bibliographic metadata extraction techniques. In *ACM international conference proceeding series*. Retrieved August 3, 2018, from www.scopus.com.

Groza, T., Handschuh, S., & Hulpus, I. (2009). *A document engineering approach to automatic extraction of shallow metadata from scientific publications* (technical report no. 2009- 06-01). Digital Enterprise Research Institute.

Guo, Z., & Jin, H. (2011). Reference metadata extraction from scientific papers. In *Proceedings of the 2011 12th international conference on parallel and distributed computing, applications and technologies* (pp. 45–49). Washington, DC, USA: IEEE Computer Society. https://doi.org/10.1109/PDCAT.2011.72.

Guo, Y., Korhonen, A., Liakata, M., Karolinska, I. S., Sun, L., & Stenius, U. (2010). Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *Proceedings of the 2010 workshop on biomedical natural language processing* (pp. 99–107). Association for Computational Linguistics.

Guo, Y., Korhonen, A., Liakata, M., Silins, I., Hogberg, J., & Stenius, U. (2011). A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. *BMC Bioinformatics, 12*(1), 69.

Gupta, S., & Manning, C. (2011). Analyzing the dynamics of research by extracting key aspects of scientific papers (pp. 1–9). In *Proceedings of 5th international joint conference on natural language processing, asian federation of natural language processing*. Retrieved November 27, 2015, from http://aclasb.dfki.de/nlp/bib/I11-1001.

Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., & Fox, E. A. (2003). Automatic document metadata extraction using support vector machines (pp. 37–48). IEEE Computer Society. https://doi.org/10.1109/JCDL.2003.1204842.

Handschuh, S., & QasemiZadeh, B. (2014). The ACL RD-TEC: a dataset for benchmarking terminology extraction and classification in computational linguistics. In *COLING 2014: 4th international workshop on computational terminology*.

Hanyurwimfura, D., Bo, L., Njogu, H., & Ndatinya, E. (2012). An automated cue word based text extraction. *Journal of Convergence Information Technology, 7*(10), 421–429. https://doi.org/10.4156/jcit.vol7.issue10.50.

Harkema, H., Roberts, I., Gaizauskas, R., & Hepple, M. (2005). Information extraction from clinical records. In *Proceedings of the 4th UK e-science all hands meeting*.

Haruna, K., Ismail, M. A., Damiasih, D., Sutopo, J., & Herawan, T. (2017). A collaborative approach for research paper recommender system. *PLoS ONE, 12*(10), e0184516. https://doi.org/10.1371/journal.pone.0184516.

Hetzner, E. (2008). A simple method for citation metadata extraction using hidden Markov models. In *Proceedings of the 8th ACM/IEEE-CS joint conference on digital libraries* (pp. 280–284). New York, NY, USA: ACM. https://doi.org/10.1145/1378889.1378937.

Hirohata, K., Okazaki, N., Ananiadou, S., & Ishizuka, M. (2008). Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the third international joint conference on natural language processing: volume-I*.

Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005). *Overview of BioCreAtIvE: critical assessment of information extraction for biology*. BioMed Central.

Houngb, H., & Mercer, R. E. (2012). Method mention extraction from scientific research paper. In *Proceedings of COLING 2012: Technical paper* (pp. 1211–1222).

Huang, Z., Jin, H., Yuan, P., & Han, Z. (2006). Header Metadata Extraction from Semi-structured Documents Using Template Matching. In *Proceedings of the 2006 international conference on on the move to meaningful internet systems: AWeSOMe, CAMS, COMINF, IS, KSinBIT, MIOS-CIAO, MONET-volume part II* (pp. 1776–1785). Berlin: Springer. https://doi.org/10.1007/11915072_84.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv:1508.01991 [Cs]. Retrieved from http://arxiv.org/abs/1508.01991.

IAO (2015): Information artifact ontology. Web ontology language, IAO. Retrieved March 28, 2018, from https://github.com/information-artifact-ontology/IAO.

Insights, E. (2013). Using citation analysis to measure research impact. *Editage Insights* (*04-11-2013*). Retrieved December 26, 2017, from http://www.editage.com/insights/using-citation-analysis-to-measure-research-impact.

Juang, B. H., & Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics, 33*(3), 251–272.

Kan, M.-Y., Luong, M.-T., & Nguyen, T. D. (2010). Logical structure recovery in scholarly articles with rich document features. *International Journal of Digital Library Systems, 1*(4), 1–23. https://doi.org/10.4018/jdls.2010100101.

Kanya, N., & Ravi, T. (2012). Modelings and techniques in named entity recognition-an information extraction task. In *IET Chennai 3rd international on sustainable energy and intelligent systems* (*SEISCON 2012*) (pp. 1–5). https://doi.org/10.1049/cp.2012.2199.

Kavila, S. D., & Rani, D. F. (2016). Information extraction from research papers based on statistical methods. In S. C. Satapathy, K. S. Raju, J. K. Mandal, & V. Bhateja (Eds.), *Proceedings of the second international conference on computer and communication technologies* (Vol. 381, pp. 573–580). New Delhi: Springer. Retrieved from April 20, 2018, http://link.springer.com/10.1007/978-81-322-2526-3_59.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization (Vol. 4, pp. 1942–1948). In *Proceedings of IEEE international conference on neural networks*. Piscataway, New Jersey: IEEE. https://doi.org/10.1109/ICNN.1995.488968.

Kern, R., Jack, K., & Hristakeva, M. (2012). TeamBeam—meta-data extraction from scientific literature. *D-Lib Magazine*. https://doi.org/10.1045/july2012-kern.

Klink, S., Dengel, A., & Kieninger, T. (2000). Document structure analysis based on layout and textual features. In *Proceedings of international workshop on document analysis systems, DAS2000* (pp. 99–111). IAPR.

Knoth, P., Anastasiou, L., Charalampous, A., Cancellieri, M., Pearce, S., Pontika, N., & Bayer, V. (2017). Towards effective research recommender systems for repositories. *ArXiv Preprint* arXiv:1705.00578.

Kondo, T., Nanba, H., Takezawa, T., & Okumura, M. (2009). Technical trend analysis by analyzing research papers' titles. In *Proceeding LTC'09 proceedings of the 4th conference on human language technology: Challenges for computer science and linguistics* (pp. 512–521). Retrieved from http://dl.acm.org/citation.cfm?id=1987773.

Kovačević, A., Ivanović, D., Milosavljević, B., Konjović, Z., & Surla, D. (2011). Automatic extraction of metadata from scientific publications for CRIS systems. *Program*. Retrieved January 12, 2016, from http://www.emeraldinsight.com/doi/full/10.1108/00330331111182094.

Kovačević, A., Konjović, Z., Milosavljević, B., & Nenadic, G. (2012). Mining methodologies from NLP publications: A case study in automatic terminology recognition. *Computer Speech & Language, 26*(2), 105–126. https://doi.org/10.1016/j.csl.2011.09.001.

Lakhanpal, S., Gupta, A., & Agrawal, R. (2015). Towards extracting domains from research publications. *Presented at the 26th modern artificial intelligence and cognitive science conference, MAICS 2015*.

Retrieved November 27, 2015, from https://ncatsu.pure.elsevier.com/en/publications/towards-extracting-domains-from-research-publications.
Lee, C. (2017). LSTM-CRF models for named entity recognition. *IEICE Transactions on Information and Systems, 100*(4), 882–887.
Li, G., Ross, K. E., Arighi, C. N., Peng, Y., Wu, C. H., & Vijay-Shanker, K. (2015). miRTex: A text mining system for miRNA-gene relation extraction. *PLoS Computational Biology, 11*(9), e1004391. https://doi.org/10.1371/journal.pcbi.1004391.
Liakata, M. (2009). Aberystwyth University—ART. Retrieved Feb 12, 2018, from https://www.aber.ac.uk/en/cs/research/cb/projects/art/.
Liakata, M. (2010). Home. Retrieved April 20, 2018, from http://www.sapientaproject.com/.
Liakata, M., Saha, S., Dobnik, S., Batchelor, C., & Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics, 28*(7), 991–1000. https://doi.org/10.1093/bioinformatics/bts071.
Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C. R., & others. (2010). Corpora for the conceptualisation and zoning of scientific papers. In *LREC*. Citeseer.
Lin, J., Karakos, D., Demner-Fushman, D., & Khudanpur, S. (2006). Generative content models for structural analysis of medical abstracts. In *Proceedings of the workshop on linking natural language processing and biology: Towards deeper biological literature analysis* (pp. 65–72). Association for Computational Linguistics.
Lin, S., Ng, J.-P., Pradhan, S., Shah, J., Pietrobon, R., & Kan, M.-Y. (2010). Extracting formulaic and free text clinical research articles metadata using conditional random fields. In *Proceedings of the NAACL HLT 2010 second Louhi workshop on text and data mining of health documents* (pp. 90–95). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved December 4, 2017, from http://dl.acm.org/citation.cfm?id=1867735.1867749.
Lopez, P. (2009). GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *International conference on theory and practice of digital libraries* (pp. 473–474). Springer.
Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *ArXiv Preprint* arXiv:1603.01354.
Mao, S., Kim, J. W., & Thoma, G. R. (2004). A dynamic feature generation system for automated metadata extraction in preservation of digital materials. In *2004 Proceedings of the first international workshop on document image analysis for libraries.* (pp. 225–232). IEEE.
Marinai, S. (2009). Metadata extraction from PDF papers for digital library ingest. In *Proceedings of the 2009 10th international conference on document analysis and recognition* (pp. 251–255). Washington, DC, USA: IEEE Computer Society. https://doi.org/10.1109/ICDAR.2009.232.
McCallum, A. K., Nigam, K., Rennie, J., & Seymore, K. (2000). Automating the construction of internet portals with machine learning. *Information Retrieval, 3*(2), 127–163. https://doi.org/10.1023/A:1009953814988.
Mejer, A., & Crammer, K. (2010). Confidence in structured-prediction using confidence-weighted models. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 971–981). Association for Computational Linguistics.
Miwa, M., & Sasaki, Y. (2014). Modeling joint entity and relation extraction with table representation. In *EMNLP* (pp. 1858–1869).
Morin, B. (2017). LibGuides: Systematic reviews: Intro. Retrieved March 27, 2018, from https://researchguides.library.tufts.edu/c.php?g=249130&p=1658802.
Mudrak, B. (2016). Scholarly publishing in 2016 | AJE | American Journal Experts. Retrieved April 2, 2018, from https://www.aje.com/en/arc/scholarly-publishing-trends-2016/.
Nasar, Z., & Jaffry, S. W. (2018). Trust-based situation awareness: Agent-based versus population-based modeling—a comparative study. In *international conference on advancements in computational sciences*. Lahore, Pakistan: IEEE.
Ni, Z., & Xu, H. (2009). Automatic citation metadata extraction using hidden Markov models. In *Proceedings of the 2009 first IEEE international conference on information science and engineering* (pp. 802–805). Washington, DC, USA: IEEE Computer Society. https://doi.org/10.1109/ICISE.2009.353.
NISO. (2004). *Understanding metadata*. 4733 Bethesda Avenue, Suite 300, Bethesda, MD 20814 USA: NISO. Retrieved from http://www.niso.org/publications/press/UnderstandingMetadata.pdf.
NSF. (2018). S&E indicators 2018 | NSF-national science foundation. Retrieved April 3, 2018, from https://www.nsf.gov/statistics/2018/nsb20181/.
Ojokoh, B., Zhang, M., & Tang, J. (2011). A trigram hidden Markov model for metadata extraction from heterogeneous references. *Information Sciences, 181*(9), 1538–1551. https://doi.org/10.1016/j.ins.2011.01.014.

Palshikar, G. K. (2013). Techniques for named entity recognition: A Survey. In *Bioinformatics: Concepts, Methodologies, Tools, and Applications* (pp. 400–426). https://doi.org/10.4018/978-1-4666-3604-0.ch022

Patil, N., Patil, A. S., & Pawar, B. (2016). Survey of named entity recognition systems with respect to Indian and foreign languages. *International Journal of Computer Applications, 134*(16), 21–26.

Peng, F., & McCallum, A. (2004). Accurate information extraction from research papers using conditional random fields. *Presented at the HLT-NAACL04*. Retrieved from October 16, 2015, from http://citeseerx.ist.psu.edu/viewdoc/summary?.

Peng, F., & McCallum, A. (2006). Information extraction from research papers using conditional random fields. *Information Processing and Management, 42*(4), 963–979. https://doi.org/10.1016/j.ipm.2005.09.002.

Projects | ISU Information retrieval group. (2017). Retrieved February 12, 2018, from https://www.datadrivenscience.iastate.edu/aflexgroup/projects.

QasemiZadeh, B., & Schumann, A.-K. (2016). The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *LREC*.

Ronzano, F., & Saggion, H. (2015). Dr. Inventor framework: Extracting structured information from scientific publications. In *Discovery science* (pp. 209–220). Springer, Cham. https://doi.org/10.1007/978-3-319-24282-8_18.

Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbühler, A., Fabry, P., et al. (2007). Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics, 76*(2), 195–200. https://doi.org/10.1016/j.ijmedinf.2006.05.002.

SemPub2015. (2015). Retrieved August 6, 2018, from https://github.com/ceurws/lod/wiki/SemPub2015.

Seymore, K., Mccallum, A., & Rosenfeld, R. (1999). Learning hidden Markov model structure for information extraction. In *Proceedings of the AAAI'99 workshop machine learning for information extraction* (pp. 37–42).

Sharnagat, R. (2014). Named entity recognition: A literature survey.

Shickel, B., Tighe, P., Bihorac, A., & Rashidi, P. (2017). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. arXiv Preprint arXiv:1706.03446.

Shuxin, Z., Zhonghong, X., & Yuehong, C. (2013). Information extraction from research papers based on conditional random field model. *TELKOMNIKA Indonesian Journal of Electrical Engineering, 11*(3), 1213–1220.

SIGKDD. (1995). Retrieved January 20, 2018, from http://www.kdd.org/.

Simoes, G., Galhardas, H., & Coheur, L. (2009). Information extraction tasks: a survey. In *Proceedings of INForum* (Vol. 2009).

Sirsat, S. R., Chavan, V., & Deshpande, S. P. (2014). Mining knowledge from text repositories using information extraction: A review. *Sadhana-Academy Proceedings in Engineering Sciences, 39*(1), 53–62.

Souza, A., Moreira, V., & Heuser, C. (2014). ARCTIC: Metadata extraction from scientific papers in pdf using two-layer CRF. In *Proceedings of the 2014 ACM symposium on document engineering* (pp. 121–130). New York, NY, USA: ACM. https://doi.org/10.1145/2644866.2644872.

Strubell, E., Verga, P., Belanger, D., & McCallum, A. (2017). Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017* conference *on empirical methods in natural language processing* (pp. 2670–2680).

Tateisi, Y., Ohta, T., Pyysalo, S., Miyao, Y., & Aizawa, A. (2016). Typed entity and relation annotation on computer science papers. In *LREC*.

Tateisi, Y., Shidahara, Y., Miyao, Y., & Aizawa, A. (2014). Annotation of computer science papers for semantic relation extrac-tion. In *LREC* (pp. 1423–1429).

Teufel, S., & Moens, M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics, 28*(4), 409–445.

Teufel, S., Siddharthan, A., & Batchelor, C. (2009). Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 3–volume 3* (pp. 1493–1502). Association for Computational Linguistics.

Tkaczyk, D., Collins, A., Sheridan, P., & Beel, J. (2018). 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* (pp. 99–108). ACM.

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P., & Bolikowski, Ł. (2015). CERMINE: Automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR), 18*(4), 317–335. https://doi.org/10.1007/s10032-015-0249-8.

Vilnis, L., Belanger, D., Sheldon, D., & McCallum, A. (2015). Bethe projections for non-local inference. arXiv:1503.01397 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1503.01397.

Wang, M., & Chai, L. (2018). Three new bibliometric indicators/approaches derived from keyword analysis. *Scientometrics*. https://doi.org/10.1007/s11192-018-2768-9.

Ware, M., & Mabe, M. (2015). The STM report: An overview of scientific and scholarly journal publishing.

Wu, J.-C., Chang, Y.-C., Liou, H.-C., & Chang, J. S. (2006). Computational analysis of move structures in academic abstracts. In *Proceedings of the COLING/*ACL *on interactive presentation sessions* (pp. 41–44). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1225403.1225414.

Yin, P., Zhang, M., Deng, Z., & Yang, D. (2004). Metadata extraction from bibliographies using bigram HMM. In *Proceedings of the 7th international conference on digital libraries: International collaboration and Cross-fertilization* (pp. 310–319). Berlin: Springer. https://doi.org/10.1007/978-3-540-30544-6_33.

Yu, J., & Fan, X. (2007). Metadata extraction from chinese research papers based on conditional random fields. In *Fourth international conference on fuzzy systems and knowledge discovery, 2007. FSKD 2007.* (Vol. 1, pp. 497–501). IEEE. https://doi.org/10.1109/FSKD.2007.394.

Zahedi, Z., & Haustein, S. (2017). On the relationships between bibliographic characteristics of scientific documents and citation and Mendeley readership counts: A large-scale analysis of web of science publications. CoRR, http://arxiv.org/abs/1712.08637.