Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Automatic summarization of scientific articles: A survey

Nouf Ibrahim Altmami [*], Mohamed El Bachir Menai

Department of Computer Science, College of Computer and Information Sciences, King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia

ARTICLE INFO

Article history:
Received 7 November 2019
Revised 11 April 2020
Accepted 21 April 2020
Available online xxxx

Keywords:
Automatic
Text summarization
Scientific article
Single-document
Multi-document

ABSTRACT

The scientific research process generally starts with the examination of the state of the art, which may involve a vast number of publications. Automatically summarizing scientific articles would help researchers in their investigation by speeding up the research process. The automatic summarization of scientific articles differs from the summarization of generic texts due to their specific structure and inclusion of citation sentences. Most of the valuable information in scientific articles is presented in tables, figures, and algorithm pseudocode. These elements, however, do not usually appear in a generic text. Therefore, several approaches that consider the particularity of a scientific article structure were proposed to enhance the quality of the generated summary, resulting in *ad hoc* automatic summarizers. This paper provides a comprehensive study of the state of the art in this field and discusses some future research directions. It particularly presents a review of approaches developed during the last decade, the corpora used, and their evaluation methods. It also discusses their limitations and points out some open problems. The conclusions of this study highlight the prevalence of extractive techniques for the automatic summarization of single monolingual articles using a combination of statistical, natural language processing, and machine learning techniques. The absence of benchmark corpora and gold standard summaries for scientific articles remains the main issue for this task.

Contents

[*] Corresponding author.
E-mail addresses: naltmami@su.edu.sa (N. Ibrahim Altmami), menai@ksu.edu.sa (M. El Bachir Menai).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2020.04.020
1319-1578/© 2020 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University.

Please cite this article as: N. Ibrahim Altmami and M. El Bachir Menai, Automatic summarization of scientific articles: A survey, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.020

1. Introduction

The advent of the Internet has generated a massive flow of information, which makes retrieval more challenging. Most scientific information is found in scientific articles, and with the expansion of research domains, it can be quite difficult for scholars to find documents relevant to their interests. Even query-based searches for some specific fields return a large number of relevant articles that far exceed human processing capabilities. Automatic summarization of these articles would be useful for reducing the time needed to review them in full and obtain the gist of the information contained within. Mainly, summaries could be generated in two ways; single-document summaries in which the task is to generate a summary from a single source and multi-document summaries in which different but related documents are summarized by comprises only the essential materials or main ideas in a document in less space.

There is a great difference between automatic multi-document summarization of generic texts and that of scientific articles. This is due to several aspects stated by Agarwal et al. (2011). First, summarizing multiple scientific articles requires the derivation of the common themes and comprehension of the document collection as a whole. Even if multiple articles take the same research direction, each article presents its own study, and different arguments are made. In contrast, the task with news articles requires summarizing different details about the same story from various reports that may contain redundant information. Second, the recent trends and updates are presented in each article from the viewpoint of its author, and each author conveys his/her own unique perspective and questions. Different viewpoints toward the same literature need to be considered and combined to result in one accurate summary of this literature. Teufel and Moens (2002) have identified three aspects that make automatic scientific article summarization different from that of generic text. First, scientific articles have a particular main structure, unlike generic text. This structure usually starts with an introduction, which states the main problem, followed by related works, methodology, experiments, and findings before finally ending with results and implications. Second, scientific articles are usually much longer than generic texts. Finally, the goal of the summary itself is not unique since researchers need to look for new contributions, findings, and proposed solutions. In addition, Bhatia and Mitra (2012) introduced another reason for research article summaries. Most of the valuable information in scientific articles is presented in a document element, 'an entity separate from the running text of the document, which either augments or summarizes the information contained in the running text.' The most common types of document elements used in scientific articles are figures, tables, and pseudocodes for algorithms; they contain the most important experimental results and ideas. These elements usually do not appear in a generic text. Gastel and Day (2016) have stated another reason to use a special summarizer for scientific articles – the written language. English is the formal language of science, but this does not mean that every scientific article should be written in English. It is best to publish papers about local or national topics of interest, such as agricultural and social sciences, in the language of those who will use the content. In addition, most scientific articles include some domain-specific keywords or ontology that indicate their subject matter; thus, using these keywords in the system could improve its performance. Finally, unlike generic text, scientific articles contain more complex concepts and technical terms (Yasunaga et al., 2019).

All the aforementioned variations motivate the distinction between summary types. Automatic *extractive* summarization generates a summary in which sentences are selected from the input article(s) and generated as they are, whereas automatic *abstractive* summarization engenders an abstract composed of rephrased sentences representing the same ideas/concepts of the source article (s). *Indicative* summaries are used to generate the idea of the text without carrying a particular content. *Informative* summaries offer a condensed version of the text. Automatically summarizing scientific articles usually requires some level of abstraction in form of informative summary (Saggion, 2011).

The main question that may arise is why the article abstract does not suffice, as it is a summary of the scientific article? There are many reasons for generating article summaries, even when the author has written an abstract. First, the abstract information usually does not include relevant content from the full text. Second, it describes the viewpoint of the author about the unique characteristic in a biased and incomplete manner (Yang et al., 2016). Third, there is no single summary that meets all the user's needs (Reeve et al., 2007). Additionally, the abstract does not reflect all of the paper's impacts and contributions (Elkiss et al., 2008) but rather what the author of the article wants to highlight. Thus, the summary generated by such a system is expected to be informative enough, cover all the main sections of the input article, and present the most important information that a reader is looking for. Finally, Yasunaga et al. (2019) have pointed to the impact factor of a scientific article on the research community. Since the significance of a paper may change over time, the summarization system should accommodate the viewpoints of other researchers (i.e., citations) as well as the major aspects highlighted by the authors of the article in the abstract.

To the best of our knowledge, no survey study of the field of automatic scientific article summarization has been published to date. There are various studies that address the problem of summarization of generic text, such as (Gambhir and Gupta, 2017; Gupta and Gupta, 2019). A comprehensive study by Gambhir and Gupta (2017) presents the different works performed in an extractive summarization field. They identified the advantages and limitations of various methods used for extractive summarization. Additionally, they covered a few abstractive and multilingual text summarization approaches. In addition, they discussed intrinsic and extrinsic methods of summary evaluation along with text

Please cite this article as: N. Ibrahim Altmami and M. El Bachir Menai, Automatic summarization of scientific articles: A survey, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.020

summarization evaluation conferences and workshops. Finally, the evaluation results of extractive summarization approaches are presented on some shared DUC datasets. Gupta and Gupta (2019) presented a comprehensive review of recent text summarization abstractive approaches. The surveyed literature was categorized according to the type of abstractive technique used. The authors also highlighted the advantages and disadvantages of various methods used for abstractive summarization along with various tools that have been used or developed by researchers for abstractive summarization. The paper also discusses the evaluation techniques being used for assessing the abstractive summaries. In contrast, this paper provides a more focused study of the scientific article summarization task and reviews the potential of technology in this field based on an examination of the literature. It covers different aspects of these studies: the proposed solutions, results, corpora used, and evaluation metrics. It also presents some advantages and limitations observed during the investigation and finally concludes with some possible future research directions. The knowledge covered in this survey would be helpful for an interested researcher to be completely aware of the main themes, directions, and results in this field. Some of the observed limitations of the surveyed studies would help in a better understanding of the cognitive basis of the task, and therefore in choosing future directions.

The rest of this paper is organized as follows. Section 2 describes the structure of a scientific article, types of scientific article summary, and automatic scientific article summarization task. Section 3 addresses the summary evaluation task by presenting the methods and corpora used in the surveyed literature. Section 4 examines prior studies in the field of scientific article summarization. A discussion of the main issues and challenges that arise in this field, as well as recommendations for future research, is presented in Section 5. Finally, Section 6 concludes the paper.

2. Background

A research article reports on original empirical and theoretical work. It can be in the social or natural sciences and within an academic field. In addition, a research article may be one of several types: it may present original research, describe the research of other scientists, or comment on current challenges and trends in the area of interest. This section presents some definitions necessary for this survey, the main structure of a scientific article, the types of summaries, and the different automatic summarization tasks.

2.1. Definitions

This section briefly introduces definitions related to the summarization of scientific articles.

Definition 1 (Article). A piece of text written for a broad audience and generally published in newspapers, magazines, journals, etc. The subjects might be of interest to the writer, or it might be related to some present issues. It is organized in an appropriate form to draw the attention of the readers. The basic outline for an article is the title; introduction, which should further address the topic under concern; the body (usually 2–3 paragraphs that are sometimes divided into sections that further discuss the topic and explain the idea); and conclusion (the ending paragraph of the article with the opinion or recommendation of the writer).

Definition 2 (Scientific article). An article that reports the methods and results of an original scientific study performed by the authors. The kind of study may vary (it could be an experiment, survey, interview, or any other kind of studies), but in all cases, raw data have been collected and analyzed by the authors, and conclusions are drawn from the results of that analysis. Section 2.2 presents the structure of a scientific article.

Definition 3 (Citation sentence). A sentence that contains explicit reference(s) to other research articles. The following is an example of a citation sentence in which three different references are cited:

Example: Evaluation is considered time consuming (Imam et al., 2013), expensive (Lin, 2004) and unstable (Lin and Hovy, 2002).

Definition 4 (Citing Author). The author who discusses or mentions another research study in his/her own work.

Example: Qazvinian and Radev are the citing authors for (Teufel and Moens, 2002), as they mention Teufel and Moens's work in their article (Qazvinian and Radev, 2008).

Definition 5 (Citing Paper). A research article that contains a direct citation to another study.

Example: the paper of (Qazvinian and Radev, 2008).

Definition 6 (Cited Paper). A research article that has been cited in another article.

Example: Teufel and Moens's article (Teufel and Moens, 2002), as it was cited in (Qazvinian and Radev, 2008).

Definition 7 (Cited Text Span). The text portion in a cited paper to which citation sentence(s) refer.

Example: the following is a text portion from (Elkiss et al., 2008):

"Of these, 2497 were cited by at least one other paper in PubMed Central. In addition we retrieved all papers in PubMed Central citing the open access subset and extracted the citing sentences".

This text portion is referenced by another portion of text from (Qazvinian and Radev, 2008) as follows:

"They conducted several experiments on a set of 2, 497 articles from the free PubMed Central (PMC) repository".

2.2. Structure of a scientific article

There is no single structure that is completely agreed upon for scientific articles, but a generic structure may be partially if not totally found in scientific articles. This structure is primarily important to facilitate communication between scientists about their results and/or findings. This format also makes the paper easy to read at different levels. Thus, it helps the reader to quickly find what they need. In the following paragraphs, our discussion focuses on the main article sections and their features, taken from Pardede (2012).

The abstract is the first section of a scientific article. It is usually between 150 and 250 words or less, and it contains an informative summary of the major aspects of the paper with no citations. It should answer the three main questions: Why was the study conducted? How was it conducted? and What conclusion was drawn? It covers the objectives of the article, materials and methods used, results, and conclusions. In addition, a set of three to five keywords are listed after the abstract for indexing purposes.

Most scientific articles have an introduction that presents necessary background information for the reader to understand the rest of the paper. This information can be scientific, historical, cultural, or even personal. The word count for this section is typically 300 to 500 words or more. It can take the form of a continuous essay or a set of paragraphs covering the problem under consideration, its background, its importance, the objectives of the paper, and an operational definition of the terms used.

The related work section is generally a summary of previous literature, and it is often presented in several paragraphs. The purpose of this section is to bring the reader up to date on the topic and to understand the research problem being studied. In addition, it is important to describe the relevant developments of past

Please cite this article as: N. Ibrahim Altmami and M. El Bachir Menai, Automatic summarization of scientific articles: A survey, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.020

studies and situate the present work within the context of the existing literature. It contains citation sentences (sentences that contain explicit references to other research articles) of prior works as well as non-citation sentences to suggest ways to fulfill basic needs for further research.

The methodology section describes the exact procedures employed by the researcher(s) in detail. This section is very important to other researchers, as it enables them to re-implement the author's work and reproduce his/her results. It usually includes the subjects used, sample preparation techniques and their origins, data collection procedures, and computer programs used.

The experimental results section presents the key outcomes of the research. It is the core of the scientific article that covers the data used in one of three forms: text, illustrations, or tables. It also reports the results of statistical tests and descriptive statistics. It is usually followed by a discussion, which is sometimes combined into the results section. The discussion explains and evaluates the findings reported and examines their implications for future research.

In some scientific articles, the final paragraph of the discussion serves as a conclusion, while some papers have a separate section for the conclusion. This section summarizes the study presented in the article, its general implications and findings, and some suggestions for future research directions. The last section of each article is the references, which presents a full bibliography for all citations in the text. There are different citation styles, and cited works may be ordered alphabetically or as they appear in the text.

2.3. Types of scientific article summary

There are two main types of scientific article summary: (1) an abstract that provides a general overview of the article and (2) a summary based on citation sentences. The first type is not an accurate scientific summary, since it states the contributions using a general form in a less focused fashion. It also describes the viewpoint of the author in a biased and incomplete manner (Yang et al., 2016). The aforementioned problems motivated the generation of citation-based summaries – i.e., the second type of summarization. A citation-based summary employs a set of citations that reference an article to create a summary of that cited article (Qazvinian and Radev, 2008; Qazvinian et al., 2013). This set of citations indicates the main contributions and findings of the article, and it contains more information and focused contributions than the abstract does (Elkiss et al., 2008). These citations can, however, be biased toward the viewpoints of the citing author(s) and may not accurately describe the contents of the reference. Furthermore, most citations address the contributions or findings of a scientific article in an incomplete form, since they do not refer to the presumptions and data used to obtain these results.

2.4. Automatic scientific articles summarization tasks

2.4.1. Single-article vs. multi-article summarization

Two types of automatic scientific article summarization can be distinguished based on the number of input articles: single-article summarization (e.g. Saggion and Lapalme, 2000; Teufel and Moens, 2002; Qazvinian and Radev, 2008; Qazvinian et al., 2010; Lloret et al., 2011; Slamet et al., 2018) and multi-article summarization (e.g. Mohammad et al., 2009; Khodra et al., 2012; Chen and Zhuge, 2014; Erera et al., 2019). The former task asks to generate a summary of an article, whereas the latter asks to summarize altogether a set of articles related to the same topic and generate a single summary. Multi-article summarization is more challenging than single-article summarization. In addition to the main issues related to multi-document summarization, such as readability and coherency, summarizing citation sentences throughout several articles remains a difficult task.

2.4.2. Related work summarization

The related work section of a scientific article is usually used to show the distinctions and points of interest of the current work compared with those of previous research works. The automatic generation of the related work section is a challenging task. It could be considered as a multi-document topic-biased summarization problem. The task is to automatically generate a related work section for a target paper by summarizing a set of related reference papers (e.g. Hoang and Kan, 2010; Hu and Wan, 2014; Chen and Zhuge, 2016). The generated summary should briefly describe each of the reference papers, show their contributions, results, advantageous, limitations, and show the relationship between the target paper and these reference papers. The task is more challenging than multi-document summarization of generic text since each scientific paper has its own characteristics, contributions, approach, and specific content to its own work. Thus, the task is not only synthesizing similar contents and removing the redundant information but also locating the particular contributions of each reference paper and organizing them into one or few paragraphs.

3. Summary evaluation

Many subjective aspects of the automatic summarization task make the evaluation of summaries generated a critical issue. What to evaluate remains unclear. Evaluation techniques have been mainly divided into two classes: *intrinsic* and *extrinsic*. The former evaluates the quality of the summaries according to specific criteria, such as sentence integrity, readability, relevance, comprehensiveness, and accuracy. The summaries are usually evaluated by users or compared with a gold standard. The problem with this technique is that there is no single 'ideal' (El-Haj et al., 2011), as one can generate different summaries for the same document; moreover, several versions of the same summary can be created using distinct phrases. In addition, depending on the summarization task itself, unique but valid summaries can be created with respect to different goals (Lloret and Palomar, 2012). Extrinsic evaluation assesses the summary according to a specific task, such as time-to-completion, success rate, and decision-making accuracy. Thus, the same assessment may vary from one system to another. Furthermore, the evaluation process itself is another issue in this regard. Evaluation is considered time consuming (Imam et al., 2013), expensive (Lin, 2004) and unstable (Lin and Hovy, 2002). Consequently, automatic and semi-automatic approaches have been proposed here. The following subsections introduce the evaluation methods used by the surveyed literature as well as the corpora used in their evaluation. Tables 1–3 summarize the evaluation

Table 1
Evaluation types and corpora used in the automatic abstract generation-based summarization approaches.

Reference	Method	Corpus	Manual Evaluation	Automatic Evaluation
Saggion and Lapalme (2000)	Extractive	Authors' own corpus	√	Recall, Precision, and F-measure = 0.22
Lloret et al. (2011)	Extractive and Abstractive	Authors' own corpus	√	ROUGE-1 = 40.20
Saggion (2011)	Transformation-based Learning	Authors' own corpus		Accuracy
Yang et al. (2016)	Extractive	AAN, Microsoft dataset		ROUGE-1, ROUGE-2
Slamet et al. (2018)	Extractive	Authors' own corpus	√	Manual

Table 2
Methods, evaluation types and corpora used in the automatic citation-based summarization approaches.*

Reference	Method	Corpus	Manual Evaluation	Automatic Evaluation
Mei and Zhai (2008)	Extractive	Authors' own corpus		ROUGE
Qazvinian and Radev (2008)	Clustering and Graph-based method	AAN		PYRAMID = 0.75
Qazvinian et al. (2010)	Extractive	AAN		PYRAMID = 0.8
Abu-Jbara and Radev (2011)	Extractive and Machine Learning	AAN	√	ROUGE-L = 0.539
Agarwal et al. (2011)	Clustering and Extractive	Authors' own corpus		ROUGE-1 = 0.5123, ROUGE-2 = 0.3303
Chen and Zhuge (2014)	Clustering and Extractive	AAN		ROUGE-1, ROUGE-2
Jaidka et al. (2014)	Extractive	Authors' own corpus based on AAN		ROUGE
Cohan and Goharian (2015)	Extractive and Machine Learning	TAC2014		ROUGE-L = 0.43, ROUGE-1 = 0.45, ROUGE-2 = 0.15
Galgani et al. (2015)	Extractive	Authors' own corpus		Proposed method based on ROUGE-1 = 0.631
Ronzano and Saggion (2016)	Extractive	TAC2014		ROUGE-2 = 0.317
Cohan and Goharian (2017)	Extractive	TAC2014		Recall, Precision, F-measure = 27 and ROUGE-1 = 53
Lauscher et al. (2017)	Extractive and Machine Learning	ACL		P, R, F = 15.0
Wang et al. (2017)	Clustering and Extractive	Authors' own corpus	√	–

Abura'ed et al. (2017)	Extractive and Machine Learning	CL-SciSumm 2016.		ROUGE-2 = 0.2985
				ROUGE-SU4 = 0.2066

Cohan and Goharian (2018)	Extractive and Machine Learning	TAC2014 CL-SciSumm dataset		Recall, Precision, F-measure = 27, ROUGE-2 = □30.7 and ROUGE-3 = □24.4

Al Saied et al. (2018)	Extractive	AQUAINT SciSumm dataset		ROUGE-2 = 0.22
				ROUGE-S = 0.18

Abura'ed et al. (2018)	Machine Learning	CL-SciSumm 2018		ROUGE = 0.29
Agrawal et al. (2019)	Extractive, Machine Learning, and Graph-based method	Authors' own corpus		P,R,F-measure = 40.66

*All these methods can be applied to summarization of generic articles unless there are no citations for the target article.

Table 3
Methods, evaluation types and corpora used in other approaches.*

Reference	Method	Corpus	Manual Evaluation	Automatic Evaluation
Teufel and Moens (2002)	Machine Learning	Authors' own corpus		Accuracy = 0.73, kappa = 0.45 and macro-F = 0.5
Filho and Pardo (2007)	Extractive	cmp-lg	√	ROUGE-1 = 0.28408
Mohammad et al. (2009)	Extractive, Clustering, and Graph-based method	AAN		nugget-based pyramid, ROUGE
Bhatia and Mitra (2012)	Machine Learning	Authors' own corpus	√	
Khodra et al. (2012)	Machine Learning	ACL-ARC		Accuracy = 94.46%
He et al. (2016)	Extractive	ACL		ROUGE-1 = 0.42, ROUGE-2 = 0.08, ROUGE-L = 0.38, ROUGE-SU4 = 0.16, Precision = 0.75
Parveen et al. (2016)	Graph-based method	PLOS DUC 2002	√	ROUGE-SU4, ROUGE-2
Yeh et al. (2017)	Machine Learning	CL-SciSumm 2016		Recall, Precision, F-measure = 0.1443

Yasunaga et al. (2019)	Machine Learning	Authors' own corpus
		CL-SciSumm 2016
			√	ROUGE
				SU4-F = 18.56
				SU4-F = 24.36

Hoang and Kan (2010)	Extractive	Author's Corpus	√	ROUGE-1 = 0.698
Hu and Wan (2014)	Machine Learning	Author's Corpus	√	ROUGE-1 = 0.47940
Widyantoro and Amin (2014)	Machine Learning	Author's Corpus		Recall, Precision, F-measure = 0.86
Chen and Zhuge (2016)	Extractive and Graph-based method	Author's Corpus		ROUGE-1 = 0.50151
Erera et al. (2019)	Extractive	Author's Corpus	√	

*All these methods can be applied to summarization of generic articles except the method presented in (Bhatia and Mitra, 2012).

approach (manual or automatic), basic methods, and corpora used by the surveyed literature.

3.1. Evaluation methods

Evaluation is the task of assessing the summary on various aspects, including quality and content, to compare different summarization systems. Suitable evaluation metrics are required to achieve this goal. There are three different approaches in this regard (Saggion and Poibeau, 2013): the *automatic* approach, such as *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) (Lin, 2004), in which the evaluation process is fully automated; manual evaluation, where a *human* is the evaluator; and the *semi-automatic approach* (i.e., *a mixed approach*), such as PYRAMID (Nenkova and Passonneau, 2004).

This section briefly describes the various evaluation approaches used in the surveyed literature. ROUGE is a software package and containing a set of metrics that is popular and widely used for evaluating automatic summarization (Lin, 2004). These metrics compare a human-generated reference or reference summary with an automatically generated summary by counting the number of overlaps between the reference and candidate summaries. It has been used in most of the surveyed literature, such as (Abu-Jbara and Radev, 2011; Agarwal et al.,

Please cite this article as: N. Ibrahim Altmami and M. El Bachir Menai, Automatic summarization of scientific articles: A survey, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.020

2011; Cohan and Goharian, 2015; Lloret et al., 2011). However, there are several ROUGE measures, including ROUGE-N (n-gram co-occurrence statistics), ROUGE-L (longest common subsequence), ROUGE-W (weighted longest common subsequence), and ROUGE-S (skip-bigram co-occurrence statistics). While it is easy to calculate these metrics, they each have their own drawbacks. First, the candidate summary is compared with a reference summary where there is no single 'ideal.' Therefore, a good summary may be penalized for including relevant sentences that are not in the reference summary. Second, the study did not measure the similarity between summary sentences and reference summary sentences. Thus, some summaries would be penalized even when they contain segments similar to the reference summary sentences.

Other automatic methods have also been adopted (e.g., Qazvinian and Radev, 2008), including the PYRAMID (Nenkova and Passonneau, 2004), a *semi-automatic* evaluation approach. It was proposed to solve the issue of diverse content being generated by different users when writing a summary. The PYRAMID approach addresses this by using multiple manual summaries to create a gold standard and exploiting the frequency of information in the manual summaries to assign importance to various facts. This approach yields a score for a summary that is equal to the sum of the summary fact weights divided by that of the ideal summary weights. This value varies between 0 and 1, and the summary with more heavily weighted facts receives a higher score. The use of pyramid scores allows researchers to detect missing information and further improve their summaries.

3.2. Corpora

In a summarization task, corpora are needed to evaluate the summarization system and compare it with other approaches. The Text Analysis Conference (TAC)[1] hosts several workshops providing many tracks addressing different areas of Natural Language Processing (NLP). The summary track was one such track from 2008 to 2011 and in 2014. The TAC2014 benchmark[2] consists of 20 topics, each containing one reference text and various articles that are included in it. These articles are published by Elsevier[3] in the biomedical domain. Each topic has four scientific summaries written by four experts in the field, and the summaries are fewer than 250 words in length. However, the dataset used by Cohan and Goharian (2015, 2017, 2018) also contains discourse facets and annotated citation texts.

The CL-SciSumm[4] dataset comprises 30 topics in total and three subsets – train, development, and test data – with one reference paper and a set of citing articles for each topic. All the papers are in Extensible Markup Language (XML) format, and the sentences have clear boundaries, as is the case in TAC. A distinguishing feature of this dataset is that the topics are annotated by only one annotator.

The ACL Anthology Network (AAN)[5], initiated by Bird et al. (2008), is a community of people who are interested in solving NLP-related problems. It consists of 'a comprehensive manually curated networked database of citations, collaborations, and summaries in the field of Computational Linguistics' (Radev et al., 2013) as well as papers published by the ACL. It has been used by most of the surveyed literature, as in Abu-Jbara and Radev (2011), Qazvinian and Radev (2008), Chen and Zhuge (2014), and Yang et al. (2016).

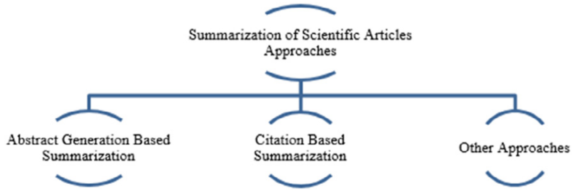
Summarization of Scientific Articles Approaches
Abstract Generation Based Summarization
Citation Based Summarization
Other Approaches

Fig. 1. Classification of summarization approaches of scientific articles.

The Microsoft dataset[6] is a collection from Microsoft Academic Search. It contains information about sentences in the article abstract, citation sentences, the authors, place of publication, and the attached paper of the citation sentences. This dataset was used by Yang et al. (2016).

The cmp-lg corpus[7] has also been used by Filho et al. (2007). Comprising 183 documents marked up in XML format, it is used as a resource for summarization, extraction, and information retrieval. The documents are scientific papers from the ACL. They cover key information for each paper (such as title, author, and date) apart from the main structural elements such as abstract, body, sections, and lists.

The PLOS[8] Medicine corpus is a set of 50 scientific articles, each with a gold standard summary associated with it. This summary has a wider perspective than the article abstract and is written by an editor of the month.

Previous datasets used in the surveyed studies either are not specifically for the automatic scientific article summarization task or are of limited size (30–50 article). As a result of these limitations, most of the existing systems are unsupervised or tuned on small data (Yasunaga et al., 2019). Thus, Yasunaga et al. (2019) motivated the creation of a large dataset (consisting of 1000 most cited papers from AAN (Radev et al., 2013)). For each target paper, they clean and keep an average of 15 citation sentences. They also create a gold standard summary for each target paper of a length equal to 151 words on average.

Fisas et al. (2016) developed a multi-layered annotated corpus of scientific papers in the domain of Computer Graphics. In this corpus, each sentence was annotated according to its role (Challenge, Background, Approach, Outcome, and Future Work). Additionally, for each citation, its purpose was specified (Criticism, Comparison, Use, Basis, Substantiation, or Neutral). All sentences in the document were graded based on their relevance for a summary, and their features, such as advantages, disadvantages, and limitations, were identified.

4. Approaches to automatic scientific article summarization

The following three subsections review state-of-the-art approaches to scientific article summarization according to the classification proposed in Fig. 1. There are mainly two classes of approaches to the summarization of scientific articles: abstract generation-based approaches and citation-based approaches. There are also other approaches for the summarization of scientific articles which focus on specific problems such as the summarization of tables, figures, and related work section.

4.1. Automatic abstract generation-based summarization

One interesting concern of various text summarization applications is the automatic generation of a research article abstract. This is a summary of the main topic and findings presented in the cor-

[1] https://tac.nist.gov/.
[2] https://tac.nist.gov/2014/BiomedSumm/.
[3] A dutch publishing and analytics company specializing in scientific, technical, and medical content.
[4] https://github.com/WING-NUS/scisumm-corpus.
[5] http://clair.eecs.umich.edu/aan/index.php.
[6] http://academic.research.microsoft.com.
[7] http://www-nlpir.nist.gov/related_projects/tipster_summac/cmp_lg.html.
[8] http://www.ncbi.nlm.nih.gov/pubmed.

Please cite this article as: N. Ibrahim Altmami and M. El Bachir Menai, Automatic summarization of scientific articles: A survey, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.020

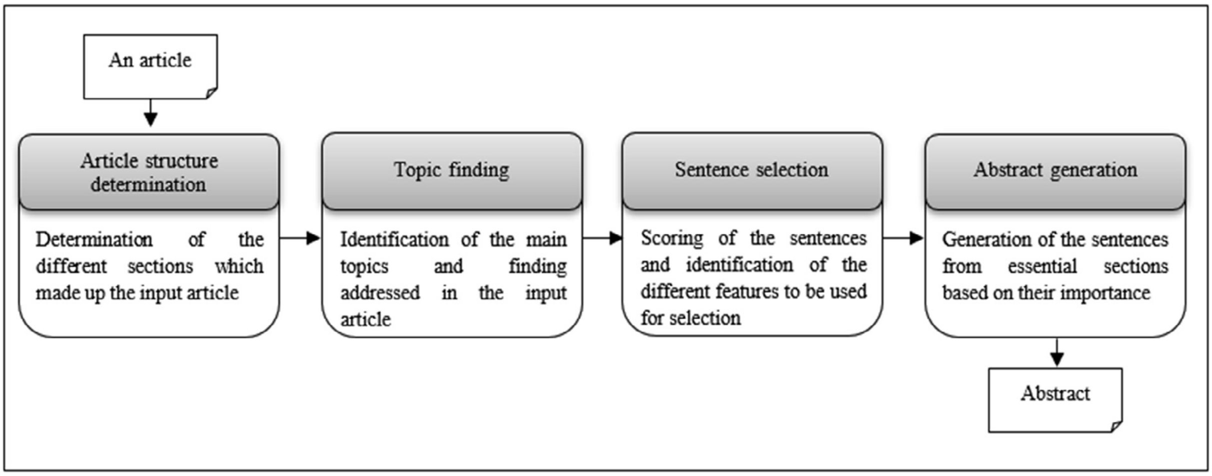
An article
Article structure determination
Determination of the different sections which made up the input article
Topic finding
Identification of the main topics and finding addressed in the input article
Sentence selection
Scoring of the sentences and identification of the different features to be used for selection
Abstract generation
Generation of the sentences from essential sections based on their importance
Abstract

Fig. 2. A Conceptual framework of abstract generation based summarization.

responding article, written by its authors. It must be included in any article presented in a journal, conference, or other contexts. The abstract can be used by researchers to obtain a general overview of the article. It can also help other automatic systems in indexing, searching, and retrieving information without accessing the whole document. While it can be generated automatically using text summarization techniques, the process is very challenging.

Fig. 2 outlines the main steps of abstract generation-based approaches. The main sections of the input article (see section 2.2) are first determined, then main topics and findings addressed in the input article are identified. Sentences are scored depending on their importance and then ranked based on their relevance. The top ranked sentences are selected to form the abstract for the input article.

Table 1 summarizes the evaluation approach (manual or automatic), basic methods, and corpora used by abstract generation-based approaches.

Saggion and Lapalme (2000) proposed an approach for automatically generating indicative and informative abstracts of scientific and technical articles called *selective analysis*, an extractive system composed of two phases. In the first phase, the system generates the indicative part of the abstract, which identifies the topics of the document. Subsequently, the system presents additional information on the reader's interest from the source text. From a corpus of source articles and their abstracts, the authors found that 72% of the abstract information was presented in the article title, the first and last sections, and the subheadings and captions of the figures and tables. In addition, they identified 55 concepts and 39 relations that form the basis for classifying different types of information elaborating the main content of a technical abstract. The concepts are classified by categories such as author, research activity, objectives, and cognitive activities. On the other hand, relations are the authors' general activities during the research and writing process, including investigation, reporting, motivation, thought, and identification. Finally, based on the abstract generated by the system, Saggion and Lapalme organized this content into indicative or informative templates, to generate an article abstract. They first selected the types of information, followed by the templates, from the target articles. After regenerating text that determines the topics of the target article, they expanded the indicative text with topic elaborations. They evaluated their proposed approach by measuring the indicativeness – the ability of indicating the fundamental content of the source document – and the acceptability, which measures the appropriateness of the automatically generated sentences compared to manual ones. They then compared their proposed system with two other summarizers: their own implementation of a word distribution-based summarizer built on computing the distribution of nouns in the text, and the commercially available Microsoft Office '97 Summarizer[9]. The acceptability evaluation showed that the quality of sentences generated with some template types is comparable to those produced by humans. With respect to the indicativeness, they found that *selective analysis* outperforms the other methods on average but not in a majority of cases.

Lloret et al. (2011) have suggested two approaches for generating research article abstracts. The first is a purely extractive summarizer (compendium$_E$), and the second is based on the extractive and abstractive technique (compendium$_{E-A}$). The extractive summarizer, compendium$_E$, relies on four main stages: 1) preprocessing (i.e., tokenization, sentence segmentation, stop words elimination, and part-of-speech tagging); 2) redundancy removal using a *textual entailment* (TE) tool (Ferrández et al., 2007); 3) sentence relevance identification, which assigns to each sentence a score that reflects its importance based on two features – *code quantity principle* (CQP) (Blake, 1992) and *term frequency* (TF) (Luhn, 1958)– and then ranks them according to their scores; and 4) summary generation, which selects the highest-ranked sentences to generate a final summary in the same order as the sentences appear within the original document. Thus, the generated summary is an extractive one. By contrast, compendium$_{E-A}$ is based on the extractive and abstractive technique. This method takes compendium$_E$ as a base and integrates an *information compression and fusion stage* between its third and fourth steps to generate an abstractive summary. New sentences are created by either combining information from two sentences or by shortening a long sentence into smaller ones. The authors evaluated their systems on a set of 50 research articles specializing in medicine, which was collected from the web. Their evaluation is based on three criteria: (i) summary information based on ROUGE scores (Lin, 2004), (ii) topics identified by dividing the sum of the keywords in the generated summary by that of the article keywords, and (iii) user satisfaction via a qualitative evaluation. Based on the results, they concluded that the compendium is useful for automatically generating a summary of research articles. Furthermore, abstractive summaries are more popular, even when they are similar to extractive summaries.

Saggion (2011) was the first researcher who investigated the transformation-based learning (TBL) method to the problem of

[9] http://www.microsoft.com/education/autosummarize.aspx.

abstract generation. His work is based on generating an abstract by a transformation of an initial text summary using several rules learned from a corpus of examples. He used a set of 219 abstracts in his experiments and set of tools such as the general architecture for text engineering (GATE) system (Maynard et al., 2002) and Weka machine learning toolkit (Geller, 2002) environment. For each abstract in the corpus, a baseline system was induced first to represent a default summary structure. Then, a decision tree was applied to induce several templates which were used later with the annotated training corpus to learn discourse correction rules using the TBL methodology. These rules were used finally to edit the initial text summary to obtain a final abstract. The experimental results were as good as the work of Saggion (2009) on classification-based predicate insertion.

Yang et al. (2016) proposed a system for an expanded abstract that describes the most important aspects of a scientific article using a data-weighted reconstruction approach. This consists of two phases: weight learning and salient sentence selection. During the first phase, semantic information from the citation sentences and social structure are considered. The authors used the target article abstract, which contains the main aspects, and the set of sentences that cite the target article to provide complementary aspects as an input to their system. First, they built a heterogeneous bibliographic network. They then identified social relations such as *paper-coauthor-paper* and *paper-cite-paper*, as well as similar semantic relations between sentences. Furthermore, they proposed a data-weighted objective function based on the learned sentence's weight and weighted reconstruction error. Thus, from the viewpoint of data reconstruction, they can detect salient sentences. The reconstruction process was performed for sentences deemed to be important by this method. They used two datasets to evaluate their proposed method: AAN[10] (Radev et al., 2013) and Microsoft.[11] They conducted several experiments and compared their proposed method with similar ones, as well as with six summarization baselines in terms of their performance. They applied ROUGE (Lin, 2004) and recall of key phrases to evaluate their system performance. Moreover, using three examples, they conducted a user and a case study to assess the quality of the generated abstract compared with the original. The results showed that, in most situations, their proposed method outperformed other methods. Thus, leveraging a sentence's weight to reconstruct the document improves the quality of the summarization.

Slamet et al. (2018) proposed a simple system that automatically generates an article abstract for the Indonesian language. Four main steps are used in their system. First, a preprocessing step (consisting of sentence extraction, case folding, tokenization, filtering, and stemming) is used to prepare the input text for the next step. Next, computing *Term Frequency-Inverse Document Frequency* (TF-IDF) (Hetami, 2015) for each term in the preprocessed text. Using cosine similarity and vector space modeling (VSM) (Hetami, 2015), the similarity between the text and the 20 keywords of the TF-IDF output are computed, and the sentences are ranked based on their similarity scores. Finally, the final abstract is compiled from the top ten sentences. The effectiveness of this proposed system was conducted by comparing manual abstracts with the output of the system. This comparison revealed that the system-generated abstract consists of three or more sentences in common with the manual abstract. This is because the author abstract (i.e., the manual abstract) contains words that are not present in the body of the article.

4.2. Automatic citation-based summarization

Citation sentences usually address the most important information from the cited article. They highlight the research problem, proposed method, reported results, contributions and even drawbacks and limitations. The set of citation sentences toward a target paper could be viewed as a short summary written by the cited researchers and presenting the impact of the target paper on the research community (Elkiss et al., 2008). One way of using these sentences is to create an article summary. This differs from the article abstract, since the citation summary represents the viewpoints of multiple researchers, whereas the abstract reflects only the authors' perspectives. Using citation sentences in the automatic summarization task for a scientific article yields a citation-based summary.

Fig. 3 outlines the main steps of the citation-based summarization approaches. A set of articles citing the input article is first retrieved, then citation sentences are extracted and their citation contexts in the input article are identified. Topic clustering is performed, and then summary sentences are selected.

Early works in automatic citation-based summarization (Abu-Jbara and Radev 2011; Elkiss et al. 2008; Qazvinian and Radev 2008) summarized the target paper by extracting a set of sentences from a set of citation sentences that cited the target paper (direct citation-based summarization). While it is worthwhile to use this set of citation sentences in the summarization process, later works (Cohan and Goharian 2015; Cohan and Goharian 2017; Mei and Zhai 2008; Qazvinian et al. 2010) pointed out some issues with using citation sentences. In citing sentences, the discussion of the target paper usually overlaps with the discussion of other cited papers or with the content of the citing paper, containing much irrelevant information. The solutions of these issues yield the generation of automatic cited text span-based summarization in which a summary of a set of cited text spans (i.e., a set of sentences in the target paper that its citing sentences refer to) is generated. This type of summary consists of words in the target paper and reflects the research community's perspectives. The following subsections review the state-of-the-art automatic citation-based summarization. Table 2 summarizes the methods, evaluation types (manual or automatic), and corpora used in the automatic citation-based summarization approaches.

4.2.1. Direct citation-based summarization

Qazvinian and Radev (2008) proposed a summarization system based on citation sentences. They first suggested a graph-based approach called C-LexRank. This system summarizes a single scientific article based on citations using community detection to extract sentences rich in information. The proposed model represents citation sentences as the vertices of a graph, in which the weighted edges represent the degree of semantic relatedness between two sentences. This weight is then computed by a similarity measure. Next, C-LexRank clusters the vertices in this graph and selects vertices (i.e., sentences) from each cluster to obtain a diverse summary. C-LexRank may effectively generate a scientific article summary using thirty sets of citation sentences from six different NLP topics in AAN (Radev et al, 2013). The researchers compared C-LexRank with different state-of-the-art baselines, such as maximal marginal relevance (MMR) (Carbonell and Goldstein, 1998) and LexRank (Erkan and Radev, 2004), and they claim that C-LexRank outperforms all other methods. They then extended their system to summarize a set of articles on the same scientific topic (Qazvinian et al., 2013) and performed experiments using question answering (QA), and dependency parsing (DP) articles. Using a nugget-based pyramid (Nenkova and Passonneau, 2004) and ROUGE (Lin, 2004), as evaluation methods, the results for the two methods and all summarization systems showed that

[10] http://clair.eecs.umich.edu/aan/index.php

[11] http://academic.research.microsoft.com

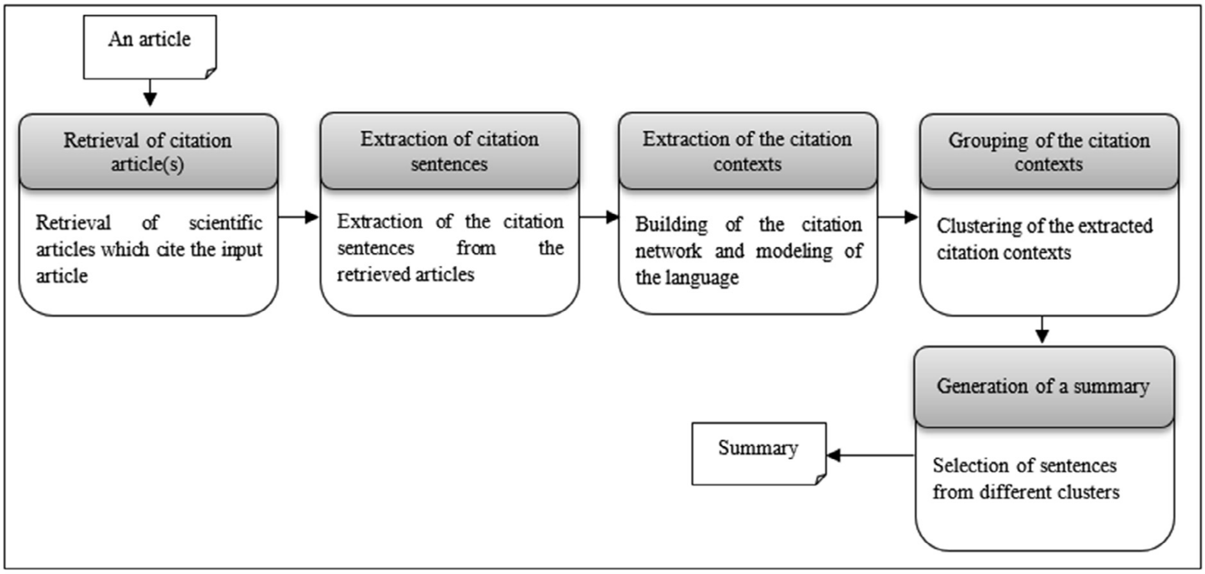
An article
Retrieval of citation article(s)
Retrieval of scientific articles which cite the input article
Extraction of citation sentences
Extraction of the citation sentences from the retrieved articles
Extraction of the citation contexts
Building of the citation network and modeling of the language
Grouping of the citation contexts
Clustering of the extracted citation contexts
Generation of a summary
Selection of sentences from different clusters
Summary

Fig. 3. A conceptual framework of citation-based summarization.

abstracts and citation sentences have unique summary-amenable information. Furthermore, citations are very useful in the creation of technical summaries from multiple documents.

Abu-Jbara and Radev (2011) noted some issues related to the readability and coherence of direct citation-based summaries, one of which is the number of scientific articles cited in the target. Including irrelevant segments causes different problems. First, the aim of automatic scientific article summarization is to reproduce the main contributions of the target article using minimal text. Thus, irrelevant sentences in the summary would needlessly increase its length. Second, the presence of these segments would alter the context of the summary and reduce readability, potentially confusing the reader. Third, the ranking algorithm would assign a low score to the relevant sentences due to these segments, even when the former cover an aspect that is not covered by any other sentence.

To tackle the aforementioned problems, Abu-Jbara and Radev (2011) included three subtasks in their proposed system: reference tagging, reference scope identification, and sentence filtering. They also addressed the ordering of summary sentences and the reference issues. In some citation sentences, the reference is not a syntactic constituent but an indication of citation. Additionally, the reference may be replaced, in some cases, with a suitable pronoun to avoid repeating author name(s) in each sentence. The authors addressed this issue with a machine learning approach. They trained a model to classify references into three classes: 'keep' to retain the reference, 'remove' to delete it, and 'replace' to substitute it with a suitable pronoun. Following this, they dealt with the appropriate number of citation sentences for a summary. The system selects summary sentences based on three criteria: (1) classification of these sentences into five categories – background, problem statement, method, results, and limitations; (2) clustering of similar sentences within each category; and (3) LexRank values (Erkan and Radev, 2004) within each cluster. They evaluated their approach on three levels. First, they separately assessed each subcomponent in their system. They then evaluated the generated summaries based on extraction quality. Finally, they ranked the summaries in terms of coherence and readability. The evaluation used an AAN dataset (Radev et al, 2013), precision, recall, ROUGE (Lin, 2004) and human judgment. The authors claimed that their system outperforms several baseline systems: random selection from citation sentences to form the summary, the MEAD summarizer (Radev et al., 2004), LexRank (Erkan and Radev, 2004), and the Qazvinian and Radev (2008) citation-based summarizer, and the remaining baselines are different versions of their proposed system for which they removed one component at a time.

4.2.2. Automatic cited text span-based summarization

Mei and Zhai (2008) did not use citation sentences directly to form a summary due to an occasional overlap between the discussion and content of the cited article with those of other cited articles (Siddharthan and Teufel, 2007). They proposed a language modeling-based approach to automatically summarize the impact of a scientific article. To accomplish this, they extracted sentences from the original article based on citations. They treated the automatic impact summarization as a retrieval problem and constructed a 'virtual impact query' of the citation sentences with a unigram language model that assigns high weight to words that reflect the impact of the paper. They then used Kullback-Leibler (KL)-divergence (Lafferty and Zhai, 2001) to score the sentences based on their impact. They used ROUGE (Lin, 2004) with manually created datasets from ACM SIGIR[12] papers to evaluate their proposed approach. Their results demonstrated the effectiveness of their approach, which outperformed several baselines from the MEAD toolkit: LEAD, MEAD-Doc, and MEAD-Doc + Cite all from (Radev et al., 2004).

Qazvinian et al. (2010) have also produced a single scientific article summarizer from citation sentences. They used a set of key phrases extracted from these sentences to represent the target article contributions. Their methodology is based on extracting a set of key phrases that covers the major information units (i.e., nuggets) using n-grams (Tomokiyo and Hurst, 2003), followed by a search for sentences with a maximum number of non-redundant key phrases. They used 25 articles from Qazvinian and Radev (2008) that had been manually annotated and highly cited in AAN (Radev et al, 2013). They adopted the pyramid evaluation method (Nenkova and Passonneau, 2004) and compared their proposed approaches with state-of-the-art baselines such as LexRank

[12] http://www.acm.org/dl

(Erkan and Radev, 2004), C-LexRank (Qazvinian and Radev, 2008), and MMR (Carbonell and Goldstein, 1998). Their approach outperformed the other baselines for n-grams of sizes 1, 2, and 3. However, it produced low variation and more stable results with respect to summary quality for equivalent n-gram sizes.

Agarwal et al. (2011) summarized several articles cited in the same target rather than the various viewpoints toward the target article using different citation sentences. They described an interactive multi-document summarizer called SciSumm for scientific articles. The input documents for this system are the collection of papers cited within the same target article. SciSumm creates a query-based summary as it captures a user's contextual needs in terms of the co-citation in the target text. The extracted segments are clustered to preserve the context in which they were cited. SciSumm comprises four modules. First, text tiling generates tiles of text relevant to citation context. The clustering module then groups these tiles into labeled clusters. A convenient and comprehensible description of each cluster is provided using these labels. Ranking is then applied to order the clusters based on relevance to the generated query. Finally, the clusters with the highest scores from the previous module are generated through summary presentation. The authors' evaluation showed that SciSumm outperforms MEAD (Radev et al., 2004) using the ROUGE metric (Lin, 2004).

Previous studies either summarized different articles co-cited in the same citation sentence (Agarwal et al., 2011) or exploited a single citation (Wan et al., 2009). Chen and Zhuge (2014) have made additional progress by taking advantage of multiple citations appearing in one paragraph or section. The main contribution of their work was to expand citations in an article by designing CFDSumm, a multi-document summarization system that exploits a set of terms that co-occur in a set of citations according to the common fact phenomenon. Their system consists of four modules. The first of these, term co-occurrence base module, uses 12,733 scientific articles from the AAN dataset (Radev et al., 2013) and the abstracts, titles, or even conclusions of these articles to compute frequently co-occurring terms. The citation processing module extracts the noun phrases by parsing citation sentences, generates n-grams (Tomokiyo and Hurst, 2003) as terms, and expands these terms based on the first module. Detecting common facts is accomplished by computing the frequent term set and then clustering the citations based on the common facts. Finally, the summarization module, which takes as input a set of references and its corresponding citation cluster, searches for articles with sentences relevant to the common fact cluster. It then removes redundancies, orders the sentences, and generates the final summary. The evaluation utilizes 13 citation paragraphs and ROUGE (Lin, 2004) as the metric, and it reveals that the proposed method performs better than the MEAD (Radev et al., 2004), SciSumm (Agarwal et al., 2011), and citation-sensitive in-browser summarizer (CSIBS) (Wan et al., 2009) methods.

Several studies have explored the merit of using citations in automatic scientific article summarization. Galgani et al. (2015) have suggested that such strategies face limitations when only the set of citing sentences is used to generate a summary. They proposed to use citances (or anchor text), the full target text, and the summary of the citing articles where available, rather than only the citances to describe the necessary information. All of these factors should be taken into consideration to obtain a global overview of the target article and, in turn, a better summary. They introduced a new trend of automatic summarization methods that combine incoming and outgoing citations along with different elements of both the citing and cited articles, in addition to the full target text. They proposed two methods to generate a summary, both of which use the same input: a target article to be summarized and a collection of its citances and citphrases extracted from connected documents. They generated the final summary by either (1) ranking the extracted text from all citances and citphrases to find common concepts over several citations and generate the final summary or (2) measuring the similarity between each sentence in the target article and the citations (either citphrases or citances) and then ranking the sentences in decreasing order. Thus, the summary sentences are those that are highly similar to the citation sentences. The underlying idea is that citations represent the main issues of the target article and can therefore be used to select the segments that represent these issues. This approach was initially developed for the legal domain to create catchphrases for case reports and subsequently applied to scientific articles. The authors evaluated their methods with ROUGE (Lin, 2004) and concluded that the performance is comparable to those of LexRank (Erkan and Radev, 2004) and C-LexRank (Qazvinian and Radev, 2008). Based on this favorable performance, they conclude that their proposed methods can be generalized to generate different types of summaries using different genres of text.

The problem with the degree of discrepancy between the reported findings in the referenced article and the citing articles has also been reported (De Waard and Maat, 2012). Citation sentences lack context from the original text. Cohan and Goharian (2015) proposed an approach to cope with this weakness of existing scientific summaries. Specifically, they extracted citation context, that is, 'textual spans in the reference articles that reflect the citation.' To perform this extraction, they model each citation sentence as an n-gram vector. They then locate the relevant text spans in the reference article through a vector space model. Candidate sentences are extracted to form the final summary based on maximum informativeness and novelty. Thus, the summary is generated using four main steps. The first is citation-context extraction: depending on the topic, these are grouped by community detection and discourse modeling. Next, elements of each citation-context group are ranked. Finally, the final summary is compiled by either iteratively selecting the top-ranked sentences or applying a greedy strategy. The researchers evaluated their methods using ROUGE-L, ROUGE-1, and ROUGE-2 (Lin, 2004) and compared the results against several well-known summarization methods. They used the Text Analysis Conference (TAC2014) dataset[13] in their assessment, which showed that they could improve existing summarization approaches; latent semantic analysis (LSA) (Steinberger and Jezek, 2004), LexRank (Erkan and Radev, 2004), MMR (Carbonell and Goldstein, 1998); and citation summary (Qazvinian and Radev, 2008).

Ronzano and Saggion (2016) investigated the use of citations toward a target paper to generate a better summary. They used the TAC2014 dataset[1] and computed the maximum ROUGE-2 (Lin, 2004) that can be produced when they generate a summary of 250 words by picking sentences from different parts of the target paper (abstract, body, citing spans) together with its citation context. For this purpose, they tried any possible combination of parts of the target paper and its citation context; then, they computed the maximum ROUGE-2 score achievable. The experimental results show the direct correlation between the obtained maximum ROUGE-2 score and the use of sentences from the citation context when the generated summary is evaluated. The best average ROUGE-2 score was obtained when the paper abstract was used as a reference summary and when the sentences to include in the final summary were chosen from the body and citation context.

Cohan and Goharian proposed another unsupervised model to solve the same issue, this time using word embedding (Bengio et al., 2003) and domain knowledge (Cohan and Goharian, 2017).

[13] https://tac.nist.gov/2014/BiomedSumm/

Their retrieval model extracts the appropriate context to capture the distinction in terminologies between the reference context in the original article and its citation text. They refer to the text spans in the original article as 'documents' and to citation sentences as the 'query.' Adding context to the citation provides a better understanding of the ideas, methods, and findings in the reference. Using TAC2014 as an intrinsic evaluation method, the authors showed that their model outperformed the baselines consistently across most of the metrics. In 2018, Cohan and Goharian presented a framework that addressed the problem of citation text inaccuracy. They first located the relevant context from the cited paper using three approaches: query reformulation as proposed by the authors, word embedding (Bengio et al., 2003), and supervised learning in which they proposed a feature-rich classifier to find the correct context for each given citation. Finally, from the faceted citations and corresponding contexts, they generated the final summary. Extensive evaluation results on TAC2014 and CL-SciSumm[14] (Jaidka et al., 2016) demonstrated that this proposal could improve some state-of-the-art approaches.

In 2017, CitationAS, an automatic tool for summary generation, was built by Wang and Zhang. It uses a set of rules to identify citation sentences and consists of three core stages. The first is clustering, in which three algorithms are used: suffix tree clustering (STC) (Zamir and Etzioni, 1999), Lingo (Osiński et al., 2004), and bisecting K-means (Steinbach et al., 2000). During this stage, citation sentences are first represented using VSM (Yang and Pedersen, 1997), and TF-IDF (Luhn, 1958) is used to calculate feature weights; similar sentences are grouped into one cluster. Next, Word2Vec (Mikolov et al., 2013), WordNet (Miller, 1998), and a combination of the two are used to generate cluster labels, following which the clusters with similar labels are merged. Finally, the clusters are sorted by size, and sentences are extracted to form the final summary using two different approaches: TF-IDF (Luhn, 1958) and MMR (Carbonell and Goldstein, 1998), which selects sentences with high scores. The authors built their own dataset based on 110,000 articles from PLOS One[15], which covers several disciplines. Through experiments, they found that using Word2Vec and WordNet separately improved performance to a greater extent than a combination of both in all clustering algorithms. One advantage of CitationAS is that the summary generated is comprehensive and representative of the topic, although it contains some redundant content.

As with most citation-based methods, Lauscher et al. (2017) proposed system also summarizes a target paper. The main components of their proposal are sentence ranking, augmentation, classification, and summarization. In experiments and for each citance, they used their learning to rank (L2R) model to rank the target paper sentences based on similarity to the citance. Several features are used during the learning stage, such as lexical, semantic, aggregate sentence embedding, similarities found using word mover's distance, and entity-based and positional features. Next, the top-ranked target article sentence is augmented with the one adjacent to it. The authors then trained a support vector machine (SVM) (Vapnik, 1995) and a convolutional neural network (CNN) (LeCun and Bengio, 1995) to classify each citation sentence according to its type (i.e., method, aim, result, implication, and hypothesis). The final summary is generated using the simple single pass clustering algorithm with word mover's distance, followed by sentence selection based on its TextRank score (Mihalcea and Tarau, 2004). The outputs are evaluated against the community and abstract gold summaries. The evaluation results show that the proposed approach outperformed those of the other participants in the CL-SciSumm shared task (Jaidka et al., 2014; Jaidka et al., 2016).

In the context of the CL-SciSumm 2017 shared task, Abura'ed et al. (2017) proposed three systems for three distinct subtasks. One of these systems is a citation-based summarizer. It is a supervised approach that produces a final summary of 250 words. This system is a modified version of Saggion et al.'s method (2016) – a trainable system that uses linear regression (Brügmann et al., 2015) to learn features and score sentences – with additional features. Several features for the reference and citing papers are combined to produce a cumulative value on which the sentences are ranked. The top-ranked sentences compose the final summary. Evaluation using ROUGE metrics (Lin, 2004) showed minor improvements compared to the mean results obtained in the CL-SciSumm 2017 shared task. In a recent paper, Al Saied et al. (2018) evaluated the use of feature maximization (FM) (Lamirel et al., 2013) in scientific paper summarization, focusing on publications using those that cited them. The proposed summarization systems are statistical, parameter-free, and language-agnostic, and they do not need additional corpora. The general framework of the proposed system is composed of five basic stages. First, the input text is preprocessed (i.e., stop-word elimination and stemming), and a word weight is computed for the set of key words in the paper title, subtitles, and human-generated abstract. Next, the sentence weights are calculated using the average of the weights of its words. The size of the final summary is determined in the third stage through a weight distribution. The summary is then generated after dealing with redundancy. The researchers assessed their systems using the DUC AQUAINT[16] corpus, which comprises unstructured data, and the results were encouraging. They also participated in the SciSumm 2016 challenge (Jaidka et al., 2016) and provided the best results to date.

In a recent work, Agrawal et al. (2019) proposed solving the task of scientific article summarization by a semi-supervised method. They believe that the contributions of any paper are best described by its aim, the method used and the final result. Depending on this and by using a k-nearest neighbor classifier (Devlin et al., 2018) and bootstrapping (Gupta and Manning, 2014), they extract these three concepts from the target paper's title, abstract and citation context (i.e., cited text spans that the citation sentences are referred to). After that, they construct a knowledge graph to graphically summarize the target paper using the extracted concepts and the citation graph. They compared their proposed approach against the work of (Gupta and Manning, 2011) and the system proposed by (Tsai et al., 2013), which ran the bootstrapping algorithm with n-gram based features.

4.2.3. Mixed automatic citation-based summarization approaches

The Computational Linguistics (CL) Summarization Pilot Task was proposed by Jaidka et al. (2014) to handle two subtasks. The first subtask addressed the problem of building a structured summary of a target paper – by extracting some information from the original paper (such as the paper's objectives, the main methodology, its results and implications), and citation-based summaries from its citing papers. Such a summarizer should automatically identify the text portion in the target paper that corresponds to the citation sentences from its cited papers and then identify the type of information described in the reference span. Three teams have participated in this shared task and submitted their runs, system descriptions and self-assessed results. The three systems were: clair-umich, MQ, and TALN.UPF from University of Michigan, Ann Arbor, USA; Macquarie University, Australia; and Universitat Pompeu Fabra, Spain, respectively. Different versions of TF-IDF

[14] https://github.com/WING-NUS/scisumm-corpus

[15] http://journals.plos.org/plosone/

[16] https://www-nlpir.nist.gov/projects/duc/data/2007_data.html

Please cite this article as: N. Ibrahim Altmami and M. El Bachir Menai, Automatic summarization of scientific articles: A survey, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.020

(Luhn, 1958) were used as baselines for the three systems. The clair-umich system implemented a supervised approach for the task of citation span identification, while MQ and TALN.UPF implemented unsupervised algorithms. The results of the participating systems were evaluated using ROUGE (Lin, 2004) and the reported results show that the clair-umich system was the best performer against MQ and TALN.UPF systems. Additionally, the team of the MQ system proposed an extractive summarizer that built an extractive summary of 250 words in length, by using the citing papers. Again, ROUGE (Lin, 2004) was used to evaluate the generated summaries, and the reported results were inconclusive regarding whether the features used by MQ system were aiding in generating better summaries or not.

Abura'ed et al. (2018) proposed several systems to participate in the 3rd shared challenge of Computational Linguistics Scientific Document Summarization[17]. This challenge proposes to address two subtasks; summarizing a scientific article by considering its citing papers (up to 250 words), automatically identifying the cited text span in the target paper, and then classifying the output in one of the following classes: Aim, Hypothesis, Implication, Results or Method. They used the CL-SciSumm 2018[18] corpus and evaluated their proposed systems using ROUGE (Lin, 2004). Regarding the first subtask, they developed a CNN to learn eighteen different scoring functions. Their proposed approach outperformed previous results reported in CL-SciSumm-17 Shared Task. While for the second subtask, they implemented supervised methods based on a CNN, as well as unsupervised systems based on word embedding representations and features computed from the linguistic and semantic analysis of the documents.

4.3. Other approaches

Teufel and Moens (2002) conducted the first study in the field of scientific article summarization. They selected informative content for the summary based on a naïve Bayes classifier (Lewis et al., 1996). For each sentence in a document, they added a rhetorical status to preserve the discourse context of the extracted segments. The main feature of their approach is that it combines discourse analysis with sentence extraction and produces content in a more principled way, especially for scientific articles. This model relies on the document structure and utilizes four dimensions of scientific articles: *problem structure*, *intellectual attribution*, *scientific argumentation,* and *attitude toward other people's work*.

In contrast to the aforementioned studies, which were specifically for developing a scientific article summarizer, Filho et al. (2007) conducted experiments to prove that a general extractive summarizer yields favorable results for scientific articles, but they made some improvements to account for the specificity of the article structure and text genre. The authors conducted their experiments using GistSumm (Pardo et al., 2003), a free language-independent extractive summarizer consisting of three main steps: (1) text segmentation to identify sentence boundaries, (2) sentence ranking, and (3) selection to generate the highest scored sentences and form the final summary. They modified GistSumm to summarize scientific articles through a simple heuristic-based approach, which allowed them to detect the text structure by searching for sentences that are not delimited by a period. They then summarized each section independently and combined the results. They conducted two experiments to test the value of their modification. The first assessed the quality of the generated summaries with respect to their textuality (i.e., coherence and cohesion) and informativity. Using 20 articles in computer science, the researchers tested the original GistSumm tool and its modified version. They found that their modification improved the informativity aspect, but textuality remained the same. In the second experiment, they used ROUGE (Lin, 2004) with 150 texts from the computation and language corpus (cmp-lg),[19] and the results showed that the modified version of GistSumm outperformed the original.

Some researchers used various document elements for different purposes, such as flow charts to describe a process, tables or plots to summarize experimental results, and pseudocode to present an algorithm. A *document element* is 'an entity, separate from the running text of the document that either augments or summarizes the information contained in the running text' (Bhatia and Mitra, 2012). Many recent efforts have sought to extract information from these elements. Liu et al. (2007) proposed a search engine called *TableSeer* to help users search for tables in digital documents. Similarly, *CiteSeerX*[20] and *Science Direct*[21] offer a table preview feature for their articles. Bhatia et al. (2010) introduced a specialized search engine for algorithms in scientific texts. Additionally, Hearst et al. (2007) proposed a search engine for figures and tables called *BioText*.

Mohammad et al. (2009) investigated the utility of both abstract and citation sentences in creating a survey of a scientific topic. They used four summarization systems in their investigation: Trimmer (Zajic et al., 2007), LexRank (Erkan and Radev, 2004), C-LexRank, and C-RR, both of which were from (Qazvinian and Radev, 2008). For input, they used text from entire papers, abstracts, and citation sentences. The surveys generated were 250 words in length and used one input at a time. Moreover, the authors utilized 10 papers in the question-answering research area and 16 from the DP set from the AAN dataset (Radev et al, 2013). The evaluation was based on the nugget-based pyramid and ROUGE (Lin, 2004). Trimmer had the best performance among all the summarizers used based on the pyramid score, while C-LexRank and LexRank were the best based on ROUGE. The results demonstrated that both abstracts and citation sentences are crucial when creating a survey for a scientific topic. Furthermore, they illustrated that citation sentences are important in survey creation for multi-document summarization but inappropriate with single-document summarization, as proven by Teufel et al. (2006).

Bhatia and Mitra (2012) generated a summary of document elements to aid in the efficient understanding of these elements. They searched for sentences in the target article that are identical in content and context to information presented in the document element. To accomplish this, they used a machine learning-based approach with a set of relevant features. Using a simple model, they measured the similarity between the candidate summary sentences and those referring to a document element. To generate succinct and useful synopses, their model attempts to balance information content with summary length. The evaluation covered various topics from 152 different publications using 290 distinct document elements. They compared two different classification methods, naïve Bayes (Bishop, 2006) and SVM (Chang and Lin, 2011), using the R-Precision metric. The former slightly outperformed the latter, but the differences were not statistically significant. In addition, the authors observed that both classifiers were most effective for figures, followed by tables and algorithms. Bhatia and Mitra (2012) also investigated different features and their effects on the classification process; they observed that feature performance based on content is worse than that of those based on context. Finally, they compared their proposed method with two state-of-the-art baselines: *Indri*[22] and reference sentences.

[17] https://www.aclweb.org/portal/content/3rd-computational-linguistics-scientific-document-summarization-shared-task

[18] https://github.com/WING-NUS/scisumm-corpus

[19] http://www-nlpir.nist.gov/related_projects/tipster_summac/cmp_lg.html

[20] http://www.citeseerx.ist.psu.edu.

[21] http://www.sciencedirect.com.

[22] http://www.lemurproject.org/indri.

However, context and content information were not included in these baselines, unlike in the experimental approach; thus, this comparison may not be completely fair. In this context, Hearst et al. (2007) and Liu et al. (2007) employed these baseline techniques, but Bhatia and Mitra observed that these methods were not useful, since the information provided was insufficient for understanding the document elements.

Khodra et al. (2012) proposed a multi-paper summarization approach that automatically summarizes a set of scientific papers that meets user specifications. It consists of three main modules: preprocessing, extraction, and presentation. This summarizer was evaluated on three aspects: accuracy, generality, and effectiveness. No comparison with other systems was provided. In contrast to all the previous studies, He et al. (2016) proposed a summarization system that studies differences across document groups, known as the differential topic model (i.e., comparative summary). The main aim of their proposal is to identify unique characteristics in order to generate group-specific topics. The design is a greedy sentence selection method, with two sentence-scoring schemas for generating the final comparative summary. It achieved significant improvements on different ROUGE metrics (Lin, 2004). In Parveen et al. (2016), a graph-based summarizer was proposed. The key to this system is the use of frequent coherence patterns, obtained through analyzing a set of biomedical article abstracts, to extract sentences. The objective was to select important, coherent, and non-redundant sentences, and this was achieved with an optimization process using mixed integer programming (MIP) (Gomory, 1960). The extracted summaries were evaluated on two datasets – PLOS Medicine (Parveen and Strube, 2015) and DUC 2002[23] – via comparison with those written by a PLOS Medicine editor,[24] ROUGE scores (Lin, 2004), and human judgment. The results were assessed against four baselines: MMR (Carbonell and Goldstein, 1998), LEAD (Radev et al., 2004), TextRank (Mihalcea and Tarau, 2004), and random. The authors also compared their model with three state-of-the-art approaches: (Parveen and Strube, 2015; Parveen et al., 2015; Radev et al., 2004). Evaluation revealed that the approach is robust, functional in both news and scientific documents of varying lengths, and superior to the baselines and state-of-the-art systems in human judgment.

Conversely, Yeh et al. (2017) proposed a system that generates citation summaries. This system classifies each sentence as a citation or non-citation one. The authors measured the similarity between each sentence in the reference article and citation sentences and then grouped it into one of two classes: cited or non-cited. After that, they applied a heuristic-based filtering strategy to refine the final summary. They used the CL-SciSumm[25] 2016 dataset and adopted various classification models, such as SVMs (Vapnik, 1995) and naïve Bayes (John and Langley, 1995), among which the L2-SVM provided the best performance across all metrics. In a recent work, Sun and Zhuge (2018) proposed a summarization system based on the semantic network. This network is built to represent the semantic link (type of relation) between the nodes of the scientific paper (i.e., sections, subsections, paragraphs, sentences, and words). Their focus was on three particular types: *is-part-of*, *similar-to*, and *co-occurrence*. The sentences were then ranked with a graph-ranking algorithm on the semantic link network constructed. The top-k ranked sentences were then selected for the final summary. The experimental results demonstrated the effectiveness of this system. In addition, the *is-part-of* relation was shown to be more helpful for short summaries than for long summaries, and it is more effective with longer papers containing more structural information. The authors also tested their model without removing stop words and observed degraded performances, although it still yielded better results than other methods.

While most automatic scientific article summarization studies are based on citation sentences, they may neglect the original goals of the author. Both citation-based types (section 4.2) were based on citation sentences, which in turn were based on the conclusion of the target paper (Yasunaga et al., 2019). Consequently, the generated summaries may not cover all the important aspects of the target paper even if they used other text sources, such as cited text spans. For this reason, Yasunaga et al. (2019) motivated the integration of both the target paper's abstract and its citation sentences. The idea behind their wok was based on the analysis of Conroy and Davis (2018), in which they found that most human-generated summaries contain a large number of terms from the paper abstract. In their work, they proposed two summarization models in which both of them make use of the original paper abstract and extracted cited text spans that the citation sentences are referred to (i.e., input). The first model then summarizes the integration of the previous input, while the second model extracts salient text from the cited text spans and then augments the abstract by this salient text. Both models are based on a neural network and evaluated using a corpus created by Yasunaga et al. (2019) for this task. The experimental results show that the generated summaries of the proposed models are more comprehensive than abstracts and traditional citation-based summaries.

Automatic generation of the related work section has been addressed in some research works, including Hoang and Kan (2010) and Chen and Zhuge (2016). Hoang and Kan (2010) proposed a system called ReWoS for automatically generating a related work section. It is a heuristic system based on a topic hierarchy tree with three main modules. The first, preprocessing, uses multiple articles related to the target paper as the input and applies a series of steps. The preprocessed sentences are then directed to the content summarization module, which may be general or specific, depending on the content and whether it describes the authors' work. Finally, the generation module produces a final summary by applying first-order traversal to the ordered node. The authors evaluated their system using ROUGE (Lin, 2004) and human judgment, in which their system outperformed two general baselines: LEAD and MEAD (Radev et al., 2004) in human evaluation. Hu and Wan (2014) proposed the automatic related work generation (ARWG) system for creating a related work section given a set of reference texts. They considered this problem one of optimization. In addition, they used the abstract and introduction sections of the target paper as well as the abstract, introduction, related work, and conclusion of its reference articles as input. First, their system used probabilistic LSA (Hofmann, 1999) to cluster sentences from both the target and reference texts into separate topic-biased groups. They then used two support vector regression models (Galanis et al., 2012) to learn the importance of each sentence. Finally, they generated the related work section of the target paper by choosing sentences from each cluster using a global optimization framework. Using ROUGE and human judgment as evaluation methods, the results showed that this proposed approach could improve the performance of LexRank (Erkan and Radev, 2004) as well as that of MEAD (Radev et al., 2004).

Chen and Zhuge (2016) used citation sentences to generate a related work section for a target article. Their system takes a target paper and set of reference articles as the input and compares the content of the target article with that of the sentences that cited the reference articles. The system then generates a citation document from this information and compares it with the abstract, introduction, and conclusion of the target article. Next, a graph of keywords extracted from the target paper and citation document is constructed. The distinguishable nodes in this graph are identified, and a final summary is generated by extracting the minimum

[23] https://www-nlpir.nist.gov/projects/duc/data/2002_data.html

[24] http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/

[25] https://github.com/WING-NUS/scisumm-corpus

Please cite this article as: N. Ibrahim Altmami and M. El Bachir Menai, Automatic summarization of scientific articles: A survey, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.020

tree that covers these nodes. A comparison of this proposed system and others based on ROUGE scores (Lin, 2004) demonstrated that it could improve four baselines – ReWoS (Hoang and Kan, 2010), ARWG (Hu and Wan, 2014), LexRank (Erkan and Radev, 2004) and MEAD (Radev et al., 2004). In contrast to previous studies, Widyantoro and Amin (2014) proposed a new approach for summarizing related work in scientific papers. Their proposal comprises two main stages. First, they extracted citation sentences by combining different methods, such as regular expression, an evidence-based approach (Powley and Dale, 2007), a co-reference system (Raghunathan et al., 2010), and the pronominal rule. Second, they categorized these citation sentences into three different classes (i.e., problem, method, and conclusion) using naïve Bayes (Kupiec et al., 1999), complement naïve Bayes (Rennie et al., 2003), and decision tree algorithms (Lin, 1999). They used precision, recall, and F-measures to evaluate system performance. Favorable results were achieved in different experiments, indicating that their system could provide a basis for the automatic summarization of related work.

Erera et al. (2019) built a system called '*IBM Science Summarizer*', which retrieves and summarizes scientific articles in the field of computer science. The *IBM Science Summarizer* was input 270,000 papers and focused on an optional user query and entities such as the scientific task, dataset, and metric. It independently summarized each section of a relevant paper returned by a search/-filtering process. If there was no query, the proposed system used the set of keywords provided by the author(s) of the paper at hand. The length of the generated summary was 10 sentences per section. The proposed system approached the task as an optimization problem in which different quality objectives (i.e., query saliency, entities coverage, diversity, text coverage, and sentence length) must be considered. It used the *cross-entropy* method (Rubinstein and Kroese, 2013) to make selections. For each paper, they generated two summaries: a section-based summary as described above, and a second summary generated by treating the entire paper as a flat text. The *IBM Science Summarizer* was evaluated by human experts, who considered three criteria: precision-oriented measure, coverage/recall, and the quality of the summary. Table 3 outlines the methods, evaluation types (manual or automatic), and corpora used in other approaches.

Finally, this study of the state of the art leads us to identify some advantages and limitations of the different approaches to the automatic summarization of scientific articles, which are briefly presented in Table 4.

5. Discussion

Owing to the relative youth of the field, research in automatic scientific article summarization has grown in the last decade. Many difficulties remain, such as the generation of abstracts using NLP techniques, the availability of labeled training and test corpora, and the scaling of collections of large documents. As shown in Fig. 1, there are mainly two classes of approaches to scientific article summarization: abstract generation-based and citation-based. The rest of the work varies from these two approaches. The abstract generation-based approaches rely on simple extractive and/or abstractive techniques; hence, researchers should pay attention to the specific content of the article abstract, which should cover different aspects of the target paper while preserving readability and coherence. On the other hand, citation-based summarization applies multiple methods ranging from the simple to the sophisticated. Some approaches search the target paper for a portion of text that matches the citation sentences. While citation sentences are a good indication of the degree of importance, it is admitted that incorporating these kinds of sentences with other methods based on whole-article summarization would improve the performance of such summarizers. Very few works in the field focus on related work summarization; most pertain to related work generation. The authors of these systems have used the target paper along with a set of reference articles in which a summary needs to be generated. All of these systems utilize extractive approaches, in which the summary sentences are extracted from the original articles and presented. Using extractive techniques with the related work sections would result in low readability. Moreover, related works contain citations sentences, an important distinction from generic text. Most existing systems do not benefit from these sentences, despite the positive findings when they are used in citation-based summarization.

Interesting observations can be made about the techniques reported thus far based on a review of the relevant literature. First, both single and multi-article summarizations have been examined. As with other fields, multi-document summarizations have recently increased in importance due to explosive growth in scientific publications and the frequent presence of pertinent information in multiple articles. A certain amount of redundancy is found in this type of summarization because the contributions from a target article may be described in multiple texts. Thus, demand is high for identifying important differences among documents. Although extractive techniques are used in single-document summarization, researchers may need to pay attention to abstractive methods, as the former tends to generate incoherent summaries (e.g., anaphora problems, semantic gaps) due to their working nature. Second, all the surveyed systems used English articles (i.e., monolingual systems), except for Slamet et al. (2018), which used an Indonesian article. Another gap is that all the studies except Bhatia and Mitra (2012) focus on summarizing texts only, whereas most of the important information in scientific articles is contained in document elements (such as tables and figures).

All the surveyed systems produced a text summary; no study adopted a visual format. Combinations of machine learning and statistical techniques are increasingly popular in scientific article summarization, whereas graph-based methods have not received more attention. Moreover, most of the surveyed studies conducted intrinsic evaluations in which the summary generated was compared with a manual reference summary, which is time consuming and expensive. The quality of the reference summary also varies depending on the individual who compiled it. Another observation is that most of the surveyed studies rely on citation sentences to generate a summary. Unfortunately, citations may not accurately reflect the information in the referenced article, since they are biased toward the viewpoint of the citing authors. In addition, most citations address the contributions or findings of a scientific article without referring to the basis of this information or the data used to obtain these results. Thus, if these citations are not present, these approaches would not work. Previous work has not explored the role of the full text and its relationship with citation texts or abstracts. Finally, most existing systems ignore the fluency of the summaries produced. For example, noisy and confusing summaries may have been generated because the cohesion, readability, diversity, and sentence order were ignored.

Despite the usefulness of the surveyed literature, it has some shortcomings that need to be addressed. In the automatic summarization of multiple scientific articles, one key step is finding relevant articles on the same topic, which is not a trivial task. Most existing systems take the set of articles to be summarized as an input. It would be helpful to devise a method to automatically retrieve and summarize multiple articles on the same topic. In this context, identifying the salient aspects of multiple scientific articles in relation to each other and to the target topic is another issue. An important problem that requires special attention here

Table 4
Advantages and limitations of the different methods applied to the automatic summarization of scientific articles.

Reference	Advantages	Limitations
Saggion and Lapalme (2000)	It is domain-independent.	The proposed system does not tackle the problem of generalization.
Lloret et al. (2011)	It applies a combination of extractive and abstractive techniques to abstract generation, which demonstrates their appropriateness and yields better summaries than purely extractive approaches.	- The summaries generated using compendiumE-A are always shorter than those with compendiumE, due to the information compression and fusion stage. Thus, the recall value is always smaller with compendiumE-A than with compendiumE.The approaches do not use the keywords from the text for summary generation. Thus, the abstract produced may not be informative enough.
Yang et al. (2016)	The amplified abstract generated is content-rich, comprehensive, and well-structured.	Using the original abstract to generate an amplified one should not be considered 'abstract generation' in its pure meaning. It is rather a summary of the article from its abstract and citation sentences.
Slamet et al. (2018)	Simple techniques are applied to abstract generation.	Using the top-ranked sentences to form an abstract for the article may cause the resulting text to suffer from a lack of readability and coherence.
Mei and Zhai (2008)	The approach demonstrates that the proposed impact summarizer successfully covers various details on the impact of the paper.	The dataset used in the experiment is very limited (i.e., only 14 papers). In addition, only one expert was employed to judge the summaries.
Qazvinian and Radev (2008)	A hierarchical agglomeration clustering algorithm is used in linear running time.	Only 25 papers were used in the experiments.
Qazvinian et al. (2010)	- It does not compute the full cosine similarity matrix for a document cluster, which is time consuming.Based on summary quality, the proposed method yields more stable results and low variation.	Only 25 papers were used in the experiments.
Abu-Jbara and Radev (2011)	- The focus is on the coherence and readability issues.The use of filtering, classification, and clustering components improve the extraction quality.	The recall score of the scope identification component was low.
Agarwal et al. (2011)	It allows users to interact with the system: the user clicks on a citation, and the system creates a query and generates a summary for this cited article.	The proposed system is interactive, but it is preferable for the system to search for the citation and create the summary automatically without user intervention.
Chen and Zhuge (2014)	The design is a multi-document summarization system that exploits the common fact phenomenon.	- The constructed terms are based on scientific abstracts from computational linguistics (i.e., they are domain-dependent). - The proposed approach is limited to citation sentences from the same paragraph.The evaluation is based on only 13 citation paragraphs from 11 papers.
- Galgani et al. (2015)	- The proposed method comprises different strategies: use of incoming and outgoing citations as well as the full article. - It can be generalized to different types of summaries using various genres of text.	- The results depend on the threshold at which the evaluation standard is set.
Cohan and Goharian (2015)	It tackles the lack of context in existing citation-based summarization approaches.	A domain-dependent (i.e., biomedical) dataset was used in the experiments.
Cohan and Goharian (2017)	It incorporates the domain knowledge such as WordNet as part of an embedding-based retrieval model.	A domain-dependent (i.e., biomedical) dataset was used in the experiments.
Wang et al. (2017)	The summary is comprehensive and reflects the topic.	The summary contains redundant information.
Lauscher et al. (2017)	Most of the reported results with the CL-SciSumm shared summarization task had a winning score.	The training data were limited.
Abura'ed et al. (2017)	It generates a short summary of 250 words using a supervised approach.	The reported results showed little improvement compared to the mean results obtained in the CL-SciSumm 2017 shared task.
- Al Saied et al. (2018)	- The proposed system is simple, parameter-free, language-independent, and does not require additional corpora. - It tackles the problem of information redundancy.	- The proposal is based on word frequency across different sections, which leads to low scores on AQUAINT corpus (i.e., unstructured data).
Hoang and Kan (2010)	The system generates a related work section for a target paper using features developed specifically for this task.	- The core idea of the system was the topic hierarchy tree, but the authors assumed it was given as input (i.e., they ignored the topic understanding module). - The set of references to be summarized were assumed to be provided by the user (i.e., the information retrieval module was not automated).The experiment was based on a limited dataset of only 10 articles.
Hu and Wan (2014)	Obvious performance improvements over various baselines were observed with both ROUGE and human judgment metrics.	Machine learning techniques require large amounts of data to be effective.
Chen and Zhuge (2016)	- It uses sentences that cite the reference paper (included in the related work section generated) along with the main text in the summarization. Thus, the proposed approach could work as both a scientific article summarizer and a related work section generator.- It improves the performance of the aforementioned methods as well as that of two other baselines.	- The main limitation is that the reference article should have citations in order to be summarized.- The dataset used was very limited in size.
Widyantoro and Amin (2014)	It successfully captures different citation styles using various techniques: regular expressions, co-reference resolution, the pronominal rule, and the evidence-based approach.	It uses external knowledge resources to calculate the features, whereby some of them depend on the training set.
Filho et al. (2007)	The results proved that a general extractive summarizer can be effective for scientific articles, but with some consideration of the specificity of these articles. This is very useful for poorer languages.	It requires more work in terms of textuality.

(*continued on next page*)

is the existence of citation sentences. Depending on publishing standards, different citation styles and writing formats present in the articles need to be summarized. An extra step needs to be taken in order to capture the variety of citations and the same reference cited in different articles. Another important aspect lies in the existence of a gold summary. Researchers typically use the article abstract as a reference (i.e., gold) summary and compare their output against this abstract. This comparison may not be completely valid due to the nature of these abstracts, which are biased toward the viewpoint of the author(s) and may not cover all the aspects that a user needs due to word count limits. Evaluation metrics also need to be adjusted for evaluating scientific article summaries; they need to select the appropriate information from the articles and compare it consistently with the appropriate correlate in the gold summary. The previous two issues are not limited to scientific papers; they are also present in single-article summarization. Finally, there are no public benchmark datasets or baseline systems available for comparison. Most of the current systems use their own dataset or collect texts from available corpora that are not specific to these tasks. The authors also compared their results with some open access systems used to summarize generic texts, and this comparison may not be entirely fair due to the specificity of the scientific article structure.

Table 4 (*continued*)

Reference	Advantages	Limitations
Mohammad et al. (2009)	This study was among the first works that investigated the use of citation texts in the automatic generation of technical surveys.	- Both citation sentences and abstracts should be used together as input to generate surveys.The dataset used in the experiments is size-limited and domain-dependent.
Bhatia and Mitra (2012)	It extracts information from the scientific article that is related to the document elements (i.e., table, figure, and algorithm).	The comparison with two other baselines is not fair, since context and content information were not utilized in these baselines.
Khodra et al. (2012)	It generates more readable summaries with flexible representation.	No comparisons with other systems were conducted to show the effectiveness of the proposed method.
He et al. (2016)	- It is able to measure sentence discriminative capability.The generated comparative summaries cover the unique characteristics of each document group.	The proposed approach was based on the frequency distribution among different groups and did not consider semantic relations.
Parveen et al. (2016)	The approach is robust, compatible with both news and scientific documents of varying lengths.	The coherence patterns used in the approach were based on a corpus of abstracts from the biomedical domain.
Yeh et al. (2017)	The approach is simple, since it tackles the summarization task as a binary classification problem (in which the system identifies whether the sentence has been cited).	- The dataset used is limited in size (10 topics for training, 10 for development, and 10 for testing).The improvement was 7.84% compared to the best performer in the CL-SciSumm 2016.
- Sun and Zhuge (2018)	- The proposed method is extensive, converges quickly, and does not require any preprocessing. - It builds a semantic network based on the extraction of semantic links from text.	- The semantic links are limited to three types: is-part-of, similar-to, and co-occurrence.
- Agrawal et al. (2019)	- It has domain-independence. - The proposed approach has scalability.	- It has low performance.
- Yasunaga et al. (2019)	- The generated summary is more comprehensive. - A large manually annotated corpus (1000 examples) is created that facilitates research on supervised approaches for scientific article summarization task.	-The target article should have citations in order to be summarized. - The neural network-based-system needs a large training set to succeed.

6. Conclusion

This paper presented a critical review of the state-of-the-art systems of summarizing scientific articles. It covered various aspects of the summarization task, including the solutions, evaluation, and corpora used in the evaluation process. It also highlighted some advantages and limitations in the literature. Our study found a high dominance of extractive techniques, single-article summarization, combinations of statistical (TF-IDF) and machine learning approaches (SVM, naïve Bayes, and clustering), and intrinsic evaluation methods (largely ROUGE metrics). The main related issues include the unavailability of training and test benchmark corpora, gold standard summaries for comparison, suitable evaluation metrics, and baseline systems needed for comparison purposes. Although graph-based methods have been successful in solving the task of multi-document summarization, they have received less attention in the field of automatic summarization of scientific articles. Most of the surveyed studies are centered around citation-based approaches and targeted a single article. Further research is needed to advance knowledge in this field by shifting from single-article to multi-article summarization and from extractive to abstractive in order to enhance the coherence and readability of the output. Deep learning approaches are also worth investigating due to their success in addressing other difficult NLP tasks.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abu-Jbara, Amjad, Radev, Dragomir, 2011. In: Coherent citation-based summarization of scientific papers. Association for Computational Linguistics, pp. 500–509.

Ahmed Abura'ed, et al., 2017. 'Lastus/Taln@ Clscisumm-17: Cross-Document Sentence Matching and Scientific Text Summarization Systems'.

Abura'ed, Ahmed, Bravo, Alex, Chiruzzo, Luis, Saggion, Horacio, 2018. 'LaSTUS/TALN + INCO@ CL-SciSumm 2018-Using Regression and Convolutions for Cross-Document Semantic Linking and Summarization of Scholarly Literature'. In: Proceedings of the 3nd Joint Workshop on Bibliometric-Enhanced Information

Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2018). Ann Arbor, Michigan (July 2018).

Agarwal, Nitin, Gvr, Kiran, Reddy, Ravi Shankar, Rosé, Carolyn Penstein, 2011. 'SciSumm: A Multi-Document Summarization System for Scientific Articles'. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations, Association for Computational Linguistics, 115–120.

Agarwal, Nitin, Gvr, Kiran, Reddy, Ravi Shankar, Rosé, Carolyn Penstein, 2011b. Towards multi-document summarization of scientific articles: making interesting comparisons with SciSumm. In: Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, Association for Computational Linguistics, pp. 8–15.

Agrawal, Kritika, Mittal, Aakash, Pudi, Vikram, 2019. Scalable, semi-supervised extraction of structured information from scientific literature. In: Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications, pp. 11–20.

Al Saied, Hazem, Dugué, Nicolas, Lamirel, Jean-Charles, 2018. Automatic summarization of scientific publications using a feature selection approach. Int. J. Digital Lib., 1–13

Filho, Balage, Paulo, Pedro, Salgueiro Pardo, T.A., das Gracas Volpe Nunes, M., 2007. Summarizing Scientific Texts: Experiments with Extractive Summarizers. In: Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007), IEEE, 520–524.

Bengio, Yoshua, Ducharme, Réjean, Vincent, Pascal, Jauvin, Christian, 2003. A Neural Probabilistic Language Model. *J. Machine Learn. Res.* 3 (Feb), 1137–1155.

Bhatia, Sumit, Mitra, Prasenjit, 2012. Summarizing figures, tables, and algorithms in scientific publications to augment search results. ACM Trans. Information Syst. (TOIS) 30 (1), 3.

Bhatia, Sumit, Mitra, Prasenjit, Lee Giles, C., 2010. Finding algorithms in scientific articles. In: *Proceedings of the 19th International Conference on World Wide Web*, pp. 1061–1062.

Bird, Steven et al., 2008. 'The Acl Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics'.

Bishop, Christopher M., 2006. Pattern Recognition and Machine Learning. Springer.

Blake, Barry J., 1992. T. Givón, Syntax: A Functional-Typological Introduction, Volume II. Amsterdam: John Benjamins, 1990. pp. Xxv+ 552.' J. Linguist. 28(2), 495–500.

Brügmann, Sören et al., 2015. Towards content-oriented patent document processing: intelligent patent analysis and summarization. World Patent Information 40, 30–42.

Carbonell, Jaime G., Goldstein, Jade, 1998. The use of MMR and diversity-based reranking for reodering documents and producing summaries. In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 335–336.

Chang, Chih-Chung, Lin, Chih-Jen, 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) 2 (3), 27.

Chen, Jingqiang, Zhuge, Hai, 2014. Summarization of scientific documents by detecting common facts in citations. Future Generation Comput. Syst. 32, 246–252.

Chen, Jingqiang, Zhuge, Hai, 2016. Summarization of Related Work through Citations. In 2016 12th International Conference on Semantics, Knowledge and Grids (SKG), IEEE, 54–61.

Cohan, Arman, Goharian, Nazli, 2015. Scientific article summarization using citation-context and article's discourse structure. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal: Association for Computational Linguistics, pp. 390–400.

Cohan, Arman, Goharian, Nazli, 2017. Contextualizing citations for scientific summarization using word embeddings and domain knowledge. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1133–1136.

Cohan, Arman, Goharian, Nazli, 2018. Scientific document summarization via citation contextualization and scientific discourse. Int. J. Digital Lib. 19 (2–3), 287–303.

Conroy, John M., Davis, Sashka T., 2018. Section mixture models for scientific document summarization. Int. J. Digital Lib. 19 (2–3), 305–322.

De Waard, Anita, Maat, Henk Pander, 2012. 'Epistemic Modality and Knowledge Attribution in Scientific Discourse: A Taxonomy of Types and Overview of Features'. In: Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, Association for Computational Linguistics, 47–55.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina, 2018. 'Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding'. arXiv preprint arXiv:1810.04805.

El-Haj, Mahmoud, Kruschwitz, Udo, Fox, Chris, 2011. 'Exploring Clustering for Multi-Document Arabic Summarisation'. In: Asia Information Retrieval Symposium, Springer, 550–561.

Elkiss, Aaron et al., 2008. Blind men and elephants: what do citation summaries tell us about a research article? J. Am. Soc. Inf. Sci. Technol. 59 (1), 51–62.

Erera, Shai et al., 2019. 'A Summarization System for Scientific Documents'. arXiv preprint arXiv:1908.11152.

Erkan, Günes, Radev, Dragomir R., 2004. Lexrank: Graph-Based Lexical Centrality as Salience in Text Summarization. J. Artif. Intelligence Res. 22, 457–479.

Ferrández, Oscar, Micol, Daniel, Munoz, Rafael, Palomar, Manuel, 2007. 'A Perspective-Based Approach for Solving Textual Entailment Recognition'. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Association for Computational Linguistics, 66–71.

Fisas, Beatriz, Ronzano, Francesco, Saggion, Horacio, 2016. 'A Multi-Layered Annotated Corpus of Scientific Papers'. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 3081–3088.

Galanis, Dimitrios, Lampouras, Gerasimos, Androutsopoulos, Ion, 2012. Extractive Multi-Document Summarization with Integer Linear Programming and Support Vector Regression. Proc. Coling 2012, 911–926.

Galgani, Filippo, Compton, Paul, Hoffmann, Achim, 2015. Summarization based on bi-directional citation analysis. Inf. Process. Manage. 51 (1), 1–24.

Gambhir, Mahak, Gupta, Vishal, 2017. Recent automatic text summarization techniques: a survey. Artif. Intell. Rev. 47 (1), 1–66.

Gastel, Barbara, Day, Robert A., 2016. How to Write and Publish a. Scientific Paper. ABC-CLIO.

Geller, James, 2002. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. SIGMOD Record 31 (1), 77.

Gomory, Ralph., 1960. An Algorithm for the Mixed Integer Problem. RAND CORP SANTA MONICA CA.

Gupta, Som, Gupta, S.K., 2019. Abstractive Summarization: An Overview of the State of the Art. Expert Syst. Appl. 121, 49–65.

Gupta, Sonal, Manning, Christopher, 2011. Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. In: *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 1–9.

Gupta, Sonal, Manning, Christopher, 2014. 'Improved Pattern Learning for Bootstrapped Entity Extraction'. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning, 98–108.

He, Lei, Li, Wei, Zhuge, Hai, 2016. Exploring Differential Topic Models for Comparative Summarization of Scientific Papers. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1028–1038.

Hearst, Marti A. et al., 2007. BioText Search Engine: Beyond Abstract Search. Bioinformatics 23 (16), 2196–2197.

Hetami, A., 2015. Perancangan Information Retrieval (IR) Untuk Pencarian Ide Pokok Teks Artikel Berbahasa Inggris dengn Pembobotan Vector Space Model. Jurnal Ilmiah Teknologi Informasi Asia 9 (1), 53–59.

Hoang, Cong Duy Vu, Kan, Min-Yen, 2010. Towards Automated Related Work Summarization. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 427–435.

Hofmann, Thomas, 1999. Probabilistic Latent Semantic Analysis. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 289–296.

Hu, Yue, Wan, Xiaojun, 2014. Automatic Generation of Related Work Sections in Scientific Papers: An Optimization Approach. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1624–1633.

Imam, Ibrahim, Nounou, Nihal, Hamouda, Alaa, Khalek, Hebat Allah Abdul, 2013. An ontology-based summarization system for Arabic documents (Ossad). Int. J. Comput. Appl. 74 (17), 38–43.

Jaidka, Kokil et al., 2014. The Computational Linguistics Summarization Pilot Task. In Proceedings of Text Ananlysis Conference, Gaithersburg, USA.

Jaidka, Kokil, Chandrasekaran, Muthu Kumar, Rustagi, Sajal, Kan, Min-Yen, 2016. Overview of the CL-SciSumm 2016 Shared Task'. In Proceedings of the Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL), 93–102.

John, George H., Langley, Pat, 1995. Estimating Continuous Distributions in Bayesian Classifiers. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., pp. 338–345.

Khodra, Masayu Leylia, Widyantoro, Dwi Hendratmo, Aminudin Aziz, E., Trilaksono, Bambang Riyanto, 2012. Automatic tailored multi-paper summarization based on rhetorical document profile and summary specification. J. ICT Res. Appl. 6 (3), 220–239.

Kupiec, Julian, Pedersen, Jan, Chen, Francine, 1999. A trainable document summarizer. Adv. Autom. Summarization, 55–60.

Lafferty, John, Zhai, Chengxiang, 2001. Document language models, query models, and risk minimization for information retrieval. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 111–119.

Lamirel, Jean-Charles, Cuxac, Pascal, Chivukula, Aneesh Sreevallabh, Hajlaoui, Kafil, 2013. A new feature selection and feature contrasting approach based on quality metric: application to efficient classification of complex textual data. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp. 367–378.

Lauscher, Anne, Glavaš, Goran, Eckert, Kai, 2017. University of Mannheim@ CLSciSumm-17: citation-based summarization of scientific articles using semantic textual similarity. In: *CEUR Workshop Proceedings*, pp. 33–42.

LeCun, Yann, Bengio, Yoshua, 1995. Convolutional networks for images, speech, and time series. The Handbook Brain Theory Neural Networks 3361 (10), 1995.

Lewis, David D., Schapire, Robert E., Callan, James P., Papka, Ron, 1996. Training algorithms for linear text classifiers. SIGIR, 298–306.

Lin, Chin-Yew, Hovy, Eduard, 2002. Manual and Automatic Evaluation of Summaries. In: Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4. Association for Computational Linguistics, pp. 45–51.

Lin, Chin-Yew, 1999. Training a selection function for extraction. In: *Proceedings of the Eighth International Conference on Information and Knowledge Management*, pp. 55–62.

Please cite this article as: N. Ibrahim Altmami and M. El Bachir Menai, Automatic summarization of scientific articles: A survey, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.020

Lin, Chin-Yew, 2004. 'Rouge: A Package for Automatic Evaluation of Summaries'. Text Summarization Branches Out.

Liu, Ying, Bai, Kun, Mitra, Prasenjit, Lee Giles, C., 2007. Tableseer: automatic table metadata extraction and searching in digital libraries. In: *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 91–100.

Lloret, Elena, Palomar, Manuel, 2012. Text Summarisation in progress: a literature review. Artificial Intelligence Rev. 37 (1), 1–41.

Lloret, Elena, Romá-Ferri, María Teresa, Palomar, Manuel, 2011. COMPENDIUM: a text summarization system for generating abstracts of research papers. In: International Conference on Application of Natural Language to Information Systems. Springer, pp. 3–14.

Luhn, Hans Peter, 1958. The automatic creation of literature abstracts. IBM J. Res. Dev. 2 (2), 159–165.

Maynard, Diana et al., 2002. Architectural elements of language engineering robustness. Nat. Lang. Eng. 8 (2–3), 257–274.

Mei, Qiaozhu, Zhai, ChengXiang, 2008. Generating Impact-Based Summaries for Scientific Literature. Proceedings of ACL-08: HLT: 816–824.

Mihalcea, Ra.da., Tarau, Paul, 2004. Textrank: bringing order into text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Mikolov, Tomas, Le, Quoc V., Sutskever, Ilya, 2013. Exploiting Similarities among Languages for Machine Translation. arXiv preprint arXiv:1309.4168.

Miller, George A., 1998. WordNet: An Electronic Lexical Database. MIT Press.

Mohammad, Saif et al., 2009. Using citations to generate surveys of scientific paradigm. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 584–592.

Nenkova, Ani, Passonneau, Rebecca, 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Hlt-Naacl 2004.

Osiński, Stanis'law, Stefanowski, Jerzy, Weiss, Dawid, 2004. Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. In: Intelligent Information Processing and Web Mining, Springer, 359–368.

Pardede, Parlindungan, 2012. Scientific articles structure. In: Scientific Writing Workshop: 16.

Pardo, Thiago, Salgueiro, Alexandre, Lucia Helena Machado Rino, Maria das Graças Volpe Nunes, 2003. GistSumm: A Summarization Tool Based on a New Extractive Method. In: International Workshop on Computational Processing of the Portuguese Language, Springer, 210–218.

Parveen, Daraksha, Strube, Michael, 2015. Integrating Importance, Non-Redundancy and Coherence in Graph-Based Extractive Summarizatio. In: Twenty-Fourth International Joint Conference on Artificial Intelligence.

Parveen, Daraksha, Ramsl, Hans-Martin, Strube, Michael, 2015. Topical coherence for graph-based extractive summarization. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1949–1954.

Parveen, Daraksha, Mesgar, Mohsen, Strube, Michael, 2016. Generating coherent summaries of scientific articles using coherence patterns. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 772–783.

Powley, Brett, Dale, Robert, 2007. Evidence-Based Information Extraction for High Accuracy Citation and Author Name Identification'. In: Large Scale Semantic Access to Content (Text, Image, Video, and Sound), Le Centre De Hautes Etudes Internationales D'informatique Documentaire, 618–632.

Qazvinian, Vahed et al., 2013. Generating extractive summaries of scientific paradigms. J. Artificial Intelligence Res. 46, 165–201.

Qazvinian, Vahed, Radev, Dragomir R., 2008. Scientific Paper Summarization Using Citation Summary Networks. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, pp. 689–696.

Qazvinian, Vahed, Radev, Dragomir R., Ozgur, Arzucan, 2010. Citation Summarization through Keyphrase Extraction. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), 895–903.

Radev, Dragomir R., et al., 2003. Evaluation Challenges in Large-Scale Document Summarization. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, Association for Computational Linguistics, 375–382.

Radev, Dragomir R., et al., 2004. 'MEAD-a Platform for Multi-document Multilingual Text Summarization'.

Radev, Dragomir R., Muthukrishnan, Pradeep, Qazvinian, Vahed, Abu-Jbara, Amjad, 2013. The ACL Anthology Network Corpus. Language Resources Evaluation 47 (4), 919–944.

Raghunathan, Karthik et al., 2010. A Multi-Pass Sieve for Coreference Resolution. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 492–501.

Reeve, Lawrence H., Han, Hyoil, Brooks, Ari D., 2007. The use of domain-specific concepts in biomedical text summarization. Inf. Process. Manage. 43 (6), 1765–1776.

Rennie, Jason D., Shih, Lawrence, Teevan, Jaime, Karger, David R., 2003. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), 616–623.

Ronzano, Francesco, Saggion, Horacio, 2016. An Empirical Assessment of Citation Information in Scientific Summarization. In: International Conference on Applications of Natural Language to Information Systems, Springer, 318–325.

Rubinstein, Reuven Y., Kroese, Dirk P., 2013. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer Science & Business Media.

Saggion, Horacio, Ronzano, Francesco, 2016. Trainable Citation-Enhanced Summarization of Scientific Articles. In: Proceedings of the Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL), 175–186.

Saggion, Horacio, Lapalme, Guy, 2000. Selective Analysis for Automatic Abstracting: Evaluating Indicativeness and Acceptability. In Content-Based Multimedia Information Access-Volume 1, Le Centre De Hautes Etudes Internationales D'informatique Documentaire, 747–764.

Saggion, Horacio, Poibeau, Thierry, 2013. Automatic Text Summarization: Past, Present and Future. In: Multi-Source, Multilingual Information Extraction and Summarization. Springer, pp. 3–21.

Saggion, Horacio, 2009. A Classification Algorithm for Predicting the Structure of Summaries. In: Proceedings of the 2009 Workshop on Language Generation and Summarisation. Association for Computational Linguistics, pp. 31–38.

Saggion, Horacio, 2011. Learning Predicate Insertion Rules for Document Abstracting. In: International Conference on Intelligent Text Processing and Computational Linguistics. Springer, pp. 301–312.

Siddharthan, Advaith, Teufel, Simone, 2007. 'Whose Idea Was This, and Why Does It Matter? Attributing Scientific Work to Citations'.

Slamet, Cepi et al., 2018. Automated Text Summarization for Indonesian Article Using Vector Space Model. In: IOP Conference Series: Materials Science and Engineering. IOP Publishing, p. 012037.

Steinbach, Michael, Karypis, George, Kumar, Vipin, 2000. A Comparison of Document Clustering Techniques. In: KDD Workshop on Text Mining, Boston, 525–526.

Steinberger, Josef, Jezek, Karel, 2004. Using latent semantic analysis in text summarization and summary evaluation. Proc. ISIM 4, 93–100.

Sun, Xiaoping, Zhuge, Hai, 2018. Summarization of scientific paper through reinforcement ranking on semantic link network. IEEE Access 6, 40611–40625.

Teufel, Simone, Siddharthan, Advaith, Tidhar, Dan, 2006. Automatic classification of citation function. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 103–110.

Teufel, Simone, Moens, Marc, 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. Comput. Linguist. 28 (4), 409–445.

Tomokiyo, Takashi, Hurst, Matthew, 2003. A Language Model Approach to Keyphrase Extraction. In: *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 33–40.

Tsai, Chen-Tse, Kundu, Gourab, Roth, Dan, 2013. Concept-based analysis of scientific literature. In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 1733–1738.

Vapnik, Vladimir Naumovich, 1995. Estimation of Dependences Based on Empirical Data. 1982. Springer-Verlag, NY.

Wan, Stephen, Paris, Cécile, Dale, Robert, 2009. Whetting the appetite of scientists: producing summaries tailored to the citation context. In: *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 59–68.

Wang, Jie, Ma, Shutian, Zhang, Chengzhi, 2017. Citationas: a summary generation tool based on clustering of retrieved citation content. Framework 7 (8).

Widyantoro, Dwi H., Amin, Imaduddin, 2014. Citation Sentence Identification and Classification for Related Work Summarization. In: 2014 International Conference on Advanced Computer Science and Information System. IEEE, pp. 291–296.

Yang, Shansong et al., 2016. Amplifying scientific paper's abstract by leveraging data-weighted reconstruction. Inf. Process. Manage. 52 (4), 698–719.

Yang, Yiming, Pedersen, Jan O., 1997. A comparative study on feature selection in text categorization. In: Icml, 35.

Yasunaga, Michihiro et al., 2019. Scisummnet: a large annotated corpus and content-impact models for scientific paper summarization with citation networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7386–7393.

Yeh, Jen-Yuan, Hsu, Tien-Yu, Tsai, Cheng-Jung, Cheng, Pei-Cheng, 2017. Reference scope identification for citances by classification with text similarity measures. In: *Proceedings of the 6th International Conference on Software and Computer Applications*, pp. 87–91.

Zajic, David, Dorr, Bonnie J., Lin, Jimmy, Schwartz, Richard, 2007. Multi-candidate reduction: sentence compression as a tool for document summarization tasks. Inf. Process. Manage. 43 (6), 1549–1570.

Zamir, Oren, Etzioni, Oren, 1999. Grouper: a dynamic clustering interface to web search results. Comput. Networks 31 (11–16), 1361–1374.

Please cite this article as: N. Ibrahim Altmami and M. El Bachir Menai, Automatic summarization of scientific articles: A survey, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.020