International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 12, No.1 (Aug-2022)

https://dx.doi.org/10.12785/ijcds/120154

Semi-Extractive Text Summarization Approach to Extract Research Objective of Academic Literature

Shubair A. Abdullah[1]

[1]*Department of Instructional Learning Technology, Sultan Qaboos University, Muscat, Sultanate of Oman*

Received 11 Feb. 2022, Revised 14 May 2022, Accepted 15 Jul. 2022, Published 6 Aug. 2022

Abstract: The task of literature review is an important for all research areas. The volume of academic literature published in the scientific journals, books and conference proceedings is increasing rapidly. It has become very difficult for researchers to analyze huge amount of published research papers and draw conclusions. Therefore, an automatic system able to extract the research objectives that recap the meaning and the definition of research from the article published is needed. The main challenge is to deal with the highly multiplicity of authors' writing style used in forming the objective sentences. A semi-extractive approach for precisely extracting research objective from the research abstract is proposed. The extraction process is performed in two tasks, search for pre-collected Key Expressions and verify the findings using pre-built General Key Phrases. An extensive evaluation has been conducted to find out how well it is performing using special dataset that includes diversified writing styles. The results showed that the method is a promising and able to extract objective sentence accurately.

Keywords: Extractive Summarization, Research Objective, Literature Review, Key Expressions, Key Phrases, General Key Phrases

1. Introduction

Conducting a literature review is an important task in writing research. The purpose of a literature review is to understand the existing research of a particular field of study, and it contributes significantly to increasing and deepening the knowledge of researchers in their fields. Recently, the literature review process has seen significant developments with the emergence of new directions in methods and purposes. In this context, an important research direction called systematic review has recently emerged that is mainly based on summarizing and reviewing the literature. A systematic review is a summary of literature that uses categorical methods to systematically search and critically appraise on a specific issue [1]. Until very recently, the researchers used to summarize the published articles by their own. They relied in most cases on the abstract part to quickly grasp the key points of a research paper, and specifically, most of them focused on the research objectives within the abstract because the objectives recap the meaning and the definition of research, and it is not possible to imagine conducting research aimlessly [2].This mechanism has become very difficult for researchers nowadays. The big challenge is to cope with the huge amount of published research papers that grows exponentially on the Internet. For example, there are various commercial publishing firms such as Springer Nature, Elsevier, Wiley-Blackwell, and Taylor Francis, who collectively own about 2,000 journals and publish thousands of papers annually [3]. To overcome this issue, automatic extraction of objective sentences is needed.

The issue of designing and developing an automatic system to extract objective sentences from the abstract part could be classified as text summarization issue. The text summarization is an automatic system that generates a condensed version of documents. It is categorized into two distinct classes, abstractive text summarization and extractive text summarization. The abstractive summarizer reforms the extracted sentences to produce the summary, and the extractive summarizer extracts the most important sentences and collects them together to produce the summary [4]. El-Kassas et. al. mentioned another class called hybrid that combines both the abstractive and extractive classes [5].

Each class of text summarization has its own challenges. Despite they have being successfully applied on short documents, the abstractive summarizers have two main challenges, determining the critical concepts in the original text and paraphrasing the concepts based on the grammar rules and the constraints of the natural language [6]. On the other hand, the main challenge of extractive summarizers is how to make the system mimic the way human experts write summaries very effectively with the existence of redundancy, spreading, and lack of semantics and cohesion in original text [7], [8].

E-mail address: shubair@squ.edu.om

With regard to the research objectives, this paper aims at proposing a semi-extractive approach to extract the objective sentence from the abstract part within research papers. The word “semi-extractive” implies the extraction process, that is, completely different from abstractive approach and similar to extractive approach. More clearly, the extraction of the objective sentences is performed in two tasks, search for pre-collected Key Expressions (KEs) and verify the finding using pre-built General Key Phrases (GKPs). These two tasks are carried out independently of the critical concepts in the original text and the redundancy, spreading, and lack of semantics and cohesion in the text within the abstract part. The aim is to overcome the highly multiplicity of authors’ writing style, which makes the search process very difficult and hinders the extraction of the objective sentences from the abstracts. For example, some authors start writing the objective sentence in the form of “this paper aims to survey ....”, while others use the form of “this paper aims at presenting survey . . . .” or the form of “this paper presents a survey . . . .”. Based on that, the key factor of the proposed method is the extraction way of objective sentences.

The contributions or this paper are represented by introducing a semi-extractive approach to extracting objective sentences with high accuracy and speed from the abstracts using pre-collected KEs and pre-built GKPs. Although the research on the text summarizations started a long time ago and achieved many successes [8], to the best of researcher’s knowledge, this paper is the first research that attempts to apply the text summarization approaches to extract objective sentences using KEs and GKPs. The proposed method will be evaluated extensively to find out how well it is performing using a carefully constructed dataset that includes diversified writing styles. The success of such research attempts will greatly facilitate writing and summarizing the literature review, whether in theses or in research papers. It also facilitates the implementation of systematic literature review, which has attracted the attention of many researchers in various fields recently.

The paper is structured as follows: Section 2 reviews the literature. Section 3 introduces the proposed method. After explaining the experiments conducted in section 4, the results of the experiments and a discussion of some observations that were made on the results are described in section 5. The conclusion and future work is explained in the section 6.

2. Literature Review

The task of text summarization is an important step to understand written texts. Increasing the number of texts dramatically, as is the situation now, makes the process of summarizing almost impossible for humans. Therefore, automation of the text summarization process has become a necessity to help humans comprehend the text content in a very short time [9]. Building a successful automatic text summarization system is a highly challenging task because it depends on the natural language and the semantic language restrictions. The building task is usually carried out in two stages, pre-processing and processing stages. The pre-processing stage aims at preparing a structured representation of the text. Several techniques have been applied to achieve this aim such as sentence segmentation, word tokenization, stop-words and punctuation marks removal [10], [11]. On the other hand, the processing stage takes the structured representation resulting from the pre-processing stage and applies summarization approaches to summarize the text. In fact, most published research focused on this stage, therefore, many techniques have developed and used for the purpose of summarizing texts with high accuracy.

In this paper, a method to automatically extract the objective of research papers from the abstract is presented. The proposed method does not differ from the methods previously presented in the literature. It involves two main stages, preparation stage and extraction stage, plus an initialization stage that aims at collecting the literature metadata. However, the method addresses some issues identified among the challenges that hinder building a successful text summarization system [5]. A recap of the important techniques and challenges that were addressed in previous research and related to this paper are presented in the following.

The string matching challenge is one of challenges that have attracted researchers. To address this challenge, text chunking techniques have been employed to split sentences in the original text into chunks of different types like verb groups and noun groups and do the matching based on chunks. For example, Maiti, S. et. al. proposed a formulation of tree-matching method to address the key issue related to evaluation of text chunking method. The method consists of two parts the first part is to find the structure error in terms of chunks error and words error, and the second part deals with the grammatical labeling. The algorithm has been evaluated using a chunker and a set of test sentences, and the results showed a promising results [12]. Another challenge is the conversion of natural language text to a complete meaning representation. The semantic parsing technique has been employed to overcome the conversion challenge. For example, the research presented by Mohamed M. Oussalah M. investigated and proposed a graph-based text summarization model using semantic role labeling and Wikipedia-based explicit semantic analysis [13]. The model involved two principal stages. The first stage involves a pre-processing task followed by semantic parsing task in addition to construction of Wikipedia index. The second stage deals with the core summarization tasks, interpreting semantic argument terms to Wikipedia concepts, computing intra-sentence similarities from Wikipedia concepts, constructing document similarity graphs, and sentence ranking and summary extraction. The results of experiments revealed a considerable performance the role-based semantic representation. The redundancy problem, which refers to incorporating repeated information in text summaries, has been addressed in the literature. Among the sematic techniques that have been tried to solve the redundancy problem is the textual entailment technique [14]. The textual entailment technique is an inference technique that attempts to infer the meaning of one text by another text, and it has been introduced as a generic semantic-based framework for text summarization [15], [16]. The LCEAS system [17] employed textual entailment technique to distinguishing between the important and unimportant sentences. In the final stage, cosine directional similarity method [18] is applied to identify non redundant sentences that must be included in the final summaries. The challenge of summarizing long text efficiently and accurately also attracted the attention. W. Shuai et al. presented EA-LTS text summarization approach that combines extractive and abstractive approaches to deal with long text. In the extraction phase, the system combines sentence vector and Levenshtein distance to conceive a hybrid sentence similarity measure. The key sentences are extracted by integrating the similarity measure into graph model. The abstraction phase constructs a recurrent neural network and devises pointer and attention mechanisms to generate summaries [19].

In summary, the research literature focused on the following issues, which are among the most curial issues that need more investigation to build an accurate automatic text summarization system: 1) String matching, 2) Conversion of natural language text to a complete meaning representation, 3) Redundancy problem, and 4) Summarizing long text efficiently and accurately. The proposed method in the research paper attempts to find solutions to all these challenges, as will be explained in the next section, which explains the proposed method in detail.

3. Proposed Method

This section presents a method proposed to automatically and precisely extract the objective of research papers from the abstract. Mainly, the method involves two stages, preparation stage and extraction stage, in addition to the initialization stage that aims at collecting the literature metadata. All stages have been implemented in Python using mid-range specifications PC with Intel i5 CPU, 8GB RAM, and Windows 10 OS. Figure 1 shows the proposed method.

A. Initialization

The purpose of initialization stage is to collect literature metadata from databases. It involves two tasks, preparing the file of research metadata and creating a text file for each abstract in the metadata file. The first task was carried out on Scopus database online platform by searching for research papers and exporting the search results to Excel CSV file, while the second task was carried out by developing a Python program (prepareAbstracts.py) to read the abstract column in the CSV file and save into separate txt file. A total of 50 research papers published in the Scopus database was targeted in this stage.

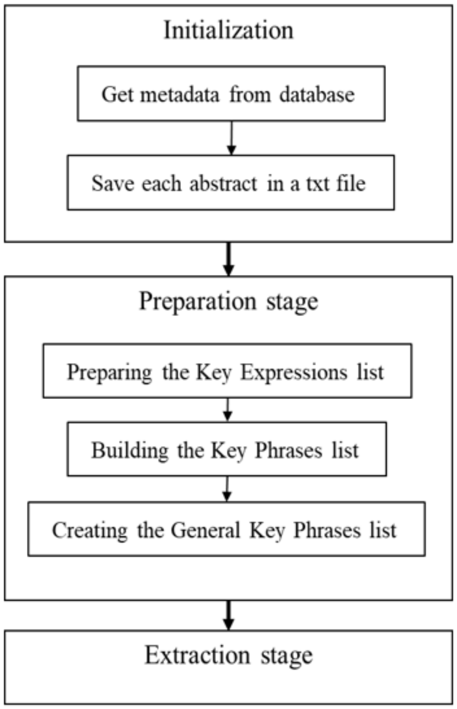

Initialization
Get metadata from database
Save each abstract in a txt file
Preparation stage
Preparing the Key Expressions list
Building the Key Phrases list
Creating the General Key Phrases list
Extraction stage

Figure 1. Proposed method to extract research objective

B. Preparation Stage

This stage involves preparations of the necessary items that will be used in extracting research objective. The work included examining and analyzing the abstracts in 50 research papers collected from the previous stage without looking at any other part of the papers. The following sections provide a comprehensive explanation of tasks of preparation stage.

1) Task-1: Preparing the Key Expressions list

This task is very important and is considered the basis for building the system. It involves collecting Key Expressions (KEs), which are expressions used to start explaining the objective of the article in the abstract part. The importance of the KEs lies in the fact that they will be used to search the abstract for the sentence used to explain the research paper objective. The KEs particular role is to solve the string pattern matching problem through collecting and mining expressions commonly used to explain the objective and using these expressions to speed up the string pattern matching task. Therefore, the greater the number of various KEs are collected, the greater the ability and accuracy of the system to extract the objective of the research. By analyzing 50 research papers from the Scopus database, 36 KEs were collected initially. The KEs collected were divided into English language sentences, and then synonymous English expressions were added to cover new set of KEs might not have collected in the examining phase. For example, the KE “the current paper” has other synonymous expressions added to it such as “the current article”, “the current study”, and “the current project”. Another example is the KE “in this research” to which other KEs have been added, such as “in this study” and “in this article”. The total number of KEs resulting from this task was increased to 228 KEs after adding the synonymous expressions.

2) *Task-2: Building the Key Phrases list*

As in the previous task, the second task was carried out by focusing on the 50 research papers from Scopus database. It sought to build the list of Key Phrases (KPs), which are phrases in English built using the KEs that have been extracted in the first task. For example, several KPs have been built from the KE “this paper” such as “this paper aims to survey”, “this paper aims at presenting survey”, and “this paper presents a survey”. Preparing the KPs list is crucial because upon finding a value of KE, the KPs will be used to determine whether this value will extract precisely the statement of research objective from the abstract or not. After completing this task, 663 KPs were built from 228 KEs.

3) *Task-3: Creating the General Key Phrases list*

The task of creating the General Key Phrases (GKPs) list is complementary to the previous tasks. It involves processing words in KPs and attaching a part of the speech mark to each word to creating GKPs. The GKPs are crucial to reduce the time needed to search for KEs and to verify its contribution to extracting the statement of research objective. It is clear and certain that each author has his own style of writing. For example, some authors start writing the objective sentence in the abstract part in the form of “this paper aims to survey ....”, while others start it in the form of “this paper aims at presenting survey . . . .” or in the form of “this paper presents a survey . . . .” or “this study focuses on the development . . . .”. The authors’ multiplicity of writing style in such a large way makes it very difficult to search for the sentence that explains the research objective in the abstract. The GKP could refer to multi KPs of different words and/or verbs. For example, the KPs “We address this problem by proposing different topology ....”, “We overcome this issue by introducing novel algorithm ...”, and “We tackle this issue by creating compressive approach . . . ” could be generalized with the GKP “pronoun verb determiner noun preposition verb+’ing’ adjective noun”. As a result of this process, 194 GKPs were created. To achieve a greater level of generalization and to reduce this large number of GKPs that may affect the performance of the system, another work was done concerning the similarity among the GPs. If there is a particular KP and another KP that is similar to it by at least 80%, then the two KPs belong to the same GKP. For example, if we have the KP 1 “a network environment is proposed” has GKP 1 “determiner noun noun verb verb (past participle)”, then KP 2 “a new network environment is proposed”, which is 86% similar to KP 1 will be represented by the same GKP 1. This step has reduced the number of GKPs to 135. This task is a form of semantic parsing forms. It attempts to convert the KPs to a general complete meaning representation. The significance of this task lies in the fact that it creates GKPs, and each of these GKPs refer to or substitutes for a large group of KPs. This naturally leads to facilitating the process of dealing with long texts and solving the problem of redundancy.

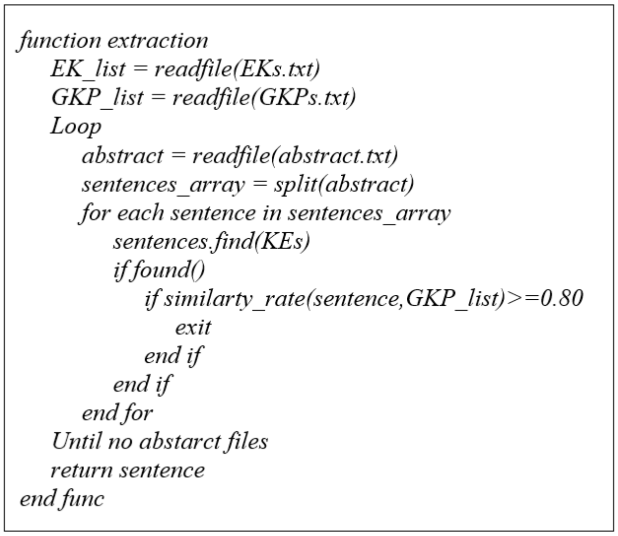
C. *Implementation of Preparation Stage*

Three different Python programs developed relying on the state-of-the-art Python packages for OS functionality, natural language processing, and data management, namely OS, NLTK, and Pandas. Table I identifies the names of programs as well as the purposes of them. As for the result of the preparation stage implementation, it can be summed up in creating the following items: 228 KEs, 663 KPs, and 135 GKPs. Table II shows examples of KEs, KPs, and GKPs.

D. *Extraction Stage*

The main goal of this stage is to extract the sentence explaining the objective of the research that the author included in the abstract. To access this sentence and extract it accurately, all paragraphs in the abstract must be firstly split into sentences. Then a process of scanning the sentences to find the potential KEs is performed. Every KE found will be verified to find out the indicative of the objective sentence. If the sentence containing the KE is at least 80% similar to one of the GKPs, the system adopts the sentence as the objective sentence and stops the scanning. Otherwise, the scanning will continue on the rest of the sentences.

The extraction stage was implemented by developing a program in Python that works mainly on the KEs and GKPs that were created in the preparation stage and uses some Python packages CSV, OS, Pandas, and NLTK. Figure 2 shows the pseudo code of the extraction stage implementation.



```
function extraction
    EK_list = readfile(EKs.txt)
    GKP_list = readfile(GKPs.txt)
    Loop
        abstract = readfile(abstract.txt)
        sentences_array = split(abstract)
        for each sentence in sentences_array
            sentences.find(KEs)
            if found()
                if similarty_rate(sentence,GKP_list)>=0.80
                    exit
                end if
            end if
        end for
    Until no abstarct files
    return sentence
end func
```

Figure 2. Pseudo Code

The above script performs the extractions of objective sentence. First, it reads the KEs and GKP lists into two

TABLE I. Implementation of preparation stage programs

Program	Packages	Purpose
prepareAbstractsText.py	OS and pandas	To read the Excel sheet of Scopus and create text file for each abstract
createKeys.py	NLTK and pandas	To create list of KEs and list of KPs
wordTagging.py	NLTK	To create list of GKPs

TABLE II. Examples of KES, KPS, and GKPS

KEs	Synonymous KEs added	KPs built from the KE	General Key Phrases
"the current paper"	"the current article"; "the current study"; "the current project"; "the current letter"; "the current review the current survey"	"the current paper presents a survey"; "the current paper aims at presenting a survey"; "the current paper aims to present a survey"	"determiner; adjective; noun; verb, present tense with 3rd person; determiner; noun" "determiner; adjective; noun; verb, present tense with 3rd person; preposition; verb gerund; determiner; noun"
"this research"	"this paper"; "this article"; "this study"; "this project"; "this letter"; "this review"; "this survey"; "this work"	"this paper aims to survey"; "this paper aims at presenting survey"; "this paper presents a survey"; "this study focuses on the development"; "this study focuses on the new"; "this study focused on the new"; "this study focused on the development"; "this paper is presenting a survey";	"determiner; noun; verb, present tense with 3rd person; infinite marker; noun" "determiner; noun; verb, present tense with 3rd person; preposition; verb gerund; noun" "determiner; noun; verb, present tense with 3rd person; preposition; determiner; adjective" "determiner; noun; verb, present tense with 3rd person; verb gerund; determiner; noun"

separate lists, and starts a loop to process each abstract. On each iteration, the abstract is broken down into sentences. The algorithm searches for KEs within each sentence in the abstract. If a match occurs, it checks the similarity between the sentence and the KGP lists. If the similarity is greater than or equal to 80%, the sentence is promoted as an objective sentence.

4. Experiments

A. Setting Up the Experiments

For the purpose of collecting data, three searches were performed on Scopus and IEEE databases for research papers. These two databases were selected because they are among the most extensive databases of publications with powerful resources for accessing different types content. Each search attempt was performed using a query string associated with a scientific field, i.e. "IPv6 Security" to search for papers published by computer engineering and computer science researchers, "Computer and Education" to search for papers published by educational researchers, and "Covid-19 Vaccine" to search for papers published by medical researchers. The goal of diversifying the searches was to obtain different styles of writing abstracts, and thus experimenting the method using the largest possible number of styles of writing the research objectives. As a result of the searches, 3 datasets containing 2296 research papers from three different sciences and published in two well-known databases were obtained.

A task of exclusion of duplicated studies has been performed by making use of the features available in MS Excel, i.e. remove duplicate in the data tab and using the Match function. The duplication was carried out at two levels, excluding the duplicates in each database separately and excluding the duplicates in the two databases together.

The datasets were used at the preparation stage that aimed at building the KEs, KPs, and GKPs, and the number of research papers was 50 research papers selected from the IPv6 security topics. The rest of research papers, 2276 were used in the process of experimenting the method. Table III shows the datasets.

B. Measurement Approach

To evaluate how well the method is performing, a special approach to measure its accuracy is followed. The approach involved assigning a value to three type of results:

1) True Positive (TP): The system has extracted a sentence and it is the objective sentence
2) False Positive (FP): The system has extracted a sentence but it is not the objective sentence
3) False Negative (FN): The system has not extracted a sentence (empty output)

TABLE III. The Datasets

Database	Topic	no. of papers
IEEE	IPv6 Security	366*
	Computer and Education	573
	Covid-19 Vaccine	181
Scopus	IPv6 Security	322
	Computer and Education	473
	Covid-19 Vaccine	381
	Total of research papers	2296

*50 used at the preparation stage

It was important to extract the objective sentences by reading the abstract to be compared with the system outputs and to calculate the values of TP and FP. This task was done only in the case of TP or FP. There was no need to read the abstracts and extract objective sentences in the case of FN because no output to compare.

The accuracy of the proposed method was calculated using the TP and FP. These two results reflect extractions done from the abstracts. However, the accuracy value does not refer precisely to the extraction of the objective sentences because the FP value is included in the calculations. Therefore, two other common metrics were used, Recall and Precision [20], [21]. The recall measures the percentage of times the method was able to extract the objective sentences correctly, and the precision measures the percentage of times the method has extracted objective sentences correctly amongst all the times that objective sentences have been extracted by the method.

The proposed method was evaluated using regression analysis. The purpose was to measure how close are the objective sentences that have been extracted, but not the objective sentences, to the actual objective sentences. The task was performed by checking the difference between the objective sentence extracted and the actual objective sentence that is written in the abstract. In this context, the extraction error (EE), which is the percentage of dissimilarity between characters of the extracted objectives and characters of the actual objectives, was calculated for the FP results only [22]. As the large percentage of dissimilarity is particularly undesirable, there is need for an appropriate metric to determine the large EE values. To for that purpose, the Root Mean Squared Error (RMSE) has been calculated [23]. The formulas used in calculation are as follows:

$$Accuracy = \frac{(TP + FP)}{n} \quad (1)$$

$$Recall = \frac{TP}{n} \quad (2)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{n=1}^{n} (EE_i)^2}{n}} \quad (4)$$

where n is the number of abstracts in the dataset

5. Experimental Results

A. Results and Discussion

This section shows the results of the experiments and discusses some observations that were made on the results during the experiments. Table IV shows the results of experiments in terms of TP, FP, and FN for papers of three topics published in IEEE and Scopus databases.

The experimental results can be divided into two parts. The first part is called “extraction part”. It contains the results of TP and FP that indicate successes in extracting a sentence from the abstract. On the other hand, the second part is called “no-extraction part” and it contains the FN results (empty outputs) that indicate fails in the extraction. The results in the extraction part confirm that the list of KEs collected contains useful KEs that may or may not lead to the correct objective sentences in the abstracts, while the results in the no-extraction part confirm that there is a shortage of KEs collection. That is because the system uses the KEs to search for the objective sentence within the abstract. If it does not find one of the KEs, it means that the collected list of KEs is insufficient. Table V shows the percentages of extractions for the TP and FP part and for the FN part (empty outputs).

The largest percentage in all experiments is for the extraction part. For example, the rate of extraction is 98% for papers of Computer and Education topics published in IEEE, and 97% for the papers of IPv6 Security topics published in Scopus. This indicates that the list of KEs is effective in searching for the objective sentences and the system is successful in extracting sentences. The lowest rates 80% and 64% were for papers of Covid-19 Vaccine topic published in both databases, IEEE and Scopus. These lower rates can be explained by the way abstracts are written in medical research papers. Some journals require dividing the abstracts into four sections, background, methods, results, and conclusions [24]. This common division in abstracts of the medical research papers caused the

TABLE IV. The experimental results in terms of TP, FP, AND FN

Database	Topic	TP	FP	FN	no. of papers
IEEE	IPv6 Security	248	23	45	316
	Computer and Education	511	50	12	573
	Covid-19 Vaccine	132	12	37	181
Scopus	IPv6 Security	297	15	10	322
	Computer and Education	359	58	56	473
	Covid-19 Vaccine	117	65	139	381

TABLE V. The experimental results in term of Rate of extractions (TP and FP) and Rate of no extractions (FN)

Database	Topic	Rate of extractions (TP and FP)	Rate of no-extractions or empty outputs (FN)
IEEE	IPv6 Security	86%	14%
	Computer and Education	98%	2%
	Covid-19 Vaccine	80%	20%
Scopus	IPv6 Security	97%	3%
	Computer and Education	88%	12%
	Covid-19 Vaccine	64%	36%

ineffectiveness of the collected KEs.

The accuracy of the proposed method was calculated to evaluate how well the method is performing in terms of extraction. Figure 3 shows the extraction accuracy for each database.

Despite the high values of extraction accuracy in both databases, the accuracy does not refer precisely to the extraction of the objective sentences because the FP value is included in the calculations. Thus, the recall and the precision measurements were calculated to evaluate ability of the method to extract the objectives correctly. Table VI shows the results of calculating the recall and precision for papers of three topics published in IEEE and Scopus databases

Regarding the recall metric, which measures the percentage of times the method was able to extract the objective sentences correctly, the results are encouraging and confirm that the method is a promising. The highest values, 89% and 92% were for papers of "Computer and Education" topic published in IEEE and "IPv6 security" topic published in Scopus. The low values of recall 73% and 46% were for the papers of "Covid-19 Vaccine" topic in IEEE and Scopus. The two values can be justified by the fact that most of the journals belonging to IEEE are interested in the field of engineering and computer science and do not follow a method of dividing the abstracts into four sections, background, methods, results, and conclusions as in most of the Scopus-indexed medical journals.

The precision metric measures the percentage of correctly excreted objective sentences among the extracted sentences by the method. Most of the values for the three topics in IEEE and Scopus databases are high. As the GKPs list is crucial in checking the KEs, the high precision

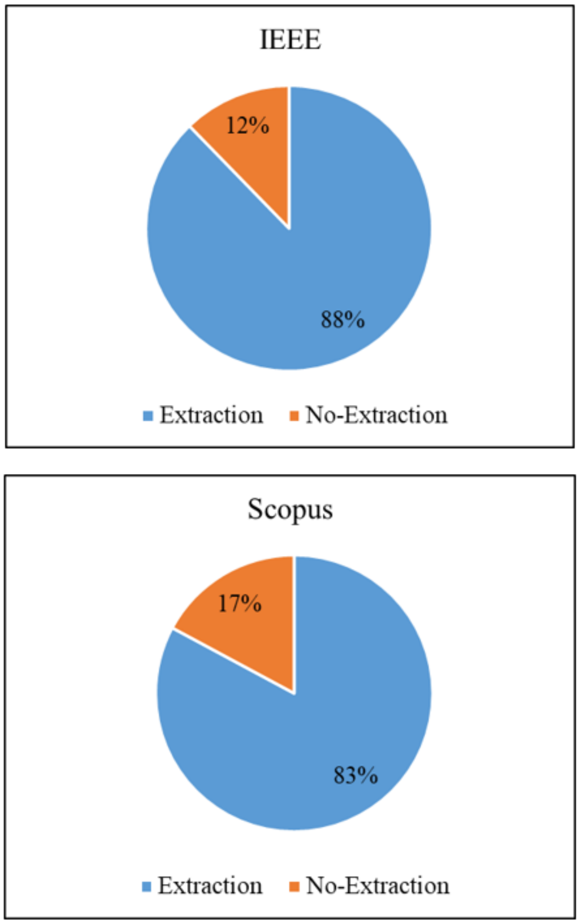


Figure 3. Extraction accuracy for each database

TABLE VI. The precision and recall calculation results

Database	Topic	Recall	Precision
IEEE	IPv6 Security	78%	92%
	Computer and Education	89%	91%
	Covid-19 Vaccine	73%	92%
Scopus	IPv6 Security	92%	95%
	Computer and Education	76%	86%
	Covid-19 Vaccine	46%	73%

values mean that the built-up GKPs list is highly effective in identifying the objective sentences out of the extracted sentences. However, the precision value 73% for the papers of “Covid-19 Vaccine” topic published in Scopus might be considered a weakness in building up the GKPs list.

To further examine the method, this paper used the regression analysis to explain the relationship between the original sentences and the extracted sentences. Focusing on the FP results only, two metrics have been calculated, the extraction error (EE) and the Root Mean Squared Error (RMSE). The EE finds the percentage of dissimilarity between characters of the extracted objectives and characters of the actual objectives, and the RMSE determines the large EE values. Table VII shows the EE and RMSE calculation results.

The EE column shows the percentages of dissimilarity between the objective sentences and the extracted objective sentences. It is clear that the percentages of dissimilarity are rather high. The RMSE gives an idea of the average dissimilarity between the objective sentences and the extracted objective sentences. The lowest RMSE value 0.43 was for papers of “Covid-19 Vaccine” published in IEEE but what weakens its significance is that the number of FPs is only 12, which is a small number. The highest RMSE values were 0.54 and 0.53 for the papers of “Computer and Education” topics published in IEEE and Scopus.

To clarify the RMSE values further, a normalization has been carried out using the following formula:

$$NormalizedRMSE = \sqrt{\frac{RMSE}{max(EE) - min(EE)}} \quad (5)$$

Normalized RMSE value is between 0 and 1, where values closer to 0 represent better extraction. Table VIII shows the RMSE and the normalized RMSE values.

The normalized RMSE column shows that some values are 1 and some are close to 1. In general, the RMSE values confirm the weakness in building up the GKPs list. However, the reason may be a weakness in the list of KEs too. For example, if there is KE like “was investigated” and the author uses this phrase in summarizing the literature within the abstract as in “... critical evaluation of these algorithms in relation to the research problem was investigated ...”, then the system will find match between the KE and this sentence. Also when comparing this sentence with the GKPs list, it is possible that there will a match that lead to incorrectly tag this sentence as an objective sentence.

B. *Summary*

It is can be concluded from the experimental results that the proposed method is a promising method in extracting objective sentence from the abstract section. However, it has some challenges that need further research and study and may determine future research directions. These challenges were mainly revealed through the experiments of medical research papers published in Scopus as well as some technical papers published in IEEE and Scoops. The challenges could be summarized in the following:

- Ineffectiveness of the collected KEs in extracting objective sentences from abstracts of medical research papers that were divided into four sections, background, methods, results, and conclusions
- Inadequacy of the collected GKPs to build a list of GKPs increased the percentage of dissimilarity between the objective sentences and the extracted objective sentences

Despite these challenges, the proposed method can be adopted for writing the literature review, whether in theses or in research papers. It can also be adopted in the implementation of systematic literature review. The results showed that correct objective sentences can be extracted with at least 50% of the entered abstract, which is an excellent ratio for summarizing research trends in a particular field while writing the literature review section.

6. Conclusion

The goal of this research paper is to introduce a semi-extractive approach to extract the objective sentence from the abstract part within research papers. The idea is inspired by the highly multiplicity of authors’ writing style, which hinders the extraction of the objective sentences from the abstracts. The big challenge for researchers is to automatic summarize the huge amount of published research papers that grows exponentially on the Internet. In particular, this paper focused on several common issues: string matching, conversion of natural language text to a

TABLE VII. The EE and RMSE calculation results

Database	Topic	FP	EE Rate	RMSE
IEEE	IPv6 Security	23	44%	0.45
	Computer and Education	50	51%	0.54
	Covid-19 Vaccine	12	40%	0.43
Scopus	IPv6 Security	15	41%	0.44
	Computer and Education	58	50%	0.53
	Covid-19 Vaccine	65	45%	0.48

TABLE VIII. The RMSE and normalized RMSE values

Database	Topic	RMSE	normalized RMSE
IEEE	IPv6 Security	0.45	0.95
	Computer and Education	0.54	1.00
	Covid-19 Vaccine	0.43	0.75
Scopus	IPv6 Security	0.44	1.00
	Computer and Education	0.53	0.92
	Covid-19 Vaccine	0.48	0.81

complete meaning representation, redundancy problem, and summarizing long texts. The approach presented includes involved two stages, preparation stage and extraction stage, in addition to the initialization stage that aims at collecting the literature metadata. The extraction stage is conducted in two tasks, search for pre-collected Key Expressions (KEs) and verify the finding using pre-built General Key Phrases (GKPs). Several experiments have been conducted to evaluate the method using 2296 research papers from two famous databases IEEE and Scopus. From the accuracy rates recorded during experiments, it could be concluded that the proposed method is a promising method in extracting objective sentence from the abstract section. However, it has some challenges that need further research and study and may determine future research directions. These challenges were mainly revealed through the experiments of medical research papers published in Scopus as well as some technical papers published in IEEE and Scoops. Despite these challenges, the proposed method can be adopted for writing the literature review, whether in theses or in research papers. It can also be adopted in the implementation of systematic literature review.

References

[1] S. Gopalakrishnan and P. Ganeshkumar, "Systematic reviews and meta-analysis: understanding the best evidence in primary healthcare," *Journal of family medicine and primary care*, vol. 2, no. 1, p. 9, 2013.

[2] M. S. Islam and S. Samsudin, "Characteristics, importance and objectives of research: An overview of the indispensable of ethical research," *Science and Technology*, vol. 33, no. 3, pp. 57–62, 2020.

[3] M. Ware and M. Mabe, "The stm report: An overview of scientific and scholarly journal publishing," 2015.

[4] M. Yousefi-Azar and L. Hamey, "Text summarization using unsupervised deep learning," *Expert Systems with Applications*, vol. 68, pp. 93–105, 2017.

[5] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, p. 113679, 2021.

[6] B. Mutlu, E. A. Sezer, and M. A. Akcayol, "Candidate sentence selection for extractive text summarization," *Information Processing & Management*, vol. 57, no. 6, p. 102359, 2020.

[7] L. Hou, P. Hu, and C. Bei, "Abstractive document summarization via neural model with joint attention," in *National CCF Conference on Natural Language Processing and Chinese Computing*. Springer, 2017, pp. 329–338.

[8] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," in *2017 international conference on computer, communication and signal processing (ICCCSP)*. IEEE, 2017, pp. 1–6.

[9] S. Song, H. Huang, and T. Ruan, "Abstractive text summarization using lstm-cnn based deep learning," *Multimedia Tools and Applications*, vol. 78, no. 1, pp. 857–875, 2019.

[10] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," *Artificial Intelligence Review*, vol. 47, no. 1, pp. 1–66, 2017.

[11] V. K. Gupta and T. J. Siddiqui, "Multi-document summarization using sentence clustering," in *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*. IEEE, 2012, pp. 1–5.

[12] S. Maiti, U. Garain, A. Dhar, and S. De, "A novel method for performance evaluation of text chunking," *Language Resources and Evaluation*, vol. 49, no. 1, pp. 215–226, 2015.

[13] M. Mohamed and M. Oussalah, "Srl-esa-textsum: A text summarization approach based on semantic role labeling and explicit semantic analysis," *Information Processing & Management*, vol. 56, no. 4, pp. 1356–1372, 2019.

[14] E. Lloret and M. Palomar, “A gradual combination of features for building automatic summarisation systems,” in *International Conference on Text, Speech and Dialogue*. Springer, 2009, pp. 16–23.

[15] A. Gupta, M. Kaur, S. Mirkin, A. Singh, and A. Goyal, “Text summarization through entailment-based minimum vertex cover,” in *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014)*, 2014, pp. 75–80.

[16] S. Lagrini, M. Redjimi, and N. Azizi, “Automatic arabic text summarization approaches,” *International Journal of Computer Applications*, vol. 164, no. 5, pp. 31–37, 2017.

[17] F. T. AL-Khawaldeh and V. W. Samawi, “Lexical cohesion and entailment based segmentation for arabic text summarization (lceas).” *World of Computer Science & Information Technology Journal*, vol. 5, no. 3, 2015.

[18] A. Sirisha and A. K. Pradhan, “Cosine similarity based directional comparison scheme for subcycle transmission line protection,” *IEEE Transactions on Power Delivery*, vol. 35, no. 5, pp. 2159–2167, 2019.

[19] S. Wang, X. Zhao, B. Li, B. Ge, and D. Tang, “Integrating extractive and abstractive models for long text summarization,” in *2017 IEEE International Congress on Big Data (BigData Congress)*. IEEE, 2017, pp. 305–312.

[20] J. R. Thomas, S. K. Bharti, and K. S. Babu, “Automatic keyword extraction for text summarization in e-newspapers,” in *Proceedings of the international conference on informatics and analytics*, 2016, pp. 1–8.

[21] R. Naidu, S. K. Bharti, K. S. Babu, and R. K. Mohapatra, “Text summarization with automatic keyword extraction in telugu e-newspapers,” in *Smart computing and informatics*. Springer, 2018, pp. 555–564.

[22] Y. Ouyang, W. Li, S. Li, and Q. Lu, “Applying regression models to query-focused multi-document summarization,” *Information Processing & Management*, vol. 47, no. 2, pp. 227–237, 2011.

[23] T. Chai and R. R. Draxler, “Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature,” *Geoscientific model development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[24] J. Yang, W. Zheng, H. Shi, X. Yan, K. Dong, Q. You, G. Zhong, H. Gong, Z. Chen, M. Jit *et al.*, “Who should be prioritized for covid-19 vaccination in china? a descriptive study,” *BMC medicine*, vol. 19, no. 1, pp. 1–13, 2021.

Shubair A. Abdullah Shubair Abdul Kareem Abdullah received his BSc degree in computer science from Basra University in 1994, and his MSc and PhD degrees in computer science from University Sains Malaysia (USM) in 2007 and 2014 respectively. He is working as assistant professor at Sultan Qaboos University, Oman, Muscat currently. His research interests include data mining, network security, and fuzzy inference systems.