Keyword Extraction Using Latent Semantic Analysis For Question Generation

G. Deena[1,2*] and K. Raja[2]

[1] Sathyabama Institute of Science and Technology, Rajiv Gandhi Salai, Chennai, 600119, India
[2] Department of Computer Science and Engineering, SRM Institute of Science and Technology Bharathi Salai, Chennai, 600089, India
[*] Corresponding author. E-mail: gdeena.08@gmail.com

Received: Dec. 10, 2021; Accepted: May 28, 2022

In-Text Mining, Information Retrieval (IR), and Natural Language Processing (NLP) dig out the important text or word from an unstructured document is coined by the technique called Keyword extraction. It helps to identify the core information about the document in specific. Instead of going through the entire document, this method helps to retrieve sufficient information instantly in a short span of time. It is essential to mine the meaningful word from the document in text analytics. The proposed system has been based on semantic relation to extracts the keyword from unstructured text documents by means of practice like Latent Semantic Analysis (LSA). In view of this method, there exists a semantic relation between the sentences available in the document and the words. Extracted text permits to signify text in a strong way and has a high preference to carry more important information about the sentences. In this regard, LSA has produced better outcomes when compared with the TF-IDF, RAKE, YAKE, and Text Rank algorithm. Consequently, the keyword extraction has been occupied in Automatic Question Generation (ACQ) to generate the Fill up the blank (FB) and Multiple Choice Questions (MCQ) with distractor set. The top five, ten keywords are involved in questionable generation. The proposed system could be implemented in the question generation system to assess the skill level of the learner.

Keywords: Natural Language Processing, Latent Semantic Analysis, Multiple Choice Questions, Keyword Extraction, Semantic relation.

http://dx.doi.org/10.6180/jase.202304_26(4).0006

1. Introduction

Using a few keywords, users may immediately get a sense of what the article is about and what it's about. Information retrieval, automated abstract generation, text tagging, and text subject classification and extraction all rely on keyword extraction. As a result, the manual extraction of keywords is a lengthy process that is labor-intensive, and the derived keywords typically differ significantly from each other due of cultural differences and thinking styles. Research into ways of automatically extracting keywords from papers has become a popular focus.

One or series of words describe the document content is known as keyword. In the internet approximately the available number of the document is crossing 1.5 billion, in such a gigantic method the indexing and ranking process of information retrieval is tough to handle and implement. Keywords are a collection of words from a document to describe the fact of the document that helps to classify the document [1] either supervised or unsupervised [2]. Keywords, which we describe as a combination with one or many terms, serve as a concise representation of the content of a document. In an ideal world, keywords would

condense the main substance of a paper. Because keywords are simple to generate, edit, recall, and distribute, terms are frequently used to describe queries inside information retrieval (IR) systems. In contradiction to algebraic signatures, keywords are not bound to any particular text collection and can be used in several corpora and inductive reasoning.

Keywords are words that precisely as well as briefly represent the content or feature of the mentioned areas in a document, and they play a significant function as a sign of vital textual information that spreads among many of the users in both hard and delicate document formats. With words, scientists can express their ideas clearly and concisely, all while focusing attention on the message they are trying to convey. keywords are vital for readers who want a rapid overview of the content, but they are also useful for those who want to find related material about a specific topic.

As a result, keyword extraction tools based on Natural Language Processing (NLP) have been developed. Scientific records are always structured, making it possible for tools for natural language processing to interact with the language's grammatical and morphosyntactic properties. Because of the word limit on Twitter, unstructured text is frequently employed in this type of medium. There is a limit of 280 characters in a twitter. As a result, language is often used in an unstructured manner. It's indeed devoid of grammatical distinctions, which might lead to confusion in the reader's mind.

The majority of the document has no proper keyword representation to describe the document. In the existing approach, the skilled or the individual requires finding out the important text from the document. It is exceptionally hard to experience all the content, materials manually; hence there is a need for good data extraction tools to give the authentic meaning of the document. That process is known as keyword extractor.

Currently, large volumes of documents are available on the internet. The retrieval of information from them is termed by the technique called keyword extraction. The Keyword is the token that articulates the sense of the whole document that based on the user query the information has been extracted from the document.

The keyword extraction technique has been applied in various applications [3] in information retrieval, text mining, and in Natural Language Processing exists like clustering of document, indexing and filtering the document, identify the topic, and generating the question automatically to assess the learner's skill level [4, 5].

Extraction has been done manually by the subject expert or semi-automatically and by a fully automated system. All because of the increase in data size over the internet day-to-day the extraction of a keyword is not feasible by the expert who needs lots of time and finds it a tedious process. Hence the automatic extraction comes with practice categorized as a supervised and unsupervised method [6].

An Author Turney [7] has introduced the supervised method of extracting the keyword. The Supervised Extraction method needs training on the dataset to extract the keyword from the document. Initially the training has to set to develop the model on the unstructured text [8]. This supervised method is domain dependent, needs to provide training to each domain is not practically feasible also is highly expensive in nature [9].

Unsupervised the method extracts the keyword without any training on the dataset which is domain-independent. This type of method is practically applied to a different set of domains. Based on the scoring and ranking value the important keyword is extracted from the document. It is applicable to a single document to extract the important word from the document or from the collection of documents [10].

There are various stages in a retrieval process like preprocessing of documents, keyword extraction, and final post-processing to select the topmost terms of the applications. A keyword term has to be a non-zero integer value and is static in nature [11]. It is easy for the user to continue the document manipulation after the retrieval of the keyword.

There are several types of extraction methods like TF-IDF, Text Rank, RAKE, and YAKE. In our Proposed System, the Latent Semantic Analysis method is used to extract the keyword for the reason that it produces the semantic relation between the term and the document. This keyword makes semantic meaning that involved generating the objective type question to assess the learner's skill level.

The rest of the paper is structured as follows, Literature review under section 2, section 3 is to discuss the different keyword extraction method, section 4 discusses the experimental result followed the section 5 gives the conclusion of the work.

2. Literature review

Keyword extraction is the most important technique in Text Mining, Information retrieval, Text Analytics, Text Classification, and in Natural Language Processing. The evolution of data on the internet needs an exact method to retrieve the keyword or text. In [12], it is suggested that the best way to sum up the document is carried via keyword extraction. The extracted term has to be sense-making, if

not meet the terms fails in the text analytics process.

The presence or absence of a designated corpus is the primary difference between these two approaches. It is necessary for supervised learning to have a manually labelled corpus, which primarily changes the task into a classification problem. Decision trees (DT), support vector machines (SVM), and naïve Bayes (NB) are commonly used.The model's accuracy and the extracted keywords are both heavily influenced by the quality of both the training corpus. Now that only a small portion of the text corpus has been labelled with keywords, every training batch must be individually labelled. Labelling is the challenge process in supervised method.

There is no requirement for a well-labeled corpus when using the unsupervised learning method, and the probabilistic ranking can be used to determine the keywords of an article. First, some feature weight signals are established based on the statistical features of the qualifying keywords, and then, in order of increasing importance from high to low, the top words with the highest weight are selected as text keywords.

The unsupervised algorithm, [13–15] the traditional method to extract keyword is given by Term Frequency and Inverse Document Function (TF-IDF) which is the statistical method that follows the systematic way to produce the refined words that describe the outline of the whole document without the support of an expert or the document creator. TF-IDF is simple and easy to implement. It lacks in semantic retrieval of text from the document. The TF-IDF is a metric for assessing the relative weight given to various words in a piece of writing. An easy method for implementing it assumes that the words are autonomous and cannot indicate the meaning of words and their place in space.

In [16], the author followed the co-occurrence technique and frequency of words in the document retrieved from the documentation and evaluation was made through the chi-square, the same procedure has been followed in [17] without any training on the corpus.

In [18], one of the unsupervised algorithm lexical methods use the Part-of-speech (POS) followed by ranking and score value to extract keyword that improves the TF-IDF. However, this part-of-speech will not extract the meaningful text for all type of domain.

Another unsupervised method is Rapid automatic keyword extractor (RAKE) which is domain and language independent [19]. In long phrases of document RAKE extraction methods are preferred where the selections of keywords are depending on the score value. YAKE algorithm given in [20], is independent of domain and language and the extraction is based on the word-net method and works for a lengthy document. Keywords containing important discussion but no commas, stop - word, or even other phrases with minimal lexical value are more common, according to RAKE's author, than those with only one or two.

In [21], the TextRank algorithm has been implemented to extract keywords where the node and edges represent the document to establish the relation between the words. Based on the well-known PageRank algorithm, the TextRank algorithm was created. A unique document's intrinsic information is taken into account while using the classic TextRank algorithm for text keyword extraction. Unsupervised methods are used to extract the important text from the document without any training on the domain. This method is applicable to the entire domain.

Further, the Supervised algorithm exists to extract the keyword where it initially requires some training to be provided with extra care to the manually created dataset or the automatically generated dataset. If the text in the new document matches with the trained dataset, then it termed as keyword else not. This is the negative aspect of this supervised method.

First supervised algorithm is KEA given in [22]. A linguistic method is incorporating that prefer the features such as Part-of-Speech, Document frequency, inner document frequency, the first occurrence of the text, and its position to extract keyword. In [23] keyword extraction has been applied in academic documents. In [24], it has been specified that reading time has been reduced.

Our paper describes and compares the unsupervised method to extract the keyword word from the document and use those words to generate the objective type question like multiple choices with the distractor and fill up the blank. Implementation has produced the Latent semantic analysis (LSA) as the precise in keyword retrieval. LSA produces a semantic relation between the sentence and the word there in the document without providing any training. Semantically related word has high impact in the various types of applications in information retrieval and in Natural Language Processing.

3. Methods of keyword extraction

In Natural Language Processing (NLP) the keyword extraction is the most important technique to identify the importance of text in the document. Supervised and Unsupervised methods are used to extract keywords from the document which describes the document in short. Unsupervised methods do not require any training dataset or any supporting material to extract important keywords

from any domain and consequently better than supervised.

Unsupervised encompasses the statistical, graph-based approaches. In statistical, TF-IDF, TextRank, RAKE, YAKE, and Latent Semantic Analysis are used to retrieve important text to describe the concept of sentences in the document. The proposed system, concludes the importance of Latent Semantic Analysis (LSA) to retrieve the meaningful text from the document, which describes the semantic relation that exists between the sentences. The core idea is to establish the relation between the meanings of the words and helps to summarize the document without changing the originality of a document. During the analysis, the small passage is considered for identifying the co-occurrence of words and all passage is collected together to summarize the document.

3.1. Document Pre-Processing

Pre-processing is the foremost step in Natural Language Processing, Information retrieval, and Text Analytics. The document consists of a collection of a word which is given in unstructured form like paragraph, passage. The unstructured document has to convert into a structured format. All the terms in the document will not make sense in the application; hence it is superior to remove the unwanted text for better results. Pre-processing is in different stages. Initially, removes the stop word, tokenize the document followed by the stemming.

a Stop-word removal
The English language has many stop-words given in the set S;
S = { 'is','was','because', the', of', on'..... }
those words are frequently occurring text which does not have any importance in keyword extraction, after the removal of stop-words, the document consists of the left out words.

b Tokenization
It is the process of extracting the phrases or sentences, text, symbols, and delimiters from the paragraph. The paragraphs are tokenized into individual sentences known as sentence tokenization. These tokens are highly utilized in the text-processing applications. The abbreviation and short-form notations are removed from the paragraph. The sentences are tokenized to get a word which is known as word tokenization.

c Stemming
It is the most important process in text analytics to convert the variant form of words into their root form. Individual words are supplied as input to this stemming process. Letters like 'es','s', 'ing', 'ed' are truncated from the text. After the pre-processing, words that are necessary for keyword extraction are available for the betterment of the output.

3.2. Index Term

Index term is a collection of words from the document to represents the topic or concept of a document. These preselected words are used to review the document in squat. in the statistical approach, basically all the noun related words are referred as keywords which is the sense-maker of the document.

$$\mathrm{D} = \{\mathrm{d0},\ \mathrm{d1},\ \mathrm{d2}\ldots\mathrm{dn}\} \text{ where } \mathrm{D} > 0$$
$$\mathrm{V} = \{\mathrm{k}_1, \mathrm{k}_2, \mathrm{k}_3 \ldots \mathrm{k}_n\} \text{ where } \mathrm{V} > 0$$

Where 'D' is the set of documents in the corpus, D = d1, d2..., dn are the individual document present in the corpus. From then, Vocabulary (V) has been generated with a unique set of words (k) from all collections. It is an essential factor in the document collection contains the individual unique term.

3.3. Term- Document matrix

The Term-Document (TD) matrix is calculated from the frequency of terms in the document, which has been represented in the matrix format. In the matrix, the column denotes the document, and the row denotes the index key terms. The cell value denotes the frequency of the word occurs in the document which establishes the relation between term. The input matrix of the cell value $\mathrm{k}_{i,j}$ which denotes the occurrence of word k_i in document d_j, is given in the matrix.

$$\begin{matrix} k11 & k12 & k13 \\ k21 & k22 & k23 \\ k31 & k32 & k33 \end{matrix}$$

3.4. Term Frequency –Inverse Document Frequency (TF-IDF)

TF-IDF is a statistical approach in Information Retrieval to summarize the content and to find out the relevancy level of the terms in the document. Term Frequency (TF) is the rate of the term; (k_i) is the weight of the term in the document; (d_j) is the Term Frequency ($\mathrm{TF}_{i,j}$) and is calculated by the raw frequency method.

$$TF(t,d) = \frac{\text{occurence of } t \text{ in } d}{\text{No.of word ind}}$$

Inverse Document Frequency (IDF) of the term equivalent in the set of documents. A frequency that shows the most selective word occurs in a single document and the slightest selective word means the frequently occurred in

all the documents. As a result, it selects the noun and their group related words from the vocabulary.

$$\mathrm{TDF(t)} = \log(\mathrm{N}/(\mathrm{df}+1))$$

Where 'df', is the count of term rate't' in the document collection 'N'.

TF-IDF is calculated by multiplying the TF with IDF that gives the score value of each term, weight has been calculated from the inverse frequency. The highest score term is more relevant to the document and the least score is unrelated terms. Utmost the noun and their related terms are retrieved which is the downside of this method which will not be suitable for a different domain.

3.5. TextRank Algorithm

TextRank (TR) algorithm is a type of graph-based method, A graph 'G' represents document text and sustain the relationship between the text and keyword that have been recognized by their score value. Primarily, all the words are considered as candidate keywords. The graph is stated as an undirected graph G= (V, E) where the 'V' denotes the vertices and 'E' represents the edges between the text contained by the size of the window (w). The node is constructed by the candidate keywords and the edges are made by of co-occurrence of the word to establish the relationship between texts in the document and are given in [24, 25]. The PageRank method is used to calculate the score of the words in the graph as given below,

$$TR(\mathrm{Vi}) = (1-c) + c * \sum \mathrm{TR(Vj)|N(Vj)|}$$

Where N (V_j) is the set of a neighbor node with Vi and V_j; 'c' is the decay factor that falls between the range of 0 to 1. Subsequently, the score has calculated for each vertex in the graph and the degree of the vertex is determined by its connection with other vertices. Arrange the vertices in the reverse order to select the topmost 'n' words and fix them as keywords. As a result, noun words are the selected keywords for the given input.

3.6. RAKE

Rapid Automatic Keyword Extraction (RAKE) is an unsupervised, domain, and language-independent to extract keywords from a single document [26, 27]. It extracts the content and noun related words as a keyword from a different domain. Content words are put up as a meaningful word that relates to another word from the document.

The file consists of word collection, delimiters of phrases to partition the word from the candidate keywords. Candidate keywords are the sequence of words as they occur between the texts without stop-words or phrase delimiters. Without the involvement of the window size, the co-occurrence graph has been constructed to find the occurrence of words in the graph. The degree of the word is the total number of words within the candidate words; next, for all the candidate words a score value is calculated by summation of an individual score of the words exist in the graph.

$$\frac{\deg(\mathrm{n}+n1)}{N}$$

deg (n+n1), the degree of a number of occurrences + the number of times it occurred as adjoin. 'N' is the frequency of that particular word. In score calculation, when the adjoining word occurs for more than two times then that word is included. 'N' topmost scored words are set keywords of the document. Due to the compound words, RAKE not able to produce accurate keywords.

3.7. YAKE

Yet Another Keyword Extraction (YAKE) is an unsupervised, language, and domain-independent [28] statistical keyword extractor to find the top most important keyword. This method will not need any corpus or vocabulary or training data to generate the keyword. During extraction, it does not depend on the Named Entity Recognizer (NER) and Part-of-Speech (POS). YAKE method encompasses the text pre-processing as the first stage followed by the feature extraction as a third stage; these features are combined with a single score to replicate the importance of each term. In the fourth, the candidate keywords generate the score values. Finally, the similarity has been measured to extract the keywords based on the score value.

The document consists of the raw data which are represented as sentences. To extract the keyword, let consider the following components, The statistical features of the document are the size of the window (w), the Neighbor node is selected based on the 'n-gram', threshold value 'θ' and remove the duplicate key term. The score values are sorted and the topmost words are extracted as a keyword using the below-given formula.

$$S(t) = \frac{\mathrm{Trc}^* \rightleftarrows \mathrm{Tpos}}{\mathrm{Tcase} + \frac{\mathrm{TFnor}}{\mathrm{Trc}} + \frac{\mathrm{Tsent}}{\mathrm{Trc}}}$$

Trc is term related to context; TF_{nor} is the normalized form Term-Frequency, T_{rc} is the term related to context, T_{sent} is the term different sentence, T_{pos} is the position of the term. the smaller the score value relevant to the keyword.

3.8. Latent Semantic Analysis

Latent Semantic Analysis is also termed as Latent Semantic Indexing (LSI) is an unsupervised, indexing method in NLP to extract the words which are semantically related to each other. In [29], Landauer suggest the relationship exists between the text document and their terms. In [30], Deerwester et al. suggest this method to examine the term-document matrix. The co-occurrence of the method exhibits the relations between the terms present in sentences of the document to encompass Singular Value Decomposition (SVD). SVD categorizes the terms and the documents semantically that represent the term-document matrix and maintain the noiseless data for effective retrieval of the result as given in [31]. The importance of the singular value corresponds to a vector representation gives the importance of pattern in the document. LSI uses SVD to capture the semantically related words that are actually hidden in the given document.

LSI method first generates the Term-Document matrix with the given document and the words from the document next generate the SVD to find the relevant word. Let consider the set of the document 'D', a number of words in the document is 'm', 'n' is the number of documents. The term-document (TD) matrix is generated by using m*n.

The number of occurrences of the word is given weight in the term-document matrix Aij, number of times the word 'i' has occurred in document 'j', where 'i' represents the row and j represents the column, The input matrix 'A' is given pre-processed and represented as a sparse matrix [32]. The original input has been reduced dimensionality without changing the content. Next, the matrix is decomposed to three different matrixes as orthogonal, diagonal, and transpose matrix,

$$A = USV^T$$

U represents the topic of the document extracted from the original, and as a result, the most relevant word is placed in the first column in the matrix. Orthogonal Matrix is given as UV, 'S' is a diagonal matrix to represent the singular scalar value in sorted order. V is the orthogonal matrix that gives the representation of the column of the input matrix as the extracted concept and the transpose matrix is given as V^T. The terms are given in Table 1.

At the final, by deleting all the 'n' larger value in the diagonal matrix dimension of the matrix has been reduced corresponds column with other two matrices. Cosine similarity is used to find the similarity between the two words.

Table 1. Different matrices.

Symbol	Description	Matrix
A	Input Matrix	m ∗ n
U	Term ∗ concept extracted	m ∗ m
S	Diagonal matrix with Scalar value	m * n
V	Sentence ∗ concept extracted	n ∗ n

4. Experimental result

4.1. Dataset

In our paper has been used a single document from the Stanford Question Answering Dataset (SQUAD) dataset is used for the implementation [33]. The dataset consists of different categories which has been considered for executing all the techniques. Same input has been used for all extraction techniques. The input contains the text data where the different unsupervised method has been implemented to extract the data. Dataset were executed in all the methods to extract the important words.

During the execution around 50 corpuses has been considered for execution. After the retrieval, those keywords are used in Automatic Question Generation. In an objective type question, the word which is more relevant to the sentences is made as a keyword that has to be replaced by the blank space [34–37]. Keyword plays a major role in the question generation system.

The experimental result shows the Latent Semantic Indexing is an exact method to extract the keyword to be applied in the application of question generation. The document describes by its size, count of the sentence, and the words and type as given in Table 2.

Python language has been used to implement the unsupervised method to extract the word and is topmost keywords from different algorithms are given in the below Table 3.

From the extracted set of TF-IDF, the majority of the keyword set are noun related terms for 1-gram, in case of the LSA the noun related terms as well the adjective words are extracted, in the rest of the algorithm keywords are extracted as 'n'-gram.

In LSA, the importance of sentences has been determined and keywords are extracted from those sentences. The weight of extracting keywords varies for each algorithm which is given in the Table 4.

From the above table, the keywords from the LSA method is semantically related to sentences, whereas other are not semantically interrelated. The extracted keyword set has been verified with the ground truth value of the SQUAD comprehension dataset in question answering gen-

Table 2. Input Document.

Reference document	No. of Sentences	No. of words	Document size	Approximately Extracted Keyword	Top most selection
SQUAD	7	134	11 KB	≈ 20	$5-10$
Abstract- 1	12	205	27 KB	≈ 30	$10-15$
Abstract-2	8	170	27 KB	≈ 25	$10-15$

Table 3. Implementation of different algorithms.

Keyword	TF-IDF	TextRank	RAKE	YAKE	LSA
K1	oxygen	diatomic oxygen gas	Diatomic oxygen gas constitutes 20	oxygen	gas
K2	element	atmospheric oxygen levels	atmospheric oxygen levels show	element	diatomic
K3	compounds	most elements	odorless diatomic gas	mass	atmosphere
K4	gas	oxide compounds	highly reactive nonmetal	number	constitute
K5	mass	oxygen	global downward trend	notably	temperature

eration. The ground truth value is the suggested answer list for the given question. The answers are the key set of the questions, hence our result compared with the ground truth value. The semantically related keywords are far better to generate objective type questions.

The extracted output has been verified by the precision, recall, and F1 Score, the precision is to check the quality and the recall is to check quantities of the output. The measure of relevancy is given by the precision and the recall is the measure of how much actually relevant result. F1 scores are the weighted average of precision and recall is given in Table 5.

4.2. Evaluation Index

The precision rate P, the recall rate R, and the F-score are all evaluation metrics used in this study. To compute any of these formulas, use the following steps, to calculate recall, divide the number of relevant documents returned by the total number of relevant documents already in existence. To calculate precision, divide the number of relevant documents returned by the search engine's total number of results.

$$\text{Precision } (P) = \frac{\text{number of correct extraction}}{\text{totalnumber of extraction}}$$

$$\text{Recall}(R) = \frac{\text{number of correct extraction}}{\text{number of relevant}}$$

$$F - \text{ Score } = \frac{2^{*}P^{*}R}{P+R}$$

From the below-specified graph, the LSA gives the superior way of extracting the keyword from the given input. It has been observed that LSA algorithm has highest score in Precision, recall and in F-score. The semantically relevant keywords are extracted and compared with a ground value of the dataset.

These keywords are highly carrying information about the sentences and they are replaced by the blank space in generating the objective type questions. Not only the noun terms carry the information but also the adjective related words carry meaningful information that those words are extracted as the keyword in LSA. Experiments on keyword extraction using the LSA approach in this research demonstrate that it has achieved some advances.

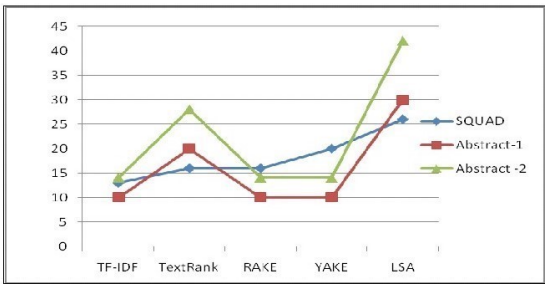

Fig. 1. Precision Value

The above Figs. 1 to 3 gives the precision, Recall, and F1-Score value of the different algorithms, comparatively, the LSA algorithm has the highest score value in all the evaluation metrics. SQUAD dataset has the ground-truth value for the predefined questions, here the five different al-

Table 4. Keywords with weight.

Keyword with weight	TF-IDF	TextRank	RAKE	YAKE	LSA
k_1w_1	oxygen (0.3582)	diatomic oxygen gas (0.122)	Diatomic oxygen gas constitutes 20 (21.333)	oxygen (0.0770)	gas (0.3613)
k_2w_2	element (0.2549)	atmospheric oxygen levels (0.1084)	atmospheric oxygen levels show (15.333)	element (0.1025)	diatomic (0.3613)
k_3w_3	compounds (0.1880)	most elements (0.1052)	odorless diatomic gas (11.0)	mass (0.1340)	atmosphere (0.268)
k_4 W_4	gas (0.1666)	oxide compounds (0.0944)	highly reactive nonmetal (9.0)	number (0.1549)	constitute (0.2508)
K_5w_5	mass	oxygen	global downward	notably	temperature

Table 5. Experimental Results.

	SQUAD			Abstract-1			Abstract-2		
	Precision (P)	Recall (R)	F-Score	Precision (P)	Recall (R)	F-Score	Precision (P)	Recall (R)	F-Score
TF-IDF	13	30	18.30	10	14	11.6	14	20	16.4
TextRank	16	38	22.51	20	28	23.3	28	80	41.4
RAKE	16	38	22.51	10	14	11.6	14	40	20.7
YAKE	20	46	27.87	10	14	11.6	14	20	16.4
LSA	26	61	39.16	30	42	35	42	80	55

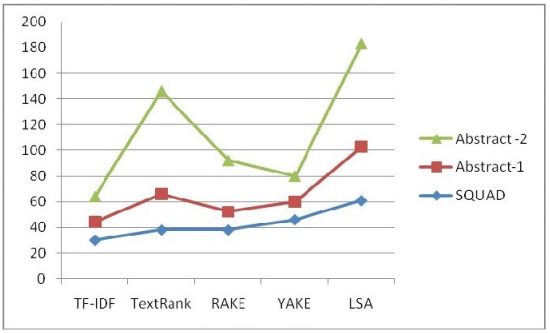

Fig. 2. Recall Value

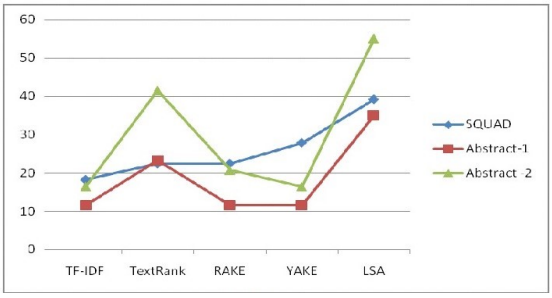

Fig. 3. F1-Score

gorithms have extracted some set of keywords, but the LSA has extracted more majorities of keywords as the ground truth value. Next, the algorithm has been executed for two sets of abstracts selected from the journal articles. In both the journals, manually the author has set five different keywords which have set as a reference for our implementation. Here, the LSA algorithm has extracted similar kind of keywords as abstracts which have been evaluated using Precision, Recall and F1-score.

Extracted keyword from the LSA algorithm has been used in the automatic question generation system to assess the skill level of the learner. Identify the sentence which holds the extracted keywords and replace that keyword with the blank space to frame the objective type questions.

5. Conclusion

Different types of unsupervised algorithms are used to extract the keyword from the document. In our paper, we have compared those algorithms using a single document and assess them through the evaluation metrics like Precision, Recall, and F-score. Keywords extracted using Latent Semantic Analysis algorithm has a higher score value in all metrics and it is semantically related to the sentences. Semantically related words are extracted as keywords; these keywords are utilized in gap-filling questions to assess the skill level of learners. In the future, a distractor set can be prepared from this keyword in generating Multiple Choice Questions (MCQ).

References

[1] S. Menaka and N. Radha, (2013) *"Text classification using keyword extraction technique"* **International Jour-**

nal of Advanced Research in Computer Science and Software Engineering *3*(12): 734–740.

[2] J. Zhao, Q. Zhu, G. Zhou, and L. Zhang, (2017) *"Review of research in automatic keyword extraction"* **Journal of software *28***(9): 2431–2449. DOI: 10.13328/j.cnki.jos.005301.

[3] D. B. Bracewell, F. Ren, and S. Kuriowa. "Multilingual single document keyword extraction for information retrieval". In: *2005 International Conference on Natural Language Processing and Knowledge Engineering*. IEEE. 2005, 517–522. DOI: 10.1109/NLPKE.2005.1598792.

[4] G. Deena and K. Raja. "A study on knowledge based e-learning in teaching learning process". In: *2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*. IEEE. 2017, 1–6. DOI: 10.1109/ICAMMAET.2017.8186686.

[5] G. Deena and K. Raja, (2018) *"The Impact of Learning Style to Enrich the Performance of Learner in E-Learning System"* **Journal of Web Engineering *17***(6): 3407–3421.

[6] A. Hulth. "Improved automatic keyword extraction given more linguistic knowledge". In: *Proceedings of the 2003 conference on Empirical methods in natural language processing*. 2003, 216–223.

[7] A. Hulth. *"Combining machine learning and natural language processing for automatic keyword extraction"*. (phdthesis). Institutionen för data-och systemvetenskap (tills m KTH), 2004.

[8] M. Litvak and M. Last. "Graph-based keyword extraction for single-document summarization". In: *Coling 2008: Proceedings of the workshop multi-source multilingual information extraction and summarization*. 2008, 17–24.

[9] L. Yao, Z. Pengzhou, and Z. Chi. "Research on news keyword extraction technology based on TF-IDF and TextRank". In: *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*. IEEE. 2019, 452–455. DOI: 10.1109/ICIS46139.2019.8940293.

[10] L. Qifei and S. Weiyu, (2018) *"Research of keyword extraction of political news based on word2vec and textrank"* **Information Research *6***: 22–27.

[11] G. K. Palshikar. "Keyword extraction from a single document using centrality measures". In: *International conference on pattern recognition and machine intelligence*. Springer. 2007, 503–510. DOI: 10.1007/978-3-540-77046-6_62.

[12] M. A. Andrade and A. Valencia, (1998) *"Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families."* **Bioinformatics (Oxford, England) *14***(7): 600–607. DOI: 10.1093/bioinformatics/14.7.600.

[13] G. Salton. *Automatic text processing*. Addison-Wesley Longman Publishing Co., 1989.

[14] C. Zhang, (2008) *"Automatic keyword extraction from documents using conditional random fields"* **Journal of Computational Information Systems *4***(3): 1169–1180.

[15] H. P. Edmundson, (1969) *"New methods in automatic extracting"* **Journal of the ACM (JACM) *16***(2): 264–285. DOI: 10.1145/321510.321519.

[16] Z. Li, D. Zhou, Y.-F. Juan, and J. Han. "Keyword extraction for social snippets". In: *Proceedings of the 19th international conference on World wide web*. 2010, 1143–1144.

[17] Y. Matsuo and M. Ishizuka, (2004) *"Keyword extraction from a single document using word co-occurrence statistical information"* **International Journal on Artificial Intelligence Tools *13***(01): 157–169.

[18] F. Liu, D. Pennell, F. Liu, and Y. Liu. "Unsupervised approaches for automatic keyword extraction using meeting transcripts". In: *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*. 2009, 620–628. DOI: 10.3115/1620754.1620845.

[19] S. Rose, D. Engel, N. Cramer, and W. Cowley, (2010) *"Automatic keyword extraction from individual documents"* **Text mining: applications and theory *1***: 1–20. DOI: 10.1002/9780470689646.ch1.

[20] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, (2020) *"YAKE! Keyword extraction from single documents using multiple local features"* **Information Sciences *509***: 257–289. DOI: 10.1016/j.ins.2019.09.013.

[21] R. Mihalcea and P. Tarau. "Textrank: Bringing order into text". In: *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004, 404–411.

[22] X. Wan and J. Xiao. "Single document keyphrase extraction using neighborhood knowledge." In: *AAAI*. ***8***. 2008, 855–860.

[23] C. Florescu and C. Caragea. "Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, 1105–1115. DOI: 10.18653/v1/P17-1102.

[24] R. Mihalcea, P. Tarau, and E. Figa. "PageRank on semantic networks, with application to word sense disambiguation". In: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. 2004, 1126–1132.

[25] L. Page, S. Brin, R. Motwani, and T. Winograd. *The PageRank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab, 1999.

[26] S. Rose, D. Engel, N. Cramer, and W. Cowley, (2010) *"Automatic keyword extraction from individual documents"* **Text mining: applications and theory *1***: 1–20. DOI: 10.1002/9780470689646.ch1.

[27] T. Pay, S. Lucci, and J. L. Cox, (2019) *"An Ensemble of Automatic Keyword Extractors: TextRank, RAKE and TAKE"* **Computación y Sistemas *23***(3): 703–710. DOI: 10.13053/CyS-23-3-3234.

[28] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt. "Yake! collection-independent automatic keyword extractor". In: *European Conference on Information Retrieval*. Springer. 2018, 806–810. DOI: 10.1007/978-3-319-76941-7_80.

[29] T. K. Landauer, P. W. Foltz, and D. Laham, (1998) *"An introduction to latent semantic analysis"* **Discourse processes *25***(2-3): 259–284.

[30] Y. Matsuo and M. Ishizuka, (2004) *"Keyword extraction from a single document using word co-occurrence statistical information"* **International Journal on Artificial Intelligence Tools *13***(01): 157–169.

[31] J. Gao and J. Zhang, (2005) *"Clustered SVD strategies in latent semantic indexing"* **Information processing & management *41***(5): 1051–1063. DOI: 10.1016/j.ipm.2004.10.005.

[32] G. Deena and K. Raja, (2019) *"Sentence selection using latent semantic analysis for automatic question generation in e-learning system"* **International Journal of Innovative Technology and Exploring Engineering** *8*(9): 86–91.

[33] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, (2016) *"Squad: 100,000+ questions for machine comprehension of text"* **arXiv preprint arXiv:1606.05250**: DOI: 10.18653/v1/d16-1264.

[34] G. Deena and K. Raja, (2019) *"Designing an Automated Intelligent e-Learning system to enhance the knowledge using machine learning techniques"* **International journal of advanced computer science and applications *10***(12): DOI: 10.14569/ijacsa.2019.0101215.

[35] G. Deena and K. Raja, (2019) *"Designing an Automated Intelligent e-Learning system to enhance the knowledge using machine learning techniques"* **International journal of advanced computer science and applications *10***(12): DOI: 10.14569/ijacsa.2019.0101215.

[36] G. Deena, K. Raja, N. B. PK, and K. Kannan, (2020) *"Developing the assessment questions automatically to determine the cognitive level of the E-learner using NLP techniques"* **International Journal of Service Science, Management, Engineering, and Technology (IJSSMET) *11***(2): 95–110. DOI: 10.4018/IJSSMET.2020040106.

[37] G. Deena and K. Raja, (2022) *"Objective Type Question Generation using Natural Language Processing"* **International Journal of Advanced Computer Science and Applications *13***(2): DOI: 10.14569/IJACSA.2022.0130263.