Extracting Key-phrase Embedding using Deep Average Network and Maximal Marginal Relevance to Enhance Information Retrieval

Alyaa abdual kahdum[1*], Wafaa AL-Hameed [2]

1College of Information Technology, University of Babylon,
alyaamerjan.sw.msc@student.uobabylon.edu.iq, Babylon, Iraq.
2College of Information Technology, University of Babylon,
it.wafaa.mohammed@uobabylon.edu.iq, Babylon, Iraq.
***Corresponding author email:** alyaamerjan.sw.msc@student.uobabylon.edu.iq, mobile: 07815659277

أستخراج تضمين العبارة الرئيسية باستخدام شبكة المتوسط العميق والصلة الهامشية الاكبر لتحسين استرجاع المعلومات

علياء عبد الكاظم[1*]، وفاء الحميد[2]

1 كلية تكنولوجيا المعلومات، جامعة بابل، alyaamerjan.sw.msc@student.uobabylon.edu.iq، بابل، العراق
2 كلية تكنولوجيا المعلومات ، جامعة بابل، it.wafaa.mohammed@uobabylon.edu.iq ، بابل، العراق

Accepted: **20/3/2024** **Published:** **30/6/2024**

ABSTRACT

Background:

Automatic keyphrase extraction (AKE) is essential to many NLP and information retrieval tasks. Extracting high-quality key phrases is difficult due to technological advancements and the exponential growth of textual data and digital sources. Unsupervised keyphrase extraction with cheap computing cost that relies on heuristic notions of phrase importance such as embedding similarities but their development necessitates in-depth subject expertise.

Materials and Methods:

This paper presents a method to obtain a semantic understanding of the query and index documents by using the embedding technique(universal Sentence encoder (USE)) while keeping the most informative using Maximal Marginal Relevance (MMR) and then scoring(an inverted index) the most documents relevant to the query vector to improve the performance of IR systems.

Results:

The proposed retrieval model implement on the (Fire2011) dataset. The final stage was evaluating the results of the baseline and the results (indexing and ranking) by using mean average precision (MAP). The result of the baseline was 0.61, while the result inverted index was 0.6277519 .

Conclusions:

In this paper, we have discussed document retrieval using keep key phrases that have informativeness properties by using maximal marginal relevance, since if we extract a fixed number of top keyphrases, redundancy hinders the diversification of the extracted keyphrases.

Keywords: Text Embedding, Universal Sentence Encoder, Deep average network, Maximal Marginal Relevance , Information Retrieval, Inverted Index.

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

INTRODUCTION

In recent years, it has been observed that data expansion has increased significantly, creating new obstacles for academics as they try to come up with creative ways to extract important information faster. Several strategies are used to obtain crucial data from the repository [1]. To handle the high data speed, a variety of methods are being investigated, from storage to information retrieval [2]. Finding the needed information without the IRS was nearly impossible because of the size of the search field. Today's computer users cannot envisage obtaining the information required daily without the aid of some sort of information retrieval tool. The most popular tool for information retrieval is a web search engine, which is used to identify information sources that are relevant to the user [3]. The collection's helpful materials are indexed by a search engine, which then looks through those indexes for relevant information [4].

Automated keyphrase extraction (AKE), which automatically selects a small collection of single or multiple words from inside the text that best summarizes a document, is essential for locating the main subjects of a text document. A keyphrase is a succinct linguistic expression that summaries the information in a long document [1]. Due to their condensed nature, these sentences can serve as a document's metadata or indicative summary, both of which can help readers identify pertinent material. The inclusion of keywords can improve the effectiveness of information searches and help readers decide whether a piece is worthwhile to read [2].Finding a single word or phrase that best encapsulates the main concepts in the text is the main objective of keyphrase extraction [3]. The main specific topic in this research is the Unsupervised Key Phrase Extraction which automatically identifying important phrases or terms from a given text corpus without relying on pre-labeled training data [4]. The benefits of unsupervised keyphrase extract methods are applicable to texts written in any language. They do not rely on language-specific rules or dictionaries, making them versatile for multilingual text analysis, can handle large volumes of text data efficiently, and allow for the identification of domain-specific terminology and key phrases that may not be present in standard dictionaries, Unlike supervised methods that require manually annotated training data, unsupervised approaches eliminate the need for expensive and time-consuming labeling efforts [5]. This makes it easier and more cost-effective to apply keyphrase extraction to new domains or when labeled training data is limited, also extracted key phrases can provide condensed representations of documents, aiding in summarizing text [6] or improving search engine results by matching user queries with relevant key phrases, and also key phrases act as important features for grouping similar documents together [7]. Finally, unsupervised key phrase extraction methods offer a flexible and efficient way to automatically extract important phrases or terms from text data, enabling various downstream applications in information retrieval, text mining, and document analysis.

Finding documents that are relevant to the query and embedding keyphrases is a typical task, and various kinds of studies have been conducted to improve information retrieval efficiency. This idea is covered in several studies. For instance, Embed Rank, an innovative unsupervised technique that makes use of sentence embeddings, may be used to extract keyphrases from single documents . the extracted keyphrases were based on the use of sentence embedding by model Sen2Vec. [8], a method for incorporating semantic relevance feedback into the information retrieval process, the type of semantic linkages between the meanings of two words serves as a measure of semantic similarity, which is based on word meaning. By using the WordNet-based similarity method, it is possible to determine how semantically similar two

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

words or sentences are.[9], a word embedding-based keyphrase extraction technique for texts. In particular, in the first build a heterogeneous text graph embedding model to combine local context information of the word graph (i.e., the local word collocation patterns) with certain key attributes of candidate words and edges of the word graph. Next, using these learned word embeddings, a unique random-walk-based ranking algorithm is developed to rate candidate words. In order to score phrases for the purpose of choosing the top-scoring phrases as key phrases, a novel and generic phrase scoring model based on word embeddings is provided.[10], two directions in this essay. The first is to suggest a brand-new technique (dubbed the MB technique) to gather unlabeled data and sort it into the proper groupings. The second is the construction of a lexical chain sentence (LCS), which differs from the conventional lexical word chain (LCW), based on words, and is based on similar semantic phrases.[11] , MDERank (Masked Document Embedding Rank) is an unsupervised keyphrase extraction approach that utilizes masked document embedding ranking. The document is mapped to its corresponding embeddings using pre-trained embedding models such as BERT or RoBERTa(embedding generation). MDERank may not be able to capture the nuances of the language and may not identify the most relevant keyphrases.[12], The data in a text archive is organized into clusters of similar content using an improved Fuzzy-CMeans clustering algorithm, which restricts the search to the closest clusters. Additionally, we use a neural network to encode the words into numerical vectors, which produces words with simulative meaning having similar vector values, providing a semantic-based clustering that depends on the context of the document rather than just the frequency of the word. [13], a new retrieval system by creating a new structure that uses the word embedding produced by the embedding layer in the skip-gram neural network during the feature extraction process in accordance with relationships. Building an indexing system based on the semantic connections between documents enables the development of a flexible retrieval system that adheres to human standards.[14], the KP-USE is an unsupervised approach for key-phrase extraction from documents. A pre-trained Universal Sentence Encoder (USE) model utilized to generate embeddings sentences vectors while the text is devised into five sections and weight each one according to how semantically close to the document. [15]

MATERIALS AND METHODS

- **Text Embedding**

  An embedding, also known as a word embedding or a phrase(sentence)embedding, is a technique used in natural language processing to represent words or phrases as a dense vector of numerical values. The purpose of embedding is to create a suitable format for machine learning algorithms to process the semantic and syntactic information about words and phrases [16,14].

  In NLP, embedding is a potent method that has transformed the discipline by allowing machine learning algorithms to handle natural language data with efficiency. Embedding allows algorithms to process text in a way that captures semantic and syntactic information, allowing them to carry out a variety of NLP tasks with high accuracy [13]. Words and phrases are represented as dense vectors [17].

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

- **Universal Sentence Encoder**

  Sentence embedding is the process of representing a sentence. as a dense vector of fixed length, usually by means of a neural network [18]. Once the sentence is encoded as a fixed-length vector, it can be used to compute similarity scores between sentences or to perform other NLP tasks. For example, cosine similarity can be used to measure the similarity between two sentence vectors, and a classification model can be trained on top of the sentence embeddings to predict the label of a given text [15].

  In order to produce high-quality embeddings for sentences in natural language, Google created the Universal Sentence Encoder (USE), a neural network-based model. It was released in 2018 [18] and is currently being utilized extensively in several natural language processing (NLP) applications.

  The Deep Averaging Network (DAN) architecture-based model and the Transformer Architecture-based model are both components of the Universal Sentence Encoder (USE). The DAN-based USE model is a more straightforward architecture that generates fixed-length sentence embeddings by averaging the input embeddings of the sentence's component words and bi-grams before feeding the data through a feed-forward deep neural network (DNN). In comparison to the transformer-based approach, this model trains more quickly and uses less computing power overall. Contrarily, the transformer-based USE model is a more intricate architecture that embeds the input text into a fixed-length embedding (512 dimensions) via a self-attention mechanism [18]. Compared to the DAN-based model, this model produces higher-quality sentence embeddings, but it takes longer to train and uses more resources.

- **Deep Averaging Network (DAN)** As the architecture was employed in this work, the Deep Averaging network (DAN) encoder will be covered in detail. The architecture proposed by Lyyer and colleagues [19,11] is used to construct a deep Averaging network (DAN) encoder. According to Lyyer and et al. authors description, the deep averaging network (DAN) contains three fundamental phases:

  A- Calculate the average vector of the embeddings related to the input tokens sequence in the first phase.
  B- Second phase: Apply one or more feed-forward layers to the average vector.
  C- Phase three is the application of (linear) classification to the representation of the final layer.

Each layer of the deep feed-forward neural network is often created with the idea that it will learn a more abstract representation of the input than the layer before it [20]. DAN was applied to the neural bag-of-words (NBOW) model.

DAN starts by working on word embeddings, and any present bi-grams in a sentence are averaged. After that, they are fed into an n-layer feed-forward deep neural network to produce 512-dimensional sentence embeddings.

- **Maximal Marginal Relevance method(MMR)**

Maximal Marginal Relevance (MMR) is a strategy used in information retrieval to rank and pick pertinent documents based on their closeness to a query while simultaneously maintaining diversity in the retrieved results. In 1998 [21], Carbonell and Goldstein published the first description of it.

- **Information Retrieval**

Information retrieval (IR) is the process of obtaining information from a collection of documents or data sources, typically in response to a user's information need. It involves searching for and retrieving relevant information based on a query or set of keywords provided by the user.

Information retrieval aims to provide the user with a set of documents or resources relevant to their query or information need. This is often accomplished using search engines, which use algorithms and techniques to rank and present the most relevant results to the user [22].

One advantage of employing semantic steps in IR is that it overcomes the constraints of approaches based on basic lexicographic word matching, i.e., simple IR models consider a document to be relevant according to a query only if the terms provided in the query appear in the document. By doing a syntactic search, semantic measurements allow the meaning of words to be considered. As a result, they are employed to improve old models [23]. The quality of the rules for phrase association influences the results. Therefore, This study presented a document ranking based on unsupervised automated keyphrase extraction(AKE) that can identify multi-word keyphrases and can handle polysemous words (words with multiple meanings) by using the context of the sentence in which the word appears.

- **Inverted Index**

In Information Retrieval (IR), an inverted index is a data structure that is commonly used to facilitate efficient full-text search and retrieval of documents based on terms or keywords. It is a key component of many search engines [24].In a traditional database, data is typically organized based on the documents or records, and each document contains its own information. However, in an inverted index, the focus is on the terms or keywords within the documents [25].

- Dictionary

In the context of information retrieval, a dictionary refers to a data structure that stores and provides access to the terms or words present in a collection of documents. It is vital in many information retrieval systems, including search engines. The dictionary facilitates efficient searching, indexing, and retrieval of documents based on user queries [26].

A dictionary is used in information retrieval (Inverted Index) Along with the term ID, the dictionary may also store additional information, such as document frequencies (the number of documents a term appears in) or term frequencies (the frequency of a term within a specific document). This information is used to build an inverted index, which maps each term to the documents that contain it. The inverted index facilitates a quick lookup of documents based on terms present in a query [25].

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

- Posting list

In information retrieval, a posting list is a data structure used to store the occurrence information of terms in a document collection or corpus. It is a fundamental component of inverted indexes, which are commonly used in search engines and other information retrieval systems.

Posting lists are used to efficiently retrieve documents containing specific terms during the search process. They allow for quick lookup of documents that contain a given term, enabling fast retrieval of relevant documents for a given query [27].

In practice, posting lists are often compressed or optimized to reduce the memory footprint and improve search performance. Techniques like delta encoding, variable-byte encoding, or even more advanced compression algorithms are applied to reduce the storage requirements of posting lists in large-scale information retrieval systems.

- **Metholodgy**

The proposed system architecture includes two main phases (Offline phase and Online phase).

A. Offline phase

In the training (offline) stage, many procedures are implemented beginning with preprocessing (Lowercasing, Tokenization, Stop Words removal, and Stemming), also extracting candidate keyphrases (noun phrases) in the full document. Following the embedding technique (DAN)is carried out for excluding the redundant keyphrases using the MMR method. Then, for indexing and ranking, through constructing an inverted index to compute and rank the most similar ones with query embedding vector and finally score the most documents relevant to the query vector.

B. Online phase

while a query is supplied by the user, many processes are conducted (cosine similarity, Keyphrase scoring, and Document scoring) to return the most relevant documents as shown in Figure(1).

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

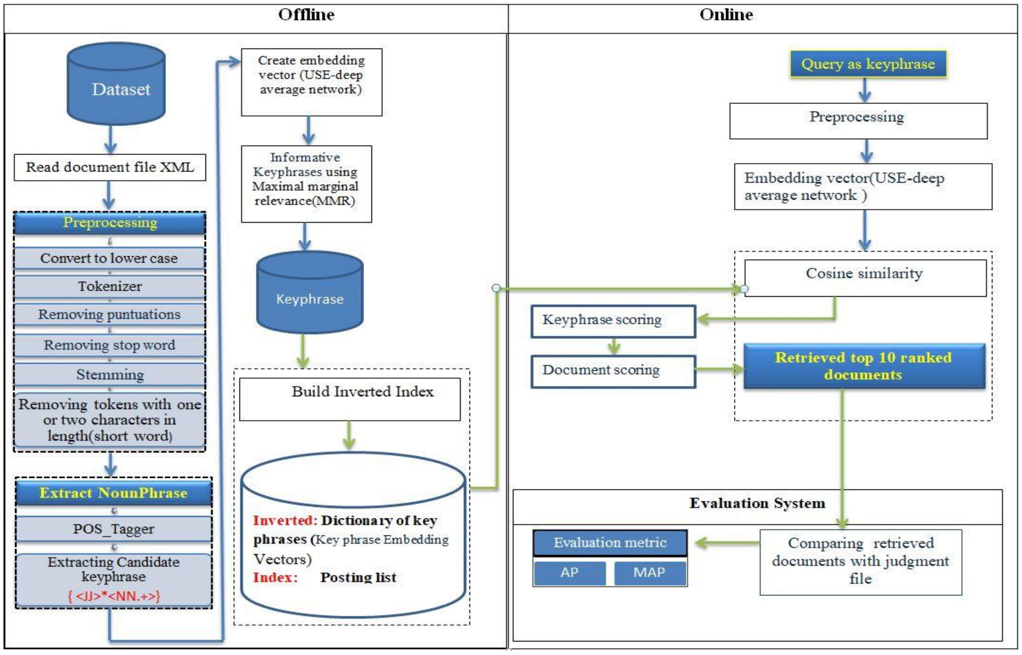

Offline
Dataset
Read document file XML
Preprocessing
Convert to lower case
Tokenizer
Removing puntuations
Removing stop word
Stemming
Removing tokens with one or two characters in length(short word)
Extract NounPhrase
POS_Tagger
Extracting Candidate keyphrase
{<JJ>*<NN.+>}
Create embedding vector (USE-deep average network)
Informative Keyphrases using Maximal marginal relevance(MMR)
Keyphrase
Build Inverted Index
Inverted: Dictionary of key phrases (Key phrase Embedding Vectors)
Index: Posting list
Online
Query as keyphrase
Preprocessing
Embedding vector(USE-deep average network)
Cosine similarity
Keyphrase scoring
Document scoring
Retrieved top 10 ranked documents
Evaluation System
Evaluation metric
AP
MAP
Comparing retrieved documents with judgment file

Figure (1):Proposed System.

- ## Description of Dataset

The data used in this system is obtained from the FIRE dataset of Forum for Information Retrieval Evaluation that is accessible on the website (http://fire.irsi.res.in/fire/static/data) and applying for access to the FIRE Information Retrieval Text Research Collection.

The suggested approach was implemented on the FIRE 2011 dataset in English. It is a category created specifically for experiments with information retrieval (IR). This collection consists of several newspaper stories on many subjects, including local news, sports, commercial, political, and medical news [28].The documents and queries in the Fire 2011 dataset are represented using XML markup language. The XML tags are used to provide a standardized format for representing and processing the text data and associated metadata. Both documents and queries are represented using specific metadata. Finally, also the dataset contains Relevance judgments: A set of relevance judgments that indicate the relevance of each document to a set of predefined queries. The relevance judgments are typically represented as a binary relevance score (relevant or not relevant) or a graded relevance score (e.g., highly relevant, somewhat relevant, not relevant).

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

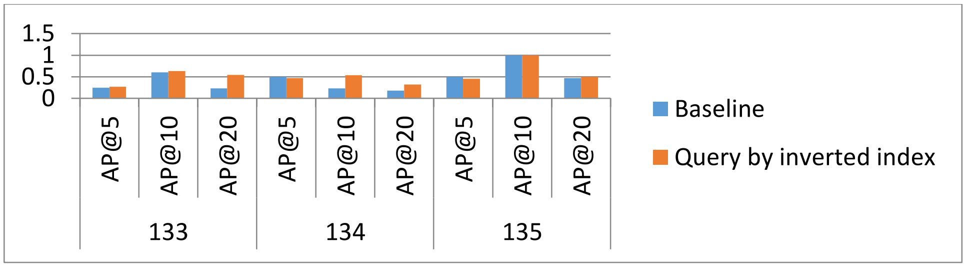
RESULTS AND DISCUSSION

The information and documents provided after using the keyphrase as a query were compared using the cosine approach, which computes the similarity of each baseline.

The cosine similarity score between the query and the documents determines the ranking documents that are retrieved; the highest scoring documents are retrieved first. Our proposed system selects several values for N, which stands for the number of documents retrieved, including 5, 10, and 20 documents retrieved for each query. After that, we used MAP to assess the outcomes and determine how many top-ranked papers should be retrieved. The average precision results for the basis and results after changing the query are displayed in **Table 1** along with the top five retrieved documents, top ten retrieved documents, and top twenty retrieved documents for each of them following the execution of Queries 133–135 from topics.

Table 1. Comparison Between Average Precision of Baseline and Average Precision of Result Query by inverted index Using Top (5,10,20) Documents.

Query	Metric	BASELINE	Query by inverted index
133	AP@5	0.25	0.269
	AP@10	0.6	0.63
	AP@20	0.23	0.543
134	AP@5	0.5	0.47
	AP@10	0.23	0.537
	AP@20	0.18	0.323
135	AP@5	0.5	0.457
	AP@10	1	1
	AP@20	0.47	0.5



Figure(2): The Average Precisions on Deferent Top n Documents.

In TABLE 2 we illustrated the results after performing the MAP metric to enter queries (133,134,135). MAP illustrates the difference between the results of query expansion and query as a key phrase. The best MAP value was used to determine the best N (top documents were chosen as relevant documents) for the retrieval.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Table 2. Result of Map Metric For Baseline And Query Modification.

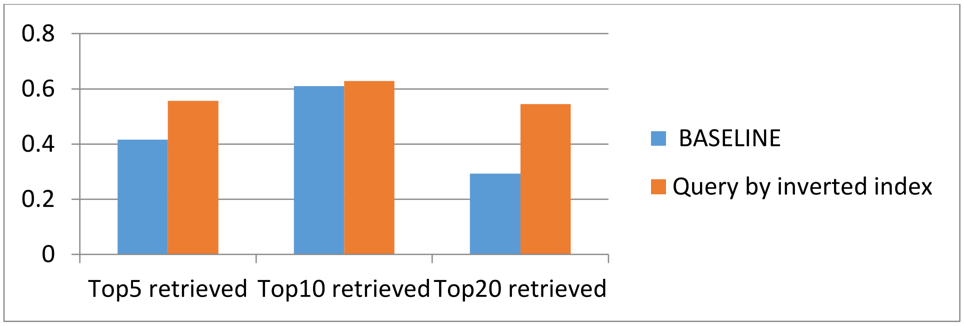
Top Ranking	BASELINE	Query by inverted index
Top5 retrieved	0.416	0.3
Top10 retrieved	0.61	0.6283
Top20 retrieved	0.2933	0.545

From the above table, we noted that:

1- At the value of N=5, we obtained very little improvements in the results when we performed path similarity from baseline [9].

2-At the values N=10, and N=20, we get better results after using the query as a keyphrase than the result obtained by using an expansion query [9].

We noted that the best result of the system was obtained when N=10 and the retrieval document by using key phrases as shown in Figure(3).



Figure(3). Results of MAP

ACKNOWLEDGMENTS:

First and foremost, praises and thanks to God, the Almighty, for His showers of blessings throughout my work to complete the research successfully.

I would like to express my deep and sincere gratitude to my paper supervisor, Dr. Wafaa Mohammed Saeed for allowing me to do research and providing invaluable guidance throughout this research. Her dynamism, vision, sincerity, and motivation have deeply inspired me. she has taught me the methodology to carry out the research and to present the research works as clearly as possible. Working and studying under his guidance was a great privilege and honor. I am extremely grateful for what he has offered me.

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Conflict of interests.

There are non-conflicts of interest.

References

[1] M. Docekal and P. Smrz, "Query-Based Keyphrase Extraction from Long Documents," no. 3, 2019.

[2] M. M. Al Hadidi, M. Alzghool, and H. Muaidi, "Keyword extraction from arabic text using the page rank algorithm," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 12, pp. 3495–3504, 2019, doi: 10.35940/ijitee.L2614.1081219.

[3] M. Zhang, X. Li, S. Yue, and L. Yang, "An empirical study of textrank for keyword extraction," *IEEE Access*, vol. 8, pp. 178849–178858, 2020, doi: 10.1109/ACCESS.2020.3027567.

[4] X. Zhu, Y. Lou, J. Zhao, W. Gao, and H. Deng, "Generative non-autoregressive unsupervised keyphrase extraction with neural topic modeling," *Eng. Appl. Artif. Intell.*, vol. 120, p. 105934, Apr. 2023, doi: 10.1016/J.ENGAPPAI.2023.105934.

[5] S. Siddiqi, "Keyword and Keyphrase Extraction Techniques : A Literature Review Keyword and Keyphrase Extraction Techniques : A Literature Review," no. January, 2015, doi: 10.5120/19161-0607.

[6] A. R. Mishra, V. K. Panchal, and P. Kumar, "Extractive Text Summarization-An effective approach to extract information from Text," in *2019 International Conference on contemporary Computing and Informatics (IC3I)*, 2019, pp. 252–255.

[7] Y. H. Farhan, S. A. Mohd Noah, M. Mohd, and J. Atwan, "Word-embedding-based query expansion: Incorporating deep averaging networks in Arabic document retrieval," *J. Inf. Sci.*, p. 01655515211040659, 2021.

[8] K. Bennani-Smires, C. Musat, A. Hossmann, M. Baeriswyl, and M. Jaggi, "Simple unsupervised keyphrase extraction using sentence embeddings," *arXiv Prepr. arXiv1801.04470*, 2018.

[9] H. M. Awad and W. M. Saeed, "Semantic Relevance Feedback Based on Local Context," in *2020 International Conference on Computer Science and Software Engineering (CSASE)*, 2020, pp. 296–301.

[10] Y. Zhang, H. Liu, S. Wang, W. H. Ip, W. Fan, and C. Xiao, "Automatic keyphrase extraction using word embeddings," *Soft Computing*, vol. 24, no. 8. pp. 5593–5608, 2020. doi: 10.1007/s00500-019-03963-y.

[11] M. Mohammed and W. Al-Hameed, "New algorithm for clustering unlabeled big data," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 24, p. 1054, Nov. 2021, doi: 10.11591/ijeecs.v24.i2.pp1054-1062.

[12] L. Zhang *et al.*, "MDERank: A Masked Document Embedding Rank Approach for Unsupervised Keyphrase Extraction," *arXiv Prepr. arXiv2110.06651*, 2021.

[13] S. K. Salsal and W. ALhamed, "Document Retrieval in Text Archives Using Neural Network-Based Embeddings Compared to TFIDF," in *Intelligent Systems and Networks: Selected Articles from ICISN 2021, Vietnam*, 2021, pp. 526–537.

[14] W. Al Hameed and S. K. Salsal, "Archiving System Optimization using Skip Gram based Neural Network as a Feature Selection," in *Journal of Physics: Conference Series*, 2021, vol. 1818, no. 1, p. 12073.

[15] L. Ajallouda, F. Z. Fagroud, A. Zellou, and E. Ben Lahmar, "KP-USE: An Unsupervised Approach for Key-Phrases Extraction from Documents," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 4, pp. 283–289, 2022, doi: 10.14569/IJACSA.2022.0130433.

[16] D. Jatnika, M. A. Bijaksana, and A. A. Suryani, "Word2vec model analysis for semantic similarities in english words," *Procedia Comput. Sci.*, vol. 157, pp. 160–167, 2019.

[17] W. Yang, L. Li, Z. Zhang, X. Ren, X. Sun, and B. He, "Be careful about poisoned word embeddings:

Exploring the vulnerability of the embedding layers in nlp models," *arXiv Prepr. arXiv2103.15543*, 2021.

[18] D. Cer *et al.*, "Universal sentence encoder," *arXiv Prepr. arXiv1803.11175*, 2018.

[19] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, "Deep unordered composition rivals syntactic methods for text classification," in *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, 2015, pp. 1681–1691.

[20] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[21] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 335–336.

[22] L. Liu and M. T. Özsu, "Encyclopedia of Database Systems Springer New York." p. 4964, 2018. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007/978-1-4614-8265-9_181

[23] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, "Semantic similarity from natural language and ontology analysis," *Synth. Lect. Hum. Lang. Technol.*, vol. 8, no. 1, pp. 1–254, 2015.

[24] A. Mallia, O. Khattab, T. Suel, and N. Tonellotto, "Learning passage impacts for inverted indexes," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1723–1727.

[25] W. Lu, X. Li, Z. Liu, and Q. Cheng, "How do author-selected keywords function semantically in scientific manuscripts?," *KO Knowl. Organ.*, vol. 46, no. 6, pp. 403–418, 2019.

[26] V. Bortnikova, I. Nevliudov, I. Botsman, and O. Chala, "Search Query Classification Using Machine Learning for Information Retrieval Systems in Intelligent Manufacturing.," in *ICTERI*, 2019, pp. 460–465.

[27] L. Zhao, X. Liu, and J. Callan, "WikiQuery-An Interactive Collaboration Interface for Creating, Storing and Sharing Effective CNF Queries.," in *OSIR@ SIGIR*, 2012, pp. 1–8.

[28] K. Park, J. S. Hong, and W. Kim, "A methodology combining cosine similarity with classifier for text classification," *Appl. Artif. Intell.*, vol. 34, no. 5, pp. 396–411, 2020.

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

الخلاصة

مقدمة:

يعد الاستخراج التلقائي للعبارات الرئيسية (AKE) ضروريًا للعديد من مهام البرمجة اللغوية العصبية واسترجاع المعلومات .يعد استخراج العبارات الرئيسية عالية الجودة أمرًا صعبًا بسبب التقدم التكنولوجي والنمو الهائل للبيانات النصية والمصادر الرقمية .استخراج العبارات الرئيسية بدون إشراف بتكلفة حوسبة رخيصة تعتمد على مفاهيم إرشادية لأهمية العبارة مثل تضمين أوجه التشابه ولكن تطويرها يتطلب خبرة متعمقة في الموضوع .

طرق العمل:

يقدم هذا البحث طريقة للحصول على فهم دلالي للاستعلام وفهرسة المستندات باستخدام تقنية التضمين (مشفر الجملة الشامل (USE)) مع الاحتفاظ بالمعلومات الأكثر استخدامًا باستخدام الحد الأقصى للملاءمة الهامشية (MMR) ومن ثم تسجيل النقاط (فهرس مقلوب معظم الوثائق ذات الصلة بمتجه الاستعلام لتحسين أداء أنظمة الأشعة تحت الحمراء.

النتائج:

تم تطبيق نموذج الاسترجاع المقترح على مجموعة البيانات .(Fire2011) كانت المرحلة الأخيرة هي نتيجة تقييم (Baseline) ونتيجة نهجي (الفهرسة والتصنيف). باستخدام متوسط الدقة (MAP) كانت نتيجة Baseline (0.61) ، بينما كانت نتيجة مع الفهرس المقلوب التي كانت 0.6277519.

الكلمات المفتاحية تضمين النص، تشفير الجملة العالمية، شبكة متوسطة عميقة، الحد الأقصى للملاءمة الهامشية، استرجاع المعلومات، الفهرس المقلوب.