A Combined Approach of Text Summarization using different Keyword Extraction Techniques

Lubna Rani Sarker
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology University

Md. Nahid Sultan
Assistant Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology University

ABSTRACT
The process of condensing and organizing a longer text is called text summarization. Summarizing lengthy documents, reports, and academic writings can be challenging. Selecting the important sentences and concepts from a text requires using a variety of text summarizing techniques, which reduces the time and effort required to read an entire article. In comparison to other cutting-edge approaches, the combined common extracted keywords employing the most popular techniques (Text Rank, Sentence Score, and Gensim Keyword Extraction) present only the important sentences that are briefer and more similar to human summary. For enhanced output summarization, a combination strategy of these cutting-edge approaches has been proposed in this thesis.

Keywords
Text Summarization, Keyword Extraction, Short and Similar, Human Summary, Combined Approach.

1. INTRODUCTION
Automatic process of condensing a set of data into a summary that only includes the most crucial or pertinent details from the original material is known as text summarizing. The raw data is mined for text, but the recovered text isn't altered in any way. Significant phrases and words found in extracted content can be used to "tag" or index a text document. Extracting text is similar to skimming, which is reading the summary, headers and subheadings, figures, the first and last paragraphs of a section, and optionally the first and last words of a paragraph, in order to decide whether or not to read the full document in detail. Another illustration of extraction is clinically significant text sequences, such as patient/problem, intervention, and result.

2. THE NEED FOR TEXT SUMMARIZATION
It can take a lot of time and effort to manually generate a summary. Such challenges are supposedly overcome by automatic text summary, which makes it simple to identify a piece of writing's main concepts. The vast majority of the data currently flooding the digital world is unstructured text data. Therefore, it is necessary to create automatic text summarization tools that make it simple for users to draw conclusions from them. At the moment, there is easy access to a vast amount of information. Implementing summarization can make texts easier to read, cut down on time spent looking for information, and allow more information to fit in a given space.

3. LITERATURE REVIEW
Summarization of text created from one or more texts that keeps the majority of the material of the real text while being less than half as long. This is known as automatic text summarizing when a machine performs it automatically. This method can be thought of as a type of compression that will ultimately cause information loss.

M. M. Haider et al. suggested a sentence-based clustering technique (K- Means) for a single document. Gensim word2vec was utilized for feature extraction. The suggested model performed the best on the numerical values, which are prioritized by the sentence scoring approach. Over text values [10], it delays. The focus of Tanwi1 et al. and associates was on creating a system that can swiftly sum up technological notions. The system ignores other words with low scores and numerical values and only outputs the highest scoring keywords that have been considered to be significant [11]. With statements that are strongly suggested by other sentences, R. Mihalcea's text summary is more likely to be informative for the provided text and will therefore receive a better grade.

4. DATASETS
Datasets has been taken in from an online repository. The first dataset is called Tennis Article. Each instance of data has a full-sized text, along with three different method generated summaries. The second dataset is called House Article. Each instance of data has a full-sized text, along with human generated summary and three different method generated summaries.

5. PRE-PROCESSING STEPS
5.1 Sentence Segmentation
The practice of breaking up a long string of text into its individual sentences is called sentence tokenization (also known as sentence segmentation). Tokenization is the process of breaking down a text string into a group of tokens. An example of a token is a word in a sentence, while a sentence is a token in a paragraph. You may consider tokens to be components. The two different kinds of word segmentation algorithms dictionary-based (DCB) and machine-learning-based (MLB) are as follows: The DCB technique segments and parses input texts using a list of keywords. The MLB method, on the other hand, uses machine learning to train a model from a corpus.

5.2 Removing Stop Words
A collection of phrases known as "stop words" is used often in all languages. Stop words in English include the words "the," "is," and "and." Stop words are used in NLP and text mining applications to eliminate superfluous keywords so that computers may concentrate on the important words.

5.3 Removing Stemming Words
Stemming can be defined as the elimination of a word's middle or the reduction of a term to its stem or root.

6. MACHINE LEARNING APPROACHES

6.1 Text Rank Algorithm

A set of sentences from a document can be extracted using Text Rank to build a document summary (either through post-processing of the extracted set of sentences, or by using the set of sentences directly as the summary).

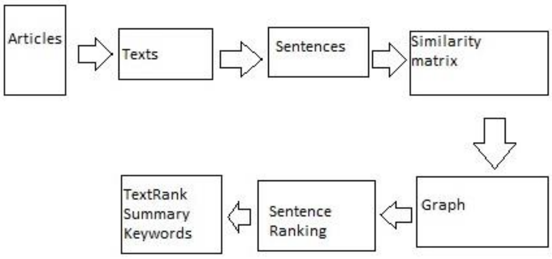

Figure 1: Diagram of Text Rank Algorithm

6.2 Text summarization using SpaCy

The Python and Cython programming languages were used to create the powerful natural language processing framework known as SpaCy. SpaCy supports deep learning processes using PyTorch and TensorFlow statistical models and is mostly utilized in the development of production software.

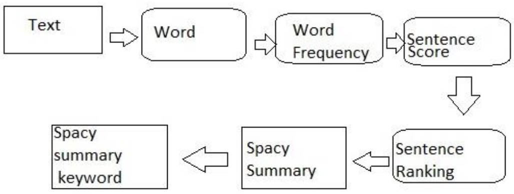

Figure 2: Dataflow Diagram using SpaCy library

6.3 Text Summarization by Keywords (using Gensim)

To find a summary, Gensim's summarization is employed. This summary is based on the TextRank algorithm, which ranks text phrases using a variant of the TextRank algorithm. Figure depicts the dataflow diagram for text summarization by keyword extraction using the Gensim package.

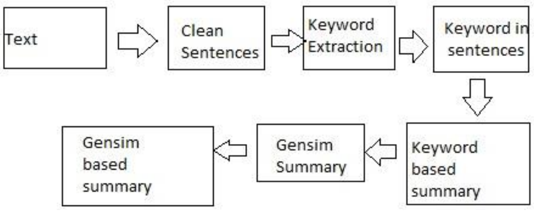

Figure 3: Dataflow diagram of Gensim Keyword summary

7. USED ALGORITHM

7.1 Combined Keywords Extraction and Summarization Method

Summary generated by TextRank algorithm, sentence score and Gensim are taken, and keywords are generated from them individually. After that we united the keywords and found a list of combined common keywords, which has been used to generate our resultant summary. The dataflow diagram of summarization in this way is shown in figure 4.

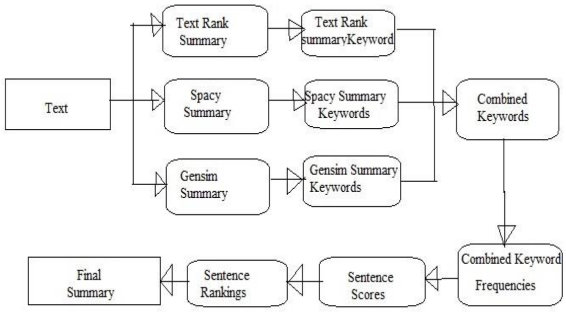

Figure 4: Dataflow diagram of proposed methodology

7.2 Summary by Combined Process.

7.2.1 Generation of key words from Text Rank summary

After ranking all the sentences from the graph representation, we rank all the output sentences and finally, we get some number of top-ranking sentences and form the final summary. Then Key words are generated from this summary.

7.2.2 Generation of key words from sentence score summary

By using Spacy library, we first find out the sentence score and get a summary from top scored sentences. Now Key words are generated from that summary.

7.2.3 Generation of key words from key word extraction using Gensim library

In this step we have used Gensim library to extract keyword and to get a final summary using those extracted keywords.

7.2.4 Generation of combined keywords

In this step we have performed an “Union” operation between the separate key words we have find out from different process to get combined keywords.

7.2.5 Counting Keyword Frequency

In this step we count the frequency of each and every keyword in the article.

7.2.6 Counting Sentence score

By adding the word frequency of very words used in a sentence we find the score of every sentence used in the article. To do so, here we used SpaCy library.

7.2.7 Generation of final summary from the combined process

Finally, we get the top scored sentences, and generate our final summary.

7.3 Accuracy Checking

We compute Precision, Recall, and F-measure values for the dataset in order to assess the effectiveness of the developed combined process of text summarization employing (Text Rank, Sentence Score, and Gensim Keyword Extraction).

The suggested combined process of summarization system's effectiveness is evaluated using the F-measure score in relation to the ROUGE-2 metrics. The best F-score for the summary can be determined using Table 3

Table 1: Average ROUGE Score calculation of proposed system summary

Average ROUGE Score	Metrics	Score in %
ROUGE-1	Recall-r	43.33
	Precision-p	71.77
	F1 Measure	54.04
ROUGE-2	Recall-r	27.70
	Precision-p	62.11
	F1 Measure	38.31
ROUGE-L	Recall-r	42.96
	Precision-p	71.16
	F1 Measure	53.57

Table 2: Comparison table of the proposed System summary F-Score with various existing methods

Source	Author	Methodology	F-Score*(in%)		
			R-1	R-2	R-L
[35]	Ramesh Nallapati, et al	SummaRuNNer (two-layer RNN-based sequence classifier)	46.60	23.10	43.03
[36]	Cengiz Harka, Ali Karcı	Karcı entropy-based summarization	49.41	22.47	46.13
[37]	Yang Liu	BERT-SUM (BERT with interval segment embeddings and inter-Sentence transformer)	43.25	20.24	39.63
[38]	Aishwarya Jadhav, et al.	SWAP-NET (Seq2Seq Model with switching mechanism)	41.60	18.30	37.70
[39]	Wafaa S. El-Kassas, et al	EdgeSumm (unsupervised graph-based framework)	53.80	28.58	49.79
	Proposed Methodology	Combined Keyword Extraction and summarization	**54.04**	**38.31**	**53.57**

The above Table shows the comparison between the implemented combined process and the existing various methods. From the results, marked with bold letter it is clear that the proposed method works better than the existing methods.

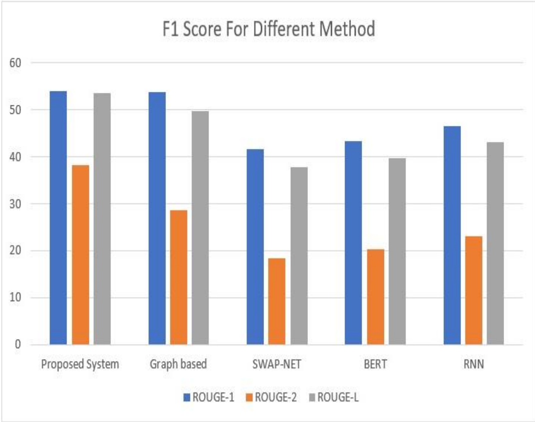

Figure 5: Comparison bar diagram of Proposed System and Existing Algorithms

8. CONCLUSIONS

Users can gain from text summarizing since it enables them to quickly extract only the information they require. Much research has been conducted in this area recently. Text summarizing is a tough procedure that requires a human-like summary to be produced. Extractive summarization is a very cohesive, less redundant and cogent way. When we tried to match the summary keywords that were generated by humans and extracted from those summaries, we discovered distinct variable results. This peculiarity and less resemblance to human-like procedures for summaries are the result of different methods' reliance on different strategies. From this comparative study we can conclude that the generated summary from the combined extracted keywords (Text Rank, Sentence Score, and Gensim Keyword Extraction) provide only the important sentences that are short and more similar to human summary than other state of the art approaches.

9. FUTURE RECOMMENDATIONS

Text synthesis is a method for shortening lengthy text paragraphs into manageable chunks. Our objective is to provide a coherent, fluid summary that only contains the main points of the text.

In the future, one should aim to summarize in the following ways:

- Text summarization in different languages with limited resources, such as Bengali.
- Text summarization with more combined strategies which results more perfection.
- Creating a system that gets concise summaries of technological topics.

10. REFERENCES

[1] Afzal M, Alam F, Malik KM, Malik GM, "Clinical Context-Aware Biomedical Text Summarization Using Deep Neural Network: Model Development and Validation," *J Med Internet Res 2020*;22(10):e19810, DOI: 10.2196/19810, PMID 33095174.

[2] P. S. Namrata Kumari1, "AUTOMATED HINDI TEXT

SUMMARIZATION TF- IDF AND TEXT RANK ALGORITHM," *JOURNAL OF CRITICAL REVIEWS*, 2020.

[3] B. Adrian, M. Megan and A. Rakshit, "Textual evidence for the perfunctorness of independent medical reviews,"*EDANZ*, pp. 2053-2062, Oct 2020.

[4] S. Tian, "Neural Abstractive Text Summarization with Sequence-to- Sequence Models," *ACM/IMS Transactions on Data Science*, pp. 1- 49,2021.

[5] W. R, "Fuzzy clustering for topic analysis and summarization of document collections," *In Advances in Artificial Intelligence.* Jan,2019.

[6] A. N and B.L.A., "An overview of text summarization techniques. In Computing Communication Control and automation (ICCUBEA," *IEEE*, pp.1-7,2016).

[7] P. D and D. Y. V, "A fuzzy approach for text mining.," *IJ Mathematical Sciences and Computing,* pp. 34-43,2015.

[8] B. Z. and H. M., "An Innovative Approach of Bangla Text Summarization by Introducing Pronoun Replacement and Improved Sentence Ranking.," *Information Processing Systems*, 2017.

[9] K. H. and I. M., "Automated Bangla text summarization by sentence scoring and ranking." *IEEE*, pp. 1-5,2013.

[10] M. M. Haider, Md. A. Hossain, and H. Rashid Mahi, "Automatic Text Summarization Using Gensim Word2Vec and K-Means Clustering Algorithm," *IEEE Region*, pp.1-25, 2 November2020.

[11] Tanwil, S. Ghosh, V. Kumar, Y. S. Jain, Mr. Avinash, "Automatic Text Summarization using Text Rank," *elsevier*, vol.06, pp. 2657, Apr 2019.

[12] R. Mihalcea, "Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization," *ACLdemo'04*, pp .20, July 2004.

[13] A. K. Yadav, M. Kumar and A.Pathre ,"Implemented Text Rank based Automatic Text Summarization using Keyword Extraction, " *Information Processing Systems*, Nov 2020.

[14] D. K. Bharti, "Automatic Keyword Extraction for Text Summarization in Multi-document e- Newspapers Articles," *European Journal of Advances in Engineering and Technology,* vol6, pp. 410-427, Jul 2017.

[15] S. K. Bharti, K. S. Babu, and S. K. Jena, "Automatic Keyword Extraction for Text Summarization," *Expert Systems with Applications*, April 2017.

[16] H. H. Chen, J. J. Kuo, and T. C. Su, "Clustering and Visualization in a Multi-lingual Multi- document Summarization System," *Springer*, pp.266-280, 2003.

[17] R. M. Alguliev, R. M. Aliguliyev, M. S. Hajirahimova, and C. A. Mehdiyev, "Maximum coverage and minimum redundant text summarization model," *Elsevier*, pp.14514-14522, Nov 2011.

[18] E. Aramaki, Y. Miura, and M. Tonoike, "TEXT2TABLE: Medical Text Summarization System based on Named Entity Recognition and Modality Identification," *Association for Computational Linguistics*, pp.185–192, Jun 2009.

[19] G. S. Erkan, and D. R. Radev, **"**LexRank: Graph-based Lexical Centrality as Salience in Text Summarization," *Journal of Artificial Intelligence Research*, pp.457-479, Jul 2004.

[20] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer", *in Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, 1995, pp. 68–73.

[21] S. A. Anam, A. M. Rahman, N. N. Saleheen, and H. Arif, "Automatic text summarization using fuzzy c-means clustering", in 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), *IEEE*, 2018, pp. 180–184.

[22] D. Das and A. Martins, "A survey on automatic text summarization. literature survey for language and statistics", *II Course at CMU*, 2007.

[23] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization", *Association for Computational Linguistics*, vol.7, pp.209-215, 2017.

[24] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques", *Journal of emerging technologies in web intelligence*, vol. 2, no. 3, pp. 258–268, 2010.

[25] B. Li, A. Drozd, Y. Guo, T. Liu, S. Matsuoka, and X. Du, "Scaling word2vec on big corpus", *Data Science and Engineering*, vol. 4, no. 2, pp. 157–175, 2019.

[26] R. A. Garcıa-Hern´andez, R. Montiel, Y. Ledeneva, E. Rend´on, A. Gelbukh, and R. Cruz, "Text summarization by sentence extraction using unsupervised learning", in Mexican International Conference on Artificial Intelligence, *Springer*, pp. 133–143, 2008.

[27] J. L. Neto, A. A. Freitas, and C. A. Kaestner, "Automatic text summarization using a machine learning approach", in Brazilian symposium on artificial intelligence, *Springer*, pp. 205–215, 2002.

[28] J. Xu and Q. Du, "A deep investigation into fasttext", in 2019 IEEE 21stInternational Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), *IEEE*, pp. 1714–1719, 2019.

[29] M. Palomar and E. Lloret,"Text summarization in progress," *Artificial Intelligence Review,* 2012.

[30] M. A. a. K. Kochut, "Semantic Tagging Using Topic Models Exploiting Wikipedia Category Network. In Semantic Computing," *Tenth International Conference on. IEEE,* 2016.

[31] L. C. S. J. A. F. a. S. S. Elena Baralis, "Multi-document summarization based on the Yago ontology," *Expert Systems with Applications,* 2017.

[32] S. P. M. A. S. S. E. D. T. J. B. G. M. Allahyari, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," *Springer*, pp.350- 362, July 2017.

[33] J. M. C. a. D. P. O'leary, "Text summarization via hidden Marcov model," *In Proceedings of the 24th annual international ACM SIGIR conference*, pp.406-407.

[34] H. P. Edmundson., "New methods in automatic extracting," *the ACM (JACM)*, pp. 264 –285. 1969.

[35] Nallapati, R.; Zhai, F.; Zhou, B. Summarunner, "A recurrent neural networkbased sequence model for extractive summarization of documents.," *In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, CA, USA, 4–9 February 2017.

[36] Hark, C.; Karcı, A. Karcı summarization: “A simple and effective approach for automatic text summarization using Karcı entropy.” *Inf. Process. Manag*. 2019, 57, 102187.

[37] Liu, Y. Fine-tune BERT for extractive summarization. *arXiv* 2019, arXiv:1903.10318.

[38] [38] Jadhav, A.; Rajan, V. “Extractive summarization with swap-net: Sentences and words from alternating pointer networks*”. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 15–20 July 2018; Volume 1.

[39] El-Kassas, W.S.; Salama, C.R.; Rafea, A.A.; Mohamed, H.K. EdgeSumm, “Graph-based framework for automatic text summarization.” *Inf. Process. Manag*. 2020, 57, 102264.