A Ranking based Language Model for Automatic Extractive Text Summarization

Pooja Gupta, Swati Nigam, Rajiv Singh

Department of Computer Science, Banasthali Vidyapith, Banasthali, India
poojagupta2291@gmail.com, swatinigam.au@gmail.com, jkrajivsingh@gmail.com

2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR) | 978-1-6654-2521-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAITPR51569.2022.9844187

Abstract—**Increased availability of the Internet and social media has created another 'world of data' comprised of text, audio and video files. It is very difficult for a user to get the accurate summary or to comprehend the relevant and important items from the available media. Additionally, readers or evaluators of these data files are interested only in the relevant content or summary to be retrieved in the less duration from the source files. Automatic text summarization (ATS) is the only way to summarize single or multiple documents to obtain relevant content from the source files. Available ATS systems generate bad summaries and take a lot of time and space for long documents due to inaccurate encoding. Therefore, in this work, we have introduced an approach for extractive text summarization using sentence ranking. Experiments have been performed over BBC and CNN news datasets and evaluated in terms of ROUGE using N-gram Language Model. The quantitative values of the metrics show the effectiveness of the proposed approach for news datasets.**

Keywords—Automatic text summarization, extractive text summarization, language model, natural language processing, ROUGE.

I. Introduction

Text summarization is aimed to summarize the documents based on their content, relevance and ranking which may be manual or automatic. As manual text summarization is very difficult to perform due to large amount of data and requires great efforts and time; therefore, automatic text summarization (ATS) has been considered for single and multiple documents [1, 2]. A generic framework for ATS has been shown in Fig. 1.

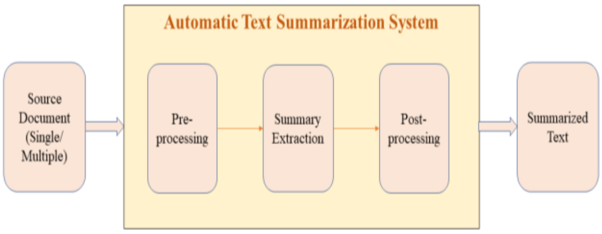

Automatic Text Summarization System
Source Document (Single/ Multiple)
Pre-processing
Summary Extraction
Post-processing
Summarized Text

Fig. 1. A general automatic text summarization system.

ATS aims to provide a short and relevant summary of the documents (single or multiple) [3, 11]. There are a number of techniques, datasets and bad summary generators available on the Internet for ATS systems. Even after almost six decades, since the start of the text summarization research, researchers are still far away from extracting the relevant summaries of scientific papers, legal documents, books, novels and web resources like websites, social media, news blogs and many more.

To solve this problem, we aim to propose a new method of sentence ranking which is better from the existing methods [5] and able to generate correct summary of the multiple documents. For generating the correct summary, we develop a new language model based on sentence ranking, which has been used for automatic text summarization.

We further calculate multiple ROUGE values for accurate summary generation. This research work focuses on automatic extractive summary generation from two datasets, BBC News articles [3] and CNN stories [4]. Both datasets are very popular and important in the context of text summarization. Experiments have been performed on them and evaluated in terms of F-score and ROUGE metrics.

Rest of the article is organized into following sections. Related work is discussed in Section II. The proposed framework is explained in Section III. Experiments and results for two public datasets are given in Section IV and Section V concludes the work.

II. Related Work

Text summarization research has started around 1958 which provided an abstractive summary of news article [9]. Since then, researchers have developed a number of techniques for text summarization and shifted towards automatic text summarization due to a huge increase of diverse documents on the Internet [10].

ATS techniques can be broadly classified into extractive, abstractive and hybrid methods [3]. Extractive text summarization is the approach of extracting or selecting the important sentences from the given input text or document to obtain the summary [6, 12, 13]. It deals with statistical features that consist of pre-processing and post-processing phases [13]. Statistical features such as frequency count, n-gram and length of sentence are used in text analysis [14]. Ranking based methods have been used to extract the scores of the sentences [15]. Highly scored sentences are used to generate summary for input documents.

Abstractive text summarization is the process of understanding the given text or document to generate meaningful summary in a few new sentences [7, 16]. It can be done using linguistic methods for generation of new and short sentences or new text which represent important information of original text. It generates a summary by extracting the semantics of the given input documents. Abstractive summarization methods can generate new sentences instead of simply filtering the sentences from the original text using different natural language processing techniques [17]. Hybrid text summarization combines extractive and abstractive techniques to get better accuracy in the result summary [8].

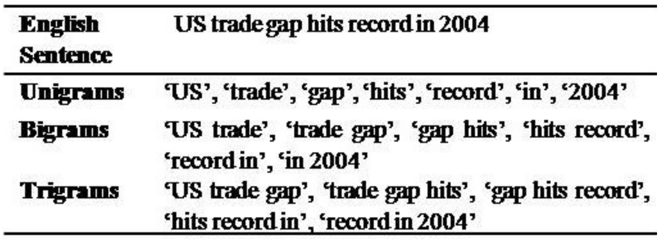
In addition to this, other machine learning approaches are used to capture the structure of the sentences in the early stage of document summarization, i.e., in the preprocessing step using language modeling. A language model can be used to calculate the scores using probability of n-gram words for given sentences to find the coherence factor of the source document (single or multiple) [18]. Language model is an essential part of natural language processing. It measures the frequency score of the sentences.

Evaluation of the text summarization methods [19] is very challenging due to the great amount of data and a diverse range of generated text summaries. Manual evaluation of summaries can be done using a human judge which includes subjective metrics such as readability, structure of the summary, grammatically correct, referential clarity, conciseness and non-redundancy [20]. Manual evaluation of summaries tends to be very time consuming and takes much efforts. Therefore, for automatic evaluation, metrics like precision, recall, F-score and ROUGE are preferred. ROUGE is an automatic evaluation method for generated summaries [21] and performs comparison with the human generated (ideal) summaries.

III. The Proposed Framework

The proposed approach is based on n-gram language model also known as baseline model. In this work, we extract unigrams, bigrams and trigrams of given datasets. Extraction of unigrams, bigrams and trigrams for a sentence is shown in Table 1.

Table 1. Extraction of n-grams



English Sentence	US trade gap hits record in 2004
Unigrams	'US', 'trade', 'gap', 'hits', 'record', 'in', '2004'
Bigrams	'US trade', 'trade gap', 'gap hits', 'hits record', 'record in', 'in 2004'
Trigrams	'US trade gap', 'trade gap hits', 'gap hits record', 'hits record in', 'record in 2004'

The proposed methodology for automatic text summarization has been divided into four major components: preprocessing, sentence generation, automatic text summarization and summary generation.

A. Preprocessing

Initially the source or input documents $(D_1, D_2, ..., D_n)$ are to be preprocessed to obtain an intermediate representation using steps shown in Fig. 2. We have broken the input documents into sentences using the term sentence tokenization. After the sentence tokenization, we have obtained the term of each sentence using the word tokenization. We have extracted the unigrams, bigrams, and trigrams of the given sentences and collected in text files. A brief description of preprocessing steps is given here.

a) Data Collection: We have collected two datasets BBC News dataset consisting of business, entertainment, politics, sport, and tech articles. We have also collected CNN news stories based on different news channels. We have created an n-gram language model for a chosen sentences at level-3.

b) Data Cleaning: For cleaning the data, we have removed all the punctuation marks and symbols, i.e. stop word removal as shown in Fig. 2. We have used total 250 sentences taken from both the datasets and then extracted the unigrams, bigrams and trigrams respectively. Statistics of generating unigrams, bigrams and trigrams for both datasets is shown in Table 2.

Table 2. Statistics of n-grams for the proposed method

Datasets	Sentences	Unigrams	Bigrams	Trigrams
BBC News	250	15270	15020	14770
CNN	250	7020	6755	6490

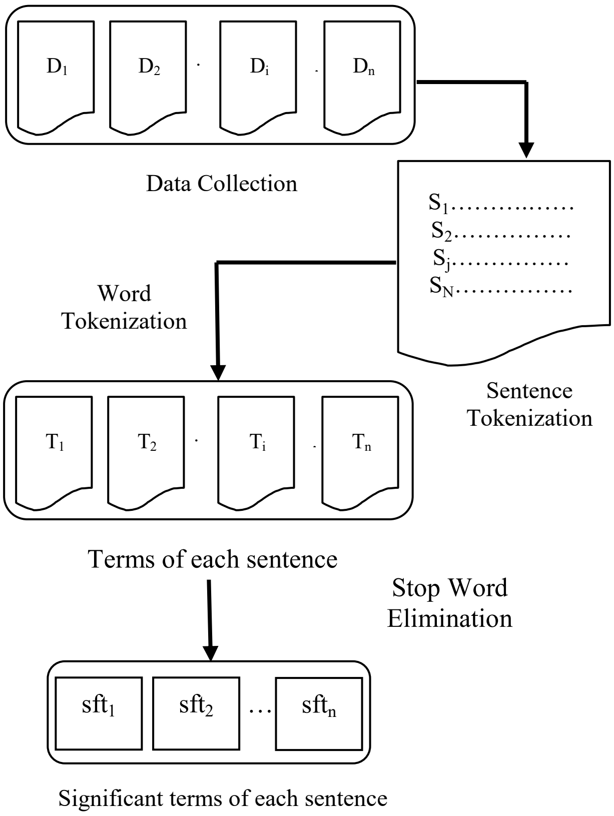


Fig. 2. Data preprocessing.

c) Sentence Tokenization: In Sentence tokenization we have to identify the start and end of a particular sentence. For example, a document D can be tokenized into sentences $(S_1, S_2, ..., S_n)$ using language modeling. The obtained sentences will be used to tokenize words, i.e., a sentence S can be tokenized into words $(\mathrm{w}_1, \mathrm{w}_2, ..., \mathrm{w}_n)$ which are required for the calculation of feature scores of separate words.

B. Sentence Generation

The preprocessed words, denoted by $w = (\mathrm{w}_1, \mathrm{w}_2, ..., \mathrm{w}_n)$ are used to compute the weights of each sentence. These weights define sum of term frequency for each word. The frequency of each word is calculated by using the probability of the occurrences of words in the sentences or documents. The probability estimation of unigram, bigram and trigram for preprocessed words $w = (\mathrm{w}_1, \mathrm{w}_2, ..., \mathrm{w}_n)$ can be calculated by equations 1, 2 and 3, respectively.

$$P(w_n) = \frac{Count(w_n)}{|V|} \tag{1}$$

$$P(w_{n-1}w_n)=\frac{Count(w_{n-1}w_n)}{Count(w_{n-1})} \tag{2}$$

$$P(w_{n-2}w_{n-1}w_n)=\frac{Count(w_{n-2}w_{n-1}w_n)}{Count(w_{n-2}w_{n-1})} \tag{3}$$

$Count$ denotes number of occurrences for unigram, bigram and trigram respectively and V denotes the number of frequency count for unigram only.

These weights are learnt for sentence generation using maximum likelihood estimation technique. After generation, sentences will be selected on the basis of coherence factor. The multiple scores for summary generation of sentences are obtained by calculating the score of each learning model. These multiple scores are used for automatic text summarization.

C. *Text Summarization*

Text summary is generated by sentence ranking as shown in Fig. 3. We have generated language model for BBC News articles as well as a CNN stories. Then we have applied the ranking algorithm to compute the probabilities of each sentence from a given document by using the proposed language model.

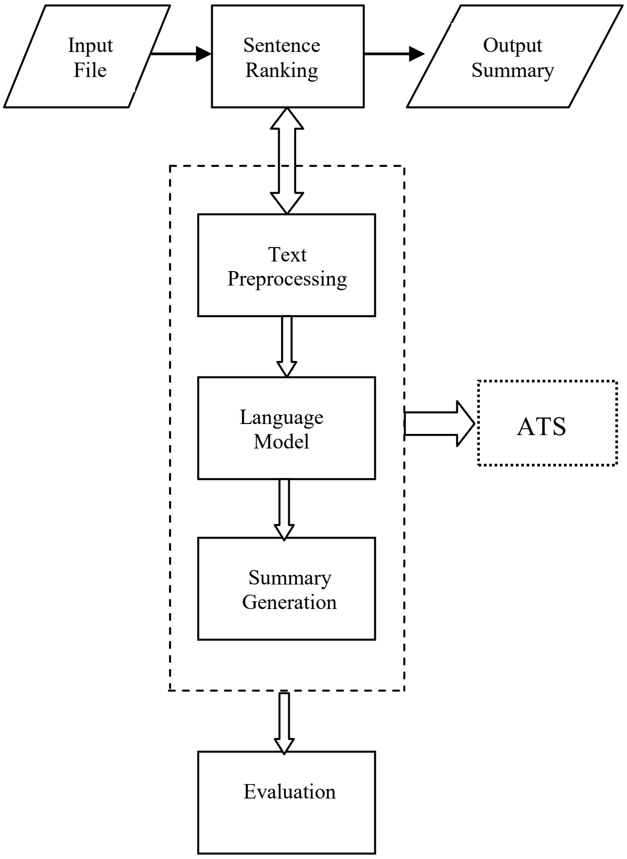


Fig. 3. Summary generation.

D. *Summary Generation*

In this step, the output scores obtained in the immediate previous step are used to generate summary of the documents. Sentence ranking has been exploited for output summaries.

The steps of the proposed approach are as follows:

Step1. Read the document and split each document into the sentences.

Step2. Read the sentences and split each sentence into the words and remove all the stop words from the sentence.

Step3. Generate Unigrams, Bigrams and Trigrams for the entire document.

Step4. Calculate frequency of the unigram, bigram and trigram on the basis of occurrences of the word present in the document.

Step5. Calculate probability of unigram, bigram and trigram using equations 1, 2 and 3 respectively, on the basis of frequency score separately.

Step6. These unigrams, bigrams and trigrams are matched with existing language model separately and matched ones are retained.

Step7. If a match is found in the existing data, then register corresponding unigram, bigram and trigram with their probabilities.

Step8. After a match has been found, calculate sum of the probabilities for each match.

Step9. Compute the average of all these probabilities.

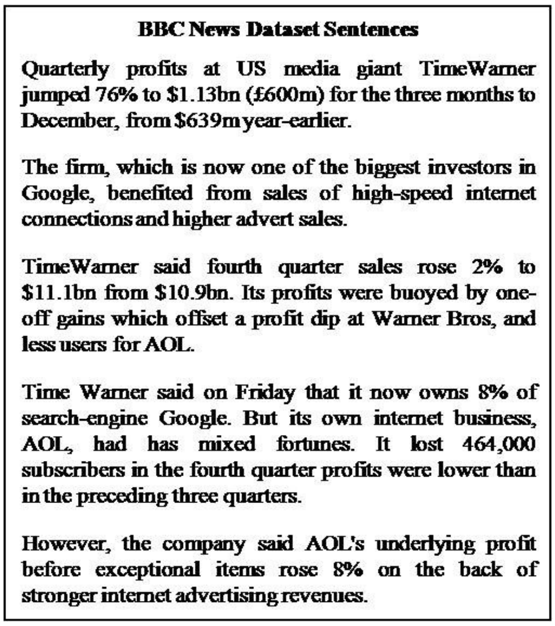
Step10. Repeat steps 1-9 for all sentences.

Step11. Sort these average probabilities of the sentences in descending order with respect to their cumulative probabilities.

Step12. Take highest probability sentences as a summary of the document.

IV. Results and Discussions

Experiments have been performed on the benchmark available datasets for single and multiple documents. Extractive summary has been produced on the basis of occurrence of the highest score of the sentence from the given datasets. We have illustrated the entire summarization process through the examples taken from both the datasets and shown in Figs. 4 and 5. Fig. 4 shows a set of sentences of BBC News dataset and their generated summary. A similar representation for CNN dataset has been shown in Fig. 5.



BBC News Dataset Sentences

Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to December, from $639m year-earlier.

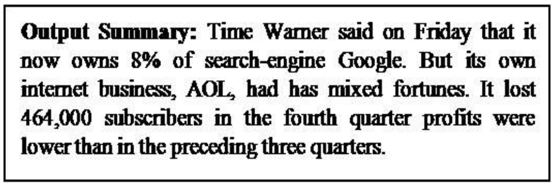
The firm, which is now one of the biggest investors in Google, benefited from sales of high-speed internet connections and higher advert sales.

TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and less users for AOL.

Time Warner said on Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters.

However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues.

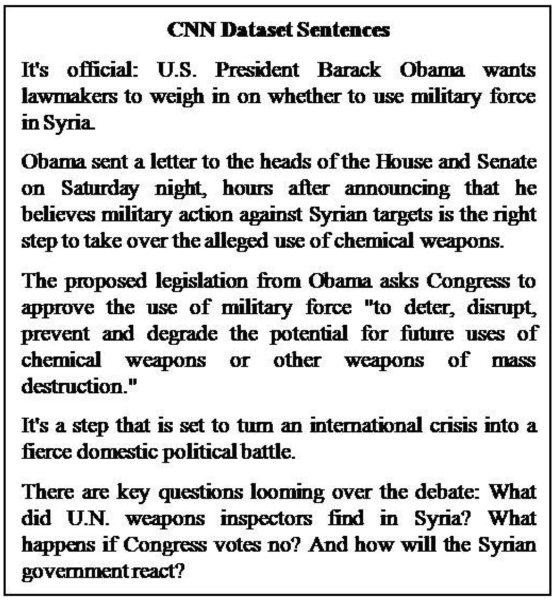
(a). Sample sentences.



Output Summary: Time Warner said on Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters.

(b). Generated summary.

Fig. 4. Summary generation for BBC News Dataset.



CNN Dataset Sentences

It's official: U.S. President Barack Obama wants lawmakers to weigh in on whether to use military force in Syria.

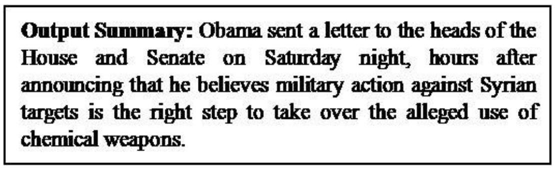
Obama sent a letter to the heads of the House and Senate on Saturday night, hours after announcing that he believes military action against Syrian targets is the right step to take over the alleged use of chemical weapons.

The proposed legislation from Obama asks Congress to approve the use of military force "to deter, disrupt, prevent and degrade the potential for future uses of chemical weapons or other weapons of mass destruction."

It's a step that is set to turn an international crisis into a fierce domestic political battle.

There are key questions looming over the debate: What did U.N. weapons inspectors find in Syria? What happens if Congress votes no? And how will the Syrian government react?

(a). Sample sentences.



Output Summary: Obama sent a letter to the heads of the House and Senate on Saturday night, hours after announcing that he believes military action against Syrian targets is the right step to take over the alleged use of chemical weapons.

(b). Generated summary.

Fig. 5. Summary generation for CNN Dataset.

Evaluation of the text summarization methods is very challenging due to the great amount of data and a diverse range of generated text summaries. Manual evaluation of summaries tends to be very time consuming and takes much efforts [22, 23]. Therefore, we have performed automatic evaluation of the proposed system on ROUGE-1, ROUGE-2 and ROUGE-3 which has been calculated using F-Score derived from precision and recall.

To evaluate the performance of the overall summarization system, we have collected 25 documents from both the datasets BBC News dataset [3] and CNN [4] for testing. These documents are tokenized into sentences by using sentence tokenization. These tokenized sentences are not part of our 250 sentences used to train the language model.

It has been observed that for BBC dataset, out of total 25 documents 11 documents and for CNN dataset, out of total 25 documents 9 documents have given the correct summaries, shown in Fig. 6. The proposed method achieves an accuracy of 44% for BBC dataset and 36% for CNN dataset.

Tables 3 and 4 show F-score of the sentences S1 to S5 for BBC News and CNN datasets, respectively. In Table 3, sentence S3 gives the highest F-score for all the ROUGE levels, and hence, generated as an output summary by the proposed method for BBC News dataset [3], shown in Fig. 4(b). From Table 4, it can be seen that sentence S2 gives the highest F-score for all the ROUGE levels, therefore, taken as an output summary for CNN Dataset [4] (shown in Fig. 5(b)). The obtained ROUGE values for both the datasets have been shown in Figs. 7 and 8. It can be easily observed from Figs. 7 and 8 that the sentences used in output summary have the highest values of different ROUGE levels.

TABLE 3. F-score of multiple ROUGE levels for BBC News Dataset

Sentences	ROUGE-1	ROUGE-2	ROUGE-3
S1	0.47	0.23	0.21
S2	0.47	0.27	0.18
S3	**0.51**	**0.30**	**0.23**
S4	0.40	0.30	0.13
S5	0.45	0.26	0.16

TABLE 4. F-score of multiple ROUGE levels for CNN Dataset

Sentences	ROUGE-1	ROUGE-2	ROUGE-3
S1	0.47	0.33	0.05
S2	**0.58**	**0.39**	**0.22**
S3	0.30	0.30	0.12
S4	0.51	0.31	0.13
S5	0.44	0.21	0.11

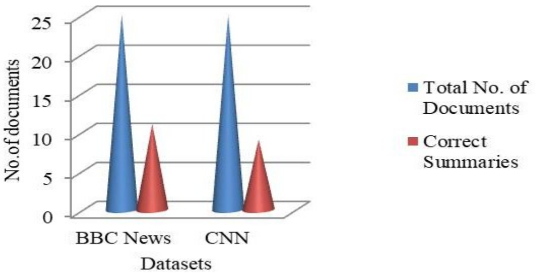


Fig 6. Evaluation of generated summaries.

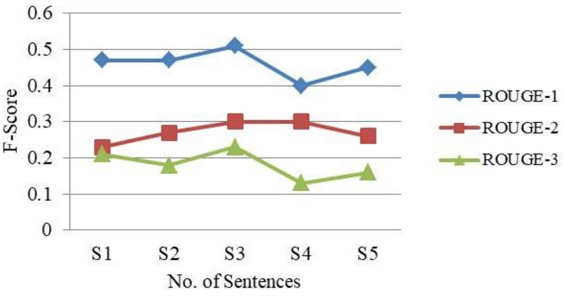


Fig. 7. Comparative analysis of ROUGE values for BBC News dataset.

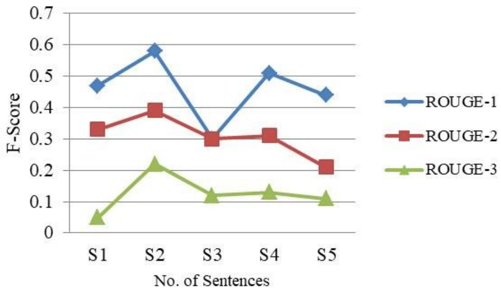


Fig. 8. Comparative analysis of ROUGE values for CNN dataset.

V. Conclusions

In this work, we have proposed an approach for extractive summary generation using sentence ranking for multiple documents. A language model has been proposed which computes multiple ROUGE values for effective summary generation for BBC News and CNN datasets. The summary generated from BBC News [3] and CNN datasets [4] have the accuracy of 44% and 36%, respectively. The results are comparable for the sentences used in the experiments for the proposed method on the basis of multiple ROUGE values. The proposed method could be further improved by using an enhanced language model with larger datasets and machine learning methods.

References

[1] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, “Automatic Text Summarization: A Comprehensive Survey,” Expert systems with applications, vol. 165, 2021.

[2] M. Gambhir and V. Gupta, “Recent automatic text summarization techniques: a survey,” Artificial Intelligence Review, vol. 47, no. 1, pp. 1-66, 2017.

[3] https://www.kaggle.com/pariza/bbc-news-summary

[4] https://www.tensorflow.org/datasets/catalog/cnn_dailymail

[5] S. M. Meena, M. P. Ramkumar, R. E. Asmitha, and G. SR. Emil Selvan, “Text Summarization Using Text Frequency Ranking Sentence Prediction,” In 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP) pp. 1-5, IEEE, 2020.

[6] R. Ferreira, L. de Souza Cabral, R. D. Lins, G. P. e Silva, F. Freitas, G. D. Cavalcanti, and L. Favaro, “Assessing sentence scoring techniques for extractive text summarization,” Expert systems with applications, vol. 40, no. 14, pp. 5755-5764, 2013.

[7] S. Gao, X. Chen, P. Li, Z. Ren, L. Bing, D. Zhao, and R. Yan, “Abstractive text summarization by incorporating reader comments,” In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6399-6406, July 2019.

[8] V. Gupta and N. Kaur, “A novel hybrid text summarization system for Punjabi text,” Cognitive Computation, vol. 8, no. 2, pp. 261-277, 2016.

[9] H. P. Luhn, “The automatic creation of literature abstracts,” IBM Journal of research and development, vol. 2, no. 2, 159-165, 1958.

[10] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, “Text summarization techniques: a brief survey,” arXiv preprint arXiv:1707.02268, 2017.

[11] M. Maybury, Advances in automatic text summarization, MIT press, 1999.

[12] N. Nazari and M. A. Mahdavi, “A survey on automatic text summarization,” Journal of AI and Data Mining, vol. *7*, no. 1, pp. 121-135, 2019.

[13] 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.

[14] M. A. Fattah and F. Ren, “GA, MR, FFNN, PNN and GMM based models for automatic text summarization,” Computer Speech and Language. vol. 23, no. 1, pp. 126-144, 2009.

[15] A. Nenkova and K. McKeown, “A survey of text summarization techniques,” In Mining text data (pp. 43-76). Springer, 2012.

[16] S. Gao, X. Chen, P. Li, Z. Ren, L. Bing, D. Zhao, and R. Yan, “Abstractive text summarization by incorporating reader comments,” In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6399-6406, 2019.

[17] N. Moratanch, and S. Chitrakala, “A survey on abstractive text summarization,” In 2016 International Conference on Circuit, power and computing technologies (ICCPCT), pp. 1-7, IEEE, 2016.

[18] R. Barzilay and L. Lee, “Catching the drift: Probabilistic content models, with applications to generation and summarization,” arXiv preprint cs/0405039, 2004.

[19] I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim, “SUMMAC: a text summarization evaluation,” Natural Language Engineering, vol. 8, no.1, pp. 43-68, 2002.

[20] I. Mani, Automatic summarization, John Benjamins Publishing, 2001.

[21] C. Y. Lin, “Rouge: A package for automatic evaluation of summaries,” In Text summarization branches out, pp. 74-81, 2004.

[22] E. Lloret, L. Plaza, and A. Aker, “The challenging task of summary evaluation: an overview,” Language Resources and Evaluation, vol. 52, no. 1,pp. 101-148, 2018.

[23] H. Kobayashi, M. Noguchi, and T. Yatsuka, “Summarization based on embedding distributions,” In Proceedings of the 2015 conference on empirical methods in natural language processing, pp. 1984-1989, September 2015.