iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.42358

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

A Review on Text Summarization Techniques

Rohit Parimoo[1], Rohit Sharma[2], Naleen Gaur[3], Nimish Jain[4], Sweeta Bansal[5]
[1, 2, 3, 4, 5]*Department of Computer Science Engineering, Inderprastha Engineering College, Ghaziabad, Uttar Pradesh*

Abstract: ***In recent years, an enormous amount of text data from diversified sources has been emerged day-by-day. This huge amount of data carries essential information and knowledge that needs to be effectively summarized to be useful.We first introduce some concepts related to extractive text summarization and then provide a systematic analysis of various text summarization techniques. In particular, some challenges in extractive summarization of single as well as multiple documents are introduced. The problems focus on the textual assessment and similarity measurement between the text documents are addressed. The challenges discussed are generic and applicable to every possible scenario in text summarization. Then, existing state-of-the-art of extractive summarization techniques are discussed that focus on the identified challenges.***

Index Terms: Summarization, Graph based summarization, Meta-heuristic based summarization, Maximal marginal relevance based summarization, Evaluation.

I. INTRODUCTION

Text summarization is an strenuous problem of Natural Language Processing (NLP) due to difficulty in interpreting every point of the text in a document. This requires a precise analysis of the text in various steps such as semantic analysis, lexical relations, named entity recognition, etc., which can be accomplished with a great deal of word knowledge only. Since it is hard to obtain the word knowledge in various aspects such as meaning of a word with respect to other content, related words, inferential interpretation, sentence generation, etc., generating abstracts as summaries have become complex. This type of summarization is classified as abstractive summarization in NLP. However, an approximation, which is classified as extractive summarization, is more flexible. In particular, system requires to identify the most relevant/significant contents of the text, extract them, order them, and return them to the user. Several aspects about a good summary have been introduced by researchers. Das and Martins (2007) have discussed three major aspects for automatic text summarization.

1) Summaries may be produced from a single or multiple documents.
2) Summaries should consist of important information.
3) Summaries should be concise.

These aspects are undoubtedly important, but a good summary should also consist of other aspects such as coverage, nonredundancy, cohesion, relevancy, and readability

II. BACKGROUND

A. Text Summarization Phases

The automatic text summarization (ATS) is a process of finding a subset of document that contains the information residing in the entire document. According to Mani (1999), a text summarization system filters the significant information from the original document to generate an abbreviated version. Generally, the summarization process can be decomposed into three phases:

- Analysis of document text to obtain text representation.
- Transformation of text representation into summary representation.
- Transfiguration of summary representation into summary text to generate summary.

The basic processing, elements, and resources, which are required to accomplish these phases are as follows.

1) *Pre-processing:* A high performance in the summarization system requires an effective pre-processing of the input text to obtain text representation. We accomplish this task of processing by employing Natural Language Tool Kit (NLTK). Here, the following steps are considered to preprocess the text.

a) *Sentence Seperation:* It is a process of recognizing the individual sentences in a document which is used as a separate unit in summarization.

b) *Stop Words Removal:* The process of stop-words removal eliminates the most frequent words occurring in a document like articles, prepositions, conjunctions, interrogations, helping verbs, etc. The stop words are removed due to their insignificant contribution in sentence extraction process.

c) *Stemming:* It is a process of converting the semantically derived term into its morpheme term. We use the Porter stemmer for English text.
d) *Part-of-Speech Tagging:* It is a process of identifying the part-of-speech words such as noun, adverb, verb, etc., in a sentence. However, the computational applications generally use more fine-grained POS tags like 'nounplural'.
e) *Keywords Extraction:* In this step, we extracte the keywords from a document. Here, all the words other than stop words are considered as keywords.
2) *Assessment of Textual Units:* The major concept which has been used in transforming the document into summary representation is text features that can be exploited to find the relevant sentences of the document. In this paper, several features are used to score the sentences such as Aggregate similarity, Bushy path, Cue phrases, Lexical relation, Named entities, Noun and verb phrases, Numerical data, Open relations, Proper noun, Sentence centrality, Sentence length, Sentence position, Sentence with title words, Sentence significance, Frequent words etc.

B. Evaluation Approaches

Evaluations are done in three stages: co-selection based evaluation (with reference summary), content based evaluation (without reference summary), and document based evaluation (with original document), which are briefly described as follows.

1) *Recall:* It is the ratio of total retrieved correct sentences to the total number of the retrieved correct sentences and no retrieved correct sentences of a document.
2) *Precision:* It is the ratio of total retrieved correct sentences to the total number of retrieved correct sentences and retrieved incorrect sentences of the document.
3) *F-score:* It measures the effectiveness of retrieval with respect to a user, which attaches β times as much importance to the recall as that of precision.
4) *Improved Rates:* We have also calculated the improved rates (IR) in the performance of the proposed methods with respect to other methods. $R = (PM - OM) / OM$ where, PM is proposed method, OM is other method, and IR is improved rates.

III. CHALLENGES

In this section, several challenges are identified during summarizing the documents in the extractive manner, which are given as follows

1) *Problem of Redundancy:* A summary is more informative as much as it contains non-redundant contents. Most of the existing approaches focus on finding relevant content from document(s) and extract them to generate the summary. But, if we investigate about the redundancy, we can cover more information in the summary. In particular, similarity measurement plays a major role in finding the redundant contents in a document. If we can precisely measure the similarity between the contents of a document, then the redundancy can be minimized in the summary.
2) *Problem of Irrelevancy:* The main aim of a summarization system is to extract relevant contents from a document that gives a quick view of the whole document. Generally, Human engineered text features are used to assess the sentences or textual units of a document. Since, it is not always feasible to incorporate all the considered features in a summary, some features may tend to create irrelevant contents in the summary. Thus, to consider all possible text features for assessment of the sentences increases complexity as well as irrelevancy. Hence, it is crucial to know which features are accountable for creating high quality summary in the given data.
3) *Problem of Loss of Coverage:* Coverage of topics of a document in the summary is an important aspect for generic text summarization. A good generic summary always reflects the information about every aspect mentioned in the document. The current summarization techniques do not focus much on coverage of topics in the generated summaries. Hence, they fail to produce good summary in case of generic summarization. This problem arises mainly in the case of multi-document summarization where the number of topics in documents are much higher than in a single document. A good summary should be readable and cohesive. By readable and cohesive mean that the contents of the summary should be conceptually related to each other.
4) *Taxonomy of Summarization Techniques:* There have been discussed a good number of works related to extractive text summarization such as Graph based methods, Maximal Marginal Relevance based methods, Meta-heuristic based methods.

IV. TECHNIQUES USED FOR TEXT SUMMARIZATION

Text summarization is broadly divided into abstractive and extractive. The brief description about each approach is discussed in following section:

A. *Abstractive Summarization*

Approach Summarizations using abstractive techniques are broadly classified into two categories: Structured based approach and Semantic based approach.

1) *Structured Based Approach:* Structured based approach encodes most important information from the document through cognitive schemes such as templates, extraction rules and other structures such as tree, ontology, lead and body phrase structure.
2) *Semantic Based Approach:* In Semantic based approach, semantic representation of document is used to feed into natural language generation (NLG) system. This method focuses on identifying noun phrase and verb phrase by processing linguistic data.

B. *Extractive Summarization*

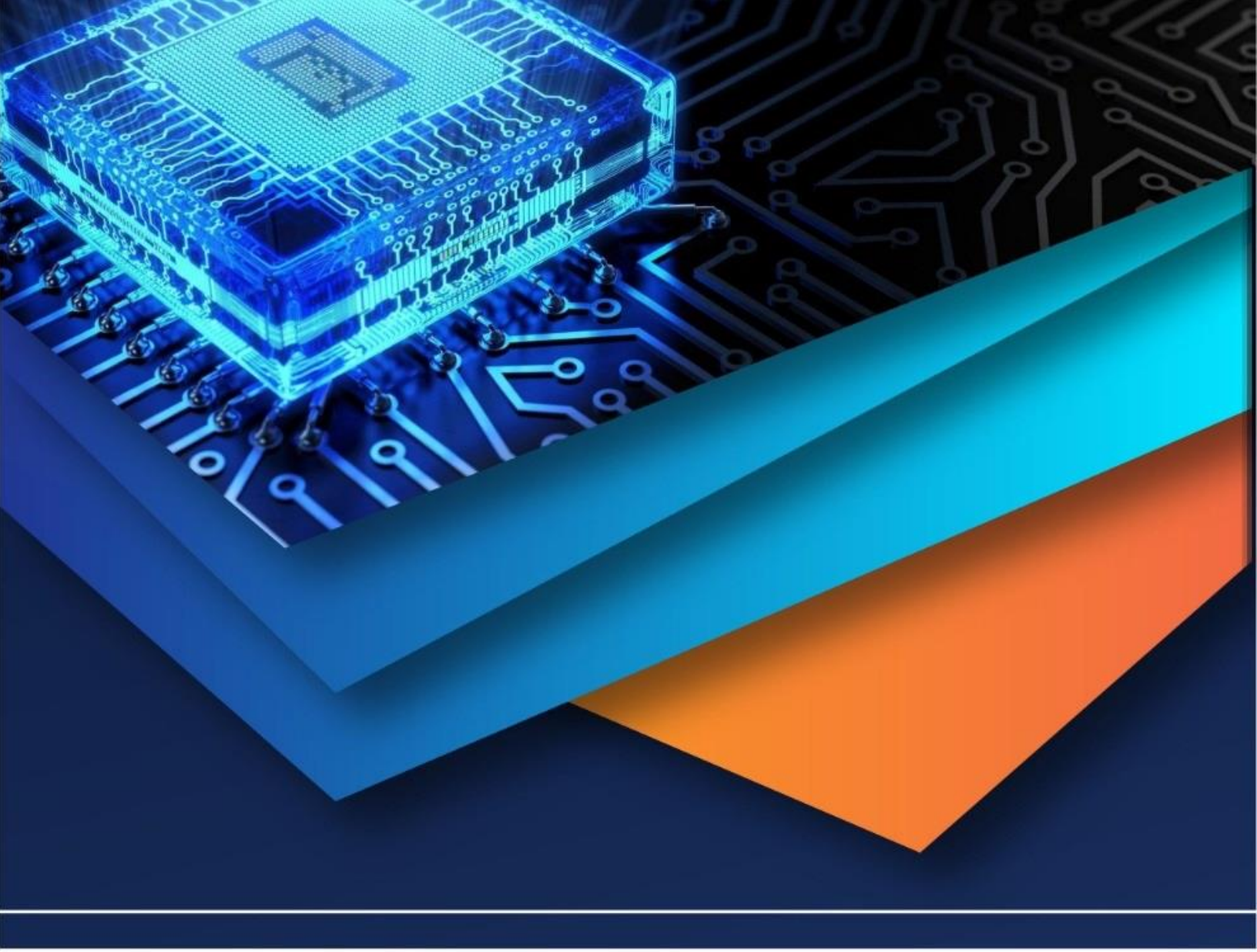
An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences.

V. CONCLUSION

Text summarization is growing as sub – branch of NLP as the demand for compressive, meaningful, abstract of topic due to large amount of information available on net. Precise information helps to search more effectively and efficiently. Thus, text summarization is need and used by business analyst, marketing executive, development, researchers, government organizations, students and teachers also. It is seen that executive requires summarization so that in a limited time required information can be processed. As abstractive summarization requires more learning and reasoning, it is bit complex then extractive approach but, abstractive summarization provides more meaningful and appropriate summary compare to extractive. Through the study it is also observed that very less work is done using abstractive methods on Indian languages, there is a lot of scope for exploring such methods for more appropriate summarization.

REFERENCES

[1] Mehdi Allahyari et al. (2017). Text Summarization Techniques: A Brief Survey. International Journal of Advanced Computer Science and Applications.
[2] Mihalcea, Rada & Rada, & Tarau, Paul & Paul. (2004). TextRank: Bringing Order into Texts.
[3] Brin and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems K. Elissa, “Title of paper if known,” unpublished.
[4] Gunawan, Dani & Sembiring, C & Budiman, Mohammad. (2018). The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. Journal of Physics: Conference Series. 978. 012120. 10.1088/1742-6596/978/1/012120.
[5] Sarica, Serhad & Luo, Jianxi. (2020). Stopwords in Technical Language Processing.
[6] Chen, Boyan & Zavarsky, Pavol & Ruhl, Ron & Lindskog, Dale. (2011). A Study of the Effectiveness of CSRF Guard. 1269-1272. 10.1109/PASSAT/SocialCom.2011.58.
[7] Lin, Chin-Yew. (2004). ROUGE: A Package for Automatic Evaluation of summaries. Proceedings of the ACL Workshop: Text Summarization Braches Out 2004. 10.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY