Multilanguage Summarizer

Siddant Bhattacharya[1], Srivats Dixit[2] , Ankita Agarwal[3]

SRMCEM,LUCKNOW

---***---

Abstract - This article provides an overview of Multilanguage Text Summarization, which is a tool that incorporates various extractive summarization techniques. These techniques operate at the sentence level, extracting and compressing sentences from documents in multiple languages. The system has been evaluated using English, Hindi, Gujarati, and Urdu documents for single-document summarization. The evaluation demonstrates that the method performs consistently well across different languages. The performance of the summarization system is measured using the F-measure score. The research study shows approximately a 2% improvement in outcomes compared to previous work.

***Key Words*:** Multilanguage Text summarization; F-measure Score, Extractive summarization

1.INTRODUCTION

One of the primary obstacles in text summarization involves accurately identifying and extracting essential concepts and information from a large volume of text. This task necessitates the utilization of advanced algorithms and techniques in natural language processing, which can be resource-intensive and pose challenges when applied to low-resource languages. Additionally, the choice between extractive and abstractive summarization relies on various factors, such as the text's nature and the intended purpose of the summary. Despite these difficulties, text summarization has made significant strides in recent years, with the introduction of novel algorithms and tools that enable summarization in diverse languages and contexts. These advancements hold the potential to enhance information accessibility for a wide range of users, including journalists, researchers, general readers, and language learners..

2. What is MULTILANGUAGE SUMMARIZATION?

2.1 Definition and types

Multilanguage Text Summarization is a technique used to create a concise version of a text document in multiple languages. It aims to extract the crucial and pertinent sentences or phrases from the original document and present them in a way that effectively communicates the main points and essential ideas of the text.

There are two primary types of Text Summarization:

- Extractive Summarization: In this approach, the essential sentences or phrases are selected from the original document and combined to form a summary. The selection process is based on the relevance of the sentences to the main themes of the text.
- Abstractive Summarization: This method involves generating a summary that may not necessarily rely on exact sentences or phrases from the original document. Instead, advanced techniques such as natural language understanding and machine learning are employed to create new sentences that capture the main ideas and themes of the original text.

3. Methodology of Multilanguage Summarization

The process of Multilanguage Text Summarization involves several steps:

- Data Collection: Relevant text documents are gathered from various sources in different languages.

- Language Identification: Machine learning techniques, such as natural language processing

(NLP), are utilized to identify the language of each text document.

- Text Preprocessing: The collected text documents undergo preprocessing to remove stop words, punctuation, and other unnecessary elements.
- Sentence Segmentation: Each text document is segmented into individual sentences.
- Summarization Method Selection: Based on the document type and desired summary output (extractive or abstractive), an appropriate summarization method is chosen.
- Text Summarization: The selected summarization method is applied to the preprocessed sentences to generate a summary.
- Translation: If required, the summary is translated into the desired language using machine translation techniques.
- Evaluation: The quality of the generated summary is assessed using metrics such as rouge score, semantic similarity, and readability.
- Refinement: The summarization process is refined through iterative testing and analysis to enhance the quality of the summary.
- Deployment: The final summary is deployed for various applications, including news aggregation, social media monitoring, and academic research.

4. Literature review

The author of this paper [1] conducted an analysis and performance comparison of three different algorithms. The paper begins by explaining various text summarization techniques. Extraction-based techniques are employed to extract important keywords that should be included in the summary. For the comparison, three keyword extraction algorithms, namely Text Rank, Lex Rank, and Latent Semantic Analysis (LSA), were utilized. The paper provides detailed explanations and Python implementations of these three algorithms. The effectiveness of the extracted keywords was evaluated using ROUGE 1. The results were compared with manually written summaries to assess the performance. Ultimately, the Text Rank Algorithm outperformed the other two algorithms, yielding better results.

The paper [2] provides a survey of several powerful Automatic Text Summarization techniques. It introduces a novel evaluation package called Recall Oriented Understudy for Gisting Evaluation (ROUGE) for assessing the quality of text summarization. The paper also presents four different measures of ROUGE: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. These measures compare the generated summary with reference summaries created by humans, enabling the evaluation of the summary's quality. The ROUGE methods are effective for automatically evaluating both single-document summaries and multi-document summaries.

. In this paper [3] author has reviewed different techniques of Sentiment analysis and different techniques of text summarization. Sentiment analysis isa machine learning approaching which machine learns and analyze the sentiments, emotions present in the text. The machine learning methods like Naive Bayes Classifier and Support Machine Vectors (SVM) are used these methods are used to determine the emotions and sentiments in the text data like reviews about movies or products. In Text summarization, uses the natural language processing (NPL) and linguistic features of sentences are used for checking the importance of the words and sentences that can be included in the final summary. In this paper, a survey has been done of previous research work related to text summarization and Sentiment analysis, so that new research area can be explored by considering the merits and demerits of the current techniques and strategies.

In the paper [4], the author introduces a system that utilizes WordNet ontology to generate abstractive summaries from extractive summaries. The system is capable of processing various document formats, including text, PDF, and Word files. The paper covers a range of text summarization techniques and provides a detailed, step-by-step explanation of the multiple document text summarization approach. To evaluate the system's performance, the author compares the experimental results with existing online extractive tools and other abstractive systems. Additionally, human-generated summaries are included in the comparison. The findings demonstrate that the proposed system yields favorable results in terms of summarization accuracy. The author also suggests future improvements for enhancing the summarization accuracy. One of the proposed methods involves comparing the system with alternative approaches to identify areas of enhancement. By incorporating additional techniques or models, the author believes the system's performance in generating abstractive summaries can be further enhanced.

The research paper [5] introduces two methods for generating generic text summaries by ranking and extracting sentences from the main text documents. The first method employs information retrieval (IR) techniques to rank sentence relevance and assign relevance scores to each sentence. The second method utilizes latent semantic analysis (LSA), specifically latent semantic indexing (LSI), to identify the semantic importance of sentences for summary creation. The author applies Singular Value Decomposition (SVD) to generate the text summary. The paper provides a step-by-step explanation of the SVD-based methods and also explores the impact of different Weighted Schemes on summary performance. The proposed methods produce generic abstractive summaries and are evaluated by comparing them to human-generated summaries. The results indicate that the proposed methods generate abstractive summaries that closely resemble human-like summaries. In the future, the author suggests exploring various machine learning techniques to further improve the quality of generic text summarization.

In the paper [6], the author introduces Text Rank, a graph-based ranking model for text processing. Text Rank is an unsupervised method that is used for keyword and sentence extraction. The approach employs a voting-based weighting mechanism to assign scores to sentences and determine their importance. The sentences are represented as nodes in a graph, and their significance is determined based on the incoming and outgoing edges from these nodes. The weight of each sentence is calculated using similarity scores between sentences. It is worth noting that Text Rank is inspired by Google's Page Rank algorithm. By applying Text Rank, the paper demonstrates that it can generate extractive summaries of the text. The results obtained from Text Rank are reported as highly effective and provide the best summarization outcomes compared to other methods.

In the research paper [7], the author introduces a graph-based method called LexRank. This approach calculates sentence scores using Eigenvector Centrality, a cosine transform weighting method. The original text is divided into sentences, and a graph is constructed with sentences as nodes. The paper provides a detailed explanation of the complete LexRank method.The results of the study demonstrate that LexRank surpasses existing centroid-based methods in terms of performance. Additionally, LexRank exhibits robustness in handling noisy data. This method is capable of generating extractive summaries of the text, summarizing the main ideas and key information contained within the document.

5. Approaches to text summarization

Based on our research findings, it is evident that extractive-based summarization implementations have shown greater success compared to abstractive-based approaches. However, even within the specific domains where these studies have been conducted, the accuracy of the extractive methods falls short of meeting the expectations of regular users. On the other hand, the research conducted on abstractive summarization indicates that while successful implementations are rare, theoretically, they have the potential to generate more coherent summaries compared to extractive methods. Despite the challenges, the prospects for achieving successful abstractive summarization implementations remain promising, with the possibility of producing summaries that are more meaningful and contextually accurate.

6. Proposed system

The proposed system for implementing the summarization technique focuses on generating concise summaries by leveraging the concepts of frequency and relevance. The initial step involves preprocessing the input document or documents to eliminate irrelevant elements. Following this, the text is tokenized, and the frequency of each word or phrase is calculated to identify the most commonly occurring terms. Relevance scores are then assigned to sentences based on the presence of important terms, facilitating the identification of highly pertinent content. These sentences are subsequently ranked according to their relevance scores, with higher-ranked sentences being deemed more significant. The top-ranked sentences are selected to form the summary, ensuring adherence to predefined length or percentage criteria relative to the original document. To enhance readability and coherence, the generated summary undergoes post-processing, refining its structure. Ultimately, the output of the system is a summary that effectively captures the primary information and core ideas from the source document(s) by employing the summarization technique. To evaluate the system's performance, metrics such as ROUGE scores can be employed to assess the quality of the generated summary in comparison to other summarization methods or summaries created by humans. In summary, the proposed system enables the practical application of the summarization technique, facilitating the extraction of essential information and the production of concise summaries from extensive textual data.

After the process of summarization, language translation plays a crucial role in enabling access to information across different linguistic boundaries. Once the summary has been generated in the source language, translation techniques such as machine translation are employed to convert the summary into the desired target language. This enables individuals who are not proficient in the source language to comprehend the key points and main ideas of the original text. Language translation post-summarization ensures that the summary remains accessible and useful to a wider range of readers, facilitating information dissemination and knowledge exchange on a global scale.

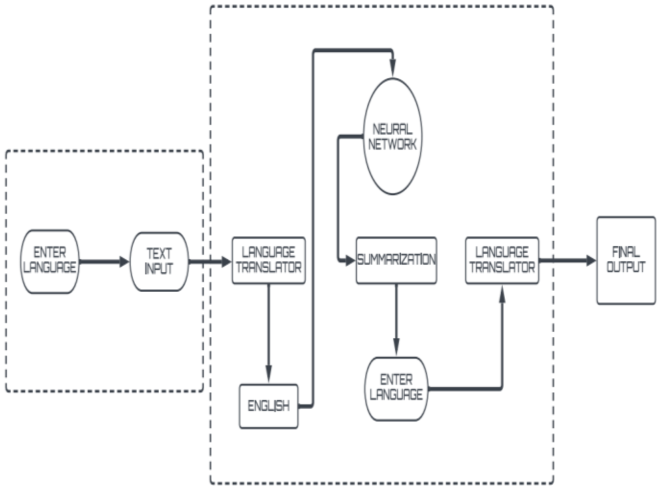

Fig -1: Block diagram of architecture

7. CONCLUSIONS

In conclusion, multilanguage translation is a vital component in facilitating effective communication and knowledge sharing across diverse linguistic communities. With the advancements in natural language processing and machine translation techniques, the process of translating summaries into multiple languages has become more accessible and efficient. Multilanguage translation after summarization enables individuals from different language backgrounds to access and comprehend essential information and key ideas, regardless of the original language of the source document. This promotes inclusivity, enhances cross-cultural understanding, and fosters global collaboration. However, it is important to continue improving translation accuracy, addressing language nuances, and considering cultural context to ensure the highest quality of multilanguage translation. By bridging language barriers, multilanguage translation plays a crucial role in promoting a more connected and inclusive world.

REFERENCES

[1] Sumitha C., Dr. A. Jaya, Amal Ganesh, "A study on Abstract Summarization Techniques in Indian Languages", Elsevier Proceeding of Computer Science, No. 87, pp.25-31, 2016.

[2] RasimAlguliev, RamizAliguliyev, "Evolutionary Algorithm for Extractive Text Summarization." Intelligent Information Management, 1, pp. 128-138, November 2009.

[3] Ronan Collobert collober, Jason Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning."

[4] Jingxuan Li, Lei Li, and Tao Li. 2012. Multidocument summarization via submodularity. Applied Intelligence 37.3: 420-430.

[5] Marina Litvak, Natalia Vanetik, Mark Last, and Elena Churkin. 2016. MUSEEC: A Multilingual Text Summarization Tool. ACL.

[6] M. Litvak, S. Kisilevich, D. Keim, H. Lipman, A. BenGur, and M. Last. 2010a. Towards languageindependent summarization: A comparative analysis of sentence extraction methods on english and hebrew corpora. In Proceedings of the CLIA/COLING 2010.

[7] Rada Mihalcea. 2005. Language independent extractive summarization. In AAAI'05: Proceedings of the 20th National Conference on Artificial Intelligence, pages 1688–1689.

[8] Dlikman. 2015. Linguistic features and machine learning methods in single-document extractive summarization. Master's thesis, BenGurion University of the Negev, Beer-Sheva, Israel. http://www.ise.bgu.ac.il/faculty/mlast/papers/ Thesisver7.pdf.

[9] Edmundson, H. P., "New methods in automatic abstracting", Journal of the Association for Computing Machinery, 16, 2, pp. 264-285, 1969.

[10] Erkan, G. and Radev, D., 2004. Lexpagerank: Prestige in multi-document text summarization. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, July.