IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

KANNADA KEYWORD-BASED TEXT SUMMARIZATION

Ms. Shwetha Kamath, Assistant Professor, MITE, Moodabidri
Tanvi Chandra, Swathi P Hegde, Sindhu D V, Sneha Devidas Gunagi
Dept. Computer Science & Engineering, MITE, Moodabidri

***Abstract*— The practice of condensing text material into a summary while retaining the major ideas is known as summarization. To best communicate the message concealed in the text, extractive summarizers operate with the provided text to extract sentences. The idea of discovering keywords and extracting sentences with more keywords than the rest is at the heart of most extractive summarization techniques. With an emphasis on essential terms, relevant words with a higher frequency than others are typically extracted to create keywords. We employed the TF (Term Frequency) model and GSS coefficients in the suggested system to extract keywords for the text ranking process. In this study, we presented a method for automatically extracting keywords from Kannada datasets for text summarising.**

INTRODUCTION

The amount of data in the world is always growing, which has substantially boosted interest in automatic summary generation. The process of text summarization comprises reducing a text file to a sentence or paragraph that conveys the key concepts. It would be beneficial if the textual file's summary or significant content could be automatically retrieved because it might be quite challenging for users to discover it inside a large text file. The two different categories of text summarization methods are extractive and abstractive. An extractive summary approach involves choosing important phrases, chapters, etc. from the initial source and appending them into a condensed version. The linguistic and statistical properties of sentences are utilized to assess their significance. Extractive algorithms select a subset of pre-existing keywords, paragraphs, or phrases from the source text to construct the summary. The extractive summarization techniques are often based on methodologies for sentence extraction and aim to capture the set of phrases that are most significant to the overall interpretation of a given material. To provide a summary that is more like what a person may make, an internal semantic structure is constructed utilizing abstractive techniques. generate. Such a summary can include terms that aren't included in the original text. A rising number of people are getting concerned about information overload as a result of the web's explosive growth. Automatic summarization is a crucial technique for addressing the problem of information overload on the internet. contain terms that aren't mentioned in the original text. A rising number of people are getting anxious about information overload as a consequence of the internet's explosive growth. Automatic summarization is a crucial technique for addressing the problem of information overload on the internet.

LITERATURE SURVEY

When compared to other languages like English, abstractive summarization research activities in Indian languages are in an early stage. This is mostly because of the variety of Indian languages and a lack of resources, including raw data and different NLP tools. The extremely few abstractive summaries work in Indian languages including Telugu, Hindi, Bengali, Kannada, and Malayalam are explained in this section.

Jancy Joseph, Asst.Professor, St. Joseph's College PilathraKannur, Kerala suggests a solution for extractive summarization for Malayalam text documents by considering the facts for text summarization are Raw Text Extraction/ Summarization Methods, Sentiment Analysis, and Named Entity Recognition. He divided the summarization process into 2 phases 1) pre-processing and 2) processing. In the pre-processing phase, he determined the sentence boundary and identified and removed the stop words. In the processing phase Sentence ranking is done based on relevant terms having the highest frequency count and the top-weighted sentence is produced using linguistic rules to generate an extractive summary.

Namrata Kumari and Pardeep Singh, National Institute of Technology Hamirpur in their paper Hindi Text Summarization using Sequence to Sequence Neural Network proposed a text summarization for the Hindi language using the seqTOseq encoder-decoder mechanism. The model utilizes Word2Vec embedding with an embedding size of 200. Adam and RMSProp are the parameters used to optimize the parameters of the network. The learning rate was set to be 0.001 along with the number of epochs while training was 50.

Jayashree. R, Srikanta Murthy. K and Sunny. K, Department of Computer Science, PES Institute of Technology, in their paper document summarization in Kannada, offered a three-phased approach to developing summaries. 1)crawling 2) indexing 3) summarization. They later tested their model using a sample news story and discovered the need to eliminate the sentence's background noise. Thus, they created an algorithm, which takes a stop word as input and discovers structurally similar terms, and adds them to the stop word list. Moreover, look for words with comparable structures to those in the primary list. They only used previously classified data; effective classifiers can provide the required categorization.

Arpita Swamy, and Srinath S, in their paper Automated Kannada Text Summarization using Sentence Features, analyzed a single document's Kannada content that may be summarised using an extraction-based technique. Sentences are graded according to various factors, including term frequency, term frequency-inverse sentence frequency, keywords, sentence length, and location within the document.

The produced summaries are assessed using the ROUGE toolkit's assessment metrics of recall, accuracy, and f-score. The performance of this suggested system is good in terms of average recall, average accuracy, and average f-score values.

Sunitha. C, Dr. A. Jaya, and Amal Ganesh worked on Abstractive summarization techniques in Indian languages. The researchers have clarified the Abstractive summarization technique, classified into two approaches structure-based approach and semantic-based approach. Many numerous methodologies are applied to these approaches.

Dhanya P. M has done comparative research on text summaries in Tamil, Kannada, Odia, Bengali, Punjabi, and Gujarati are taken for purpose of comparison. Vishal Gupta works on an automated Punjabi text extraction summarization system.

Aakash Sinha, Abhishek Yadav, and Akshay Gahlot suggested a model which is built on a neural network that comprises one input layer, one hidden layer, and one output unit. The document is delivered to the input layer, mathematics are undertaken in the hidden units and an outcome is created just at the final layer. They employed ROUGE for all assessment reasons.

Ashok Urlana, Nirmal Surange, Pavan Baswani, Priyanka Ravva, and Manish Shrivastava, in their paper TeSum: Human-Generated Abstractive Summarization Corpus for Telugu designed a pipeline that crowd-sources summarization data and then aggressively filters the content via fully automated and incomplete expert evaluation. Using this pipeline, they developed a Telugu Abstractive Summarization dataset (TeSum), and it was verified using human assessment based on sampling. They also offer starting points for a variety of models that are frequently used for summary. A few newly published datasets for summarization scraped the online material based on the presumption that the publishers would provide the summary along with the article.

PROPOSED SYSTEM

In this project, we condensed a lengthy text into a section while keeping the most crucial lines to convey the text's core idea. To begin summarising the text, we tokenize the sentences where the NLP allows us. Now that the text document has been cleaned up, conjunctions, adverbs, and full stops have been eliminated. After deleting the stop words, count the frequency of each word in the text file that is left, and then eliminate all the low-frequency words. By using our technique, we arrive at the GSS coefficient, which is then utilized to rate the text. Then choose the most frequently occurring keywords. The selected sentences with the highest frequency of keywords are then summarised.

1) Sentence Tokenization

Tokenization is one of the first tasks in any NLP pipeline. Simply said, tokenization is the process of breaking down the primary text into tokens, which are distinct collections of words or phrases. If the text is broken up into words, it is termed "Word Tokenization," and if it is separated into sentences, it is referred to as "Sentence Tokenization." The letter "space" is commonly used for word tokenization, whereas "full stops," "exclamation marks," and "newlines" are generally used for sentence tokenization.

2) Text Cleaning

It is the process of removing unneeded or extraneous material from a corpus of texts. Eliminate any extraneous letters, punctuation, numerals, and symbols that don't contribute to the sense of the text. Text cleaning is a necessary pre-processing step for all-natural language processing (NLP) applications, such as text categorization, sentiment analysis, and language translation. By cleaning the text data, we may reduce noise, improve the data's quality, and simplify it for computers to understand and analyze the language.

3) Removing the stop words

They have regularly used words that are omitted from summary processes because they are less important in the conclusion of the document. In Kannada words like (and), (it), ಇದೆ (is), etc. are frequently used stop words in sentences.

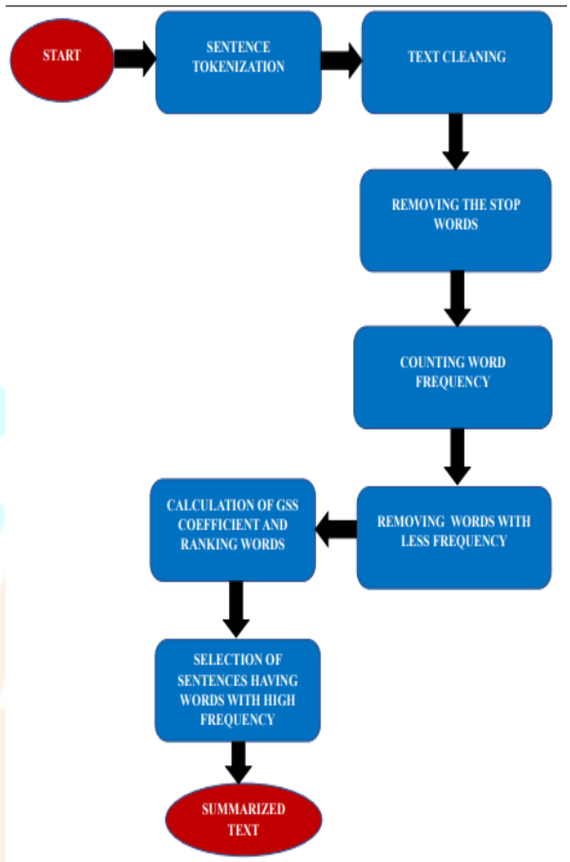

Figure 3.1 Block diagram of the summarization processes

4) Counting Word Frequency

The term frequency describes the number of times a phrase appears in a document. TF(t) = Wt / N is the formula for calculating the term frequency. Wt is the number of occasions a term, t appears in the document, and N is the entirety of the words in it.

5) Removing the words with less frequency

The words which have less TF score are removed.

6) Calculation of GSS coefficient and ranking words

GSS coefficients which evaluate a term's importance within a particular category are calculated. In our case, the GSS co-efficient, a feature selection method, determines the relevance metric. The definition of
f (u, v) is: f (u, v) = p(u, v) * p(u', v) * p(u, c'), where p(u, v) is the probability that a document contains the keyword u and falls under category v. P(u', v') represents the probability that an article does not include u and does not relate to v. P(u', v) represents the probability that a document belongs to v but excludes u. The likelihood that a document includes u but does not relate to v is expressed as p(u, v'). which in turn is used to rank the text through an algorithm

7) Selection of sentences

The keywords that have the highest frequency are selected. And also the sentences which have keywords with the highest frequency are summarized.

RESULT AND DISCUSSION

The model is tested for three categories of documents. Those three categories are Sports, Cinema, and Astrology. Evaluation is done using Recall-Oriented Understudy for Gisting Evaluation (ROUGE). Precision and F-Measure are two parameters that ROUGE takes into account to determine how similar two summaries are. Evaluation of summarization is difficult because the summarization perspective differs for each human being. So for all the categories, one human summary is used.

Tests Cases	Precision	F-Score
Sports	0.54	0.67
Cinema	0.64	0.78
Astrology	0.56	0.67

Table 4.1 ROUGE-1 Precision and F-Measure values

Tests Cases	Precision	F-Score
Sports	0.52	0.60
Cinema	0.59	0.72
Astrology	0.55	0.68

Table 4.2 ROUGE-2 Precision and F-Measure values

Table 4.1 depicts the scores of the ROUGE-1 evaluation which tests the overlaps of unigrams (words) between the system summary and the real summary. Whereas Table 4.2 is the ROUGE-2 evaluation result which examines the overlap of bigrams. In both the ROUGE evaluation, the values are closer to one. This indicates the high similarity between model-generated summaries and human-generated summaries.

CONCLUSION

It is more crucial than before to have an automatic summarization system that can lessen the human effort required and information overload due to the Internet's phenomenal surge in data. A good summary is intended to keep important lines that encapsulate the text's main ideas and minimize repetition to provide an information-rich summary. The relevance, coverage, and variety of each phrase are still not sufficiently captured by existing text summarization approaches and definitions of representative qualities. This proposed study presents a single document extractive summary method that makes use of Kannada text summarising methods. The recommended approach is scored-based.