International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031(Online)
www.ijlret.com‖ Volume 1 Issue 6‖November 2015 ‖ PP 08-13

Summarization of Malayalam Document Using Relevance of Sentences

Ajmal E.B
Department Of CSE
Ilahia College of Engineering And Technology
Muvattupuzha, India

Rosna P Haroon
Department Of CSE
Ilahia College of Engineering And Technology
Muvattupuzha, India

Abstract—Text summarization is an emerging technique for finding out the summary of the text document. Text summarization is nothing but summarizing the content of given text document. Text summarization has got so uses such as Due to the massive amount of information getting increased on internet; it is difficult for the user to go through all the information available on web. Summarization techniques need to be used to reduce the users time in reading the whole information available on web. In this paper propose a Malayalam text summarization system which is based on MMR technique with successive threshold. Here the sentences are selected based on the concept of maximal marginal relevance. The key idea is to use a unit step function at each step to decide the maximum marginal relevance and the number of sentences present in the summary would be equal to the number of paragraphs or the average number of sentences present in the text document, which can be achieved by using successive threshold approach. We apply MMR approach on Malayalam text summarization task and achieve comparable results to the state of the art.
Keywords- Maximum Marginal Relevance, Successive Threshold, Unit step function.

I. INTRODUCTION

Text summarization has become an important and timely tool for assisting and interpreting text information in todays fast-growing information age. It is very difficult for human beings to manually summarize large documents of text. There is an abundance of text material available on the internet. However, usually the Internet provides more information than is needed. Therefore, a twofold problem is encountered: searching for relevant documents through an overwhelming number of documents available, and absorbing a large quantity of relevant information. The goal of automatic text summarization is condensing the source text into a shorter version preserving its information content and overall meaning.

A summary can be employed in an indicative way as a pointer to some parts of the original document, or in an informative way to cover all relevant information of the text. In both cases the most important advantage of using a summary is its reduced reading time.

A good summary system should reflect the diverse topics of the document while keeping redundancy to a minimum. Summarization tools may also search for headings and other markers of subtopics in order to identify the key points of a document. Microsoft Words AutoSummarize function is a simple example of text summarization.

Text Summarization methods can be classified into extractive and abstractive 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.

An Abstractive summarization attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document.

This paper focuses on extractive text summarization of Malayalam documents. Even though a lot of text summarization systems have been developed for summarizing documents in various languages, there is no such well performing system for Malayalam. The existing systems having high computational cost and time and

also the storage capacity.To address the issues of computational cost time and storage capacity, here proposes a text summarization system that works on the concept of maximal marginal relevance between the sentences or the words. The key idea is to use a unit step function at each step to decide the maximum marginal relevance and the number of sentences present in the summary would be equal to the number of paragraphs or the average number of sentences present in the input text document, which can be achieved by using successive threshold approach.

Malayalam is the official language of Kerala and there are around 33 million people who speak Malayalam. There is a vast amount of online data available in Malayalam and more than 30,000 articles are available in Malayalam Wikipedia. This warrants us to develop tools that can be used to explore digital information presented in Malayalam and other native languages. In this paper, we propose the MMR based Malayalam Text Summarization with Successive Thresholds.

Malayalam Text Summarization with Successive Thresholds. This paper is organized in five sections.Section II reviews the related works. Section III describes proposed scheme. In section IV,we highlight the evaluation of the proposed scheme. Conclusion and future work is described in section V.

II. RELATED WORK

Attempts to automatically summarize documents started early since 1958. The method based on word frequencies by Luhn is one of the oldest but still relevant method. This method measures the importance of a sentence based on the presence of keywords (most frequently occurring words in a document otherthan the stopwords) in the sentence. Text summarization method by Ed-mundson used cue words, title words, and sentence location for determining the sentence weights [4]. Text summarization for Malayalam documents by Rajina Kabeer and Sumam Mary Idicula [1].

Graph theoretical approaches to summarization represents a document as an undirected graph, in which the nodes represent the sentences in the document. Two nodes in the graph are are connected if the cosine similarity of the sentences corresponding to the nodes is above some particular threshold. The sentences corresponding to the nodes with the highest cardinality or in other words the sentences which are more similar to other sentences in the document are considered important and are included in the summary [8]. Methods based on Co-reference chains and Lexical chains are based on the semantic structure of the document.

Semantic graph based approaches extracts semantic triplets (Subject-Object-Predicate triplets) from each of the sentence in the document. These triplets are used to generate a semantic graph of the document. A sub-graph of the semantic graph is selected using machine learning techniques and the sentences corresponding to the sub-graph are extracted into the summary [2].

the summarization process as a classification problem Le. To classify the sentences in a document as summary sentences and non-summary sentences based on the features that they possess [4]. Nave Bayes method, Neural networks and Hidden Markov Model (HMM) are some of the machine learning approaches used for text summarization.

Information extraction by abstractive text summarization for Telugu language by Jagadish S Kallimani, Srinivasa KG and EswaraReddy B [7], Tamil document summarization using semantic graph method by Banu M, Karthika C, Sudarmani P and Geetha T.V [14], Text extraction for an Agglutinative Language by Sankar K, VijaySundar Ram R and Sobha Lalitha Devi which was used for summarizing Tamil documents [8], Keyword extraction based summarization of categorized Kannada text documents by Jayashree.R, Srikanta Murthy.K and Sunny.K [12] and Bengalitext summarization by sentence extraction by Kamal Sarkar [6] are some text summarization works done for Indian languages.

III. PROPOSED SCHEME

In the proposed method, a single-document input is summarized based on the concept of maximal marginal relevance between the sentences or the words. The key idea is to use a unit step function at each step to decide the maximum marginal relevance and each word meaning is calculated with the help of a dictionary, finally the number of sentences present in the summary would be equal to the number of paragraphs or the average number of sentences present in the input text document, which can be achieved by using successive threshold approach.

A. Maximal Marginal Relevance

The key idea in this technique is to use a unit step function at each step to decide the maximum marginal relevance. The automatic summarization process may contain following steps.

1.Input a document to be summarized.
2.Now the document is traversed and eliminates the words that are not useful (stop word removal).
3.Starting with the starting position of the sentence until the document finishes.
4.Identify the most important word/sentence (by meaning) with the help of a Malayalam dictionary.
5.Using the unit step function we can calculate the relevant information required.

The unit step function used in the algorithm is given as uk+1=arg max (sim1(ui,Q)-max(sim2(ui,uj)))
Where
Q: the user input document
ui: the most important word/sentence
uj: the remaining sentences in the document
U: the selected list of sentences.
6. The process may be terminated once an appropriate number of words or sentences is in U. Which can be achieved by using successive threshold approach.

B. Successive Threshold Approach

The concept behind this approach is that the number of sentences present in the summary would be equal to the number of paragraphs or the average number of sentences present in the input text document. That is initially count the total number of paragraphs and sentences in the given text document, if the total number of paragraphs in the input text document is meet a threshold value then take the value of n as number of paragraphs otherwise take n as average number of sentences in the input document. After applying all the preprocessing steps and the MMR technique to select the relevant information or the sentences from the document. Then count the total number of sentences say it is m, if m is equal to n, then these are the sentences finally included in the summary. Else, repeat the steps of MMR technique until the m value will be equal to n.
As an example consider the following input text shown in figure 1 and the corresponding output obtained for the text using proposed method is shown in figure 2.

കൊല്ലം: ജനസമ്പർക്ക പരിപാടിക്കെതിരായ വിമർശനങ്ങൾ കേട്ട് ഇതിൽ നിന്ന് പിന്മാറില്ലെന്ന് മുഖ്യമന്ത്രി ഉമ്മൻ ചാണ്ടി. സാധാരണക്കാരുടെ പ്രശ്നങ്ങൾ പരിഹരിക്കാനുള്ള ശ്രമമാണ് നടത്തുന്നത്. വളരെ ന്യായമായ കാര്യങ്ങളിൽ വേഗത്തിൽ തീരുമാനങ്ങൾ എടുക്കാൻ കഴിയുന്നതാണ്. എന്തൊക്കെ വിമർശനം ഉണ്ടായാലും ജനങ്ങൾക്ക് വേണ്ടി ജനപക്ഷത്ത് നിന്ന് അവരുടെ പ്രശ്നങ്ങൾ പരിഹരിക്കും. കൊല്ലത്ത് ജനസമ്പർക്ക പരിപാടി ഉദ്ഘാടനം ചെയ്ത് സംസാരിക്കുമ്പോഴാണ് അദ്ദേഹം ഇക്കാര്യങ്ങൾ പറഞ്ഞത്. ശാസ്താംകോട്ട കായലിന്റെ സംരക്ഷണത്തിന് ആദ്യം ചെയ്യേണ്ടത് ജില്ലയിലേക്ക് വെള്ളം കൊടുക്കാൻ മറ്റൊരു ശ്രോതസ് കണ്ടെത്തി അത് സജ്ജമാക്കുകയാണ്. കല്ലടയാറിൽ, കടപുഴയിൽ ബണ്ട് കെട്ടി, അവിടുത്തെ വെള്ളം ജില്ലയിലേക്ക് വിതരണം ചെയ്യാൻ 19 കോടി രൂപയുടെ പദ്ധതി തയ്യാറാക്കി ധനകാര്യ വകുപ്പിലേക്ക് അനുമതിക്കു അയച്ചിട്ടുണ്ടെന്നും അദ്ദേഹം അറിയിച്ചു. ആലപ്പാട് പാക്കേജിൽ പെടുത്തി സുനാമി ദുരിതാശ്വാസ പ്രവർത്തനങ്ങളുടെ ഭാഗമായി നിർമ്മിച്ച വീടുകളുടെ അറ്റകുറ്റ പണികൾക്കും, കുടിവെള്ളം, സീവേജ് തുടങ്ങിയ സൗകര്യങ്ങൾക്കും വേണ്ടി 10 കോടി രൂപ അനുവദിച്ചു. കൊല്ലം കരുനാഗപ്പള്ളി ഭാഗത്ത് 2000 കുടുംബങ്ങൾക്ക് വേണ്ടി നിർമ്മിച്ച ഫ്ലാറ്റുകളിലെ സീവേജ്് സൗകര്യം ഒരുക്കുവാനുള്ള 7 കോടി രൂപയുടെ പദ്ധതി തയ്യാറാക്കിയത് മാസങ്ങൾക്കുള്ളിൽ നടപ്പിലാക്കും. ഇവിടത്തെ കടൽ തീരം സംരക്ഷിക്കുന്നതിനു വേണ്ടിയുള്ള 11 കോടി രൂപയുടെ പദ്ധതി പൊതു മേഖല സ്ഥാപനങ്ങളായ കഞ്ഞ്ളഘ, ഗങ്ങഘ ചേർന്ന് വഹിക്കും.

അഷ്ടമുടി കായലും, തങ്കശ്ശേരി കടലോരവും, തെന്മലയും ചേർത്ത് ഒരു ടൂറിസം സർക്യൂട്ട് രൂപീകരിക്കും. ഇവിടെ 5 കോടി രൂപ ചെലവു വരുന്ന ഒരു വാട്ടർ സ്പോർട്ട് പദ്ധതിയും തുടങ്ങും. കൊല്ലത്തിന്റെ വളരെ കാലമായുള്ള ആവശ്യമാണ് ഒരു കോടതി സമുച്ചയം. അതിന് പണം അനുവദിച്ചു. എവിടെ സ്ഥാപിക്കണം എന്ന കാര്യത്തിൽ മാത്രമാണ് ഇനി തീരുമാനം ആവാനുള്ളത്. കൊല്ലത്തിന്റെ ചുമതലയുള്ള മന്ത്രി ഷിബു ബേബി ജോൺ കളക്ടറുമായി കൂടിയാലോചിച്ച് ഒരു മാസത്തിനകം തീരുമാനം എടുക്കും. കൊട്ടാരക്കരയിൽ കേന്ദ്രീയ വിദ്യാലയം സ്ഥാപിക്കുവാനായി 5 ഏക്കർ ഭൂമി അനുവദിക്കുമെന്നും അദ്ദേഹം പറഞ്ഞു. ഈ ഘട്ടത്തിലെ അഞ്ചാമത്തെ ജനസമ്പർക്ക പരിപാടിയാണ് കൊല്ലത്ത് നടക്കുന്നത്. ഇതു വരെയുള്ള ജില്ലകളിൽ നിന്നുയർന്നു വന്ന ഒരു പ്രശ്നം ഹിമോഫീലിയ രോഗികൾക്ക് കാരുണ്യ ഫണ്ടിൽ നിന്നും അനുവദിച്ചിട്ടുള്ള രണ്ടു ലക്ഷം രൂപ മതിയാകുന്നില്ല എന്നാണ്. ഹിമോഫീലിയ രോഗികൾക്ക് ആജീവനാന്തം മരുന്ന് കഴിക്കേണ്ടതാണ്, അവരുടെ ആവശ്യം തികച്ചും ന്യായമാണ്. ഈ പരിപാടിക്കിടയിൽ തന്നെ ഹിമോഫീലിയ രോഗികൾക്ക് അനുവദിക്കേണ്ട തുകയുടെ പരിധി ഉയർത്താൻ വേണ്ടി നിയമ ഭേദഗതി വരുത്തി, അവർക്കുള്ള മരുന്നുകൾ ആജീവനാന്തം സൗജന്യമായി കൊടുക്കുവാനുള്ള തീരുമാനം എടുത്തിട്ടുണ്ടെന്നും മുഖ്യമന്ത്രി പറഞ്ഞു

Figure 1. Input Text Document

കല്ലടയാറിൽ, കടപുഴയിൽ ബണ്ട് കെട്ടി, അവിടുത്തെ വെള്ളം ജില്ലയിലേക്ക് വിതരണം ചെയ്യാൻ 19 കോടി രൂപയുടെ പദ്ധതി തയ്യാറാക്കി ധനകാര്യ വകുപ്പിലേക്ക് അനുമതിക്കു അയച്ചിട്ടുണ്ടെന്നും അദ്ദേഹം അറിയിച്ചു. കൊല്ലം കരുനാഗപ്പള്ളി ഭാഗത്ത് 2000 കുടംബങ്ങൾക്ക് വേണ്ടി നിർമ്മിച്ച ഫ്ലൂറ്റുകളിലെ സ്വീവേജ്് സൗകര്യം ഒരുക്കുവാനുള്ള 7 കോടി രൂപയുടെ പദ്ധതി തയ്യാറാക്കിയത് മാസങ്ങൾക്കുള്ളിൽ നടപ്പിലാക്കും. ഇവിടത്തെ കടൽ തീരം സംരക്ഷിക്കുന്നതിനു വേണ്ടിയുള്ള 11 കോടി രൂപയുടെ പദ്ധതി പൊതു മേഖല സ്ഥാപനങ്ങളായ കഞ്ഞേഘ, ഗങ്ങഘ ചേർന്ന് വഹിക്കും. ആലപ്പാട് പാക്കേജിൽ പെടുത്തി സുനാമി ദുരിതാശ്വാസ പ്രവർത്തനങ്ങളുടെ ഭാഗമായി നിർമ്മിച്ച വീടുകളുടെ അറ്റകുറ്റ പണികൾക്കും, കുടിവെള്ളം, സ്വീവേജ് തുടങ്ങിയ സൗകര്യങ്ങൾക്കും വേണ്ടി 10 കോടി രൂപ അനുവദിച്ചു.

Figure 2. Output Summary

IV. EVALUATION

As shown in the below table is the different parameter evaluation of the existing method. The method is implemented on 6 different dataset of different sizes and various parameters such as precision, recall and F-measure is calculated.

Table I
PARAMETER EVALUATION EXISTING METHOD

Dataset	Precision	Recall	F- measure
Dataset1	0.485	0.525	0.530
Dataset2	0.5309	0.5765	0.5665
Dataset3	0.5807	0.6635	0.5978
Dataset4	0.6679	0.7814	0.7449
Dataset5	0.656	0.7756	0.7645
Dataset6	0.772	0.7901	0.7801

Table II shows parameter evaluation of the Proposed method. The method is implemented on 6 different dataset of different sizes and various parameters such as precision, recall and F-measure is calculated.

Table II
PARAMETER EVALUATION PROPOSED METHOD

Dataset	Precision	Recall	F- measure
Dataset1	0.535	0.565	0.543
Dataset2	0.5407	0.5805	0.5785
Dataset3	0.5917	0.6743	0.6537
Dataset4	0.6779	0.7896	0.7549
Dataset5	0.673	0.7826	0.7775
Dataset6	0.852	0.8910	0.8018

V. CONCLUSION AND FUTURE WORK

The text summarization provides the summary of the text document. Here an efficient technique of text summarization is proposed. The proposed method is works on the concept of maximal marginal relevance between the sentences or the words. The key idea is to use a unit step function at each step to decide the maximum marginal relevance, and the number of sentences present in the summary would be equal to the number of paragraphs or the average number of sentences in the input text document, which can be achieved by using successive threshold approach. Analysis shows that proposed method is more accurate. More quality parameters are generated by incorporate another methods is future work.

ACKNOWLEDGMENT

The authors would like to thank HOD, Prof. Rosna P Haroon, Department of Computer Science and Engineering, Ilahia College of Engineering And Technology for her moral and technical support.

REFERENCES

[1]. Rajina Kabeer, SumamMary Idicula, Text Summarization for Malayalam Document- an Experience, International Conference on Data Science and Engineering(ICDSE), 2014.
[2]. Jurij Leskovec, Natasa Milic-Frayling,Marko Grobelnik, Extracting Summary Sentences Based on the Document Semantic Graph, Microsoft Research,2007.
[3]. Elena Lloret, Text Summarization : An Overview,University of Ali-cante,2007.
[4]. Vishal Gupta, Gurpreet Singh Lehal, A Survey of Text Summarization Extractive Techniques, Journal of Emerging Technologies in Web Intelligence, 2010.
[5]. Dipanjan Das, Andre F.T. Martins, A Survey on Automatic Text Summarization, Language Technologies Institute, Carnegie Mellon Univer-sity,2007.
[6]. Kamal Sarkar, Bengali Text Summarization by Sentence Extraction, Jadavpur University.
[7]. Jagadish S Kallimani, Srinivasa KG, Eswara Reddy B" Information Extraction by an Abstractive Text Summarization for an Indian Rgional Language Natural Language Processing and Knowledge Engineering (NLP-KE), 2011 7th International Conference, Nov 2011.
[8]. Sankar K, VijaySundar Ram R and Sobha Lalitha Devi ,Text Extraction for an Agglutinative Language ,May 2011.
[9]. Martin Hassel, Evaluationof AutomaticText Summarization: A practical implementation, 2004.

[10]. Bindu.M.S, Sumam Mary Idicula, A Hybrid Model For Phrase Chunk-ing Employing Artificial Immunity System And Rule Based Methods International Journal of Artificial Intelligence Applications(IJAIA), Vol.2, No.4,October2011.
[11]. Rajeev RR, RajendranN, Elizabeth Sherly, A Suffix Stripping Based Morph Analyser For Malayalam Language, Proceedings of 20th Kerala Science Congress, 2005.
[12]. Jayashree.R, SrikantaMurthy.K,Sunny.K, Keyword extraction based summarizationof categorised Kannada text documents International Journal on Soft Computing (IJSC) Vol.2, No.4,November2011
[13]. Aysun Guran, Eren Bekar, Selim Akyokus, A Comparison of Feature and Semantic-Based Summarization Algorithms for Turkish International Symposium on Innovations in Intelligent Systems and Applications, Kayseri Cappadocia,TURKEY,21- 24 June 2010.
[14]. BanuM,Karthika C, Sudarmani P and Geetha T. V, Tamil Document Summarization Using Semantic Graph Method International Conference on Computational Intelligence and Multimedia Applications, 2007.