An Extractive Multi-document Summarization System for Malayalam News Documents

Manju K[1], David Peter S[2], and Sumam Mary idicula[3]
{manju@mec.ac.in[1], davidpeter123@gmail.com[2] , sumam@cusat.ac.in[3]}

College of Engineering,Cherthala, Alappuzha, Kerala, India[1], Cochin University of Science and Technology, Kochi, India[2],Cochin University of Science and Technology, Kochi, India[3]

Abstract. The flooding of digital data necessitates the need for a system that can take information from multiple documents and provide it in a summarized form. Due to the unavailability of automatic tool for summarizing Malayalam documents, this work serves as an introduction. In this work, we have investigated on an extractive multi document summarizer for Malayalam language which uses a sentence scoring technique. An online Malayalam Wordnet is used in the work for semantic similarity checking. Sentence score is calculated based on the features selected for each sentence. Feature selection is done by considering the heuristic measures like sentence length, sentence position, presence of numerical data, existence of proper noun in a sentence, term frequency-inverse document frequency in the documents. Top ranking sentences are selected as initial summary. Then cosine similarity measure is applied to remove redundancies and the summary is generated as per the length specified. Experimental results demonstrates the effectiveness of the proposed system on the data set selected as bench mark.

Keywords: Multidocument Summarization, Malayalam Language, Sentence Scoring, Extractive, Heuristic measures, Word Net.

1 Introduction

With the enormous growth of online information, it has become humanly in-feasible to efficiently separate useful information from such a huge mass of data. This necessitates the need to develop tools that can process and extract relevant information. One solution to this information overload problem is offered by using efficient text summarization techniques. Text summarization is a method that aims to generate a condensed version of one or more textual documents by extracting the most significant content from it. In this age of Internet, text summarization has to play an important role as it can be used to get summary of related contents from different links. When considering newspaper websites the news related to the same incident will be published differently. Multi document summarization system helps to summarize these articles to get an essence of the incident.

Text summarization method can be classified into Extractive and Abstractive summarization. In extractive summarization, summary is generated by choosing significant sentences from the original document while in abstractive summarization, summary is generated by formulating new sentences according to the documents. Depending on the number of documents simultaneously analyzed, text summarization is classified as single and multidocument summarization. Multidocument summarization can either be generic or query dependent. Generic summarization system extracts main ideas from the text collection whereas query dependent summaizattion system selects sentences with respect to the query given by the user.

This paper uses a sentence scoring method along with Wordnet for generating the extractive summary for multiple Malayalam newspaper articles which are similar in topic. Even though there are several methods previously proposed for English and other foreign languages, there is no complete system for Indian Languages especially Malayalam. The morphological richness and agglutinative nature of Malayalam language accounts for the very few attempts made to summarize Malayalam documents. Malayalam is one among the 22 scheduled languages of India. It is the official language in the state of Kerala and in the Union territories of Lakshadweep and Puduchery. Malayalam belongs to the Dravidian language family and is spoken by approximately 33 million people. Since a vast amount of online information related to different topics are available in Malayalam, it is difficult for the users to find the desired information quickly. Following were the challenges faced while processing Malayalam Language:

- No upper or lower case for Malayalam letters like English.
- The same word can appear with inflectional and morphological variations in sentences.
- Same concept expressed using synonyms in different sentences.
- Unavailability of a freely and publicly available corpora.

The proposed method addresses these issues while generating the extractive summary. We have followed a simple and effective method for scoring sentences which does not require a training phase nor a deeper semantic analysis of the sentence. Wordnet is an online lexical reference system in which Malayalam nouns, verbs, adjectives and adverbs are organized into synsets. In this system wordnet is used to obtain semantically related words.

The paper is organized as follows: section 2 discusses the related work in the area of extractive summarization which takes sentence specific features; section 3 discusses the overview of the proposed system; section 4 discusses the results and discussion and section 5 concludes the paper.

2 Related Work

In this section we will have a study on the work done in the area of extractive text summarization. The very first work on summarazation was by LUHN[1] in 1958 which was based on frequency of words in a document. The sentences that contain those frequent words were important than other sentences in the document and were chosen as part of the summary. In 1958 Baxendal[2] took sentence location as a scoring criterion along with word frequency to calculate the sentence score. H.P Edmundson[3] in 1969 included two more features title word and cue words for determining the sentence weight. In 2001 MEAD[4] a centroid based summarization model was introduced where all documents were modeled as bag of words. Nobata.et.al[5] in 2001 used sentence location, sentence length, TF/IDF, headline and query as score functions to extract significant sentences. Vasudeva Varma.et.al[6] in 2005 considered sentence level features and word level features for scoring the sentences.Abuobieda.et.al[7] in 2012 developed a pseudo genetic based model for text summarization.They used title feature, sentence length, numerical data and thematic words for scoring sentences. Rafael et.al[8] analysed the scoring features using ROUGUE evaluation matrix. Mendoza et.al[9]addresses the summarization problem as a binary optimization problem and used sentence position, sentence length, title similarity, cohesion and coverage as target function for sentence scoring. The ideas obtained from these works have been the source of motivation and the inputs gathered from the related methodologies facilitated in designing the layout for the proposed summarization system for Malayalam.

3 System overview

The overall architecture of the proposed system is shown in Fig:1. The summarizing system takes multiple Malayalam documents coming under a single topic as input. The input documents are in text file format. These files are subjected to preprocessing where the stopwords get eliminated and then stemming is performed on the sentences. Now the sentences are scored and top ranking sentences are taken as the significant sentences for the summary. The system assigns score to each sentence based on a set of features. Features include the sentence position, sentence length, number of numeric data in the sentences, number of proper nouns in each sentence, and the term frequency inverse document frequency for each sentence. Word net is used to perform the semantic similarity checking which affects the metric for the score of each sentence.

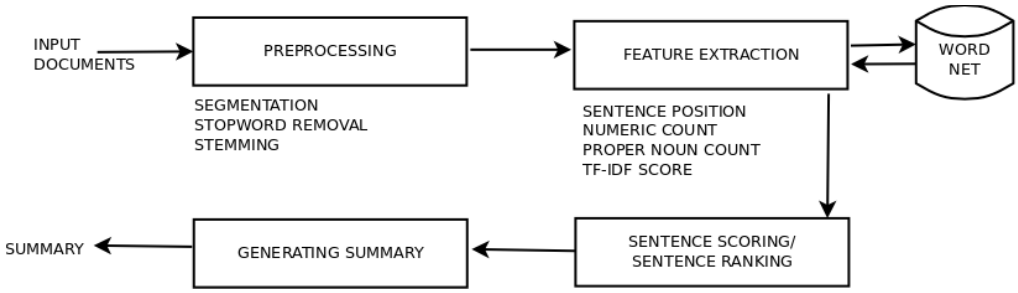


Fig.1: Process flow of the generic multidocument summarizer.

3.1 Pre-processing phase

Following steps are performed in the pre-processing phase.

- The segmentizer module of the system breaks each document into sentences.
- Remove stop words like *'athe','avan','ithu', etc.* which does not contribute to the understanding of the main idea presented in the text.
- The words in each sentence gets converted into its root form which is literally called stemming.

3.2 Sentence Scoring

Scoring of sentences to extract most relevant sentences from the input document is done by taking the weighted average of the features identified. Features like sentence position, sentence length, number of numeric data, number of proper nouns in each sentence are extracted from the segmentized text prior to stopword removal and stemming. The TF-IDF feature is extracted from the preprocessed text.

3.2.1 Sentence Position

Sentence Position is the position of a sentence in a document and the value is normalized to a scale of 0 and 1. It is calculated as per the equation,

$$PositionF = (maxpos - curpos + 1) / maxpos \quad (1)$$

where *maxpos* is the maximum number of sentences in the document and *curpos* is the position of the sentence in the document.

3.2.2 Sentence length

The Sentence length feature is defined as

$$SentlengthF = N * Len(S_{i,k}) / Len(Doc_k) \quad (2)$$

where $Len(S_{i,k})$ is the length of sentence *i* in k[th] document and N the total sentences in Doc_k

3.2.3 Numeric value as a feature

The sentence containing numerical data is relevant as it indicates event related attributes like time of occurrence, population, death toll, statistical data, etc., and is most probably included in the summary. The score is calculated as the ratio of number of numerical data in the sentence to length of sentence.

$$NumF = \frac{No.\ of\ numerical\ data \in S_i}{Length\ of\ sentence\ S_i} \quad (3)$$

3.2.4 Proper Noun Count

To obtain the proper noun count in sentence, the sentences are tagged using the POS tagger[11]. Here the tagger takes tokenized text and outputs parts-of-speech tagged text. From the tagged text we can obtain the count of proper nouns in each sentence.

$$Proper\ NounF = \frac{No.\ of\ propernouns \in S_i}{Length\ of\ sentence\ S_i} \quad (4)$$

3.2.5 Tf-Idf Score

The goodness of a sentence is usually represented by the importance of the words present in it. TF-IDF is a simple but powerful heuristic for ranking the sentence according to their importance. A Vector Space model is built at the sentence level by grouping all the sentences of the documents. Now for scoring the sentences, we determine the TF-IDF of each sentence in a document. The Tf-Idf calculation is done on the preprocessed text.

$$Tt - Idf\left(S_i\right) = Tf_{(t,i)} * Idf_t \quad (5)$$

where is the number of times the term *t* occurs in the sentence and gives the information about the number of sentences in which the term *t* appears.

$$Idf_t = \log\left(\frac{N}{N_t}\right) \quad (6)$$

where *N* is the total sentences in a document D and is the number of sentences in a document D in which the term *t* occurs. Taking the sum of TF-IDF of each term *t* in the sentence, we get the TF-IDF score of each sentence in the document. Since longer sentences will be having more no.of terms, we apply L2 Normalization to get the resultant score.

The terms identified for the vector space model are called the keywords of the vocabulary. In order to improve the quality of keyword selection, Parts Of Speech(POS) tagging is considered. POS tagging is the process of annotating the terms in the text with its parts of speech based on its definition and the context in which the term is used. The terms which belong to the Noun category and Verb Category are retained in the vocabulary. Before finalizing the vocabulary, it is checked for the existence of synonyms. This is done by comparing the terms with the Word Net for malayalam[10]. Word Net is an online lexical reference in which Malayalam nouns, verbs, adjectives and adverbs are arranged into synonyms set or synset, each representing one underlying concept. A synset is a set of synonyms, and two words are said to be synonymous if their mutual substitution does not change the meaning of a sentence in the given context. The system interacts with the Word Net to get the synset-id of each term. The words with common synset-id are conceptually similar and only one word from this set is retained in the vocabulary. Making use of word net while constructing the term frequency matrix, improves the chance of a sentence in the summary.

3.3 Summary Generation

The sentence score is calculated by taking the linear weighted combination of all features. The overall score of a sentence *S* based on the features will be,

$$Score\left(S\right) = \sum_{i=1}^{n} w_i * F_i \quad (7)$$

Now we have a key,value pair consisting of sentences and its corresponding scores from all documents. Before ranking the sentences we performs the redundancy elimination using *cosine-similarity* measure. The similarity between two sentences , is calculated as

$$Simi(S_i, S_j) = \frac{S_i * S_j}{\sqrt{S_i^2} * \sqrt{S_j^2}} \quad (8)$$

+The similarity score will be between the values 0 and 1. 0 denotes the sentences are dissimilar and 1 denotes the sentences are similar. Taking T as the threshold, the dissimilar sentences and their corresponding scores are selected. Now the sentences are sorted in descending order based on their score. According to the compression ratio, the summary length is found. Now the summary is generated till the summary length is reached.

4 Experimental Results and Discussion

Text summarization in Malayalam Language is in its infancy, no standard data set is available. To test the performance of the system, three types of document sets coming under different domains, each with two articles that are related taken from prominent news paper websites were used. The articles extracted were saved as text files in UTF-8 format. The human generated summary for each document set was used as the reference summary for evaluation. We are following an intrinsic evaluation scheme by comparing the system generated summary with a reference summary. If the compression ratio is 70%, the summary length will be 30% of the length of the largest file. The summarization system selects representative sentences from these input documents to form an extractive summary. The common information retrieval metrics, precision and recall are used to evaluate the new summary.

4.1 Precision, Recall, F-Measure

Precision and Recall is determined on the basis of the system generated summary and the reference summary(human summary).

Recall is the fraction of sentences chosen by the person, that were also correctly identified by the system.

$$Recall = \frac{system : human\,choice\,overlap}{sentences\,chosen\,by\,human}$$

Precision is the fraction of system sentences that were correct.

$$Precision = \frac{system : human\,choice\,overlap}{sentences\,chosen\,by\,system}$$

F-measure is defined as the composite measure of Precision and Recall.

$$F_1 Score = \frac{2 * P * R}{P + R}$$

The system was tested and analyzed on the data set selected with different compression ratio. As the sentences are scored based on sentence related features, the increase in number of proper-nouns in a sentence can make the sentence to be included in the summary, even though it has not much relevance. Incorporating a word net check before finalizing the keywords improves the term frequency score, there by the TF-IDF score which makes the sentences significant in the summary. As the number of sentences in the system generated and ideal summary is same the precision and recall measure values will be the same. The reference summary is human generated depending on person to person there will be change and this can affect the performance of the system. If the input documents are of reasonable length then the system gives a comparable result for all compression ratio.

Table.1: Performance analysis measured using F-Score

Data Set	Evaluator 1			Evaluator 2		
	Compression Ratio			Compression Ratio		
	70%	50%	30%	70%	50%	30%
SET1	0.5	0.39	0.5	0.37	0.54	0.5
SET2	0.33	0.42	0.46	0.35	0.36	0.45
SET3	0.63	0.5	0.75	0.5	0.45	0.38

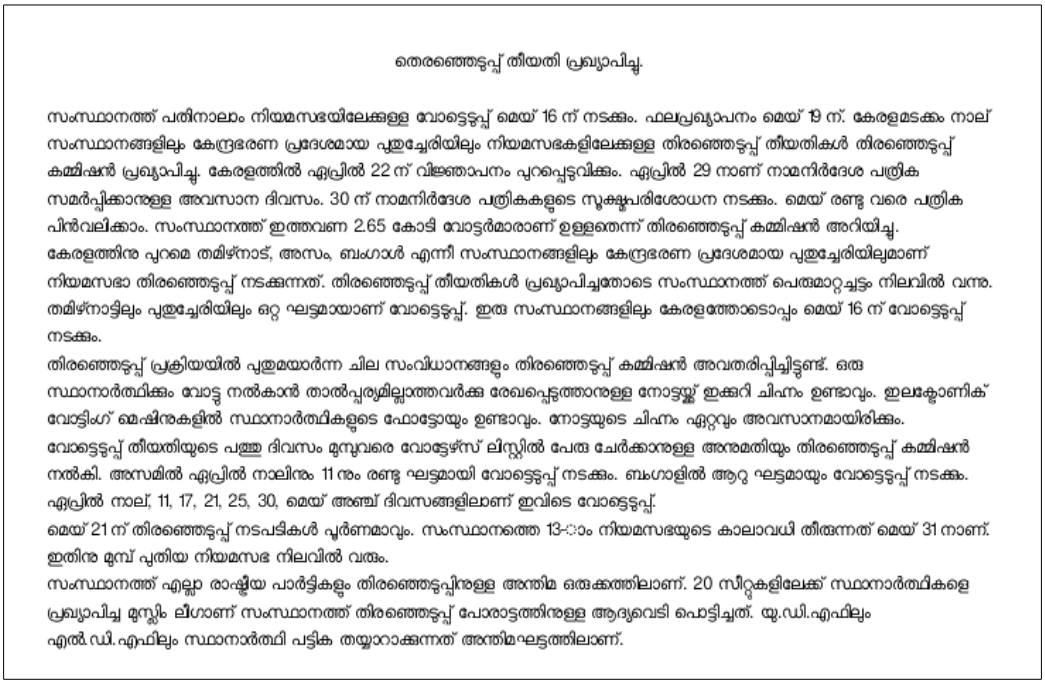
4.2 Example

The system takes two related articles as input as in Fig.2 and Fig.3. With 50% compression ratio the summary generated by the system is as in Fig:4.

തെരഞ്ഞെടുപ്പ് തീയതി പ്രഖ്യാപിച്ചു

കേരളം അടക്കം അഞ്ച് സംസ്ഥാനങ്ങളിലെ തെരഞ്ഞെടുപ്പ് തീയതി പ്രഖ്യാപിച്ചു. കേരളത്തിൽ മേയ് 16 നാണ് വോട്ടെടുപ്പ്. വോട്ടെണ്ണൽ 19 ന്. തെരഞ്ഞെടുപ്പ് വിജ്ഞാപനം ഏപ്രിൽ 22. നാമനിർദേശപത്രിക സമർപ്പണം 29 വരെ. സൂക്ഷ്മപരിശോധന ഏപ്രിൽ 30 ഉം പിൻവലിക്കാനുള്ള അവസാന തീയതി മേയ് രണ്ടും ആണ്. തമിഴ്നാട്ടിലും പുതുച്ചേരിയിലും ഒറ്റ ഘട്ടമായി മേയ് 16 നും അസമിൽ രണ്ട് ഘട്ടങ്ങളായും (ഏപ്രിൽ 4, ഏപ്രിൽ 11), പശ്ചിമ ബംഗാളിൽ ആറ് ഘട്ടങ്ങളായും (ഏപ്രിൽ 4, ഏപ്രിൽ 11, ഏപ്രിൽ 17, ഏപ്രിൽ 21, ഏപ്രിൽ 30, മേയ് 5) വോട്ടെടുപ്പ് നടക്കും. അഞ്ചിടത്തെ വോട്ടെണ്ണൽ 19 ന് ഒരുമിച്ച് നടക്കുമെന്നും മുഖ്യ തെരഞ്ഞെടുപ്പ് കമീഷണർ നസീം സെയ്ദി വാർത്താസമ്മേളനത്തിൽ അറിയിച്ചു. കേരളത്തിൽ 2.56 കോടി വോട്ടർമാരുണ്ട്. അഞ്ച് സംസ്ഥാനങ്ങളിലായി ആകെ 17 കോടി വോട്ടർമാർ. ഇത് ആദ്യമായി നിഷേധ വോട്ടിന് പ്രത്യേക ചിഹ്നവും സ്ഥാനാർഥികളുടെ ചിത്രവും വോട്ടിങ് മെഷീനിൽ ഏർപ്പെടുത്തിയിട്ടുണ്ട്. വോട്ടർമാരുടെ ചിത്രം പതിച്ച സ്ലിപുകൾ ഇലക്ഷോറൽ ഓഫീസർമാർ വീടുകളിൽ വിതരണം ചെയ്യും. ഭിന്നശേഷിയുള്ളവർക്ക് വോട്ട് ചെയ്യാൻ പ്രത്യേക സൗകര്യം ലഭ്യമാക്കും. വോട്ടെടുപ്പിനായി കേരളത്തിൽ 21,000 പോളിങ് സ്റ്റേഷനുകൾ ഒരുക്കും. ഡൽഹിയിൽ മുഖ്യ തെരഞ്ഞെടുപ്പ് കമീഷണറുടെ അധ്യക്ഷതയിൽ ചേർന്ന സന്പൂർണ യോഗമാണ് തെരഞ്ഞെടുപ്പ് തീയതി സംബന്ധിച്ച തീരുമാനമെടുത്തത്. ഇന്നു മുതൽ മാതൃകാ പെരുമാറ്റച്ചട്ടം നിലവിൽ വന്നതായി മുഖ്യ തെരഞ്ഞെടുപ്പ് കമീഷണർ അറിയിച്ചു. സാമൂഹ്യ വിരുദ്ധ പ്രവർത്തനങ്ങൾ തടയുന്നതിന് ഫലപ്രദമായ നടപടികൾ സ്വീകരിക്കും. അഞ്ച് സംസ്ഥാനങ്ങളിലെ എല്ലാ ജില്ലകളിലും അഞ്ച് വീതം കേന്ദ്ര തെരഞ്ഞെടുപ്പ് നിരീക്ഷകരെ നിയമിക്കും. അസമിൽ ഒന്നാം ഘട്ടം-61 സീറ്റ്, രണ്ടാംഘട്ടം-65 സീറ്റ്, ബംഗാളിൽ ഒന്നാംഘട്ടം-18 സീറ്റ് (ഏപ്രിൽ 4), 31 സീറ്റ് (ഏപ്രിൽ 11), രണ്ടാംഘട്ടം-56 സീറ്റ്, മൂന്നാംഘട്ടം-62 സീറ്റ്, നാലാംഘട്ടം-49 സീറ്റ്, അഞ്ചാം ഘട്ടം-53 സീറ്റ്, ആറാംഘട്ടം-25 സീറ്റ്, തമിഴ്നാട്ടിൽ-234 സീറ്റ്, പുതുച്ചേരി-30 സീറ്റ്, കേരളം-140 എന്നിങ്ങനെയാണ് വിവിധ തീയതികളിൽ വോട്ടെടുപ്പ് നടക്കുന്ന നിയോജക മണ്ഡലങ്ങൾ.

Fig. 2. Input file 1



തെരഞ്ഞെടുപ്പ് തീയതി പ്രഖ്യാപിച്ചു.

സംസ്ഥാനത്ത് പതിനാലാം നിയമസഭയിലേക്കുള്ള വോട്ടെടുപ്പ് മെയ് 16 ന് നടക്കും. ഫലപ്രഖ്യാപനം മെയ് 19 ന്. കേരളമടക്കം നാല് സംസ്ഥാനങ്ങളിലും കേന്ദ്രഭരണ പ്രദേശമായ പുതുച്ചേരിയിലും നിയമസഭകളിലേക്കുള്ള തിരഞ്ഞെടുപ്പ് തീയതികൾ തിരഞ്ഞെടുപ്പ് കമ്മിഷൻ പ്രഖ്യാപിച്ചു. കേരളത്തിൽ ഏപ്രിൽ 22 ന് വിജ്ഞാപനം പുറപ്പെടുവിക്കും. ഏപ്രിൽ 29 നാണ് നാമനിർദേശ പത്രിക സമർപ്പിക്കാനുള്ള അവസാന ദിവസം. 30 ന് നാമനിർദേശ പത്രികകളുടെ സൂക്ഷ്മപരിശോധന നടക്കും. മെയ് രണ്ടു വരെ പത്രിക പിൻവലിക്കാം. സംസ്ഥാനത്ത് ഇത്തവണ 2.65 കോടി വോട്ടർമാരാണ് ഉള്ളതെന്ന് തിരഞ്ഞെടുപ്പ് കമ്മിഷൻ അറിയിച്ചു.

കേരളത്തിനു പുറമെ തമിഴ്നാട്, അസം, ബംഗാൾ എന്നീ സംസ്ഥാനങ്ങളിലും കേന്ദ്രഭരണ പ്രദേശമായ പുതുച്ചേരിയിലുമാണ് നിയമസഭാ തിരഞ്ഞെടുപ്പ് നടക്കുന്നത്. തിരഞ്ഞെടുപ്പ് തീയതികൾ പ്രഖ്യാപിച്ചതോടെ സംസ്ഥാനത്ത് പെരുമാറ്റച്ചട്ടം നിലവിൽ വന്നു. തമിഴ്നാട്ടിലും പുതുച്ചേരിയിലും ഒറ്റ ഘട്ടമായാണ് വോട്ടെടുപ്പ്. ഇരു സംസ്ഥാനങ്ങളിലും കേരളത്തോടൊപ്പം മെയ് 16 ന് വോട്ടെടുപ്പ് നടക്കും.

തിരഞ്ഞെടുപ്പ് പ്രക്രിയയിൽ പുതുമയാർന്ന ചില സംവിധാനങ്ങളും തിരഞ്ഞെടുപ്പ് കമ്മിഷൻ അവതരിപ്പിച്ചിട്ടുണ്ട്. ഒരു സ്ഥാനാർത്ഥിക്കും വോട്ട നൽകാൻ താൽപര്യമില്ലാത്തവർക്ക രേഖപ്പെടുത്താനുള്ള നോട്ടയ്ക്ക് ഇക്കുറി ചിഹ്നം ഉണ്ടാവും. ഇലക്ടോണിക് വോട്ടിംഗ് മെഷീനുകളിൽ സ്ഥാനാർത്ഥികളുടെ ഫോട്ടോയും ഉണ്ടാവും. നോട്ടയുടെ ചിഹ്നം ഏറ്റവും അവസാനമായിരിക്കും.

വോട്ടെടുപ്പ് തീയതിയുടെ പത്തു ദിവസം മുന്പുവരെ വോട്ടേഴ്സ് ലിസ്റ്റിൽ പേരു ചേർക്കാനുള്ള അനുമതിയും തിരഞ്ഞെടുപ്പ് കമ്മിഷൻ നൽകി. അസമിൽ ഏപ്രിൽ നാലിനും 11 നും രണ്ടു ഘട്ടമായി വോട്ടെടുപ്പ് നടക്കും. ബംഗാളിൽ ആറു ഘട്ടമായും വോട്ടെടുപ്പ് നടക്കും. ഏപ്രിൽ നാല്, 11, 17, 21, 25, 30, മെയ് അഞ്ച് ദിവസങ്ങളിലാണ് ഇവിടെ വോട്ടെടുപ്പ്.

മെയ് 21 ന് തിരഞ്ഞെടുപ്പ് നടപടികൾ പൂർണമാവും. സംസ്ഥാനത്തെ 13-ാം നിയമസഭയുടെ കാലാവധി തീരുന്നത് മെയ് 31 നാണ്. ഇതിനു മുന്പ് പുതിയ നിയമസഭ നിലവിൽ വരും.

സംസ്ഥാനത്ത് എല്ലാ രാഷ്ട്രീയ പാർട്ടികളും തിരഞ്ഞെടുപ്പിനുള്ള അന്തിമ ഒരുക്കത്തിലാണ്. 20 സീറ്റുകളിലേക്ക് സ്ഥാനാർത്ഥികളെ പ്രഖ്യാപിച്ച മുസ്ലിം ലീഗാണ് സംസ്ഥാനത്ത് തിരഞ്ഞെടുപ്പ് പോരാട്ടത്തിനുള്ള ആദ്യവെടി പൊട്ടിച്ചത്. യു.ഡി.എഫിലും എൽ.ഡി.എഫിലും സ്ഥാനാർത്ഥി പട്ടിക തയ്യാറാക്കുന്നത് അന്തിമഘട്ടത്തിലാണ്.

Fig. 3. Input file 2

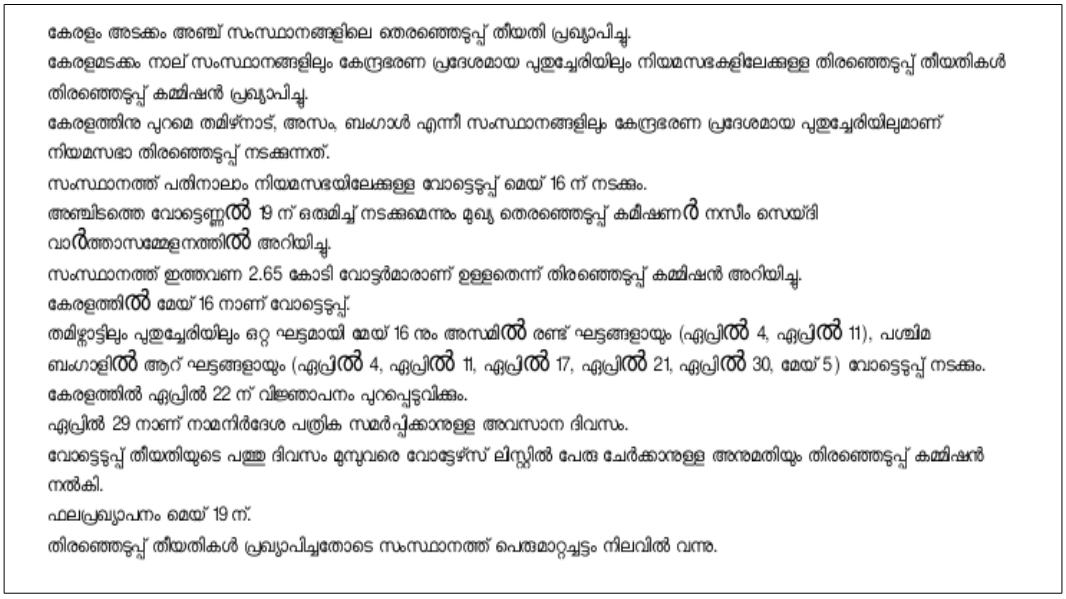

കേരളം അടക്കം അഞ്ച് സംസ്ഥാനങ്ങളിലെ തെരഞ്ഞെടുപ്പ് തീയതി പ്രഖ്യാപിച്ചു.
കേരളമടക്കം നാല് സംസ്ഥാനങ്ങളിലും കേന്ദ്രഭരണ പ്രദേശമായ പുതുച്ചേരിയിലും നിയമസഭകളിലേക്കുള്ള തിരഞ്ഞെടുപ്പ് തീയതികൾ തിരഞ്ഞെടുപ്പ് കമ്മിഷൻ പ്രഖ്യാപിച്ചു.
കേരളത്തിനു പുറമെ തമിഴ്നാട്, അസം, ബംഗാൾ എന്നീ സംസ്ഥാനങ്ങളിലും കേന്ദ്രഭരണ പ്രദേശമായ പുതുച്ചേരിയിലുമാണ് നിയമസഭാ തിരഞ്ഞെടുപ്പ് നടക്കുന്നത്.
സംസ്ഥാനത്ത് പതിനാലാം നിയമസഭയിലേക്കുള്ള വോട്ടെടുപ്പ് മെയ് 16 ന് നടക്കും.
അഞ്ചിടത്തെ വോട്ടെണ്ണൽ 19 ന് ഒരുമിച്ച് നടക്കുമെന്നും മുഖ്യ തെരഞ്ഞെടുപ്പ് കമീഷണർ നസീം സെയ്ദി വാർത്താസമ്മേളനത്തിൽ അറിയിച്ചു.
സംസ്ഥാനത്ത് ഇത്തവണ 2.65 കോടി വോട്ടർമാരാണ് ഉള്ളതെന്ന് തിരഞ്ഞെടുപ്പ് കമ്മിഷൻ അറിയിച്ചു.
കേരളത്തിൽ മേയ് 16 നാണ് വോട്ടെടുപ്പ്.
തമിഴ്നാട്ടിലും പുതുച്ചേരിയിലും ഒറ്റ ഘട്ടമായി മേയ് 16 നും അസമിൽ രണ്ട് ഘട്ടങ്ങളായും (ഏപ്രിൽ 4, ഏപ്രിൽ 11), പശ്ചിമ ബംഗാളിൽ ആറ് ഘട്ടങ്ങളായും (ഏപ്രിൽ 4, ഏപ്രിൽ 11, ഏപ്രിൽ 17, ഏപ്രിൽ 21, ഏപ്രിൽ 30, മേയ് 5) വോട്ടെടുപ്പ് നടക്കും.
കേരളത്തിൽ ഏപ്രിൽ 22 ന് വിജ്ഞാപനം പുറപ്പെടുവിക്കും.
ഏപ്രിൽ 29 നാണ് നാമനിർദേശ പത്രിക സമർപ്പിക്കാനുള്ള അവസാന ദിവസം.
വോട്ടെടുപ്പ് തീയതിയുടെ പത്തു ദിവസം മുമ്പുവരെ വോട്ടേഴ്സ് ലിസ്റ്റിൽ പേരു ചേർക്കാനുള്ള അനുമതിയും തിരഞ്ഞെടുപ്പ് കമ്മിഷൻ നൽകി.
ഫലപ്രഖ്യാപനം മെയ് 19 ന്.
തിരഞ്ഞെടുപ്പ് തീയതികൾ പ്രഖ്യാപിച്ചതോടെ സംസ്ഥാനത്ത് പെരുമാറ്റച്ചട്ടം നിലവിൽ വന്നു.

Fig. 4. Summary

5 Conclusion

Multidocument summarization system has a lot of significance in this fast growing digital world as human find difficulty in manually summarizing multiple documents. The proposed method discusses extraction based summary generation for multiple Malayalam documents coming under a single topic, by considering certain sentence specific features. Our approach is domain independent, even though we have illustrated with news articles. From the analysis we can say that this is one of the simplest and effective method for summarizing multiple documents. The usage of word net for finalizing the keywords in the vector space model has increased the term frequency score. This method does not need more semantic knowledge. As a future enhancement after considering relevant number of sentence related features, rather than taking the highest score sentence as summary sentence, we can use genetic algorithm based approach for summary generation. Lack of cohesion is an important drawback of extractive based multidocument summarization, improving cohesion in summary generation can be taken as a future work.

References

[1] Luhn H P. The Automatic Creation of Literature Abstracts.*IBM Journal of Research and Development* Vol-2(2):p.159-165 (1958)

[2] Baxendale P. Machine-made index for technical literature - an experiment.*IBM Journal of Research Development* Vol-2(4):354-361 (1958)

[3] Edmundson H P. New methods in automatic extracting. *Journal of the ACM* Vol-16(2):264-285 (1969)

[4] Radev D R.,Jing H., Stys M. and Tam D. Centroid-based summarization of multiple documents. *Information Processing and Management* Vol-40:919-938 (2004)

[5] Nobata C., Sekine S., Murata M., Uchimoto K., Utiyama M., Isahara H. Sentence extraction system assembling multiple evidence. *In Proceedings of the Second NTCIR Workshop Meeting* (2001)

[6] Jagadeesh J., Prasad Pingali, Vasudeva Varma. Sentence Extraction Based Single Document Summarization.*Workshop on Document Summarization, IIIT Allahabad.* (2005)

[7] Abuobieda A., Salim N., Albaham A., Osman A., Kumar Y. Text summarization features selection method using pseudo genetic-based model. *International Conference on Information Retrieval Knowledge Management* pp.193-197 (2012)

[8] Rafael Ferreira, Luciano de Souza Cabral, Rafael Dueire Lins, Gabriel Pereira e Silva, Fred Freitas, George D.C. Cavalcanti, Rinaldo Lima, Steven J. Simske, Luciano Favaro. Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications.* 0957-4174, 40 (14), pp. 5755â“5764 (2012)

[9] Mendoza M. Bonilla S. Noguera C. Cobos C. Leon E. Extractive single-document summarization based on genetic operators and guided local search *Expert Syst. Appl.* Vol-41 (9), pp. 4158â“4169;2014

[10] Wordnet for Malayalam, *http://malayalamwordnet.cusat.ac.in*

[11] Malayalam POS Tagger, *www.iiitmk.ac.in/MalayalamPOSTagger*