A Study on Different Text Summarization Methods in Dravidian Languages

Rahul Raj M
M.Tech scholar, Dept. of CSE
Ilahia College of Engineering & Technology
Muvattupuzha, India

Rosna P Haroon
Asst. Professor, Dept. of CSE
Ilahia College of Engineering & Technology
Muvattupuzha, India

Abstract— Recent studies shows that there are several text summarization systems are being developed in Indian languages. Within that more researches are taking place in South Indian Languages which fells in the category of Dravidian language family. This paper accumulates some of the most accepted text summarization systems in Dravidian languages.

Keywords—Dravidians, N-gram, centroid, root word, Laten Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA)

I. INTRODUCTION

'Dravidian' languages are a category of one of the most powerful language families found in India, which contains the languages like Malayalam, Tamil, Telugu, Kannada, Kodagu, Badaga, Byari, Tulu etc [1]. All these languages are spoken in the southern part of India and some of the parts of Pakistan. Predominantly the south India is considered as the home of 'Dravidian' languages, whose meaning is the language of 'Dravidians'.

Dravidian language family evolved as the subfamily of the broad Indo-European language family. Because of this reason Influence of Sanskrit can be found in these languages in both grammatically and vocabulary. Out of these languages Malayalam, Tamil, Telugu, Kannada is having more than 30 million speakers. Rapid growth in the field of natural language processing takes place mainly in these four languages due to large amount of speakers.

Up to an extent Grammatical elements, vocabulary, alphabet, the use of symbols are similar for these four languages due to the reason that they belong to same family. For example we can see that the set of alphabets are common for both Telugu and Kannada. Similarly some word phrases are common for both Malayalam and Tamil. The comparative study of techniques used for text summarization will help to enhance prevailing methodologies; even we can propose a multi lingual text summarizer with cost efficiency and high accuracy.

The similarity of these languages makes preprocessing of them alike. Main preprocessing steps for Dravidian languages include selection of candidate terms, filtering, vector space model, and ranking. Candidate term selection includes methods like tokenization,
N-gram approach linguistically oriented. Filtering is the process of vocabulary pruning which reduces the vocabulary by exempting unnecessary terms. Lemmatization, stemming, stop word removal is some of the techniques used for filtering. Mathematical representation of input sentences as vector is achieved by Vector space modeling (VSM). Ranking is the method increases importance of the sentence in the summarization process. Frequency count, inverse frequency count is some of the methods used for ranking [3].

II. RELATED WORKS

A. *Tamil*

Tamil is one of the oldest Dravidian languages evolved before three thousand years back. According to the 2001 census of India around 62 million speakers found in Indian subcontinent. The neighboring country Sri Lanka has Tamil speakers with an amount of 20 percentage of total population of the country [4]. Tamil is a

language with 12 vowels and 18 consonants can form more than 200 compound characters. The main text summarization works in Tamil listed as below.

(i) Summarizer using Centroid approach

This method proposed a new approach on summarizing contents of a news paper. The exaggerated news writing style is not suitable for a reader who has less time to spend for reading. In such a context this method gains importance. The algorithm used for this is based on a centroid method. Content of news paper is represented in multi dimensional vector space. The system works in the assumption that the contents with higher relevance will be near to the centroid of the vector space [8].

(ii) Tamil Text Compaction System

This implements a morphological based approach. Main steps are input processing, identification of type, extraction of compact word and output processing. Input processor is the morphological analyzer receives input and removes suffixes if present and delivers the root word (RW). The RW can fell in to two categories abnormal word (AW) which includes abbreviation and acronyms and normal words (NW) which are the common Tamil words. The NW is searched in the binary search tree and corresponding compact word will be retrieved. For AW a hashed technique is used for retrieving compact word. The compact words are passed to Tamil morphological generator and text compaction is done.

(iii) Summary generation for cricket match

A summary generation system for a particular cricket match in Tamil is the proposed system. The user should specify the url of the webpage to be summarized. A modified version of apriory algorithm will be used to find association rules from feature vector. Using correlation of variance plotting interestingness of the match can be determined [10].

(iv) Template based multilingual summary generation

A tourism oriented template based tourism oriented summary system for both Tamil and English. Every natural language is converted in to a universal networking language (UNL) which doesn't have any language barriers. The seven tourism based templates are, god, food, flaura and fauna, boarding facility, transport facility, place and distance. Semantics are used to avoid ambiguities in context. Example, the word "bat" has meaning cricket bat also a mammal can be avoided by using semantic constraints. Knowledge is extracted from the input and converts into UNL graph. Such UNL graph is generated for each template. Conversion of UNL to natural language is done by UNL dictionary corresponding to Tamil and English [12].

(v) Summarization Using Laten Dirichlet Allocation

This is a multiple Tamil text document summarization system, which uses the concept of Laten Dirichlet Allocation (LDA) for document modeling. It follows a hierarchy of data representation in which includes collection of documents called cluster break down in to collection of topics with a probability distribution. The cluster with most probability density will have the most relevance. Topics again divided into words with probability density. After preprocessing, clustering of documents takes place. If the selected domain is spots the clusters will be cricket, foot ball, tennis etc. Topic modeling is done for sentence relevance analysis and it will be based on query given by the user. Example: certain period of time. Sentence scoring is done as a result of LDA followed by the redundancy elimination and sentence ranking text summary will be generated.

(vi) Summarization: Semantic graph method

Advantage of this system is that Language-Neutral Syntax (LNS) is the system used for performing semantic analysis of the input here. Subject-Object-Predicate triples created here for each sentence and using that a semantic graph will be generated. After applying semantic normalization (for reducing noises and redundancies) a support vector machine (SVM) used for extracting relevant sentences as summary [21].

B. Kannada

Kannada is the official language of the south Indian state Karnataka has 35 million native speakers and 9 million worldwide speakers gives rise to 44 million total speakers according to the survey on "Top 30 Languages by Number of Native Speakers" carried out by vistawide.com [5]. Kannada has 49 basic letters including vowels and consonants with which can derive other compound characters.

(i) Summarization using keyword extraction

This paper proposed an extractive text summarizer for Kannada using keyword extraction method. Combination of GSS (Galavotti, Sebastiani, Simi) coefficients and IDF (Inverse Document Frequency) methods along with TF (Term Frequency) used for keyword extraction and summarizations. First step is the creation of data set by crawling the web by Wget tool. TF and IDFis calculated for each word in the document. For this

multi document summary system the GSS calculation is the process of identification of relevant terms for a particular topic. IDF gives importance of a word in the particular document. A ranking method is applied to sentences and highest ranked sentences are displayed in the output [13].

(ii) Summarization using Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a method used for semantic analysis of the input document. It creates semantic relationships among the contents of the sentences. LSA is an automated technique which uses only mathematical concept for semantic relation generation with a matrix like representation, unlike other systems LSA doesn't use used defined dictionaries or trained examples. Singular Value Decomposition (SVD) is used for summarization purpose. Dimensions of the sentence vectors which are principal and mutually orthogonal are found by SVD. The orthogonality removes redundancy in the output [14].

(iii) sArAmsha - A Kannada abstractive summarizer

This is an abstractive summarizer developed in Kannada. Abstractive summarizing methods are very common in English and Japanese languages but that branch is still to be explored more in India languages. sArAmsha uses Part of Speech Tagging (POS), stemming and Named Entity Recognition as preprocessing steps. If else rules contains grammatical constrains to omit redundant terms and irrelevant sentences called Information Extraction (IE) rules are used for abstraction. Sentence based abstraction templates are used to prepare summary [15].

(iv) Artificial neural network approach to text summarization

System implements machine learning by an artificial neural network. Namely a back propagation network is trained on a collection of text to do summarization [16]. The back propagation always improves the quality of training neural network because of the reason that it will support error calculations at hidden and output layers.

(v) Federated Document Summarization: Probabilistic Approach

This paper proposes an m-length extractive summarizer having m sentences in the extracted output. It combines the concepts of both similarity ranking and naïve Bayesian classifier. Jaccard's similarity score and conditional probability method are used for ranking purpose. There are two types of ranking is done in this system. First is called text ranking based on the similarity of sentences. Ranking is done by creating similarity matrices. Next one is the Naïve Bayesian ranking in which ranking carried out by conditional probability method. Select common sentences from these two ranked sets which goes to output and the remaining m-common sentences are selected from two sets based on the rank [17].

C. Malayalam

Malayalam is one of the most complex Indian languages in terms of grammar and vocabulary. It has high influence of Sanskrit and Tamil, contains 51 alphabets (Previously it was 56 and 5 of them were omitted). Malayalam is the official language of south most Indian state Kerala with more than 31 million speakers. It is also the official language of Indian union territory Lakshadweep situated in the Arabian Sea [6]. Other than India, Malayalam has considerable number of speakers in Middle East, USA and European countries. Some of the text summarization methods implemented in Malayalam are consolidated here.

(i) Text summarization for Malayalam using sentence extraction

A single document summarizer system for Malayalam based on ranking of sentences is implemented here. Feature score and Google page ranking formula are used in this. System starts with preprocessing sentences with tokenization followed by POS tagging.

After performing stop word removal stemming is performed by LALITHA (A light weight Malayalam stemmer using suffix stripping). The sentence scoring is carried out by five steps. Calculation of position score is the process of assigning score for each sentences based on their position in the document and with respect to position in the paragraph. Length score is calculated using a formula in which the sentences with more words will have higher scores. It is based on a convention that longer sentences will contain important information. Sum of position score and length score will give surface score of the sentence. TF-ISF(term frequency-inverse sentence frequency)score is calculated so as to understand the information content of the word, which is the inverse count of number of occurring of a particular word. Again there is an assumption that, terms occurring in most of the sentences are less important than the terms occurring in fewer sentences. TF-ISF calculation leads to topic similarity score (TSS) calculation which calculates the similarity of each sentence with the topic of the input. Sum of surface score and content score gives intermediate score (IS). Similarities of sentences are calculated using semantic relationship among them. In sentence ranking phase ranking is done. Sentences are ranked in the decreased order and the summary will contain first k- sentences [18].

(ii) Summarization using relevance of sentences

This system makes use of maximal marginal relevance between the sentences or the words for summary creation. Maximal marginal approach is used to find the relevance of each word throughout the input document with used of a Malayalam dictionary. A unit step function is used for this purpose. Successive threshold approach is used to decide the number of sentences to be present in the output. It may be the number of paragraphs present in the input or the average number of sentences present in the document [19].

(iii)Extractive Malayalam Document Summarization Based on Graph Theoretic Approach

This is an extractive summarizer based on graph theoretic approach. The graph is plotted in such a way that the nodes represent sentences and an edge is connected if two sentences shares common words. Similarity is found by cosine similarity. This representation allows making sub graphs and the summarization is achieved by extracting sub graphs, also query specific summaries can be obtained from these sub graphs [23]. The sentences with high cardinality will be placed in the output.

<<sumam miss paper>>

D. Telugu

Telugu is the official languages of two Indian states Telangana and Seemandhra. According to 2001 census of Indian government there are more than 75 million speakers all around the globe speaks this language. Telugu has 12 primary vowels, 23 primary consonants, and 10 aspirated consonant symbols which were migrated from Sanskrit with several derived letters [7]. Main text summarization techniques experimented in Telugu is given below.

(i) Telugu - English Dictionary Based Cross Language Query Focused Multi-Document Summarization

This method preferable for a multi-lingual query answering system in which user gives multiple queries and system responds with corresponding answer. Summarizer has importance in the query answering phase in which condensed content will reveal exact answer to user. Query processing stars with user gives input to the system with a collection of questions. After carrying out a collection of preprocessing steps root word in Telugu are obtained and these are translated in English using Telugu- English bilingual lexicon to find all possible translation words. Relevance based language modeling is used to rank topics in terms of probability of producing query from that. There is no training samples are available for this method, there for we use a probability based classifier. Using Hyperspace Analogue to Language (HAL) and the joint probability expression final score of each sentence can be calculated. The highest scored sentences will be present in the answer [20].

(ii) Summarization for classification

This is a summarizer implemented for reducing the complexity of SVM classifier. Ranking of sentences are carried out by the use of term frequency calculation, sentence score calculation which uses the concept of sentence location, centrality etc. A weight based ranking mechanism is used for extracting contents [22].

IV.CONCLUSION

This paper focuses on different text summarization mechanisms prevailing in Dravidian languages namely Tamil, Telugu, Kannada and Malayalam. Relevance of this literature study is to make awareness about the state of development of text summarization techniques in Dravidian languages. This paper not arguing that here is a perfect collection of all the methods in text summarization due to the exponential growth of experiments natural language processing especially in Indian languages. Anyway this is a reference to the aspirant in text summarization who can explore more in this area.

ACKNOWLEDGMENT

We would like to thank all the faculty members and students of Ilahia college of Engineering and technology for their immense support. Also we like to express gratitude towards friends and family and all other good hearts for their motivation and support to this work come true.

REFERENCES

[1] Bhadriraju Krishnamurti “The dravidian Languages”, Cambridge University Press,2003.
[2] vishal gupta, a survey of text summarizers for indian languages and comparison of their performance, Journal of emerging technologies in web intelligence, vol. 5, no. 4, November 2013.
[3] M. Hanumanthappa1, M Narayana Swamy and N M Jyothi,” Automatic Keyword Extraction from Dravidian Language”, Dept of Computer Science, Bangalore University, IJISET-International journal of Innovative science and technology, Vol.1, Issue 8, October 2014
[4] Danielle Devore, Kathryn Jones, Genee Kim, Jessie Mailhes, Dr. Rahul Chakraborty ,“Tamil Language and Culture Guide”
[5] Jennifer Torres B.S, Brooke Rutland B.S,Texas State University class of 2012,Rahul Chakraborty, Ph. D., CCC-SLP, “Kannada Manual: Language and Culture”

[6] Bianca Moncada, B.S, Kayla Marsh, B.S,Texas State University Class of 2012,Academic Advisor: Rahul Chakraborty, "Malayalam Manual: Language and Culture"
[7] Jessica Jimenez, B.S, Halya Lenard, B.S.,"The Langauge and Culture of"
[8] Syed Sabir Mohamed, Research Scholar, Faculty in Computer Science & Engineering, Sathyabama University, Chennai, India ,Shanmugasundaram Hariharan, Department of Computer Science and Engineering, TRP Engineering College, Tiruchirappalli, India, "A Summarizer for Tamil Language Using Centroid Approach"
[9] N.M..Revathi, G.P.Shanthi, Elanchezhiyan.K, T V Geetha, Ranjani Parthasarathi & Madhan Karky Tamil Computing Lab (TaCoLa), "An Efficient Tamil Text Compaction System ", College of Engineering Guindy, Anna University, Chennai.
[10] J. Jai Hari Raju, P. Indhu Reka, K.K Nandavi, Dr. Madhan Karky,"Tamil Summary Generation for a Cricket Match " Tamil Computing Lab (TaCoLa), College of Engineering Guindy, Anna University, Chennai.
[11] Subalalitha C.N, E.Umamaheswari, T V Geetha, Ranjani Parthasarathi & Madhan Karky, "Template based Multilingual Summary Generation ", Tamil Computing Lab (TaCoLa) College of Engineering Guindy Anna University, Chennai.
[12] N. Shreeya Sowmya1, T. Mala,"Tamil Document Summarization Using Laten Dirichlet Allocation ",Department of Computer Science and Engineering, Anna University, Guindy, Chennai.
[13] Jayashree.R , Srikanta Murthy.K and Sunny.K,"document summarization in kannada using keyword extraction ",Department of Computer Science, PES Institute of Technology, Bangalore, India, David Bracewell, et al. (Eds): AIAA 2011,CS & IT 03, pp. 121–127 , 2011.
[14] Geetha J.K., Deepamala N, Dept. of CSE, RVCE, Bangalore, India "Kannada text summarization using Latent Semantic Analysis", IEEE, Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference, on 10-13 Aug. 2015, 1508 – 1512.
[15] Embar, V.R. ; Dept. of Comput. Sci., M.S. Ramaiah Inst. of Technol., Bangalore, India ; Deshpande, S.R. ; Vaishnavi, A.K., "sArAmsha - A Kannada abstractive summarizer", Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on 22-25 Aug. 2013, 540 – 544, IEEE.
[16] Jayashree R, Srikantamurthy K, Basavaraj S Anami, Vijay M and Bharath B N Department of Computer Science PES Institute of Technology Bangalore, India , "An Artificial Neural Network approach to Text Summarization for the South Indian Language of Kannada", Hybrid Intelligent Systems (HIS), 2013 13th International Conference on 4-6 Dec. 2013, 45 – 48,IEEE.
[17] Ranganatha S. Assistant Professor, Govt. Engineering College, Hassan, Karnataka, IndiaVinay S. K.Student, PES Institute of Technology, Bangalore, Karnataka, IndiaBhargava H. S. Student, Govt. Engineering College, Hassan, Karnataka, India,"Federated Document Summarization Using Probabilistic Approach for Kannada Language", International journal of innovative research & development, Vol 3 Issue 1, pages 228-233.
[18] Renjith S R, Sony P, Dept.of Computer Science, College of Engineering Cherthala Kerala, India" An automatic text summarization for Malayalam using sentence extraction", IRF International Conference, 14th June 2015, Chennai, India, ISBN: 978-93-85465-35-2.
[19] Ajmal E.B, Rosna P Haroon, Department Of CSE Ilahia College of Engineering And Technology Muvattupuzha, India," Summarization of Malayalam Document Using Relevance of Sentences", International Journal of Latest Research in Engineering and Technology (IJLRET)volume 1 Issue 6‖November 2015 ‖ PP 08-13.
[20] Prasad Pingali, Jagadeesh Jagarlamudi, Vasudeva Varma Language Technologies Research Centre (LTRC), International Institute of Information Technology (IIIT), Hyderabad, India," Telugu - English Dictionary Based Cross Language Query Focused Multi-Document Summarization".
[21] Banu, M. Anna University Chennai Karthika, C. ; Sudarmani, P. ; Geetha, T.V. , "Tamil Document Summarization Using Semantic Graph Method ", Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on (Volume:2), 128 – 134, IEEE.
[22] Vishnu Murthy. G,Dr. B. Vishnu Vardhan Mekala Sreenivas, P.Vijaypal Reddy, "Text Classification using Text Summarization– A case study on Telugu Text", International Journal of Advanced Research in Computer Science and Software Engineering 3(7), July - 2013, pp. 1399-1403.
[23] Ajmal E.B, Rosna P Haroon, Department Of CSE Ilahia College of Engineering And Technology Muvattupuzha, India, "***An*** Extractive Malayalam Document Summarization Based on Graph Theoretic Approach",IEEE, August 2015.