International Journal of Applied and Advanced Scientific Research (IJAASR)
Impact Factor: 5.655, ISSN (Online): 2456 - 3080
(www.dvpublication.com) Volume 3, Issue 1, 2018

ISSUES IN MALAYALAM TEXT SUMMARIZATION

D. K. Kanitha*, D. Muhammad Noorul Mubarak & S. A. Shanavas***
* Department of Linguistics, University of Kerala, Kariyavattom, Thiruvananthapuram, Kerala
** Department of Computer Science, University of Kerala, Kariyavattom, Thiruvananthapuram, Kerala

Cite This Article: D. K. Kanitha, D. Muhammad Noorul Mubarak & S. A. Shanavas, "Issues in Malayalam Text Summarization", International Journal of Applied and Advanced Scientific Research, Volume 3, Issue 1, Page Number 201-204, 2018.

Abstract:
Text Summarization is the process of creates an abridged version of the original text and it covers overall idea about the document. The human summarization requires lot of time and effort. At the same time summarization system produce summary within a short span of time. It generates summaries or abstracts of large documents. Many techniques have been developed for summarization of text in various languages. The techniques may be language dependent or independent. Some techniques may be varies from its discourse structure. The summarization methods can be classified as extractive and abstractive. The abstractive method requires language processing tools. The extractive summarization depends on statistical and linguistic tools. This paper mainly concentrated some of the issues faced by the Malayalam text summarization. The Malayalam summarization faces some difficulties for creating a fruitful summary.

Key Words: Natural Language Processing, Automatic Text summarization & Methods of Automatic Text Summarization

1. Introduction:

Due to the information revolution electronic documents are becoming a principle media of information and thousands of documents are available from the internet. The search engine retrieve a heap of information, some pages are relevant and some pages are irrelevant. It consumes time for the user to check out all pages. For the process of speed up searching, the summary of a document is remarkable. The technology of automatic summarization is very useful in this context. Now the technological development in Malayalam is enormous. People commonly used their mother tongue for communication and interaction with the system. That is the importance of Natural Language Processing (NLP). It is a field of computer science and linguistics concerned with the interactions between computers and natural languages. NLP is very attractive method of human-computer interaction. Computational linguistics is the applied field of linguistics, which related to artificial intelligence dealing with acquisition and production of natural languages. Text Summarization is the sub field of Natural Language Processing. It is the process of condensing the source text into shorter version preserving its information content and overall meaning. Text summarization is a technique, where a text is entered into the computer and returns the summary of a text. The technique has begins in 50's and wide scope in recent years.

Now the user can get abandon text materials in Malayalam. Some text may be 100 or more pages. Within these read every page and conclude the abstract is time consuming. At the same time the graft will appear in a short paragraph it is fascinating and user can identify the idea within a short time. Some of the uses of text summarization are given below:

Summarize the news to SMS for mobile phones.
Summarize the medical data for doctors.
Search the information in foreign language the user get a translated abstract of summarized document.
Summarize the legal documents.
Summarize the media reports

Text summarization methods can be classified into extractive and abstractive summarization (Hovy and Lin, 1997). Abstractive text summarization understands the original text and re-tells it in few words. Same way as the human summarizer create summary. The abstractive summarization is a tedious task because the natural language generation techniques are used for generating summary. Extractive text summarization extracts important sentences and paragraphs from the original document and concatenated them into shorter form. Statistical, heuristic and linguistic methods are used for extractive text summarization. The extractive summarization is simpler than abstractive summarization. So the extractive summarization methods are widely used in automatic text summarization. This paper focuses some issues of text summarization especially generate a virtuous summary for Malayalam documents. The road map of rest of the paper is organized as follows. Section-2 gives a review on existing summarization methods. Section-3 shows the issues of Malayalam text summarization. Section-4 shows the evaluation methods of text summarization. Section-5 concludes the graft.

2. Models of Text Summarization:

Commonly summarization systems follow extractive and abstractive summarization methods. The earlier systems focus on domain dependent and concentrated on scientific articles. Later the systems are domain

independent and summarize any type of articles. Microsoft Word's Auto Summarize function is a simple example of automatic text summarization. In 1958 Luhns developed a summarization system for summarize scientific articles. The sentences are ranked on the basis of word frequency and phrase frequency. The stop word removal and stemming the high frequency word held sentences are selected for summary sentences. The main drawback of this system was duplication in summary sentences.

Baxendale (1958) proposed a method for sentence extraction such as document title held sentences, first and last sentences of each paragraph of a document. The author proposed that the newspaper articles the first sentences are high chance to include in summary. But in technical papers the last sentence or concluding sections are high chance to include in summary. Lin and Hovy (1997) claimed that Baxendale position method is not a suitable method for sentence extraction in different domains. Because the discourse structure of a sentence varies from different domains. The main disadvantage of this system was it is domain related summarization system. Edmundson (1969) system considered four parameters to generate the summary. The parameters are cue phrases, keywords, title words and location. The main drawback of this system was duplication in summary. Barzilay and Elhadad (1997) proposed a lexical chain method to score the sentences. The concept of lexical chain was first introduced in Morris and Hirst, 1991. The lexical chain links the semantically related terms within different parts of document. Barzilay and Elhadad used a wordnet to construct the lexical chains.

Dalianes (2000) developed a system named as SweSum. This system was the first web based automatic text summarizer for Swedish news articles. SweSum is also available for Danish, Norwegian, English, Spanish, French, Italian, Greek, Farsi and German Texts. The SweSum used client/ server architecture. The web client input the original text and accepts the summarized text. The web server accepts the source text and performs tokenizing, keyword extraction and sentence ranking. The sentences are scored using statistical, linguistic and heuristic method. The weighting approaches are position, numerical value, font based feature etc. Score of each word in the sentence is calculated and then find the sentence score. SweSum shows better result in query based text summarization. Conroy and O'Leary (2001) applied Hidden Markov model for sentence extraction. The system determined the probability of inclusion of a sentence in summary depend on whether the previous sentence is related to next sentence. Radev et.al., (2004) proposed a system MEAD which computes the score of sentence based on some features such as similarity to centroid, position of sentence, sentence length, etc.

Farisum (2004) system followed SweSum architecture for sentence extraction. It is a web based summarizer for Persian. The Farisum used the same architecture of SweSum but one difference was it does not use any lexicon. The statistical method latent semantic analysis (LSA) identifies semantically similar sentences. Latent semantic analysis is a statistical method for language processing and finds semantic similarity between words and texts (Landauer and Dumais, 1997). It identifies the conceptual meaning of words and the similarity of sentences. (Steinberger, J. and Jezek, K. 2004). LSA is an unsupervised learning method for finding vector space semantic representation from a source document. The words close in meaning will occur in same contextual space. LSA used a Mathematical technique named as Singular Value Decomposition (SVD) to find the semantically similar sentences. It shows the semantic similarity of words and sentences. After the input matrix creation then compute the SVD matrix. LSA is also known as Latent semantic Indexing. The LSI finds underlying meaning or concepts of input document. Yihong Gong and Xin Liu (2002) suggest LSA based algorithms of text summarization. Commonly a mathematical matrix Singular Value Decomposition is used for ranking the sentences. The sentences are ranked on its conceptual space and top ranking sentences are selected as summary sentences. Now day's statistical algebraic methods are widely used in information retrieval and text summarization.

Azmi and Al-Thanggam (2012) proposed a model based on extractive technique for create summary in Arabic language. It proposed an algorithm based on Rhetorical Structure Theory and create summary. After create the summary sentences are ranked and highest ranking sentences are selected as summary. Gupta (2013) proposed a hybrid algorithm for Hindi and Punjabi Text Summarization. This method finds the feature score of sentences and high scored sentences are collected for summary.

3. Issues of Malayalam Text Summarization:

Malayalam is a Dravidian language it is one of the 22 official languages of India and was designated a classical language in India in 2013. It is used by around 36 million people. It is spoken mainly in the south west of India, particularly in Kerala, the Laccadive Islands, and also in Bahrain, Fiji, Israel, Malaysia, Qatar, Singapore, UAE and the UK.

The main features of Malayalam are:

- ✓ It is an agglutinative in nature.
- ✓ This is a syllabic alphabet in which all consonants have an inherent vowel.
- ✓ The structure of sentences is simple, compound and complex.
- ✓ The morphology of language is inflectional, derivational and compounding.
- ✓ The main word classes are Noun, Verb, Adjectives, Adverbs, Postpositions and Conjunctions.

Recently the numerous Malayalam documents are available from net. But finding the relevant data from various web pages is heavy task. Reading every pages and find relevant data is time consuming.

Commonly the summarization systems depend on abstractive or extractive techniques. Malayalam text summarization uses the abstractive methods it require heavy language processing tools. So suggest a proper method for abstractive based Malayalam text summarization is difficult. Same way the issues are very high. Some of the issues faced by the text summarization are:

- **Named Entity Recognition (NER):** The NER identifies the names of some special entities such as person, place, organization etc. In English the NER is easy. But find the Named Entity is difficult in Malayalam documents.
- **Co-Reference Resolution:** Understand the idea behind the text requires co-reference resolution. English co-reference resolution easy handle because it has only limited co-referents. But Malayalam some cases the co-references are omitted then built the sentences. To tackle this problem requires sufficient language processing tools.
- **Parts of Speech Tagging:** Assigning parts of speech to a given word is called POS tagging. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories. Sometimes a word in Malayalam may be treated as both noun and verb it generates an ambiguity. Disambiguation is the difficult problem in text summarization. Sometimes it is overcome by checking the preceding word or following word.
- **Word Sense Disambiguation:** Commonly the words in Malayalam have different meaning in different context. Summarization method needs the accurate meaning of each word.
- **Suffix Stripping:** Malayalam suffix stripping removes the inflected form of noun or verb. The words in Malayalam have large number of inflections. It is tedious task to remove the suffix from each word.
- **Parsing:** Parsing or syntactic analysis is a computer program that recognize or analyses the text using the rules of formal grammar. The accurate Malayalam parser need the proper grammar implementation, so building the grammar is computationally complex task.
- **Anaphora Reference Resolution:** Anaphora reference resolution that determines relationships between hierarchically related entities such relationships include a pronoun like hers, his, their, mine etc refers to a previously mentioned named entity. No special tool for identifying the relationship. The recognition of relation between the entities is very difficult in Malayalam.
- **Collocation:** Collocation means two or more words frequently occur in a sentence by chance. Sometimes it may be noun phrase or verb phrase. Frequency is the simple method used for finding word collocation.
- **Polysemy:** The same word has different meaning in different context. The Malayalam text summarization polysemy is difficult problem.
- **Tokenizing:** Tokenization or word segmentation is the process of splitting the given text into small units called tokens. The tokens may be words or number or punctuation mark. NLTK's tokenizer is sufficient for tokenize polysynthetic languages with specified word boundary (Eg: English, French or European Languages). Considering the word boundaries and split the sentences into words couldn't get a complete word. Compound words are also used for constructing sentences. Proper tokenize a text or sentences in Malayalam requires linguistic tools.
- **Compound Word:** A compound word consists of more than one lexeme. The agglutinative language like Malayalam the most words are created by joining morphemes. So high chance in formulate compound word and a proper compound word splitter is necessary.
- **Word and Sentence Boundary Identification:** It is difficult Malayalam documents because proper identification of word and sentence boundary is possible by the implementation of grammar.
- **Length of Summary:** The length of summary is difficult. Nobody can determine the actual length of summary.

4. Evaluation Methods:

The summary evaluation (Hovy, E.H. and Lin, C.Y.1999) either manually or automatically is a tedious task. The main difficulties are:

- ✓ There is no fair system for evaluating the summary systems.
- ✓ The system generated summary is different from human summary.

Hence the evaluation of summary is necessary for any summarization system. There is no single evaluation scheme to evaluate all aspects of summary. So combination of evaluation methods are used for evaluate summary. Mainly two methods are used for summary evaluation such as intrinsic and extrinsic evaluation. (Spark Jones and Galliers 1995 Mani and Maybury 1999). The intrinsic evaluation predicts the quality of summary based on content and co-selection measures. The co-selection measures are Precision, Recall and F-score. The content based measures are cosine similarity and unit overlap. The extrinsic evaluation predicts the quality of summary based on related task of summarization.

5. Conclusion:

The text summarization systems many follow extractive, abstractive or hybrid methods. But in the case of creating a Malayalam document summary it is faced lot of problems. Some systems in agglutinative languages follow the extractive based methods. The pure extractive based method is used the semantics minimal for creating the summary. The statistical and linguistic based methods are used more semantically related sentences are generated as summary. But the linguistics analysis of the Malayalam text is computationally complex. The language analysis and creating a summary it is faced numerous issues. When user develops a system for Malayalam text summarization recognize the issues and generate a proper tool for understand the semantics. The extractive based summarization methods which understand the semantics of sentences that is suitable for summarizing articles in Malayalam.

6. References:

1. Luhn, (1958). The automatic creation of literature abstracts. IBM Journal of Research Development, 2(2):159–165.
2. Baxendale, P. B. (1958). 'Machine-made index for technical literature: an experiment'. IBM Journal, 354–361.
3. Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.
4. E. Hovy and C-Y Lin. (1997). Automated Text Summarization in SUMMARIST. in Proceedings of the Workshop of Intelligent Scalable Text Summarization, July.
5. Barzilay, R., & Elhadad, M. (1997). Using lexical chains for text summarization. In Proceedings of the ACL'97/EACL'97 workshop on intelligent scalable text summarization (pp. 10–17), Madrid, Spain.
6. Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284.
7. Hovy, E.H. and Lin, C.Y. (1999). Automated Text Summarization in SUMMARIST. In I. Mani and M. Maybury (eds), Advances in Automated Text Summarization. Cambridge: MIT Press, pp. 81–94
8. Mani, I & Maybury, M. T. (Eds.). (1999). Advances in automated text summarization. Cambridge, MA: The MIT Press.
9. Hahn.U, and Mani.I. (2000).The challenges of automatic summarization. Computer 33: 29-36.
10. Gong.Y., and Liu.X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of ACM SIGIR. New Orleans, USA.
11. Steinberger, J. and Jezek, K. (2004). Using Latent Semantic Analysis in Text Summarization and Summary Evaluation. Proceedings of ISIM '04, pages 93-100.
12. Das.D., and Martins.A.F.T. (2007). A Survey on Automatic Text Summarization. Literature survey for Language and Statistics II, Carnegie Mellon University.
13. Gupta,V., and Lehal.G.S. (2010). A Survey of Text Summarization Extractive Techniques. Journal Of Emerging Technologies In Web Intelligence, VOL. 2, NO. 3.
14. En.wikipedia.org/wiki/Malayalam