ISSN:0975-9646

Malayalam Text Summarization Using Graph Based Method

Kanitha D K,
Computational Linguistics, Department of Linguistics, University of Kerala,

D. Muhammad Noorul Mubarak,
Department of Computer Science, University of Kerala, Kariavattom, Thiruvananthapuram

S.A. Shanavas,
Dept.of Linguistics, University of Kerala, Thiruvananthapuram.

***Abstract*— Automatic text summarization system generates summaries or abstract of large documents. Many techniques have been developed for summarization of text in various languages. One of the most commonly used statistical methods is graph theoretic approach. The sentences are represented as nodes and the relation is represented as edges. The cardinality of a graph shows the importance of sentences. The graph based algorithm is sufficient for agglutinative language like Malayalam. The algorithms are evaluated on Malayalam news articles and performances are compared using precision, recall and f-measures.**

***Index Terms*— Automatic Text summarization, Malayalam text summarization, Graph theoretic approach.**

I. INTRODUCTION

Due to the exponential growth of information in Internet thousands of documents are available from the web. The search engines retrieve heap of web pages with bundle of data user find the appropriate or significant information. 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.

Natural Language Processing (NLP) 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.

Uses of Automatic text summarization

- 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.

Text summarization methods can be classified into extractive and abstractive summarization (Hovy and Lin, 1997) [4]. Abstractive text summarization, it understands the original text and re-tells it in few words. The generated summary may be the new sentences which show the overall content of the document. Linguistic methods are used for abstraction. The abstractive summarization is a tedious task because understanding the meaning of sentences and it require natural language processing tools. The extractive summarization method selects the importance sentences from source document and concatenate into shorter form. Extractive summarization is simpler than abstractive summarization. Now most of the systems follow extractive text summarization method. Statistical, heuristic and linguistic methods are used for extractive text summarization.

This paper focuses on graph theoretic approach to generate a virtuous summary for Malayalam documents. The road map of this paper is organized as follows. Section-2 gives a review on existing summarization methods especially concentrated on extractive methods. Section-3 shows the graph based algorithm and how ranking the documents. Section-4 shows the experimental results. Section-5 concludes the graft.

II. RELATED WORKS

In literature most of works have been concentrated on the sentence-extraction method. This review mainly focuses on statistical method used for sentence scoring.

1. Luhn's Method (1958)

Sentences are ranked on the basis of word frequency and phrase frequency. After performing the stop word removal and stemming the high frequency word included sentences are selected for summary sentences. The high scored sentences are selected for summary. It gives the summary of same topic or context. The main drawback of this system was duplication in summary sentences.

2. Baxendale (1958)

Baxendale proposes a straight forward method for sentence extraction such as document title, first and last sentences of a document or each paragraph. He argued that newspaper articles the first sentences are high chance to include in summary. But in technical papers the last sentence or concluding sections are having 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. The discourse structure of a sentence varies from different domains. The main disadvantage of this system was it is domain related.

3. Edmundson (1969)

Now many current automatic text summarization systems follow Edmunson's method. He considers four parameters to generate the summary. The methods are cue phrases, keywords, title words and location. The main drawback of this system was duplication in summary.

4. Barzilay and Elhadad(1997)

This summarization approach proposes a lexical chain method to score the sentences. The concept of lexical chain was 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.

5. SweSum(Dalianis 2000)

SweSum was the first web based automatic text summarizer for Swedish. It summarizes Swedish news text in HTML based text format on the World Wide Web. SweSum is also available for Danish, Norwegian, English, Spanish, French, Italian, Greek, Farsi, and German Texts. It uses statistical, linguistic and heuristic methods to obtain the summary sentences.

The SweSum architecture uses client/ server application. The web client input the original text and accepts the summarized text. The web server accepts the source text and performs tokenizing, scoring, keyword extraction and sentence ranking. The sentences are scored using statistical, linguistic and heuristic method. Such as position, numerical value, font based feature etc. Score of each word in the sentence is calculated and then find the sentence score. A threshold is preset and generate the desired summary with some statistical information such as number of words, frequent keywords etc. The query based text summarization SweSum shows better result.

6. Conroy and O'Leary (2001)

It applies Hidden Markov model for sentence extraction. The system states the probability of inclusion of a sentence in summary depend on whether the previous sentence is related to next sentence.

7. MEAD (Radev et.al., 2004)

This system computes the score of sentence based on some features such as similarity to centroid, position of sentence, sentence length, etc.

8. Farisum(2004)

This system follows SweSum architecture for sentence extraction. It is a web based summarizer for Persian. The Farisum uses the same architecture of SweSum but one difference was it does not use any lexicon.

Microsoft Word's Auto Summarize function is a simple example of automatic text summarization. Text summarization methods include statistical, linguistics and heuristics approaches. Tf-idf is an example of corpus based approach. Position and title method is an example of heuristics approach. Lexical chain method is an example of discourse structure approach. Now a days a new approach Latent Semantic Analysis is widely used in information retrieval and text summarization.

9. Yihong Gong and Xin Liu(2002)

LSA (Latent Semantic Analysis) based algorithms are suggested for text summarization. A mathematical matrix Singular Value Decomposition is used for Latent Semantic Indexing. First create a term by sentence matrix (tf-idf). Columns represent the sentences and row represents the terms. After finding the term matrix calculate SVD. The SVD of term by sentence matrix is defined as:

$$A = U\Sigma V^T$$

'U' is an m × n column-orthonormal matrix whose columns are called left singular vectors; $\Sigma = \text{diag}(\sigma 1, \sigma 2, \ldots, \sigma n)$ is an n × n diagonal matrix, whose diagonal elements are non-negative singular values sorted in descending order, and V = [vij] is an n × n orthonormal matrix, whose columns are called right singular vectors. The V^T matrix is used for sentence selection algorithm.

10. TextRank (Mihalcea and Tarau 2004)

Sentences are represented as edges and the relation is represented as edges. The score for each vertex is computed on the basis of link between the terms within the sentences.

11. Hybrid approach (LSA + Cluster, R. Yang, Z. Bu, and Z. Xia (2012))

The authors proposed the text summarization method based on LSA and Cluster based method. The sentences are ranked using LSA then the cluster method is used for sentence selection.

12. Grap based (Sankar K, Vijay Sundar Ram R and Sobha Lalitha Devi, 2011)

The authors proposed a graph based algorithm for summarizing articles in Tamil. This approach each vertex represents a sentence and edges show the connectivity between sentences. Vertices of the graph are first marked with sentence weights and edges are marked with Levenshtein similarity weights. Average of all levenshtein similarity weights of all edges connected to a vertex is calculated to find out the vertex weights. The sentence rank is the average of sentence weight and vertex weight. Sentence weight is the sum of all affinity weights of all words in the sentence. Affinity weight of a word is calculated as the sum of number of occurrences of the word

in the document divided by the total number of words in the document. Levenshtein similarity weight two sentences is calculated as by the difference between the max length of two sentences and Levenshtein distance of two sentences divided by the max length of two sentences.

13. LexRank

Erkan and Radev (2004) proposed LexRank which is a summarization system for multiple documents where the semantically similar are represented as connection between the nodes. The important sentences selected on the basis of a random walk on the graph.

14. GRAPHSUM

Baralis et al. (2013) proposed a summarizer based on graph model which represents correlations among multiple terms by discovering association rules.

15. E-mail summarization using graph method

Carenini et al. (2008) proposed a summarization method for summarizing email conversations. A graph is built with the conversation involving a few emails in which nodes represent conversations and edges represent replying relationship between two nodes. The vertex weight is assigned to each node and finds the rank of conversation.

16. Ferreira et al. (2014)

The authors suggested a graph based clustering algorithm for sentences. The sentences are represented as vertex and the relation is based on the four distinct relations such as semantic similarity, statistical similarity, discourse relations and co-reference resolution.

17. Graph-based Extractive summarization (Parveen and Strube 2015)

This approach doesn't depend on any parameter and training data as it is an unsupervised technique and summary being coherent and good quality.

III. PROPOSED METHODOLOGY FOR MALAYALAM TEXT SUMMARIZER

Malayalam is a Dravidian language used predominantly in the state of Kerala, India. 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.

Malayalam has a rigid and vast grammar structure. Computationally understand the language structure, identify the meaning of sentence, and extract the relationship and implementing the grammar is a tedious task. Now a day's 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 it is time consuming. An efficient summarizer handles this task efficiently.

The graph theoretic approach extracts the semantically similar sentences. The similarity is determined by the different similarity scoring approaches like cosine similarity, longest common sub sequences, Levenshtein similarity etc. The relation of node is represented by the feature score of sentences.

3.1 Architecture of proposed system

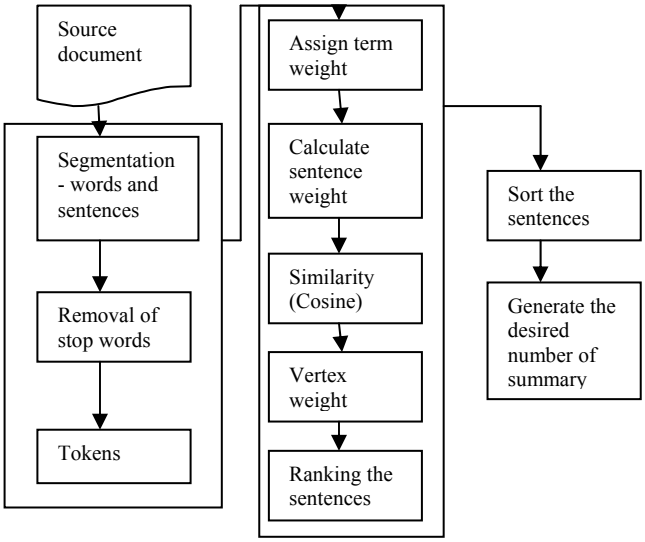

Figure1: System architecture

3.2 Algorithm for Malayalam Text summarization

1. Input the .txt files.
2. The sentence and word tokenizer () split into sentences and words.
3. Removes the characters such as (,) . ! etc,
4. Content words can be compared to stop words list. If the word is included in the stop word list move to next word.
5. If it is not a stop word placed in word dictionary and also keeps the sentence number.
6. Calculate the affinity weight of sentences
7. Calculate the sentence weight.
8. Calculate cosine similarity between the sentences
9. Calculate the vertex weight.
10. Rank the sentences on the basis of vertex weight and sentence weight.
11. Preset a threshold and extract the desired number of sentences.

Ranking of sentences using the graph theoretic method as explained below:

Affinity weight of

S1=(1/47+2/47+4/47+1/47+1/47+2/47+2/47+1/47+1/47+1/47)= 0.32

S2=(1/47+2/47+2/47+1/47+1/47+1/47+1/47+1/47+2/47+4/47+1/47+1/47+1/47+1/47)=.4

S3:

(1/47+1/47+1/47+1/47+4/47+1/47+1/47+1/47+1/47+1/47+1/47+1/47+2/47+2/47+1/47+1/47+1/47)=.4

S4: (2/47+2/47+2/47+4/47+1/47+1/47+1/47)=.26
Sentence weight of
S1= 0.32/10=0.032
S2=0.4/10=0.04
S3=0.04
S4=0.26/10=0.026
Similarity (s1,s2)= 14-6/14=0.6
Similarity(s1,s3)=16-8/16=0.5
Similarity(s1,s4)=10-8/10=0.2
Similarity(s2,s3)=16-13/16=0.1
Similarity (s2,s4)=14-13/14=0.07
Similarity(s3,s4)=16-12/16=0.3
Rank of Sentences are:
S1=0.032+.6=0.63
S2=0.04+6=0.64
S3=0.04+0.5=0.54
S4=0.02+0.3=0.32

S1: മനുഷ്യന് ആദ്യമായി ചന്ദ്രനില് കാലുകുത്തിയിട്ട് ഈ ജൂലായ് 20ന് 40 വര്ഷം തികയുന്നു.
S2: 1969 ജൂലായ് 20ന് അമേരിക്കന് ബഹിരാകാശ സഞ്ചാരി നീല് ആംസ്ട്രോങ് ആദ്യമായി ചന്ദ്രനില് കാലുകുത്തുന്ന മനുഷ്യനെന്ന ബഹുമതി നേടി.
S3:ഭൂമിയില്നിന്നു രണ്ടരലക്ഷം നാഴിക അകലെയുള്ള ചന്ദ്രനില് ആളെയിറക്കി സുരക്ഷിതമായി ഭൂമിയില് തിരിച്ചെത്തിക്കുക എന്ന ദൗത്യവുമായി അമേരിക്ക അപ്പോളോ പദ്ധതിക്കു രൂപം നല്കുന്നത്.
S4:അപ്പോളോ - അമേരിക്ക ആദ്യമായി ചന്ദ്രനില് മനുഷ്യനെ ഇറക്കിയ ഉപഗ്രഹം.

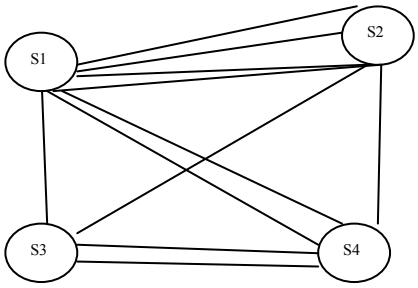

Figure 2: Graphical representation of the sentences.

IV. EVALUATING SUMMARY IN AUTOMATIC TEXT SUMMARIZATION

Text summarization technique creates summary or extract of a text. The summarization technique has been developed for many years but recent years the wide use of Internet there have been great mobility in summarization techniques. The summary evaluation either manually or automatically is a tedious task.

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 some related task. The proposed summarizer is evaluated the quality of summary on the basis of precision, recall and f-scores measures. Precision score shows the fraction of the sentences chosen by the humans and selected by the system are correct. Recall score shows the fraction of the sentences chosen by humans is recognized by the machine. F-measure is computed by combining recall and precision.

Source files	Precision	Recall	F-score
Math_articles_essays _Sum1	0.40	0.50	0.44
Math_articles_sports _Sum2	0.40	0.80	0.53
Math_articles_travel _Sum3	0.47	0.58	0.52
Math_articles_tech _Sum4	0.60	0.60	0.60
Math_articles_heal _Sum5	0.56	0.33	0.42
Math_articles_bus _Sum6	0.63	0.54	0.58

Table1: Result of system summary compare with human summary

The result shows that generated summaries 51% of the sentences are semantically similar with human generated summaries.

V CONCLUSION

The rate of information growth in Malayalam documents in WWW needs an efficient and accurate summarization system. The abstractive summarization requires heavy computational models for language generation. Such a situation the extractive text summarization produces the summary within the limited time. The performance of graph based extractive summarization method shows good result in summarizing Malayalam documents. The result shows that graph based algorithms perform well and obtain the satisfactory results.

REFERENCES

1. Luhn, “The automatic creation of literature abstracts”, IBM Journal of Research Development, 2(2):159–165, 1958.
2. P. B. Baxendale, “Machine-made index for technical literature: an experiment”, IBM Journal, 354–361, 1958.
3. T. K. Landauer, and S. T. Dumais, “A solution to Plato’s problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge”, Psychological Review, 104, 211-240. 1997.
4. E. Hovy and C-Y Lin, “Automated Text Summarization in SUMMARIST”, Proceedings of the Workshop of Intelligent Scalable Text Summarization, July 1997.
5. R. Barzilay and 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. and 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.L. Yu, J. Ma, F. Ren, and S. Kuroiwa (2007) "Automatic text summarization based on lexical chains and structural features," in Proceedings of the Eighth International IEEE ACIS Conference.
14.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.
15.R. Yang, Z. Bu, and Z. Xia. (2012). "Automatic summarization for Chinese text using affinity propagation clustering and latent semantic analysis," Web Information Systems and Mining. Lecture Notes in Computer Science, vol. 7529, pp. 543-550.
16.Sankar K, Vijay Sundar Ram R and Sobha Lalitha Devi, Text Extraction for an Agglutinative Language, Problems of Parsing in Indian Languages, May 2011 Special Volume.
17.M. Banu, C. Karthika, P Sudarmani and T.V. Geetha, "Tamil Document Summarization Using Semantic Graph Method", Proceedings of International Conference on Computational Intelligence and Multimedia Applications, pp. 128-134, 2007.
18.S. Kumar, V. S. Ram and S. L. Devi, "Text Extraction for an Agglutinative Language," Proceedings of Journal: Language in India, pp. 56-59, 2011.
19.Erkan G, Radev D (2004) LexRank: graph-based lexical centrality as salience in text summarization. J Artif Intell Res 22:457–479
20.Baralis E, Cagliero L, Mahoto N, Fiori A (2013) GRAPHSUM: discovering correlations among multiple terms for graph-based summarization. Inf Sci 249:96–109. doi:10.1016/j.ins.2013.06.046
21.Carenini G, Ng RT, Zhou X (2007) Summarizing email conversations with clue words. In: Proceedings of the 16th international conference on World Wide Web. ACM. pp 91–100
22.Carenini G, Ng RT, Zhou X (2008) Summarizing emails with conversational cohesion and subjectivity. ACL.
23.Ferreira R, de Souza Cabral L, Freitas F et al (2014) A multi-document summarization system based on statistics and linguistic treatment. Expert Syst Appl 41:5780–5787. doi:10.1016/j.eswa.2014.03.023
24.Parveen D, StrubeM(2015) Integrating importance, non-redundancy and coherence in graph-based extractive summarization. In: Proceedings of the 24th international conference on artificial intelligence. AAAI Press. pp 1298–1304 8:353–361
25.En.wikipedia.org/wiki/Malayalam
26.pypi.python.org/pypi/sumy/0.3.0/