International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Automated Kannada Text Summarization using Sentence Features

Arpitha Swamy, Srinath S

Abstract*: *There is a growing requirement for the text summarization due to the difficulty of managing exponential increase of information accessible on the World Wide Web. Text summarization is a process to extract the contents in the original text to the shorter form which provides important information to the user. The summarizer presented in this paper produces the extractive summaries of Kannada text documents. The proposed summarizer system considers five features to determine the important sentences in the document. The features used are Term Frequency, Term Frequency-Inverse Sentence Frequency, Keywords feature, Sentence length and Sentence position. The value of each feature is computed and score for each sentence in the document is the average of all the feature score values. The sentences with the top scores are selected to be included in the extractive summary. The results of the proposed model are evaluated using ROUGE toolkit to measure the performance based on F-score of generated summary. Experimental studies on custom-built dataset with 50 Kannada text documents shows significantly better performance in producing extractive summaries as compared to human summaries.

Index Terms*: *Inverse Sentence Frequency, Natural Language Processing, ROUGE, Term Frequency, Text summarization.

I. INTRODUCTION

Automatic summarization is a method to compress the contents in the original document into shorter form. The shorter form is called the summary and it should contain the vital information of the original document. Due to the large amount of textual data accessible in the form of online documents, there is a requirement for an automatic tool to produce summary of large documents so that it saves effort of users in finding the information they are interested in. The summary of document is created by pre-processing the document then extracting the features followed by sentence scoring step to find out which sentences are important and crucial in the document.

Text summarization techniques are mainly classified as extractive summarization techniques and abstractive summarization techniques. Extractive techniques generates the summary of the document by identifying important sentences using some word and sentence based statistical and linguistic features. In Abstractive techniques, summary is created by selecting and reordering the words in the original document or by adding some new words that are not exist in the original document through linguistic analysis of the text. Text summarization systems are also categorized as single-document summarization systems and multi-document summarization systems based on the count of documents passed as input. Depending on the purpose, summarization systems are classified as generic, domain specific, or query-based.

There are many summarization systems available to summarize the documents in English language and they are producing summaries with satisfactory accuracy. But in Indian languages, there is no accurate and complete document summarization system to produce the summary. Therefore, developing an automated text summarization system for Indian languages can help readers understanding large documents and provides the essential information about the document content in less time. We proposed an approach to extractive text summarization in one of the Indian languages –Kannada. Kannada language is spoken mainly in the Karnataka state of country India. The work presented in this paper generates the extractive summary of Kannada text document using Sentence features.

The flow of the paper is ordered as follows: We discuss the developed text summarization method for Kannada documents in section II. The experimental results obtained by the proposed summarization method are illustrated in section III. Finally, conclusion of the paper and directions for future work is briefed in section IV.

II. PROPOSED METHOD

Over the past years, only little research works has been carried out to solve the problem of Kannada text document summarization in Natural Language Processing [1][2][3][4][5][6][7][8]. Our work proposed the Kannada document summarizer, an application of Natural Language Processing (NLP) extracts the important information from the text document. There are mainly two techniques in automatic summarization - text extraction and text abstraction. The extraction method produces summary by selecting the important words, phrases or sentences from the input document. An abstraction method creates the summary by adding some new words that are not present in the input document. There are mainly three basic steps to produce the summary: pre-processing, feature extraction and the summary generation [9]. The proposed automatic text summarization model for Kannada documents
is illustrated in Fig. 1.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Arpitha Swamy*, Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, JSS Science & Technology University, Mysuru, India**.**
Srinath S, Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, JSS Science & Technology University, Mysuru, India.

A. Pre-processing

In the process of Kannada document summarization, some pre-processing operations are carried out before the sentence scoring algorithm is executed. The pre-processing function prepares the document for ranking of sentences and the generation of summary.

The pre-processing operations performed on the documents are:

a) Tokenization – A document is the arrangement of sentences and every sentence consists of group of words. Each word is treated as a token. Tokenization splits the document into sentences and then sentences into individual words.

b) Stop words removal – The commonly occurring words are called stopwords, which have less importance in the conclusion of document, are discarded for summarization process. In Kannada words like ಮತ್ತು (and), ಅದು (it), ಇದೆ (is), etc. are frequently used stopwords in sentences.

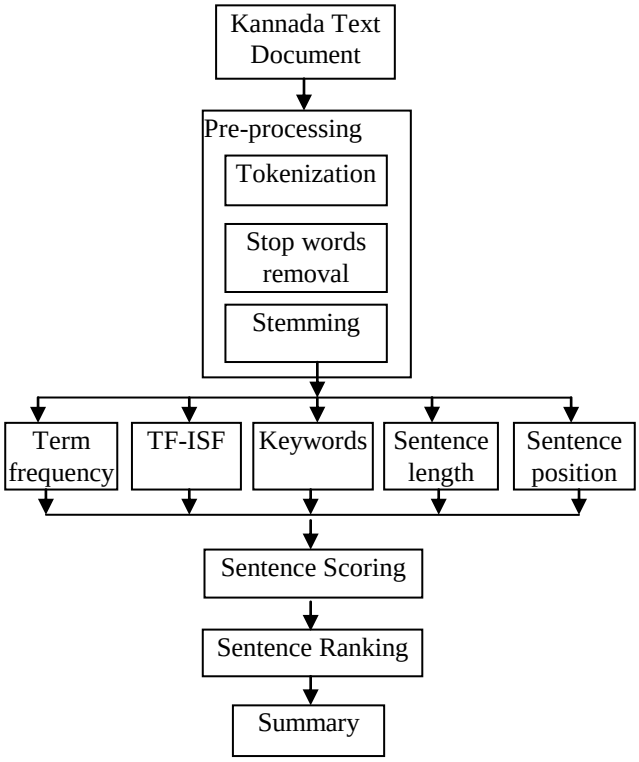

Fig. 1: Text summarization process in the proposed system

c) Stemming – The document may contain many words with same root but in different forms. All such words are converted to their root form for ease. Transformation of words into their canonical forms is done by the stemming algorithm. For example, the words ಆನೆಗೆ ಆನೆಯ ಆನೆಗಳ ಆನೆಯನ್ನು ಆನೆಗಳಿಗೆ should be converted into their original form ಆನೆ. All inflected words are converted into their root form through stemming operation using a predefined suffix list in this work.

B. Feature Extraction & Sentence Scoring

After an input document is pre-processed, the sentences in the document are assigned scores based on five important features: Term frequency, Term frequency-Inverse sentence frequency, Keywords, Sentence length and Sentence position in the document.

a) Term Frequency (TF) – Term Frequency is a measure to determine how frequently a term appears in the document. The term frequency is calculated as in (1)

$$TF(t) = N_t / TW \quad (1)$$

Where N_t represents the count of a term t occurring in the document and TW is the total count of terms present in the document.

The score of each sentence is computed as the sum of the value of the term frequency of all word present in that sentence.

b) Term Frequency-Inverse Sentence Frequency (TF-ISF) – This numerical statistic determines the importance of a word in the sentence of the document. The value of TF-ISF increases proportionally with the number of times that a word occurs in a sentence, but is offset by the frequency of the word in the document. This measure helps to control the fact that some words are more frequent than others. The inverse sentence frequency is calculated as in (2)

$$ISF(t,d) = log(N/ n_t) \quad (2)$$

Where, n_t is the total count of sentences containing the word w, N is the total count of sentences in the document d.

The term frequency-inverse sentence frequency (TF-ISF) feature score is given by combining the term frequency and inverse sentence frequency as in (3)

$$TF\text{-}ISF(t) = TF(t)*ISF(t,d) \quad (3)$$

The value of each sentence is calculated as the sum of the tf-isf scores of every word present in that sentence.

c) Keywords – Keywords are the words that occur frequently in a document and they are probably associated with the subject of the document. The 10 percent of total words which are frequently occurring in the document are considered as keywords. The score for this feature is computed as the ratio of the total number of times a keyword appeared in a sentence over the total number of times a keyword appeared in the document as given in (4).

$$Keywords\ score = K_s / K_d \quad (4)$$

Where K_s represents the total number of times a keyword appeared in the sentence and K_d is total number of times a keyword present in the document.

The score for each sentence is computed as the sum of the scores of the keywords present in that sentence.

d) Sentence length –The sentence length feature is used to filter out the short sentences present in the document. The score for the sentence length feature is computed as the ratio of length of sentence to the length of the longest sentence in the document as in (5).

$$Sentence\ length = L_s / L_{ls} \quad (5)$$

Where L_s is the length of sentence and L_{ls} indicates the length of longest sentence in the document. The length of a particular sentence is given by counting the total number of words present in the sentence [10].

e) Sentence Position – The sentence position in the

document has a substantial effect on the document content. The positional value of sentences is determined by assigning the highest value to the first sentence of the document and assigning the lowest value to the last sentence of the document [11]. The sentence positional value is determined using the formula as in (6)

$$Sentence\ Position = (T_s - SP) / T_s \quad (6)$$

Where, T_s is the total count of sentences present in the document and SP represents the actual positional value of a sentence in the document.

For example, consider there are 5 sentences in the document, the score of this feature for each sentence is computed as

First sentence = (5-0) /5, second sentence = (5-1)/5, third sentence = (5-2)/5, fourth sentence = (5-3)/5 and fifth sentence = (5-4)/5.

The final score for every sentence s of the document is computed as the average of all the feature score values (Term frequency, Term frequency-Inverse sentence frequency, Keywords, Sentence length and Sentence position in the document) for that particular sentence and is given by (7)

$$Total\_Score(s) = [TF(t) + TF\text{-}ISF(t) + Keywords\ score + Sentence\ length + Sentence\ Position] / N_f \quad (7)$$

Where N_f is the total number of features used to score the sentence.

C. Sentence Ranking & Summarization

The sentences of the document are ranked based on the computed total scores of all sentences and the summary is created by selecting the n top ranked sentences from the document. The number of sentences (n) required in the summary is specified by the user. The sentences selected for the summary are reordered to retain the order same as in the original document.

D. Summarization Algorithm

The sentence features based summarization algorithm takes the Kannada text document as input and produces summary with required number of sentences as specified by the user.

Input: Kannada text document and the number of sentences (n) needed in the summary

Output: Extractive Summary with appropriate number of sentences

Start

Step 1: Read the input text document

Step 2: Split the document into sentences

Step 3: For each sentence

a) Tokenize the sentence into words

b) Remove stopwords

c) Join the remaining meaningful words into sentence

Step 4: Compute the values for the features - Term frequency, Term frequency-Inverse sentence frequency, Keywords, Sentence length and Sentence position for each sentence in the document as

a) Term frequency: $TF(t) = N_t / TW$

b) Term frequency-Inverse sentence frequency:

$TF\text{-}ISF(t) = TF(t)*ISF(t,d)$

c) Keywords: $Keywords\ score = K_s / K_d$

d) Sentence Length: $Sentence\ length = L_s / L_{ls}$

e) Sentence Position: $Sentence\ Position = (T_s - SP) / T_s$

Step 5: Find the total score for each sentence. It is computed as average of all feature scores for each sentence

$$Total\_Score(s) = TF(t) + TF\text{-}ISF(t) + Keywords\ score + Sentence\ length + Sentence\ Position / N_f$$

Step 6: Select the top n sentences with the highest scores

Step 7: Sort the sentences to preserve the order of selected sentences in the summary same as in the original document

Step 8: Extract the selected sentences from the original document to form summary.

Stop

III. RESULTS & DISCUSSION

The dataset is created by collecting 50 documents belonging to different categories from Kannada Webdunia website and documents are saved in the text files using the Unicode standard UTF-8 format. The five categories chosen are Astrology, Business, Cricket, Politics and Sandalwood. The statistics of dataset considering 10 documents in each category for evaluation is given in the Table I.

Table I: Dataset Statistics

Category	Total Number of Sentences	Total Number of Words
Astrology (A)	76	749
Business (B)	53	874
Cricket (C)	82	1093
Politics (P)	60	822
Sandalwood (S)	81	942

The proposed system is evaluated against documents of five different categories chosen from the dataset. Only one human summary for each document is considered for evaluation. The generated system summary is evaluated by comparing it to the human summary using the ROUGE toolkit [12]. There are different ROUGE measures - ROUGE1, ROUGE2, ROUGEL, and ROUGES etc. We have used ROUGE1 measure to evaluate the system generated summaries.

The generated system summaries are evaluated using ROUGE toolkit in terms of evaluation measures - Recall, Precision and F-score and their values for 10 documents in each category is as shown in the Table II, Table III and Table IV respectively. Precision can be defined as the ratio of count of common sentences present in both system and model summaries over the total count of sentences present in the system summary. Recall is defined as the ratio of number of common sentences present in both system and model summaries and the total count of sentences present in the model summary. F-score is defined as a composite measure that combines recall and precision. It is calculated as the harmonic average of recall and precision. The notation D1, D2 ...D10 in the table represents the document numbers.

Table II: ROUGE Recall Scores for 10 documents in five categories

Article	Categories				
	A	B	C	P	S
D1	0.8	0.54	1.00	0.56	0.53
D2	0.82	0.63	0.78	0.13	0.23
D3	0.45	0.75	0.30	0.46	0.27
D4	0.72	0.60	0.62	0.80	0.50
D5	0.61	0.58	1.00	0.71	0.78
D6	0.21	0.66	0.49	0.50	0.72
D7	0.34	1.00	0.55	0.43	0.78
D8	0.55	0.61	0.79	0.63	0.50
D9	0.57	0.14	0.64	0.56	1.00
D10	0.77	1.00	1.00	1.00	0.14
Average	**0.58**	**0.65**	**0.72**	**0.58**	**0.55**

Table III: ROUGE Precision Scores for 10 documents in five categories

Article	Categories				
	A	B	C	P	S
D1	1.00	0.68	0.52	0.44	0.51
D2	1.00	1.00	1.00	0.07	0.35
D3	0.64	0.47	0.55	0.44	0.63
D4	0.60	0.46	0.62	1.00	0.61
D5	1.00	0.50	1.00	1.00	1.00
D6	0.36	1.00	0.52	0.46	1.00
D7	0.46	0.63	0.62	0.59	1.00
D8	0.38	0.56	0.69	0.65	0.53
D9	0.70	0.12	1.00	0.45	1.00
D10	0.61	1.00	1.00	1.00	0.09
Average	**0.67**	**0.64**	**0.75**	**0.61**	**0.67**

Table III: ROUGE F-score values for 10 documents in five categories

Article	Categories				
	A	B	C	P	S
D1	0.89	0.60	0.68	0.50	0.52
D2	0.90	0.77	0.87	0.09	0.28
D3	0.53	0.58	0.39	0.45	0.38
D4	0.65	0.52	0.62	0.89	0.55
D5	0.75	0.54	1.00	0.83	0.88
D6	0.27	0.79	0.51	0.48	0.84
D7	0.39	0.77	0.58	0.50	0.87
D8	0.44	0.59	0.74	0.64	0.52
D9	0.63	0.13	0.78	0.50	1.00
D10	0.68	1.00	1.00	1.00	0.11
Average	**0.61**	**0.63**	**0.72**	**0.59**	**0.59**

The average values of recall, precision and f-score for all documents in each category is as shown in the Fig. 2. The proposed method has good precision scores in all categories.

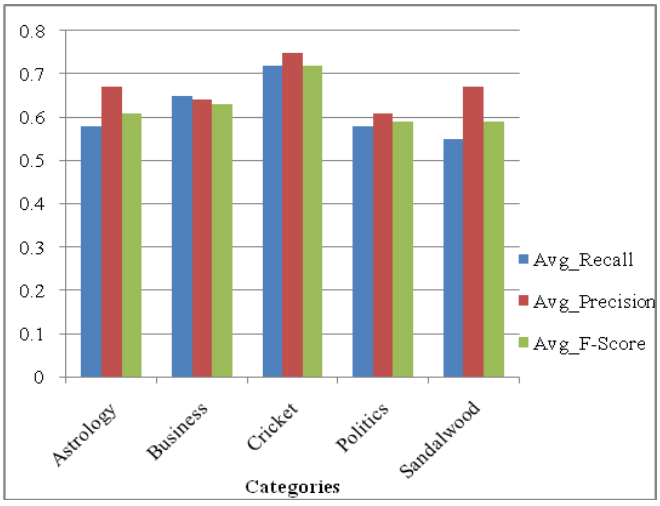

Fig. 2: Average values of recall, precision and f-score for each category of documents

IV. CONCLUSION & FUTURE WORK

An extraction based Kannada text summarization method for a single document is discussed in this paper. The different features used to score sentences are Term frequency, Term frequency-Inverse sentence frequency, Keywords, Sentence length and Sentence position in the document. The generated summaries are evaluated using ROUGE toolkit with recall, precision and f-score evaluation measures. The performance of this proposed system is good in terms of average recall, average precision and average f-score values. In the future, the performance of the proposed system may further be enhanced by considering more statistical and linguistic features in the process of sentence scoring and sentence ranking.

REFERENCES

1. Kallimani, J.S. and Srinivasa, K.G., 2010, August. Information retrieval by text summarization for an Indian regional language. In *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010)* (pp. 1-4). IEEE.
2. Jayashree, R., Murthy, S.K. and Sunny, K., 2011. Keyword extraction based summarization of categorized Kannada text documents. *International Journal on Soft Computing*, *2*(4), p.81.
3. Jayashree, R., Murthy, S. and Anami, B.S., 2012, November. Categorized Text Document Summarization in the Kannada Language by Sentence Ranking. In *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)* (pp. 776-781). IEEE.
4. Kallimani, J.S., Srinivasa, K.G. and Reddy, B.E., 2012. Summarizing news paper articles: experiments with ontology-based, customized, extractive text summary and word scoring. *Cybernetics and Information Technologies*, *12*(2), pp.34-50.
5. Jayashree, R., Murthy, K.S. and Anami, B.S., 2013, December. An artificial neural network approach to text document summarization in the Kannada language. In *13th International Conference on Hybrid Intelligent Systems (HIS 2013)* (pp. 45-48). IEEE.
6. Ranganatha, S., Vinay, S.K. and Bhargava, H.S., 2014. Federated Document Summarization Using Probabilistic Approach for Kannada
7.
8. Language. *International Journal of Innovative Research & Development*, *3*(1), pp.228-233.
9. Jayashree, R., Murthy, K.S. and Anami, B.S., 2014. Hybrid methodologies for summarisation of Kannada language text

documents. *International Journal of Knowledge Engineering and Data Mining*, *3*(1), pp.82-114.
10. Geetha, J.K. and Deepamala, N., 2015, August. Kannada text summarization using Latent Semantic Analysis. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1508-1512). IEEE.
11. Rautray, R., Balabantaray, R.C. and Bhardwaj, A., 2015. Document summarization using sentence features. *International Journal of Information Retrieval Research (IJIRR)*, *5*(1), pp.36-47.
12. Suanmali, L., Salim, N. and Binwahlan, M.S., 2009, May. Feature-based sentence extraction using fuzzy inference rules. In *2009 International Conference on Signal Processing Systems* (pp. 511-515). IEEE.
13. Efat, M.I.A., Ibrahim, M. and Kayesh, H., 2013, May. Automated Bangla text summarization by sentence scoring and ranking. In *2013 International Conference on Informatics, Electronics and Vision (ICIEV)* (pp. 1-5). IEEE.
14. Lin, C.Y., 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

AUTHORS PROFILE

Arpitha Swamy was born in Mandya, Karnataka, India in 1989. She pursued Bachelor of Engineering degree in Computer Science and Engineering in 2010 and recieved Master of Technology degree in Computer Network Engineering in 2012 from Visvesvaraya Technological University, Belgaum, Karnataka, India. She is currently pursuing Ph.D. degree in Computer Science and Engineering from JSS Science & Technology University, Mysuru, Karnataka, India.

From 2011 to 2012, she was an Intern in Citrix R&D India Private Limited, Bangalore, Karnataka. She joined as a Lecturer in Government Polytechnic College under Department of Technical Education, Bangalore, Karnataka in 2012. She has 6 years of teaching experience and 1 year of research experience.

Ms. Arpitha Swamy is a member of Indian Society for Technical education (ISTE), New Delhi, India from 2018. She has presented a conference paper in IEEE International Conference on New Trends in Engineering & Technology. Her research area includes Information Retrieval, Natural language processing mainly working on Text summarization problem.

Dr. Srinath S was born in Chamrajanagar, Karnataka, India in 1974. He completed Bachelor of Engineering degree in Computer Science and Engineering in 1995, pursued Master of Technology degree with 1st Rank in 2002 and received Ph.D. degree from Mysore University in 2015.

He is an Associate Professor in the department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, JSS Science and Technology University, Mysuru, Karnataka, India. He has organized many training programs which were sponsored by different government agencies. He has also carried out the research projects sponsored by the VTU, AICTE and MHRD (Government of India).

Dr.Srinath S has more than 20 scientific publications in national and international journals including conference proceedings. His research area includes Pattern Recognition, Natural language processing and Image Processing.