Comparative Study of Text Summarization in Indian Languages

Dhanya P.M
Department of Computer Applications
Cusat
Kochi

Jathavedan M
Department of Computer Applications
Cusat
Kochi

ABSTRACT
Text Summarization has been an area of interest since many years .There are a lot of summarization methods in foreign languages like English, Chinese [9], [10], Arabic[11], Korean [12], Persian [13] etc. Recently some methods have been developed for Indian languages also. This paper presents a comparison of various text summarization methods seen in Indian languages. Summarization techniques in Tamil, Kannada, Odia, Bengali, Punjabi and Gujarathi are taken for the purpose of comparison. Sample text consisting of three sentences is taken as an example and we try to find out the summary sentences using all the eight methods.

General Terms
Summarization. Document.

Keywords
Feature, Weight, Score.

1. INTRODUCTION
The world of documents containing text is huge and expanding every day in the World Wide Web. The majority of the data is in the form of natural language text. Most often we keep on reading lengthy documents to get some useful information. The need for an automatic Text summarizer has increased much due to the abundance of documents in the internet. Text summarization is a method of representing huge document/documents in a condensed form without affecting the meaning of the text .So it is a method of presenting the most important content of a document to the reader from a set of unstructured document/documents. The data available in Internet is unstructured when compared to other databases. So retrieving useful information itself is a challenge. The main objective of Text summarization is to identify the main theme of the document and the sub themes in it and to create a concise matter. This helps to save time, and storage space**.**

2. TEXT SUMMARIZATION IN INDIAN LANGUAGES
Languages in India can be divided into Indo-Aryan languages and Dravidian languages. Indo-Aryan languages include Hindi-Urdu, Assamese, Bengali, Gujarati, Marathi, Punjabi, Rajasthani , Sindhi, Oriya etc. Dravidian languages include languages like Malayalam, Tamil, Telugu, Kannada etc. Though Malayalam and Telugu are Dravidian in origin, over eighty percent of their lexicon is borrowed from Sanskrit. Works in Text summarization is done in Indian languages like Tamil, Kannada, Gujarathi, Punjabi, Bengali, Oriya etc.

2.1 Summarization in Tamil
2.1.1 Method 1
The method[1] proceeds by creating a complete graph of the entire document, where each vertex represents a sentence and edges show the connectivity between sentences. Vertices of the graph are first marked with sentence weights SW_i and edges are marked with Levenshtein similarity weights LSW_i. Average of all levenshtein similarity weights of all edges connected to a vertex is calculated to find out the vertex weights VW_i. The average of sentence weight and vertex weight is calculated to find out the rank of a sentence $Rank_i$. Sentence weight SW_i 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/ total number of words in the document. Levenshtein similarity weight between two sentences is calculated as

(Max length of two sentences- Levenshtein distance of two sentences) / Max length of two sentences.

Example: - Consider the following paragraph under the topic Indian railway as an example and calculate the affinity weights of all the words.

2.1.1.2 Indian Railway
[1]The railways in India are the largest rail web in Asia and the world's second largest under one management. [2] However officially, the first train in India (and in Asia) was flagged off on April 16, 1853, a Saturday, at 3:35 pm between Mumbai and Thane, a distance of 34 kms. [3] The importance of the day can be gauged from the fact that the Bombay government declared the day as a public holiday.

For simplicity the sentences have been numbered. The distinct words in the above text are "railway", "India", "largest", "rail", "web", "Asia", "world", "second", "management", "officially", "train", "flagged", "April", "Saturday", "Mumbai", "Thane", "distance", "importance", "day", "gauged", "fact", "Bombay", "Government", "declared", "public", "holiday". There are total 26 words in the above text .The affinity weights of each and every word are obtained as follows. AW ("India") = 1/ 26, where 1 is the number of additional occurrences of the word "India". AW ("Asia") = 1/26, AW (largest) = 1/26. For all other words AW is obtained

to be zero. Now the Sentence weights for all the three sentences can be calculated as

SW (1) = 1/26 + 1/26 + 1/26 = 3/26 = 0.12.

SW (2) = 1/26 + 1/26 = 2/26 = 0.08 and

SW (3) = 0.

Now we plot the complete graph of the document as

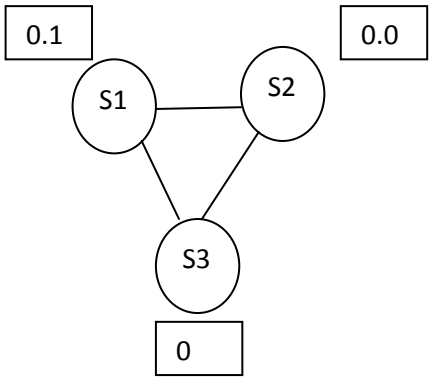

Fig 1: Graph of the above document

The number of words in sentences one, two and three are ten, eleven and ten, So their lengths are also ten, eleven and ten.

LSW (S1, S2) = Maxlen (S1, S2) - LD (S1, S2) / Maxlen (S1, S2) =11 - 9/ 11 = 0.18.

LSW (S1, S3) = 10-10/10 = 0

LSW (S2, S3) = 10-10/10 = 0

VW (1) = 0.18 + 0 = 0.18

VW (2) =0.18 + 0 = 0.18

VW (3) = 0+0 = 0

Rank 1= 0.12 + 0.18 = 0.30

Rank 2 = 0.08 + 0.18 = 0.26

Rank 3 = 0

Here the highly ranked sentence is Sentence 1.

2.1.2 Method 2

Here[2] also sentence scoring techniques are used which are then used to select the highly ranked sentences. A positional score is given to each and every sentence in the text, which can be calculated as $PS_i = (1 - Pos_i) / N$ where Pos_I is the position of the ith sentence in the text and N is the total number of sentences in the text. Take the text mentioned in method 1 and calculate the positional score of each and every sentence. $PS_1 = 1- 1/3 = 2/3$. $PS_2 = 1-2/3 = 1/3$. $PS_3 = 1-3/3 = 0$. Paragraph scoring of the sentences is done using the formula $Para\ PS_i$ = 1-Sentence position in the paragraph/Number of sentences in the paragraph. Since the above text has only one paragraph, the paragraph scoring has the line score values. Now each sentence is given a length score as number of words in each sentence/ Total number of words in the text. Thus we get,$LS_1 = 10 / 26$. $LS_2 = 11 / 26$ and $LS_3 = 10 / 26$.

The surface score of the sentences is calculated as the sum of all the three scores discussed above. $Surf_1 = 2/3 + 2/3 + 10/26 = 1.71$. $Surf_2 = 1/3 + 1/3 + 11/26 = 1.09$. $Surf_3 = 0 + 0 + 10/26 = 0.38$.Term frequency score of a sentence is calculated as sum of the frequencies of all the words in a particular sentence / Total number of words in a sentence. Using this formula we get $TFS_1 = 8 + 2 / 10 = 1$. $TFS_2 = 11/11 = 1$. $TFS_3 = 8 + 2/ 9 = 10/10 = 1$. Now calculate the Topic Similarity Score as number of common words in the topic and the Sentence / log (number of words in the sentence) + log (number of words in the topic). Thus obtain $TSS_1 = 2/ (\log 10 + \log 2) = 1.54$.$TSS_2 = 1 / (\log 11 + \log 2) = 0.76$. $TSS_3 = 0 / (\log 10 + \log 2) = 0$.Now calculate the intermediate score of each and every sentence $IS_1 = Surf_1 + TFS_1 + TSS_1 = 1.71 + 1 + 1.54 = 4.25$. $IS_2 = 1.09 + 1 + 0.76 = 2.85$. $IS_3 = 0.38 + 1 + 0 = 1.38$.Here the first and the second sentences are highly scored sentences than the third sentence. Final Page Rank Score is calculated using the Page Rank Formula. The following table shows the values of all the parameters and the final score of a sentence.

Table 1. Features and their values

	Features						
	PS_i	Para PS_i	LS_i	$Surf_i$	TFS_i	TSS_i	IS_i
S_1	0.67	0.67	0.38	1.71	1	1.54	4.25
S_2	0.33	0.33	0.42	1.09	1	0.76	2.85
S_3	0	0	0.38	0.38	1	0	1.38

2.2 Kannada

2.2.1Method 1

In Kannada[3] ,sentences are scored based on line score and sentence score. Line score = 1/ line number x 10, So we get $Line\ score_1 = 1/ 1 \times 10 = 10$, $Line\ score_2 = ½ \times 10 = 5$. $Line\ score_3 = 1/3 \times 10 = 3.33$. Sentences with numerical values get more score, so more score for sentence two. Sentences which contain keywords are given high scores and so all the sentences get high scores. Now calculate the sentence score as the sum of word scores. Thus $Sentence\ score_1$ = railway(1) + India(2) + largest(2) + rail(1)+ web(1) + Asia(1) + Second (1) + world(1) + management (1) = 11 . $Sentence\ score_2$ = officially (1) + train (1) + India (2) + Asia (2) + flagged (1) + April (1) + Saturday (1) + Mumbai (1) + thane (1) + distance (1) = 12. $Sentence\ score_3$ = importance (1) + gauged (1) + day (2) + fact (1) + Bombay (1) + government (1) + public (1) + holiday (1) = 9. The above line score and the sentence score are combined together to get the final score of a sentence. So in this method first two sentences score more.

Table 2: Features and their values

	Features				
	Line	**Numeri**	**Keyword**	**Sentence**	**Final**

	score	cal values	score	score	score
S_1	10	4	10	11	35
S_2	5	0	11	12	28
S_3	3.33	0	10	9	22.33

2.2.2 Method 2

In method 2[4] , two word lists namely GSS list and TFIDF lists are maintained and the sentences which contain those words are scored more. Calculate the TFIDF of all the words and create the list. TF= frequency of a term in a document / number of terms in a given document.IDF= Log 10 (N / n) where, N is the total number of documents.

TF(India)=2/26
TF(largest)=2/26
TF(web)=1/26
TF(Asia)=2/26 and
TF(rail)= 1/26

and so on. So words with more TF scores are India, Asia and largest. Since there is only one document, there is no importance for IDF values .Sentences which contain these TF words will be scored more and so we select the first two sentences as the summary. The second sentence score more.

Table 3: TF scores

	Sum of TF scores
Sentence $_1$	0.46
Sentence $_2$	0.50
Sentence $_3$	0.38

2.3 Odia Summary

According to this method[5] weight of a word $_i$ = frequency of the term/ total number of terms in the document. The weight of a sentence will be the sum of the weights of all the words in the sentence / number of terms in the sentence. Again take the paragraph mentioned above as an example. Weight (India) = 2/ 26, weight (Asia) = 2/26, Weight (Largest) = 2/26 and weight (day) = 2/26. All other words have the weight 1/26. Weight of first sentence = 0.46. Weight of the second sentence = 0.50.Weight of the third sentence = 0.383 .The results show that the first two sentences as the highly scored ones and the second sentence gets the highest score. The results are same as that of the Kannada method 2.

2.4 Bengali Summarization

In Bengali summarization[6] thematic terms and positional scores are considered. So first calculate the TF scores of all the words in the sentences .Thus obtain the weight of first sentence as 0.46, 0.50, and 0.383. Here second sentence get the highest score. Positional score is calculated as 1/√line number. Positional score of the first sentence = 1/√1=1, where as for the second and third are 1/√2, 1/√3 respectively. The final score is obtained as

$\alpha \times S_k + \beta \times P_k$.

2.5 Punjabi Summarization

In Punjabi[7] many features are used. Sentence length = number of words / number of words in the maximum length sentence. The values for the first, second and third sentences are 10/11, 11/11 and 10/11. Thus second sentence gets the maximum score. Considering the number feature second sentence get the maximum score.After finding the feature values for all other features ,the results show , the maximum weight sentence as sentence one or sentence two. Since the method uses a regression model direct results with the above sample document is not possible. This is the only method which uses weight learning.

2.6 Gujarathi Summarization

The method[8] employs calculation of Information content of a sentence which uses two constants α and β which depend on the corpus, so direct results are not possible in this case. The method also deals with coherence between the sentences. If there is less coherence between two sentences in the summary, then sentence which lies between them in the original text is also added to the summary in order to increase the coherence. This step is not seen in other techniques.

3. COMPARISON OF VARIOUS TECHNIQUES

From the above sections it can be seen that Text summarization is implemented only in certain Indian languages and works are yet to be produced in other languages. Most of the methods have selected a set of features based on which they rank the sentences. Even though some of the features are common, they have selected some features specific to the language like the Punjabi English Noun feature. Now obtain a graphical comparison of the set of features being used. The graph in Fig 2: shows the number of features used by various languages for text summarization. Punjabi method uses the maximum number of features which is ten and odia uses the least number of features which is one. The accuracy of the method depends on the number of features and the contribution of that feature towards summary. The above discussed methods shows a recall scores of 0.45, 0.48, 0.43, 0.66, 0.42, 0.412, 0.42, 0.82 respectively. A detailed comparison is shown in Table 4.

Table 4: Comparison of techniques

Indian languages			
language	**Features**		
Tamil Method 1	Text ranking based on parameters	Parameters used are affinity weight, sentence weight, levenshtein similarity	Unsupervised graph based method
Tamil Method 2	Text ranking	Parameters used are position, length, topic similarity, term frequency etc.	Unsupervised non graph based method
Kannada method 1	Text ranking	Parameters used are position score, numerical values, keywords	Supervised method-uses a dictionary of keywords
Kannada method 2	Text ranking	Parameters used are TF-IDF GSS coefficients	Multi document,Documents in four different categories are considered
Odia	Text ranking	Frequency of the term	Single document, non graph based
Bengali	Text ranking	Thematic terms whose TFIDF > threshold, sentence length.	Unsupervised method and non graph based
Punjabi	Text ranking	Sentence length, TF-ISF, Number feature, English Punjabi Nouns, Cue phrases, Title keyword	Weight learning using regression
Gujarathi	Theme feature vector	Finds Information content of a sentence	Unsupervised and non graph based.Deals with low coherence

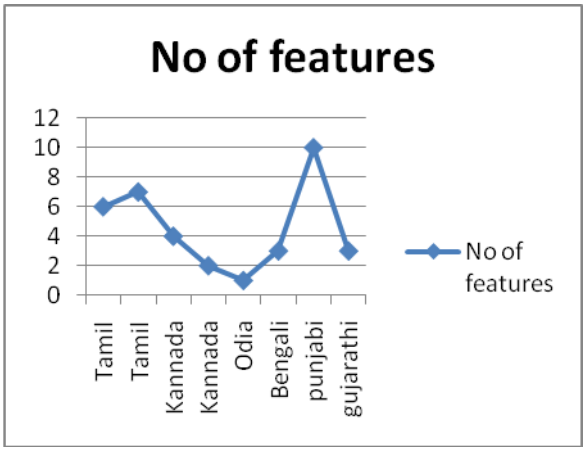

Fig 2: Features

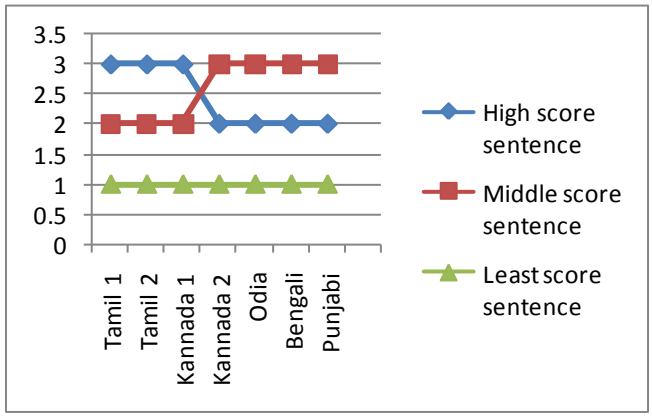

Fig 3: Sentence scoring

4. SCORE COMPARISON

Now compare the results of various methods discussed above. From the graph in Fig 3,it is seen that two methods in Tamil and Kannada method 1 select the first sentence as the highest scored one among the sentences taken to the summary, while Kannada method 2 , Odia , Bengali and Punjabi selects the methods create the summary with the first two second sentence as the maximum scored one. All the methods give same results in the selection of least scored sentences and all the sentences in the sample paragraph.

5. CONCLUSION

Here eight different languages are selected and their summarization methods are compared. It is noticed that the methods are not language specific. So the same set of sentences in English are used for comparing all the methods. Most of the methods are based on scoring sentences depending on features .Feature selection plays a key role in the summary generation. Not all the features are given equal weight. The weight of a feature depends on the contribution of the feature towards the summary. In almost all methods

testing is done by comparing the results with results of human summarizers.

6. REFERENCES

[1] Sankar K, Vijay Sundar Ram R and Sobha Lalitha Devi, Text Extraction for an Agglutinative Language , Problems of Parsing in Indian Languages, M a y 2 0 1 1 Special Volume.

[2] Krish Perumal, Bidyut Baran Chaudhuri, Language Independent Sentence Extraction Based Text Summarization, Proceedings of ICON-2011: 9th International Conference on Natural Language Processing.

[3] Jagadish S KALLIMANI, Srinivasa K, G**,** Information Retrieval by Text Summarization for an Indian Regional Language, 2010 IEEE .

[4] Jayashree.R1, Srikanta Murthy.K2 and Sunny.K1, Document summarization in kannada using keyword extraction, CS & IT-CSCP 2011.

[5] R. C. Balabantaray, B. Sahoo, D. K. Sahoo, M. Swain , Odia Text Summarization using Stemmer, International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868, Volume 1– No.3, February 2012

[6] Kamal Sarkar Bengali text summarization by sentence extraction, Proceedings of International Information Management(ICBIM-2012),NIT Conference on Business and Durgapur, PP 233-245.

[7] Vishal Gupta, Gurpreet Singh Lehal, Features Selection and Weight learning for Punjabi Text Summarization, International Journal of Engineering Trends and Technology- Volume2 Issue2- 2011.

[8] Alkesh Patel, Tanveer Siddiqui, U. S. Tiwary , A language independent approach to multilingual text summarization, RIAO2007, Pittsburgh PA, USA, May 30- June 1(2007).

[9] Lei Yu1, Jia Ma1, Fuji Ren1,2, Shingo Kuroiwa1, Automatic Text Summarization Based on Lexical Chains and Structural Features, Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing,0-7695-2909-7/07 ,2007 IEEE.

[10] Xinlai Tang, Study on Chinese Text Summarization Based on Extracting Sentences from Subtopics, World Congress on Software Engineering,, 256 – 259,Volume-1,2009.

[11] Aqil M. Azmi, Suha Al-Thanyyan, A text summarizer for Arabic, Computer Speech and Language, Pages 260-273 ,Volume 26 Issue 4, August 2012 .

[12]Jae-HoonKim,Joon-HongKim, Korean text summarization using an aggregate similarity**,**In proceedings of IRAL '00 Proceedings of the fifth international workshop on on Information retrieval with Asian languages , Pages 111-118,2000.

[13] Mehrnoush shamsfard, Tara akhavan, parsumist: A Persian Text Summarizer, 978-1-4244-4538-7/09/$25.00 ©2009 IEEE.