KEYWORD EXTRACTION BASED SUMMARIZATION OF CATEGORIZED KANNADA TEXT DOCUMENTS

Jayashree.R[1], Srikanta Murthy.K[2] and Sunny.K[1]

[1]Department of Computer Science, PES Institute Of Technology, Bangalore, India
`jayashree@pes.edu, sunfinite@gmail.com`
[2]Department of Computer Science, PES School Of Engineering, Bangalore, India
`srikantamurthy@pes.edu`

ABSTRACT

The internet has caused a humongous growth in the number of documents available online. Summaries of documents can help find the right information and are particularly effective when the document base is very large. Keywords are closely associated to a document as they reflect the document's content and act as indices for a given document. In this work, we present a method to produce extractive summaries of documents in the Kannada language, given number of sentences as limitation. The algorithm extracts key words from pre-categorized Kannada documents collected from online resources. We use two feature selection techniques for obtaining features from documents, then we combine scores obtained by GSS (Galavotti, Sebastiani, Simi) coefficients and IDF (Inverse Document Frequency) methods along with TF (Term Frequency) for extracting key words and later use these for summarization based on rank of the sentence. In the current implementation, a document from a given category is selected from our database and depending on the number of sentences given by the user, a summary is generated.

KEYWORDS

Summary, Keywords, GSS coefficient, Term Frequency (TF), IDF (Inverse Document Frequency) and Rank of sentence

1. Introduction

With the growth of the internet, a large amount of data is available online. There is a demanding need to make effective use of data available in native languages. Information Retrieval [IR] is therefore becoming an important need in the Indian context. India is a multilingual country; any new method developed in IR in this context needs to address multilingual documents. There are around 50 million Kannada speakers and more than 10000 articles in Kannada Wikipedia. This warrants us to develop tools that can be used to explore digital information presented in Kannada and other native languages. A very important task in Natural Language Processing is Text Summarization. Inderjeet Mani provides the following succinct definition for summarization is: *take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's application needs*[14].There are two main techniques for Text Document Summarization: extractive summary and abstractive

summary. While extractive summary copies information that is very important to the summary, abstractive summary condenses the document more strongly than extractive summarization and require natural language generation techniques. Summarization is not a deterministic problem, different people would chose different sentences and even the same person may chose different sentences at different times, showing differences between summaries created by humans. Also, semantic equivalence is another problem, because two sentences can give the same meaning with different wordings. In this paper, we present an extractive summarization algorithm which provides generic summaries. The algorithm uses sentences as the compression basis. Keywords/phrases, which are a very important component of this work, are nothing but expressions; single words or phrases describing the most important aspects of a given document. The list of keywords/phrases aims to reflect the meaning of the document. Guided by the given keywords/phrases, we can provide a quick summary, which can help people easily understand what a document describes, saving a great amount of time and thus money. Consequently, automatic text document summarization is in high demand. Meanwhile, summarization is also fundamental to many other natural language processing and data mining applications such as information retrieval, text clustering and so on [11][2].

2. Literature Survey

Previous work on key phrase extraction by Letian Wang and Fang Li [3] has shown that key phrase extraction can be achieved using chunk based method. Keywords of document are used to select key phrases from candidates. Similarly, another approach by Mari-SannaPaukkeri et al[2] selects words and phrases that best describe the meaning of the documents by comparing ranks of frequencies in the documents to the corpus considered as reference corpus. The SZETERGAK system by Gabor Berend[1] is a frame work that treats the reproduction of reader assigned keywords as a supervised learning task. In this work, a restricted set of token sequences was used as classification instances. One more method of You Ouyang[4] extracted the most essential words and then expanded the identified core words as the target key phrases by word expansion approach. A novel approach to key phrase extraction proposed by them consists of two stages: identifying core words and expanding core words to key phrases. The work of automatically producing key phrases for each scientific paper by Su Nam Kim et al[5] has compiled a set of 284 scientific articles with key phrases carefully chosen by both their authors and readers, the task was to automatically produce key phrases for each paper. FumiyoFukumoto[6] present a method for detecting key sentences from the documents that discuss the same event. To eliminate redundancy they use spectral clustering and classified each sentence into groups each of which consists of semantically related sentences. The work of Michael .J . Paul et al[7] use an unsupervised probabilistic approach to model and extract multiple viewpoints in text. The authors also use Lex rank, a novel random walk formulating to score sentences and pairs of sentences from opposite view points based on both representativeness of the collections as well as their contrast with each other. The word position information proves to play a significant role in document summarization. The work of You Ouyang [8] et al illustrates the use of word position information, the idea comes from assigning different importance to multiple words in a single document .Cross Language document summary is another upcoming trend that is growing in Natural Language Processing area, wherein the input document is in one language , the summarizer produces summary in another language. There was a proposal by Xiaojun Wan et al [9] to consider the translation from English to Chinese. First the translation quality of each English sentence in the document set is predicted with the SVM regression method and then the quality score of each sentence is incorporated into the summarization process; finally English sentences with high translation scores are translated to form the Chinese summary. There have

been techniques which use A* algorithm to find the best extractive summary up to given length, which is both optimal and efficient to run. Search is typically performed using greedy technique which selects each sentence in the decreasing order of model score until the desired length summary is reached [10]. There are two approaches to document summarization, supervised and unsupervised methods. In supervised approach, a model is trained to determine if a candidate phrase is a key phrase. In unsupervised method graph based methods are state-of-the art. These methods first build a word graph according to word co occurrences within the document and then use random walk techniques to measure the importance of a word [12].

3. Methodology: The methodology adopted by us can be described in four major steps:

3.1. Crawling

The first step is creating the Kannada dataset. Wget , a Unix utility tool was used to crawl the data available on **http://kannada.webdunia.com**. Data was pre-categorized on this web site.

3.2. Indexing

Python was the language of choice. The indexing part consisted of removing HTML mark up; English words need not be indexed for our work. Beautiful Soup is a python HTML/XML parser which makes it very easy to scrape a screen. It is very tolerant with bad markup. We use Beautiful Soup to build a string out of the text on the page by recursively traversing the parse tree returned by Beautiful Soup. All HTML and XML entities (ಅ :ಅ , < : <) are then converted to their character equivalents. Normal indexing operations involve extracting words by splitting the document at non-alphanumeric characters, however this would not serve our purpose because dependent vowels (ಾ, ಾ etc.) are treated as non-alphanumeric, so splitting at non-alphanumeric characters would not have worked for tokenization. Hence a separate module was written for removing punctuations.

3.3. Keyword Extraction

Documents in five categories were fetched: sports, religion, entertainment, literature, astrology. The next step is to calculate GSS coefficents and the Inverse Document Frequency (IDF) scores for every word (in a given category in the former case) in a given document belonging to these categories. Every word in a given document has a Term Frequency (TF), which gives the number of occurrence of a term in a given document, defined by:

TF= frequency of a term in a document / number of terms in a given document.

Similarly, the IDF of a document is given by the formula,

IDF=Log_{10}(N/n), where 'N' is the total number of documents indexed across all categories. and 'n' is the number of documents containing a particular term.

Hence TF and IDF are category independent. The GSS coefficients which evaluate the importance of a particular term to a particular category are calculated. GSS(Galavotti-Sebastiani-

Simi) co-efficient [13] is a feature selection technique used as the relevance measure in our case. Given a word w and category c it

is defined as:

f(w , c)= p(w , c)* p(w' , c') -p(w ', c) * p(w , c')

where, p(w, c) is the probability that a document contains word 'w' and belongs to category 'c'

p(w' , c') is the probability that a document does not contain 'w' and does not belong to 'c'

p(w' , c) is the probability that a document does not contain 'w' and belongs to 'c'

p(w , c') is the probability that a document contains 'w' and does not belong to 'c'.

GSS coefficients give us words which are most relevant to the category to which the documents belong. IDF gives us words which are of importance to the given documents independently. Thus using these two parameters to determine relevant parts of the document helps in providing a wholesome summary**.**

3.4. Summarization

Given a document and a limit on the number of sentences, we have to provide a meaningful summary. We calculate the GSS coefficients and IDF of all the words in the given document. If the document is already present in our database, GSS coefficients and IDF values are already calculated offline. These values are then multiplied by the TF of the individual words to determine their overall importance in the document. We then extract top n keywords from each of the lists (GSS coefficients and IDF). Then sentences are extracted from the given document by retrieving Kannada sentences ending with full stops. Due care is taken to see that full stops which do not mark the end of a sentence (ಡಹ. etc.) are not considered as split points. Each of these sentences is then evaluated for the number of keywords it contains from the GSS and IDF lists as follows:

Rank of sentence = Number of keywords contained by the sentence from both the lists

_____

Total number of sentences in the document

The top 'm' sentences are then returned as document summary, where 'm' is the user specified limit.

The same concept can be extended to paragraphs to provide a possibly more meaningful context than individual sentences particularly in the case of large documents.

3.5. Tests

The following lists were obtained by running our algorithm with a sports article on cricket as input with n= 20:

GSS co-efficient list:

> ಐಪಿಎಲ್, ಆಧ್ಯತೆ, ನನ್ನ, ಆಡುವುದು, ದೇಶಕ್ಕಾಗಿ, ಕ್ರಿಕೆಟ್, ಉತ್ತಪ್ಪ, ಸುದ್ದಿಗಳಿಗೆ, ಲೀಗ್, ಪ್ರತಿನಿಧಿದೇಸುವುದೇ, ಪ್ರೀಮಿಯರ್, ನಿರ್ಣಾಯಕ, ತಂದದಲ್ಲಿ, ಇಂಡಿಯನ, ವಿಶ್ವಕಪ್, ಏಕದಿನ, ಕರ್ನಾಟಕ, ಭಾರತ, ನಲ್ಲಿ, ಸಂಭವನೀಯರ

IDF list:

> ಐಪಿಎಲ್, ಕ್ರಿಕೆಟ್, ವಿಶ್ವಕಪ್, ನಲ್ಲಿ, ಮತ್ತಪ್ಪ , ಮೊದಲ, ಭಾರತ, ಕ್ರೀಡಾ, ಜಗತ್ತು , ಲೇಖನಗಳು, ಕ್ರಿಕೆಟಿಗರು , ಏಕದಿನ, ಅಂಕಿಅಂಶ, ಟಿ , ಟಿಕರ್, ಶೋಧಿಸು, ಸಹ, ಇದನ್ನು, ಮುಖ್ಯ, ಪುಟ

The following lists were obtained with a blog entry about another blog on films (category: literature) as input with n = 20:

GSS co-efficient list:

> ಈ, ಲೇಖನ, ವೆಬ್, ಕವನ , ಬ್ಲಾಗ್ , ವಿವಿಧ , ಸಾಹಿತ್ಯ , ವಾರದ, ದುನಿಯಾ, ಅವರು, ನಮ್ಮ, ಮತ್ತು, ಎಂದು, ಕಥೆಗಳು , ಖ್ಯಾತ, ಪುಟ, ಸಾಹಿತಿಗಳು, ಅವರ, ನಾವು , ಎಂದು

IDF list:

> ಕೃಷ್ಣ , ಪಿಚ್ಚರ್ , ವೆಬ್ , ಬ್ಲಾಗ್ , ವಾರದ , ದುನಿಯಾ , ಇಲ್ಲಿ , ಲೇಖನ , ಸಿನಿಮಾ , ಅವರು , ಈ , ವಿಕಾಸ , ಮಚ್ಚು , ಕೌರ್ಯ , ಭೂಗತ , ಪತ್ರಕರ್ತ, ಬ್ಲಾಗಿನ , ಕಾಣುತ್ತಿದ್ದ , ಎತ್ತಿ , ಬ್ಲಾಗು

As evident , these lists contain stop words and hence, stop word (noise) removal is essential.

3.6. Noise Removal

Stop words which do not give meaning to the document considered noise and hence should not be evaluated as keywords. To remove stop words we have implemented an algorithm which takes a list of stop words prepared manually as input and finds structurally similar words and adds them to the stop word list.

Some of the words in our primary list of stop word which is created and maintained manually are:

> ನನ್ನ, ನಿನ್ನ, ಇದು, ಅದು, ಯಾಕೆ, ಹೇಗೆ, ಆದರೆ, ಮತ್ತು, ಹೋಗು, ನೀನು, ನೀವು , ಇತ್ತು , ಮಾಡು , ಈ , ಆ , ಅಲ್ಲಿ, ಎಲ್ಲಿ, ಹಾಗೂ, ಎಂಬ, ಅಥವಾ ,ನಲ್ಲಿ , ಇಲ್ಲ , ಬಾ , ಏನು, ಆಗದೆ , ತಾನು , ಇವರಿಗೆ , ಅಂದರೆ , ಈಗ , ಅಂಥ

3.6.1 Finding Structurally similar words for Primary list words

Example: Consider the word ' ಯಹಕ್ೆೀ '

When split into individual Unicode characters,

it becomes : ಯ (U+0CAF) + ◌ಾಹ (U+0CBE) + ಕ (U+0C95) +◌ಾ◌ೆ(U+0CC6). The vowel sound at the end is not considered as an alphanumeric character. So our similarity module does the following in order :

1. Fuzzy search for words which contain the unmodified word at the beginning.
2. Strip all the non-alphanumeric characters at the end of the word and then fuzzy search for words which contain the modified word at the beginning.

A sample of stop words that were obtained by our algorithm :

ಯಾಕೆಂದರೆ , ಯಾಕೋ, ಯಾಕಿಷ್ಟ , ಯಾಕ್,ಯಾಕಾಗಬಾರದು , ಯಾಕೂಬ್ etc.

For the personal pronoun 'ನನನ' , some of the words obtained are:

ನನ್ನಿಂದ , ನನ್ನನ್ನ, ನನ್ನವಳು, ನನ್ನಜ್ಞನಿಗೊಂದಾನೆಯಿತ್ತು , ನನ್ನೊಂದಿಗೇ, ನನ್ನಾಸೆ etc.

For the verb 'ಬರಿ', the similar words were:

ಬರಿಸುವ, ಬರಿಯ, ಬರಿದೆ, ಬರಿಗಾಲಲ್ಲೇ, ಬರಿಗಣ್ಣಿಗೆ, ಬರಿಗ್ಯೆಲಿ, ಬರಿಗಾಲಿನಲ್ಲಿ, ಬರಿಗಾಲಲ್ಲಿ, ಬರಿಗ್ಯೆಯಲ್ಲಿ, ಬರಿದಾಗಿದ, ಬರಿಸಿಕೊಳ್ಳಿ, ಬರಿಸುವುದಿಲ್ಲ, ಬರೆಯಲು, ಬರಹದ, ಬರೆದಿರುವ, ಬರುವೆನು, ಬರಬಹುದು, ಬರೆಯುವುದಿಲ್ಲ, ಬರೆದಿದ್ದ, ಬರೆಯಿರಿ, ಬರಲೇ, ಬರುವರು, ಬರಹಗಾರರಿಗಾಗಿ, ಬರುತೀರ, ಬರೆಯಲಾರರು, ಬರೆದುಕೊಂಡ, ಬರೆಯಬಲ್ಲರು, ಬರ್ತಿಲ್ಲ etc.

As evident, though some words have semantic relationship to the primary stop word, a lot of words have no such relationship and further work needs to be done to find methods which will prevent such words from being penalized as stop words. Starting with a basic list of stop words, this program can be used to find structurally similar words and semantically unrelated words can be manually removed from the stop words list.

3.7.Results

Evaluation-I gives the result of a manual evaluation of the summarizer with three different human summaries across various categories. Three different people were asked to create reference summaries for random documents in each category. The same documents were then fed to the program and the limit was kept at m=10.The number of sentences common between the two summaries gives the relevance score; the average of three scores is shown for each document.

Evaluation-II consisted of a single human reference summary for a data set consisting of 20

documents across different categories. This human reference summary was compared to the summary generated by the machine.For Evaluation III and Evaluation IV, we considered 20 documents, across categories: Literature, Entertainment, Sports, and Religion. Two independent reviewers were asked to summarize the documents (not Abstractive), the results are shown in Table 3 and Table 4.

Evaluation V and Evaluation VI given in Table 5 and Table 6 shows the number of common sentences between human summary and Machine summary. Wherein machine summary is generated for the same set of documents across same categories considered for Evaluation III and Evaluation IV. The maximum limit on the number of sentences to be given by the machine is the number of sentences obtained by human summarizers for a given document.

Evaluation VII and Evaluation VIII consists of the average of the human summary and machine summary, means the number of sentences common between the human summary and machine summary is taken into consideration and average is taken.

Table 1. Evaluation-I Results.

Category	Human 1 Score	Human 2 Score	Human 3 Score	Average
Literature	0.4	0.5	0.2	0.37
Entertainment	0.5	0.2	0.6	0.43
Astrology	0.8	0.7	0.7	0.73
Sports	0.4	0.4	0.4	0.40

Table 2.Evaluation-II Results.

Category	Score
Literature	0.7
Entertainment	0.8
Astrology	0.8
Sports	0.76

Table 3.Evaluation-III

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Literature	30/32	34/37	13/18	32/37	20/26	13/17	19/23	20/22	29/35	12/17
Entertainment	11/16	02/04	14/18	13/19	10/16	04/09	19/25	06/12	09/15	07/13
Sports	11/17	07/11	08/14	14/16	12/13	10/15	14/20	07/13	02/05	07/15
Religion	13/15	08/11	03/03	08/11	13/17	18/20	06/08	04/06	09/10	19/21

Table 4.Evaluation- IV

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Literature	07/12	12/14	25/28	01/01	19/22	31/35	19/24	19/22	17/20	11/14
Entertainment	06/11	04/08	05/08	25/27	01/03	06/12	06/12	06/12	09/14	03/08
Sports	20/22	11/17	04/09	10/16	05/09	10/22	10/19	13/28	41/50	42/52
Religion	11/13	06/08	10/13	01/01	10/12	16/20	12/15	06/07	01/01	12/13

Table 5.Evaluation –V

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Literature	03/07	06/12	18/25	01/01	11/19	27/31	11/19	11/19	10/17	08/11
Entertainment	03/06	02/04	02/05	18/25	01/01	02/06	02/06	02/06	03/09	01/03
Sports	16/20	08/11	02/04	04/10	01/05	06/10	06/10	09/13	32/41	34/42
Religion	08/11	02/06	04/10	01/01	04/10	10/16	08/12	02/06	01/01	05/12

Table 6.Evaluation VI

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Literature	03/07	05/12	18/25	01/01	10/19	22/31	10/19	10/19	11/17	07/11
Entertainment	02/06	01/04	01/05	17/25	01/03	02/06	02/06	02/06	02/09	01/03
Sports	15/20	08/11	01/04	05/10	01/05	05/10	05/10	07/13	32/41	35/42
Religion	08/11	01/06	04/10	01/01	04/10	10/16	05/12	01/06	01/01	05/12

Table 7. Evaluation VII

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Literature	0.42	0.50	0.72	100	0.81	0.87	0.81	0.81	0.58	0.72
Entertainment	0.50	0.50	0.4	0.72	100	0.34	0.34	0.34	0.34	0.34
Sports	0.8	0.72	0.50	0.40	0.20	0.6	0.6	0.70	0.78	0.80
Religion	0.72	0.34	0.4	100	0.4	0.62	0.67	0.34	100	0.42

Table 8. Evaluation VIII

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Literature	0.25	0.42	0.72	100	0.52	0.70	0.52	0.70	0.64	0.63
Entertainment	0.34	0.25	0.2	0.68	0.34	0.34	0.34	0.34	0.23	0.34
Sports	0.75	0.72	0.25	0.5	0.2	0.5	0.5	0.53	0.78	0.83
Religion	0.72	0.16	0.4	100	0.4	0.62	0.41	0.16	100	0.41

Below is a snap shot of the summarizer output and the actual document in that order.

--> ಅವರೀಗ ವಿಶ್ವ ಟೆಸ್ಟ್ ಕ್ರಿಕೆಟ್‌ನಲ್ಲಿ ಅತ್ಯಧಿಕ ರನ್ನುಗಳ ಸರದಾರ

--> ಇದನ್ನು ಕ್ರಿಕೆಟ್ ಚರಿತ್ರೆಯ ಮೈಲಿಗಲ್ಲು ಎನ್ನಿ ಸುವರ್ಣಾಕ್ಷರಗಳಲ್ಲಿ ಬರೆದಿಡಬೇಕಾದ ಕ್ಷಣ ಎನ್ನಿ

--> ಸಿಡಿಲ ಮರಿಗೆ ಸಾಟಿ ಯಾರೂ ಇಲ್ಲ ಎಂಬುದು ಅಷ್ಟೇ ದಿಟ ,   ಅವರೀಗ 12 ಸಾವಿರ ಕ್ರಿಕೆಟ್ ಟೆಸ್ಟ್ ರನ್ನುಗಳನ್ನು ದಾಟಿದ ಏಕೈಕ ಆಟಗಾರ

--> ಲಾರಾ ಅವರು 131 ಟೆಸ್ಟ್ ಪಂದ್ಯಗಳಲ್ಲಿ ಒಟ್ಟುಗೂಡಿಸಿದ್ದ 11,953 ರನ್ನುಗಳನ್ನು ತಮ್ಮ 151ನೇ ಪಂದ್ಯದಲ್ಲಿ ಹಿಂದಿಕ್ಕಿದ ಸಚಿನ್, ಇದೀಗ ಅಗ್ರಮಾನ್ಯ ರನ್ ಸರದಾರ

--> ಈ ರನ್ನುಗಳ ಶಿಖರವೇರುವ ಕ್ಷಣಗಳಿಗಾಗಿ ಅವರು ಸಾಕಷ್ಟು ಕಾಯಬೇಕಾಯಿತು ಎಂಬುದು ನಿರ್ವಿವಾದ

--> ಈ ಹಿಂದಿನ ಶ್ರೀಲಂಕಾ ಸರಣಿಯಲ್ಲೇ ಆವರು ದಾಖಲೆ ಮುರಿಯುತ್ತಾರೆ ಎಂದು ಎಲ್ಲರೂ ಭಾವಿಸಿದ್ದರು

--> ಆದರೆ ಅಲ್ಲಿ ನಡೆದ ಮೂರು ಟೆಸ್ಟ್ ಪಂದ್ಯಗಳಲ್ಲಿ ಸಚಿನ್ ಬ್ಯಾಟಿನಿಂದ ಹೊರಬಂದದ್ದು ಕೇವಲ 95 ರನ್ನುಗಳು ಮಾತ್ರ

--> ಆಸೀಸ್ ವಿರುದ್ಧದ ಮೊದಲ ಟೆಸ್ಟ್ ಪಂದ್ಯ ನಡೆದ ಬೆಂಗಳೂರಿನಲ್ಲಿ 49 ರನ್ ಗಳಿಸಿ ಔಟ್ಟಾಗಿದ್ದ 35ರ ಹರೆಯದ ಸಚಿನ್‌ಗೆ ಈ ದಾಖಲೆ ಮಾಡಲು ಬೇಕಾಗಿದ್ದದ್ದು ಕೇವಲ 15 ರನ್ನುಗಳು

--> ಶುಕ್ರವಾರದ ಶುಭದಿನ ಮೊಹಾಲಿಯ ಪಂಜಾಬ್ ಕ್ರಿಕೆಟ್ ಅಸೋಸಿಯೇಶನ್ ಮೈದಾನದಲ್ಲಿ ಈ ಕ್ಷಣಗಳಿಗಾಗಿ ಕಾತರದಿಂದ ಕಾಯುತ್ತಿದ್ದ ಮೊಹಾಲಿ ಕ್ರಿಕೆಟ್ ಅಭಿಮಾನಿಗಳನ್ನು ಈ 35ರ ಹರೆಯದ ವಂಡರ್ ಕಿಡ್ ನಿರಾಶೆಗೊಳಿಸಲಿಲ್ಲ

--> ಅದೆಷ್ಟೋ ಬಾರಿ ಕಾಂಗರೂಗಳನ್ನು ಚಚ್ಚುತ್ತಲೇ, ಕಾಡುತ್ತಲೇ ಇದ್ದ ಸಚಿನ್, ಅದೇ ಬಲಾಢ್ಯ ತಂಡದ ವಿರುದ್ಧ ಈ ಸಾಧನೆ ದಾಖಲಿಸಿದ್ದು ಕೂಡ ಇಲ್ಲಿ ಗಮನಾರ್ಹ

3.8. Summarization of unclassified documents

The main purpose of classification in our work is to make efficient use of feature selection techniques such as GSS co-efficient for summarizing documents. The same feature selection techniques can be used to train classifiers thereby improving the efficiency of classifiers.

When an unclassified document is given as input, the stop word removal method described above is applied. It is then submitted to the classifier which generates relevance scores depending on the number of GSS co-efficient keywords the particular document contains. Such a method will ensure that the classifier does not train itself on and is not misguided by frequently occurring but non-relevant terms to a category. Once a document is classified, the procedure outlined above can be followed.

If all the values returned by the classifier across all categories fall below the threshold, TF-IDF can be used as the fallback for summarization.

4.Conclusion

It is evident that evaluating summaries is a difficult task, as it is not deterministic. Different people may choose different sentences and also same people may choose different sentences at different times. Sentence recall measure is used as the evaluation factor. Although humans can also cut and paste relevant information of a text, most of the times they rephrase sentences when necessary, or they join different related information into one sentence. Hence paraphrasing and rephrasing are the issues in summarization. However in our test, sentences were selected as a whole both by the machine and human summarizers

Most Evaluation systems in Summarization require at least one reference summary, which is the most difficult and expensive task. Much effort is needed in order to have corpus of texts and their corresponding summaries.

The Classifier helps determine the relevant category. Hence any document given as input is able to produce the summary. There is no standard stop word list for Kannada, or methods to do that. Hence a given procedure in this work can be used as a stop word removal tool using structurally similar word removal technique. The summarizer can be used as a tool in various organizations such as Kannada Development Authority, Kannada SahityaParishath etc.

References

[1] Gabor Berend,Rich´ardFarkasSZTERGAK : Feature Engineering for Keyphrasextraction,Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pp 186–189,Uppsala,Sweden, 15-16 July 2010.

[2] Mari-SannaPaukkeri and TimoHonkela,'Likey: Unsupervised Language-independent KeyphraseExtraction,Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pp 162–165,Uppsala, Sweden, 15-16 July 2010.

[3] Letian Wang, Fang Li, SJTULTLAB: Chunk Based Method for Keyphrase Extraction, Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010,pp 158– 161,Uppsala, Sweden, 15-16 July 2010.

[4]   You OuyangWenjie Li Renxian Zhang,'273. Task 5.Keyphrase Extraction Based on Core Word Identification and Word Expansion',Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pp 142–145, Uppsala, Sweden, 15-16 July 2010.

[5]   Su Nam Kim,ÄOlenaMedelyan,~ Min-Yen Kan} and Timothy BaldwinÄ,'SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles',Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pp 21–26,Uppsala, Sweden, 15-16 July 2010.

[6]   FumiyoFukumotoAkina Sakai Yoshimi Suzuki,'Eliminating Redundancy by Spectral Relaxation for Multi-Document Summarization',Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL 2010, pp 98–102,Uppsala, Sweden, 16 July 2010.

[7]   Michael J. Paul,ChengXiangZhai,RoxanaGirju,'Summarizing Contrastive Viewpoints in Opinionated Text',Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp 66–76,MIT, Massachusetts, USA, 9-11 October 2010.

[8]   You Ouyang ,Wenjie Li Qin Lu Renxian Zhang, 'A Study on Position Information in Document Summarization', Coling 2010: Poster Volume, pp 919–927, Beijing, August 2010.

[9]   Xiaojun Wan, Huiying Li and JianguoXiao,'Cross-Language Document Summarization Based on Machine Quality Prediction',Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp 917–926,Uppsala, Sweden, 11-16 July 2010.

[10]   Ahmet Aker Trevor Cohn,'Multi-document summarization using A* search and discriminative training',Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp 482–491,MIT, Massachusetts, USA, 9-11 October 2010.

[11]   Hal Daum´e III*,DanielMarcu*,'Induction ofWord and Phrase Alignments for Automatic Document Summarization', Computational Linguistics, 31 (4), pp. 505-530, December, 2006 .

[12]   ZhiyuanLiu,Wenyi Huang, YabinZheng and MaosongSun,'AutomaticKeyphrase Extraction via Topic Decomposition',Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp 366–376,MIT, Massachusetts, USA, 9-11 October 2010.

[13]   L. Galavotti, F. Sebastiani, and M. Simi,'Experiments on the use of feature selection and negative evidence in automated text categorization' Proc. 4th European Conf. Research and AdvancedTechnology for Digital Libraries,SpringerVerlag,pp.59-68,2000.

[14]   Automatic Summarization by Inderjeet Mani, John Benjamins Publishing Co.

Authors

Jayashree.Ris an Asst.Professor in the Department of Computer Science and Engineering, at PES Institute of Technology, Bangalore. She has over 15 Years of teaching experience. She has published several papers in international conferences. Her research area of interest is Natural Language Processing and Pattern Recognition.

Dr.K.Srikanta Murthy is a Professor and Head, Department of Computer Science and Engineering, at PES School of Engineering, Bangalore. He has put in 24 years of service in teaching and 5 years in research. He has published 8 papers in reputed International Journals; 41 papers at various International Conferences. He is currently guiding 5 PhD students .He is also a member of various Board of Studies andBoard of Examiners for different universities. His research areas include Image Processing ,Document Image analysis, Pattern Recognition, Character Recognition ,Data Mining and Artificial Intelligence.

Sunny.Kis a final year Computer Science undergraduate student.