Journal of Network and Innovative Computing
ISSN 2160-2174 Volume 1 (2013) pp. 139 - 153

A New Approach to Summarization in the Kannada Language by Sentence Ranking

Jayashree R[1], Srikantamurthy K[2] and Basavaraj S Anami[3]

[1] Department of Computer Science and Engineering, PES Institute of Technology, Bangalore, India,
jayashree@pes.edu

[2] Department of Computer Science and Engineering, PES Institute of Technology, Bangalore, India,
srikantamurthy@pes.edu

[3] Department of Computer Science and Engineering, KLE Institute of Technology,
Hubli, India
anami_basu@hotmail.com

Abstract: **Text summarization aims at producing quick and concise summary from a document and is considered central to Information Retrieval (IR) systems. In this paper, we have presented a sentence ranking based method for Kannada language text summarization. Each word in a Kannada document is assigned a weight and the weight of the sentence is computed as the sum of weights of all words present in the sentence. We have chosen the first 'm' sentences by arranging them in the descending order of their weights. The data used for testing is devised from the documents available in Kannada web portal called Kannada webdunia.**

In this methodology, keywords are extracted from Kannada language documents by combining the feature extraction techniques, namely, TF (Term Frequency) and Inverse Document Frequency (IDF). The stop words are removed by using a technique developed which finds structurally similar words in a document. The methodology is compared with the key word extraction based summarization [18]. The results are satisfactory.

Keywords: *Summary, Keywords, GSS coefficient, TF, IDF, Ranking, Word weight, Sentence.*

I. Introduction

The vast amount of information available online makes the need for finding useful information in an efficient and effective way, more obvious. The growing demand for better Information Retrieval (IR) techniques has given rise to lot of research work. There is also 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 [14] provides the following succinct definition for summarization: *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.* 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 does require Natural Language generation techniques, which condense the information in a way similar to human summarization. Summarization is a non 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 to be addressed, because two sentences can give the same meaning with different wordings.

Summarization is a difficult concept because we have to capture the contents of the document as a whole, and also capture its important content in a concise way. We have to reduce the content of the document by selection or through generalization. Again there could be two variations; text extraction and fact extraction. In this work, we have developed methodologies to provide summaries based on text extraction.

There has been an extensive literature about summarization techniques. Marina Litvak et al (2008), have proposed 'Graph-Based Keyword Extraction for Single-Document summarization'. This is an interesting approach suggested wherein they introduce two approaches: Supervised and unsupervised for the cross-lingual keyword-extraction. This key word extraction is to be used as a first step in extractive summarization of text documents.

Gabor Berend et al (2010) have developed a frame work that treats the reproduction of reader assigned keywords as a supervised learning task, In SZETERGAK system, a restricted set of token sequences was used as classification instances.

Another approach by Mari-Sanna Paukkeri et al (2010), 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.

II. Literature Survey

A work on key phrase extraction by Letian Wang and Fang Li, (2010) has shown that key phrase extraction can be achieved using chunk based method. Keywords of document are used to select key phrases from candidates.

Another method proposed by You Ouyang et al.(2010) 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 (2010) 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.

Fumiyo Fukumoto et.al (2010) presents 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 (2010) uses 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 et al (2010) 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 (2010) 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 which is mentioned in the work of Ahmet Aker Trevor Cohn (2010).

Vishal Gupta et al (2012) , have worked on Automatic Punjabi Text Extractive Summarization system. The system developed retains sentences based on statistical and linguistic features.

The work of Zhiyuan Liu et al (2010) demonstrates 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.

According to Haiqin Zhang et al (2010), summaries based on user preferences are quite useful. A good summary should change according to the interests and preferences of the user. For a given document, they first extract user annotations and their contexts and construct a new keyword set together with the original keyword of the text. Then they weigh sentences according to keywords and produce summaries.

A novel Integer Linear Programming (ILP) formulation is used to generate summaries taking into consideration content coherence, sequence of sentences etc to get a better summary which was stated by Hitoshi Nishikawa et al (2010).Size limitation becomes a bottleneck for content coherence which creates interest in improving sentence limited summarization techniques.

The first appearance of the word is important and the importance decreases with the ordinal positions of appearances. The notion here is that the first sentence or a word in a paragraph is very important which may not be true, because it depends on the writing style, some may prefer to give background first and then keep conclusive sentences at the end .This was proved by You Ouyang et al (2010).

As was mentioned in the Literature, there is a need to improve Sentence limited summarization. Hence, in this work we have developed a sentence limited summarizer based on sentence ranking.

The algorithm uses sentences as the compression basis. We generate summaries for documents collected from a Kannada portal (http://www.kannadawebdunia.com) using sentence ranking approach, where in we assign weights to the words in a sentence for a given document after removal of stop words and calculate the score of the sentence (ranking), which is nothing but the summation of all weights of words contained in a sentence. If the score of a sentence S1 is greater than the score of sentence S2, then choose S1.The process is repeated for all sentences in a given document, the sentence limit is 10 in this work. The work carried out by Jayashree.R et al (2011) ‘ Text Document Summarization in the Kannada Language using Key word Extraction ‘ focus on summarization based on key word weights where in key words in a given sentence are assigned weights using GSS feature extraction method. The

results of the work are used here for comparison with our current approach.

The rest of the paper is organized as follows: section –II describes the methodology, section –III highlights Results and Discussion and finally section-IV is about the conclusion of this work.

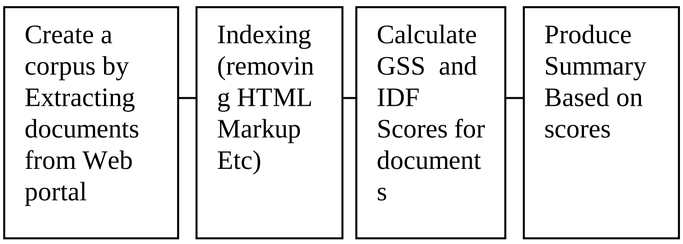
III. Methodology

The sample document presented below is an input document to the summarizer developed by us.

Fig2.1.A sample document input to the summarizer pertaining to Entertainment:

ಕರಾವಳಿಯ ಕುವರಿ, ಬಾಲಿವುಡ್ ನಟಿ, ಮಾಜಿ ವಿಶ್ವಸುಂದರಿ, ಬಚ್ಚನ್ ಸೊಸೆ ಐಶ್ವರ್ಯಾ ರೈ ಮದುವೆಯಾದ ಮೇಲೆ ಖಾಸಗಿ ಜೀವನದಲ್ಲಿ ಬಚ್ಚನ್‌ರ ಮುದ್ದಿನ ಸೊಸೆಯಾಗಿ, ಅಭಿಷೇಕ್‌ನ ಪ್ರೀತಿಯ ಹೆಂಡತಿಯಾಗಿ ತನ್ನ ವೃತ್ತಿ ಬದುಕನ್ನೂ ಅಷ್ಟೇ ವ್ಯಾವಹಾರಿಕವಾಗಿ ಪಕ್ವತೆಯಿಂದ ಮುಂದುವರಿಸಿಕೊಂಡು ಹೋಗುತ್ತಿರುವ ಗಟ್ಟಿಗಿತ್ತಿಇಂಥ . ಐಶ್ವರ್ಯಾ ರೈ ನಿಧಾನವಾಗಿ ತಾನೀವರೆಗೆ ಸಹಿ ಹಾಕಿದ ಪ್ರಾಜೆಕ್ಟುಗಳನ್ನೆಲ್ಲ ಮುಗಿಸಿ ಅಮ್ಮನಾಗಿಬಿಡುವ ಬಯಕೆಯಿದೆಆದರೂ . ಆಕೆ ಈವರೆಗೆ ಸಹಿ ಹಾಕಿದ ಪ್ರಾಜೆಕ್ಟು ಮುಗಿಯಬೇಕೆಂದರೆ ಇನ್ನೂ ಒಂದೆರಡು ವರ್ಷವಾಗಬೇಕುಅಷ್ಟು ಬ್ಯುಸಿ . ಈಕೆ. ಇಂತಿಪ್ಪ ಐಶ್‌ಳ ಬಾಲ್ಯದ ದಿನಗಳು ಹೇಗಿದ್ದವು ಗೊತ್ತಾಥೇಟ್ ಎಲ್ಲಾ . ಮ ವರ್ಗದಮಧ್ಯ ಮಕ್ಕಳಂತೆಯೇಆಕೆ ಓದುತ್ತಿದ್ದ !, ಟಿವಿ ನೋಡುತ್ತಿದ್ದ ದಿನಗಳನ್ನು ಆಕೆಯ ಬಾಯಿಯಿಂದಲೇ ಕೇಳಿದರೆ ಚೆನ್ನ. "ಅಪ್ಪ ಹೆಚ್ಚಾಗಿ ನೌಕೆಯಲ್ಲೇ ಇರುತ್ತಿದ್ದರು ಬಹುತೇಕ ಬಾಲ್ಯದ ದಿನಗಳು ಅಮ್ಮ ಹಾಗೂ ನನ್ನ ಅಣ್ಣನ ಜೊತೆಗೆ ಕಳೆದುಹೋಯಿತು . .ಸಣ್ಣವಳಿದ್ದಾಗಲೇ ನಾನು ಕನಸುಗಾರ್ತಿ ನನ್ನದೇ ಕನಸುಗಳ ಲೋಕದಲ್ಲಿ ನಾನಿರುತ್ತಿದ್ದೆಸ್ವಲ್ಪ ಸೂಕ್ಷ್ಮ ಸ್ವಭಾವದವಳಾದ ನಾನು . ಯಾವಾಗಲು ನನ್ನದೇ ವಯಸ್ಸಿನವರ ಜೊತೆಗಿರುತ್ತಿದ್ದುದು ಕಡಿಮೆ . ಮಾವ, ಅತ್ತೆಯ ಮಕ್ಕಳೊಂದಿಗೆ ಬೆರೆಯುವುದಕ್ಕಿಂತ ಹೆಚ್ಚು ಮಾವ ಅತ್ತೆಯವರೊಂದಿಗೇ ಬೆರೆಯುತ್ತಿದ್ದೆ. ನನ್ನ ಅಪ್ಪ ಅಮ್ಮ ಯಾವತ್ತೂ ನನ್ನನ್ನು ಮುಕ್ತವಾಗಿ ಅಭಿಪ್ರಾಯ ಹಂಚಿಕೊಳ್ಳಲು ಪ್ರೇರೇಪಿಸುತ್ತಿದ್ದರು."

We have extracted documents from four different categories: *Literature, Sports, Entertainment, and religion.* The methodology adopted by us can be best described by using four major steps, which is shown using schematic block diagram shown below.



The first step is creating a corpus from the data extracted from the portal. 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.

The second step is 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. Documents in four categories were fetched: sports, religion, astrology and entertainment.

The third step is to calculate GSS coefficients and the Inverse Document Frequency (IDF) scores for every word (in a given category in the latter case). Every word in a given document has a Term Frequency (TF), which gives the number of occurrence of a term in a given document; Term Frequency and Inverse document Frequency are defined by the following formulae:

$$TF = \frac{\text{frequency of a term in a document}}{\text{Number of terms in a given document}} \quad \ldots\ldots (1)$$

$$IDF = Log_{10}(N / n) \quad \ldots\ldots (2)$$

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. Also 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 which is used as the relevance measure in our case.

Given a word ‘w’ and category ‘c’ it is defined as:

$$f(w, c) = p(w, c) * p(w', c') - p(w', c) * p(w, c') \quad \ldots\ldots (3)$$

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 word ‘w’ and does not belong to category ‘c’.
p (w' , c) is the probability that a document does not contain word ‘w’ and belongs to category ‘c’.
p (w , c') is the probability that a document contains word ‘w’ and does not belong to category ‘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. Hence using these two parameters which determine relevant parts of the document provides us a wholesome summary.

The fourth step is summarization: Given a document and a limit on the number of sentences, the algorithm has to provide a meaningful summary. The algorithm calculates the GSS coefficients and IDF of all the words in the given document, if the document is 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. Instead finding out how many key words a sentence contains, the algorithm considers all words in a sentence.Each word is assigned weight; the word weight formula is given by,

*Word weight = term frequency * inverse document frequency+ GSS * term frequency……(4)*

Hence the weight of a sentence is the sum of weights of all words. Then we choose top scoring 'm' sentences, i.e. if the score of sentence S1 is greater than the score of S2, we choose S1.Then top 'm' 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 are not considered as split points. Each of these sentences is then evaluated for the number of keywords it contains from the list as follows:

$$\text{Rank of sentence} = \frac{\text{sum of values of words in the sentence}}{\text{Total number of words in a sentence}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

The algorithm was tested on one document each belonging to Sports, Entertainment, Religion and Literature category. The GSS coefficients and IDF list obtained are as below with value n=20:

GSS co-efficient list for category sports:

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

IDF List for category sports:

ಕ್ರಿಕೆಟ್, ವಿಶ್ವಕಪ್, ನಲ್ಲಿ, ಮತ್ತಷ್ಟು , ಮೊದಲ, ಭಾರತ, ಕ್ರೀಡಾ, ಜಗತ್ತು , ಲೇಖನಗಳು, ಕ್ರಿಕೆಟಿಗರು , ಏಕದಿನ, ಅಂಕಿ,ಅಂಶ, ಟಿಕರ್, ಶೋಧಿಸು, ಸಹ, ಇದನ್ನು, ಮುಖ್ಯ,ಪುಟ,

GSS Coefficient List for category entertainment:

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

IDF List for category entertainment:

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

GSS Coefficient List for category Religion:

ಧರ್ಮ, ಗುರು, ಧರ್ಮದ, ಕುರಿತು, ಉತ್ಸವಗಳು, ಮುಖ್ಯ, ಪುಟ, ತಮ್ಮ, ಸಿಖ್, ಮತ್ತಷ್ಟು, ಗುರುಗಳು, ದಶ, ಎಂದು, ಕೃಷ್ಣ, ಕಕರ್, ತೀರ್ಥಕ್ಷೇತ್ರಗಳು, ನಾನಕ್, ವಾಣಿ, ಪಂಚ, ದಾಸ್

IDF Coefficient List for category entertainment:

ಗುರು, ಹರ್, ಸಿಖ್, ತಮ್ಮ, ತೇಗ್, ಬಹದ್ದೂರ್, ಹಿಂದೂಗಳ, ದಾಸ್, ಗುರುಗಳು, ದಶ, ಧರ್ಮ, ಅನುಭವಿಸಿದರು, ಹೇಳಿಕೆಯಿಂದ, ಶಿಕ್ಷೆಯನ್ನು, ಒಂಬತ್ತನೇ, ಹುಟ್ಟುಹಾಕಿದರು, ಕಂಕಣಬದ್ದರಾಗಿದ್ದರು, ಪಟ್ಟಣವನ್ನು, ಆನಂದಪುರ್,

IV. Noise Removal

Another objective of this work was to look at dimensionality reduction techniques and their application for Kannada language documents. A document is more often seen as vector of features, there is an unrealistic requirement to the classifier in terms of time required for classifying this large feature vector. One way of reducing dimensionality is to ensure that words which are considered as noise (high frequency) should not be evaluated as keywords. To remove stop words we implemented an algorithm, which takes as input, a stop word which is entered manually and finds structurally similar words and adds these words to the stop word list.

Some of the words in our primary list of stop words, which are 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

Algorithm for finding structurally similar words:

The following are simpler steps of the algorithm: Consider the word, ' ಯಾಕೆ' (Why)

When split into individual Unicode characters, it becomes: (U+0CAF) + (U+0CBE) + (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, for example, the word , 'ಯಾಕೆ' (why) .

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, that is the word, 'ಯಾಕೆ' gets *modified to* 'ಯಾಕ '(English equivalent of 'why').

A sample of stop words that were obtained by our algorithm for the word 'ಯಾಕೆ' is shown below:

ಯಾಕೆಂದರೆ, ಯಾಕೋ, ಯಾಕಿಷ್ಟು, ಯಾಕ್ಟ್,ಯಾಕಾಗಬಾರದು , ಯಾಕೂಬ್ etc.

For the personal pronoun 'ಬರಿ '(write), some of the words that were obtained using the algorithm for finding structurally similar words are given below:

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

As evident, though some words have semantic relationship to the primary stop word, a lot of such 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 list of basic stop words, this program can be used to find structurally similar words and semantically unrelated words can be manually removed from the stop word list.

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.

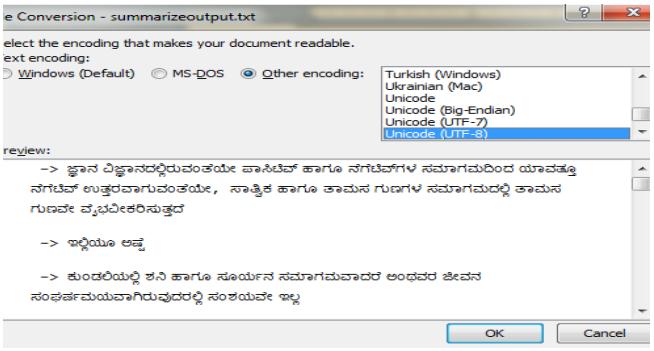
When an unclassified document is given as input, the stop word removal method described above is applied. It is then submitted to the classifier which generates relevance scores depending on the number of GSS co-efficient keywords the particular document contains. Such a method will ensure that the classifier 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.

V. Results and Discussions

Results below show the comparison of three summaries; summary using word weight age [17], summary using sentence ranking (current approach) and human summary. Human summary refers to the manual summary obtained by native language speakers. We requested 9 such native language speakers to generate summaries for the documents pertaining to the Literature, Entertainment, Sports and Religion categories. The documents were chosen randomly from the database. We compared human summary with machine summary using word weight age approach (**which is Machine summary1 Vs Human summary),** Machine summary using sentence ranking approach with human summary which is (**Machine summary2 Vs human summary**) and finally, two Machine summaries (**which are Machine summary 1 Vs Machine summary2).**

We chose 7 random files in the following categories: Religion, Entertainment, Sports and Astrology.

The Screen shot of the summary generated for a document in Astrology is shown below:



The summary generated is based on the limit given by the user, as it is sentence limited summarizer. The point to be noted is that, extractive summary may not be as effective as abstractive summary. We also want to highlight that paraphrasing and rephrasing may be difficult to achieve. The results mentioned below illustrate that.

The summary generated by machine, both by using sentence ranking and keyword extraction based method is compared with expert summary (human) .The tables below show the number of common sentences between machine summary and human summary. For example, if the number of common sentences is 2 between machine summary and human summary for a sentence limit of 10, then 2/10 gives a score of 0.2

Initially, 7 documents pertaining to different categories are considered, the results are shown below.

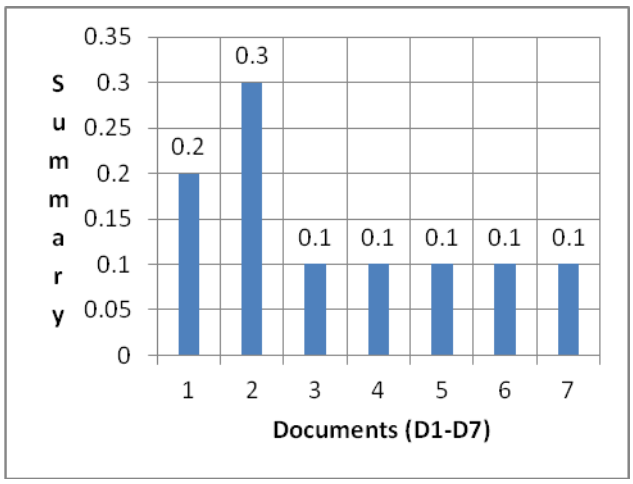


Fig 1.1 Graph showing Machine Sumary1 Vs Human Summary (Sports)

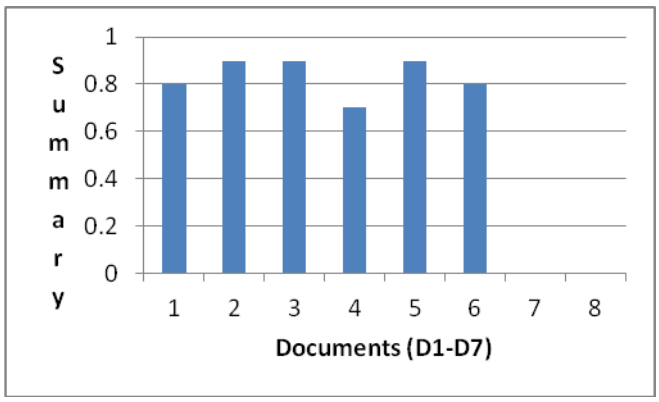


Fig 1.2 Graph showing Machine Sumary2 Vs Human Summary (Sports)

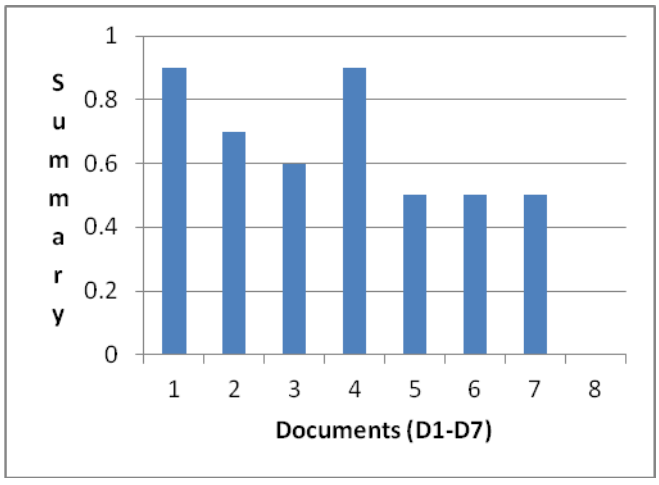


Fig 1.3 Graph of Machine Summary1 Vs Machine Summary2 (Sports)

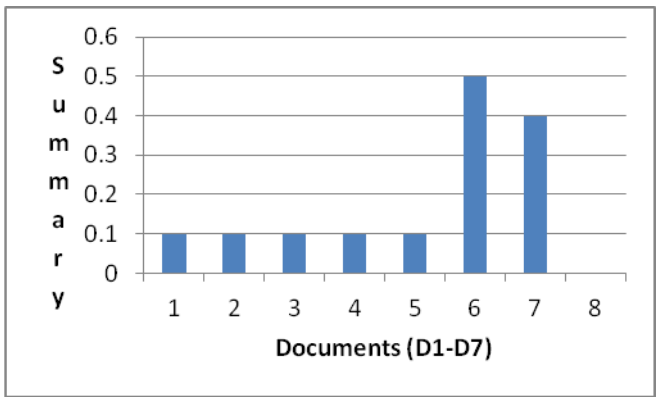


Fig 1.4 Graph of Machine Summary1 Vs Human Summary (Entertainment)

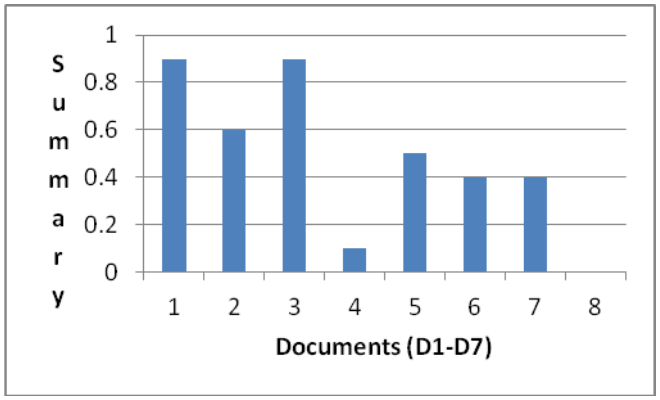


Fig 1.5 Graph of Machine Summary1 Vs Machine Summary2 (Entertainment)

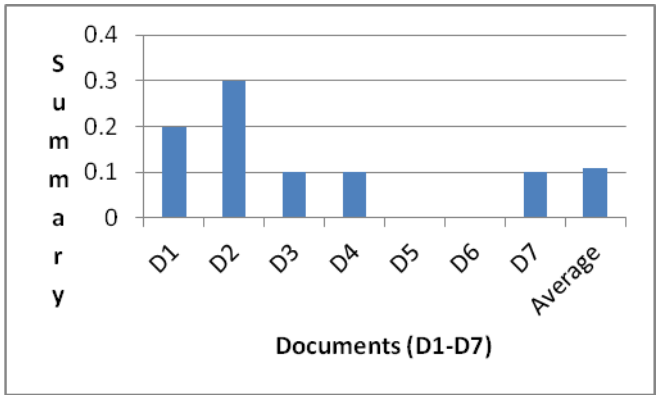


Fig 1.6 Comparison of Machine Summary1 VS Human summary (Religion)

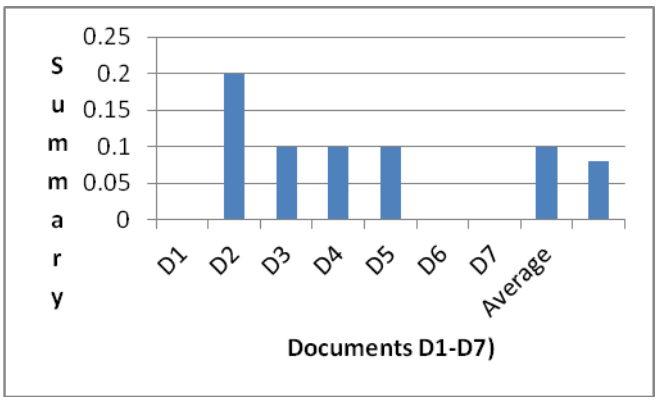


Fig 1.7 Comparison of Machine Summary2 VS Human summary (Religion)

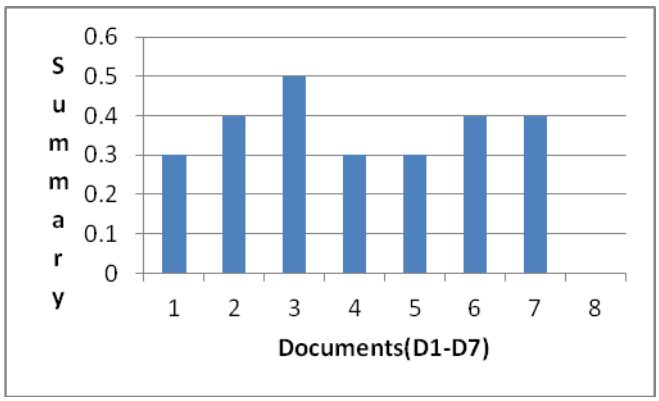


Fig 1.10 Comparison of Human (Expert1) Vs Machine summary2 (Literature)

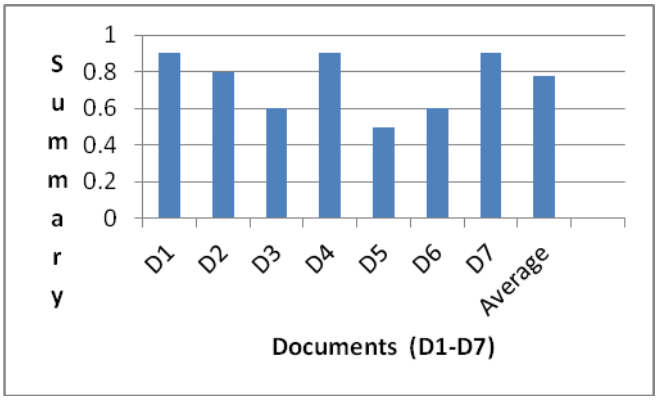


Fig 1.8 Comparison of Machine Summary1 Vs Machine summary2 (Religion)

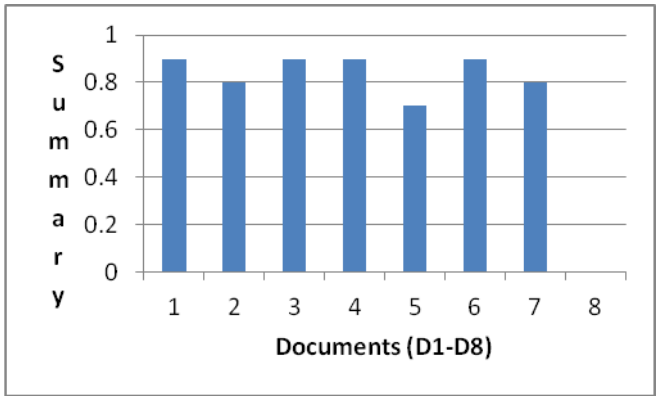


Fig 1.11 Comparison of Machinesummary1Vs Machine summary2 (Literature)

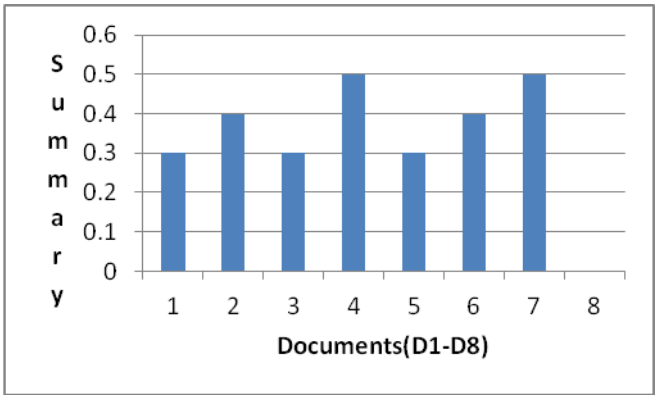


Fig 1.9 Comparison of Human Vs Machine summary 1 (Literature)

The results mentioned below are obtained by comparing machine generated summaries with that of the human (called as experts, native language speakers). summaries

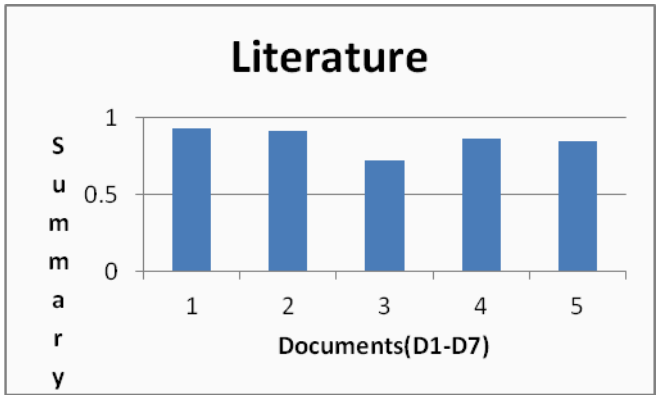


Fig 1.12 comparison of Human summary (Expert1) Vs Machine summary1 (Literature)

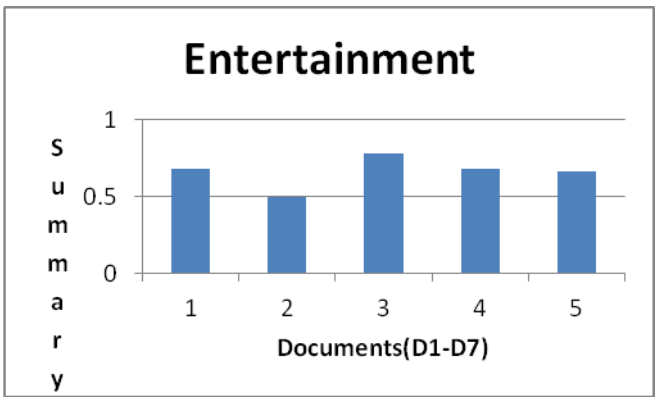


Fig1.13 comparison of Human (Expert1) summary Vs Machine summary1 (Entertainment)

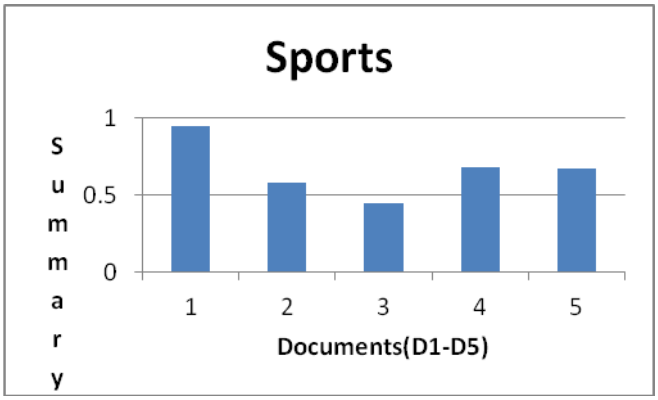


Fig1.14 comparison of Human (expert1) summary Vs Machine summary1 (Sports)

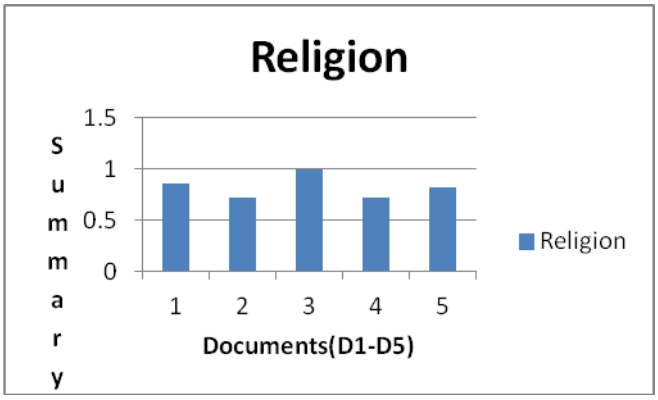


Fig 1.15 comparison of Human (expert1) summary Vs Machine summary1 (Religion)

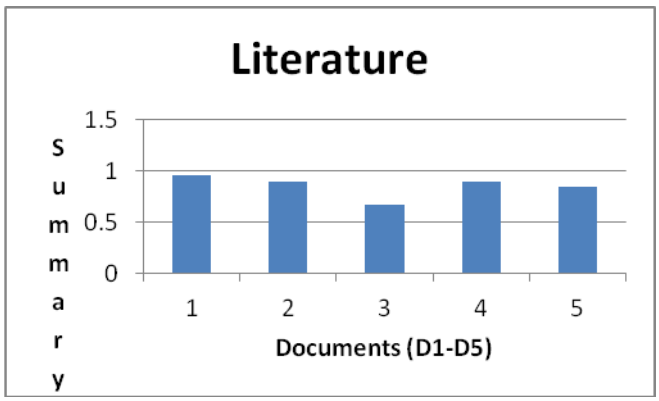


Fig 1.16 comparison of Human (expert2) summary Vs Machine summary2 (Literature)

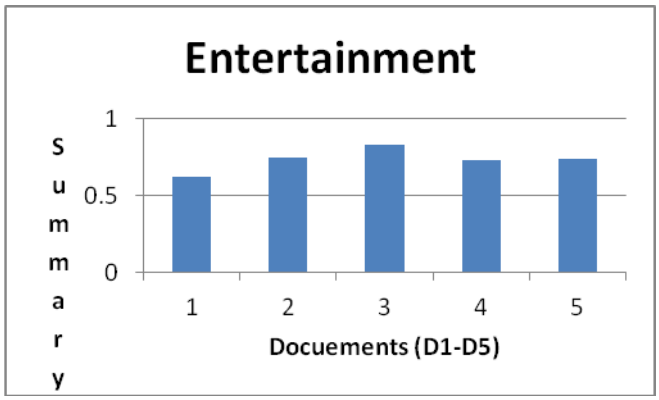


Fig 1.17 comparison of Human (expert2) summary Vs Machine summary2 (Entertainment)

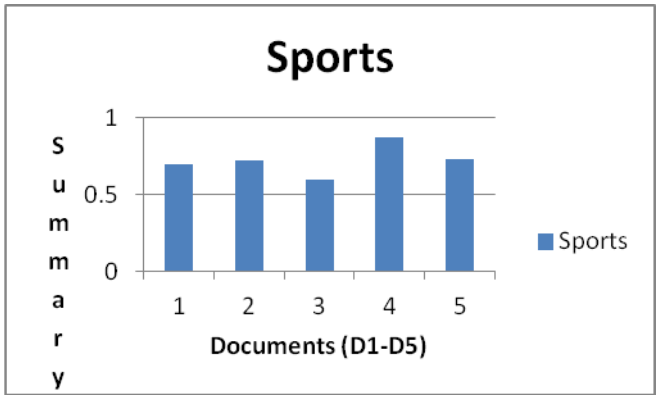


Fig 1.18 comparison of Human (expert2) summary Vs Machine summary2 (Sports)

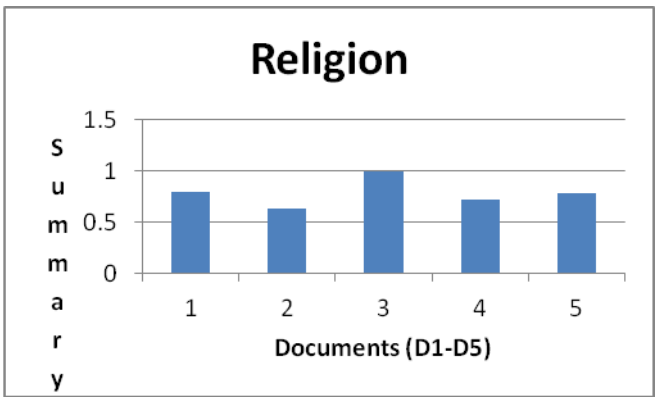


Fig 1.19 comparison of Human (expert2) summary Vs Machine summary2 (Religion)

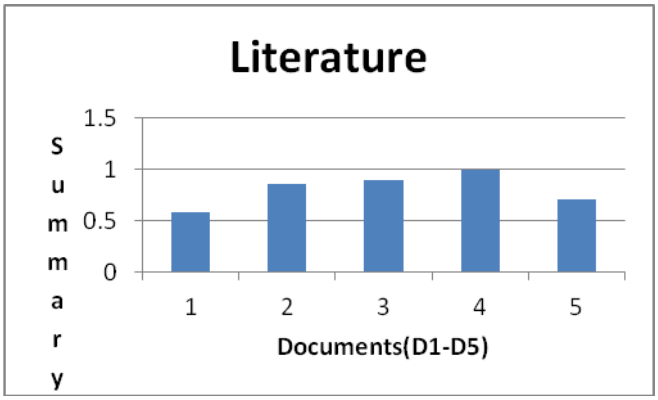


Fig 1.20 comparison of Human (expert3) summary Vs Machine summary1 (Literature)

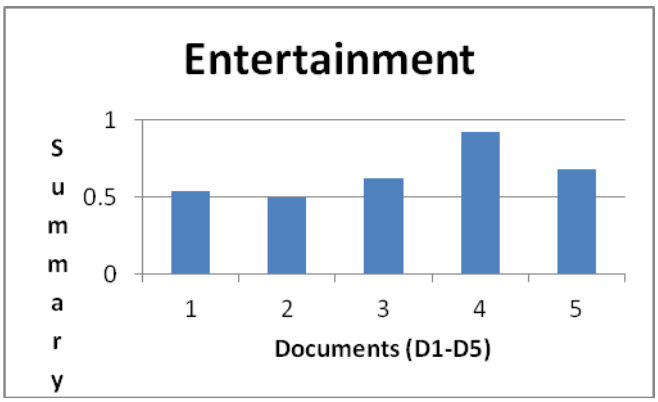


Fig 1.21 comparison of Human (expert3) summary Vs Machine summary1 (Entertainment)

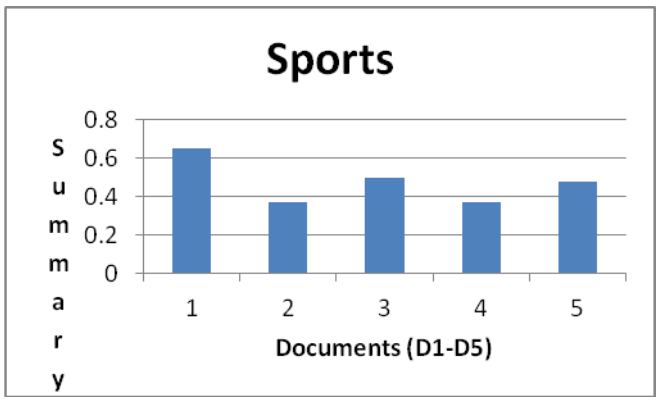


Fig 1.22 comparison of Human (expert3) summary Vs Machine summary1 (Sports)

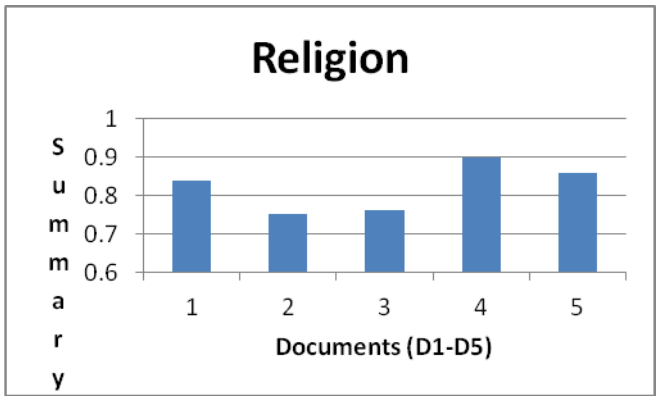


Fig 1.23 comparison of Human (expert3) summary Vs Machine summary1 (Religion)

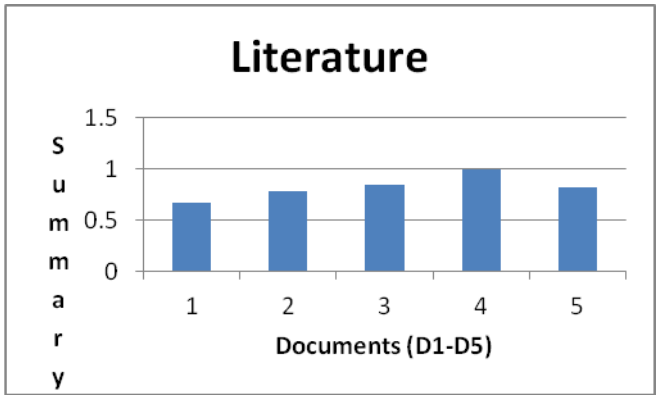


Fig 1.24 comparison of Human (expert4) summary Vs Machine summary2 (Literature)

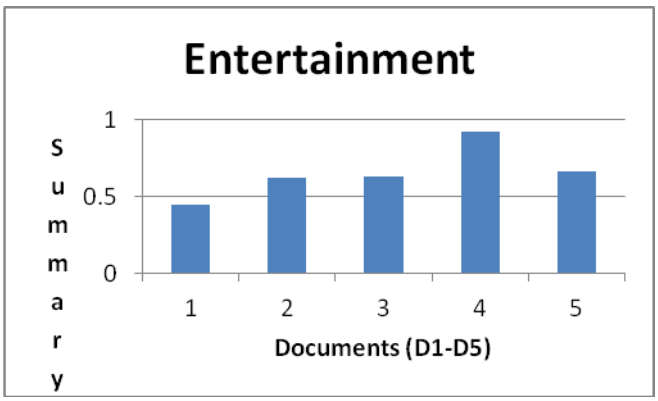


Fig 1.25 comparison of Human (expert4) summary Vs Machine summary2 (Entertainment)

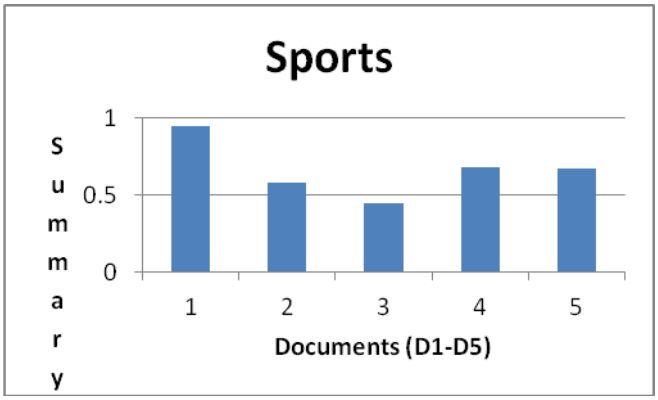


Fig 1.26 comparison of Human (expert4) summary Vs Machine summary2 (Sports)

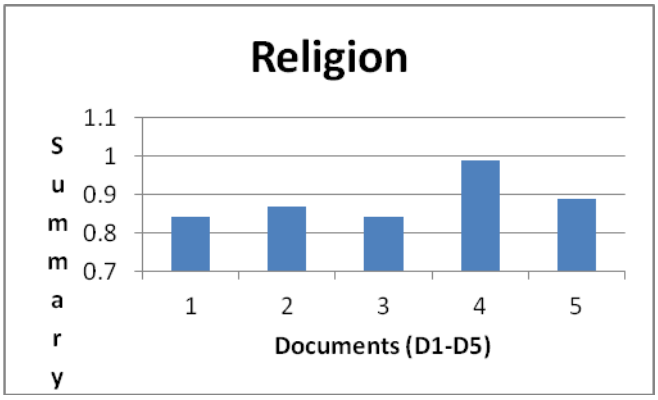


Fig 1.27 comparison of Human (expert4) summary Vs Machine summary2 (Religion)

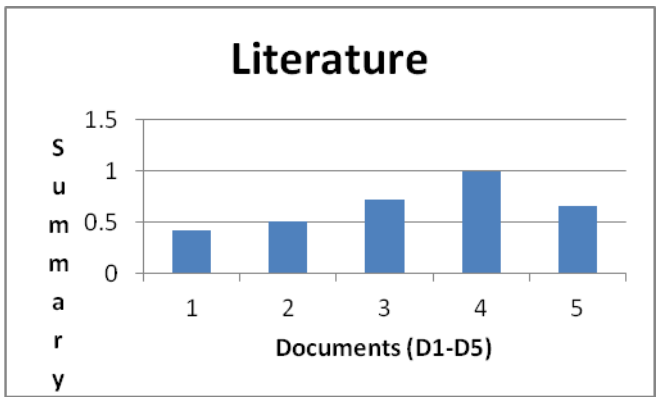


Fig 1.28 comparison of Human (expert5) summary Vs Machine summary1 (Literature)

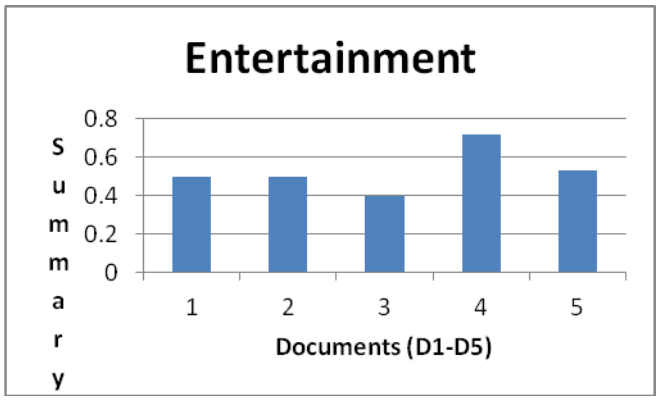


Fig 1.29 comparison of Human (expert5) summary Vs Machine summary1 (Entertainment)

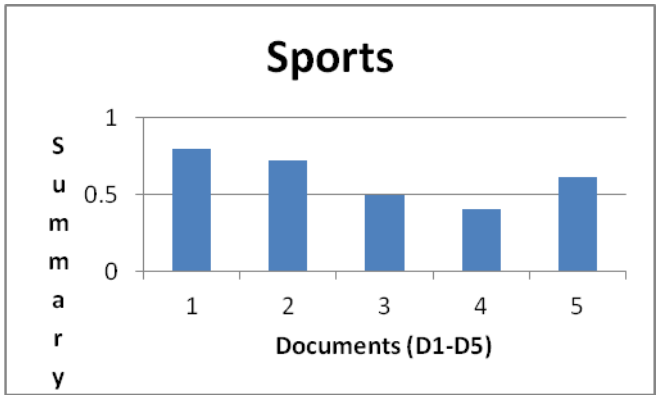


Fig 1.30 comparison of Human (expert5) summary Vs Machine summary1 (Sports)

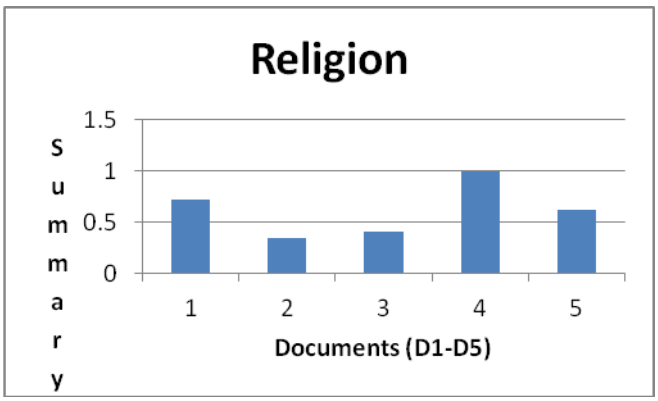


Fig 1.31 comparison of Human (expert5) summary Vs Machine summary1 (Religion)

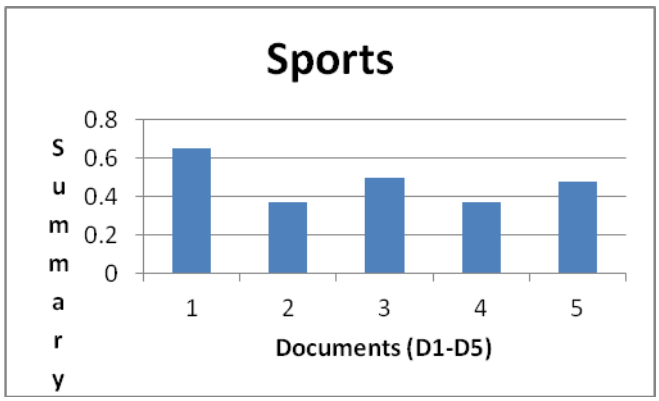


Fig 1.34 comparison of Human (expert6) summary Vs Machine summary2 (Sports)

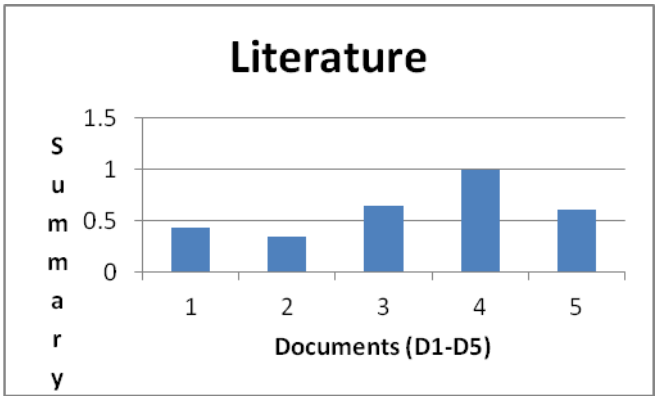


Fig 1.32 comparison of Human (expert6) summary Vs Machine summary2 (Literature)

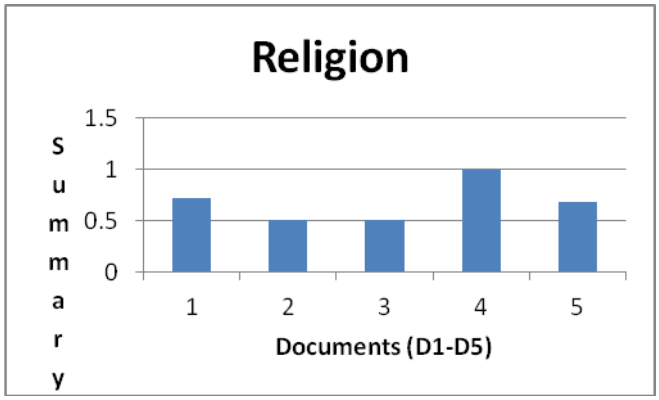


Fig 1.35 comparison of Human (expert6) summary Vs Machine summary2 (Religion)

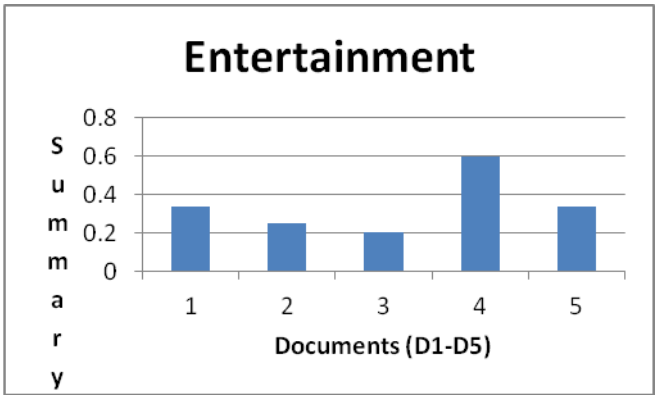


Fig 1.33 comparison of Human (expert6) summary Vs Machine summary2 (Entertainment)

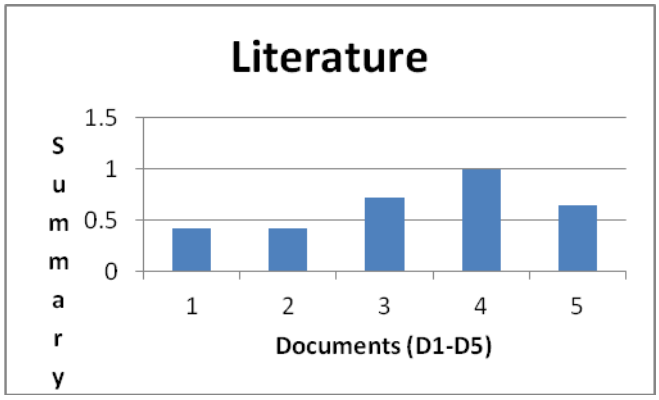


Fig 1.36 comparison of Machine Summary1 Vs Machine summary2 (Literature)

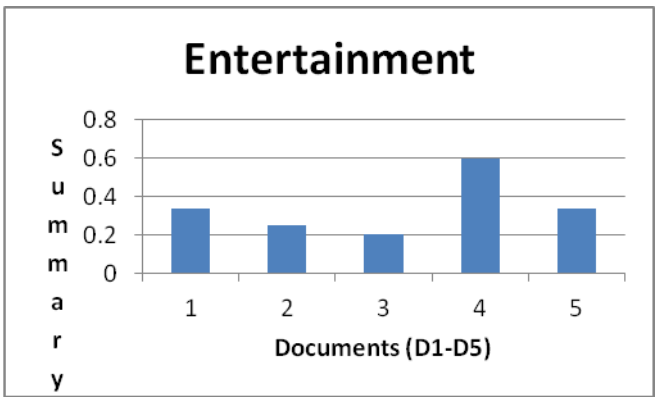


Fig 1.37 comparison of Machine summary1 Vs Machine summary2 (Entertainment)

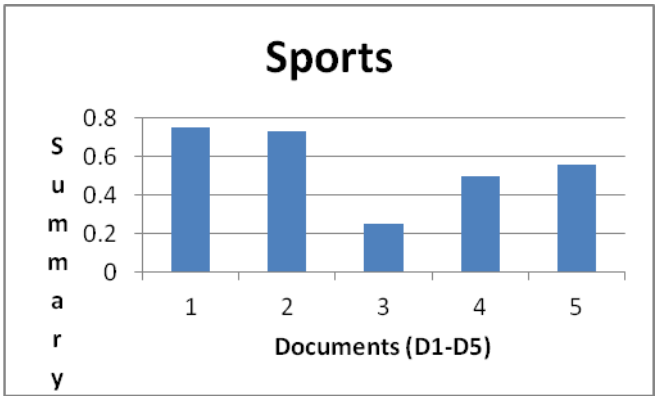


Fig 1.38 comparison of Machine Summary1 Vs Machine summary2 (Sports)

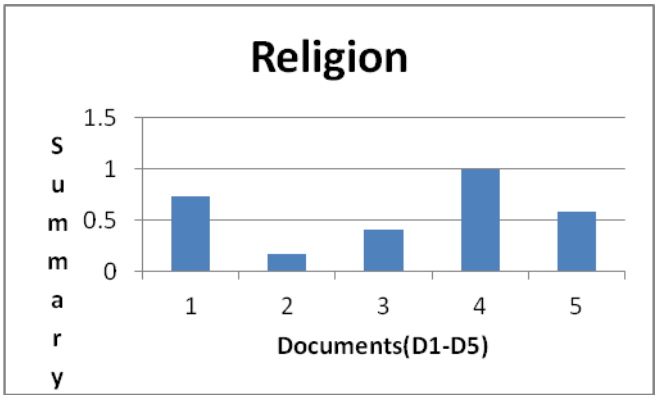


Fig 1.39 comparison of Machine Summary1 Vs Machine summary2 (Religion)

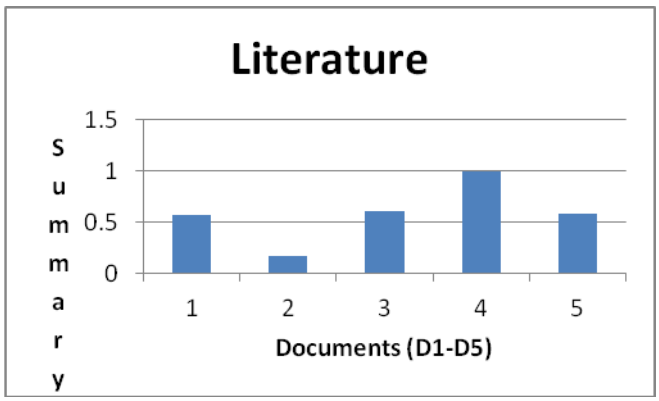


Fig 1.40 comparison of Human summary (expert7) Vs Machine summary1 (Literature)

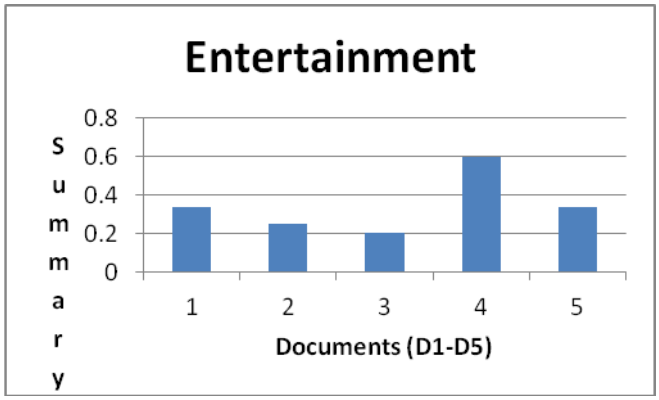


Fig 1.41 comparison of Human (expert7) summary Vs Machine summary1 (Entertainment)

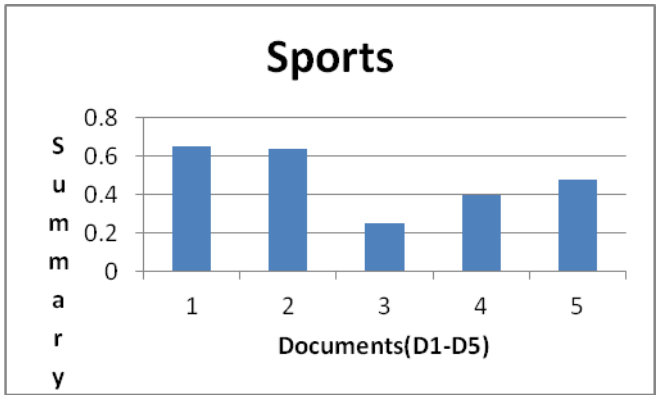


Fig 1.42 comparison of Human(expert7) summary Vs Machine summary1 (Sports)

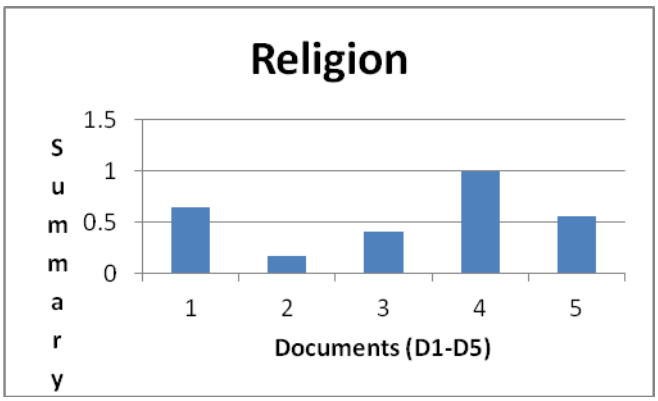


Fig 1.43 comparison of Human (expert7) summary Vs Machine summary1 (Religion)

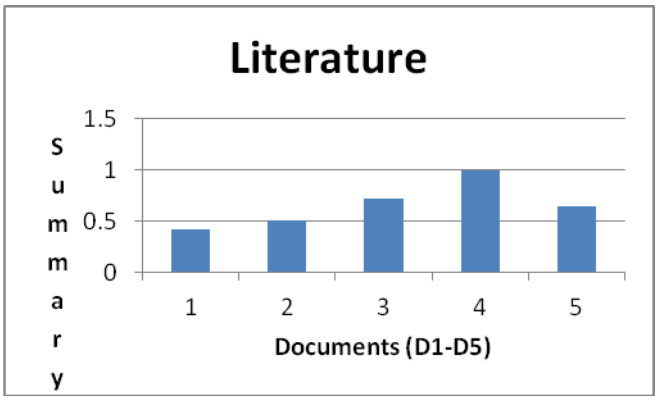


Fig 1.44 comparison of Human (expert8) summary Vs Machine summary2 (Literature)

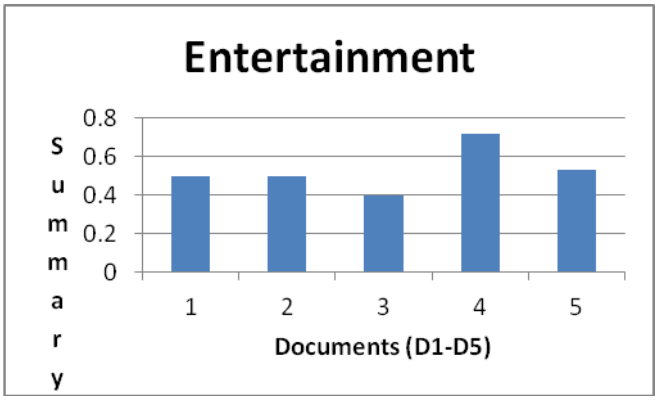


Fig 1.45 comparison of Human (expert8) summary Vs Machine summary2 (Entertainment)

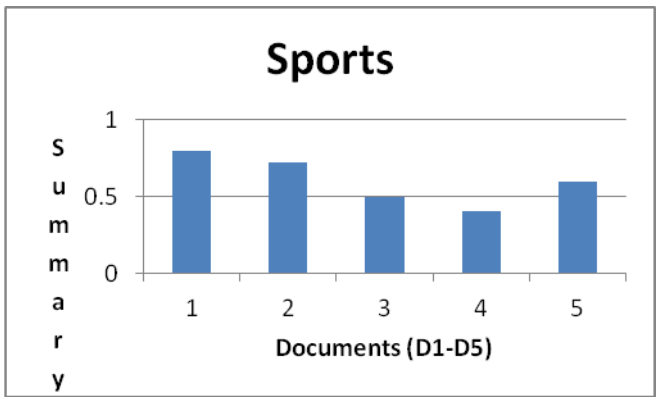


Fig 1.46 comparison of Human (expert8) summary Vs Machine summary2 (Sports)

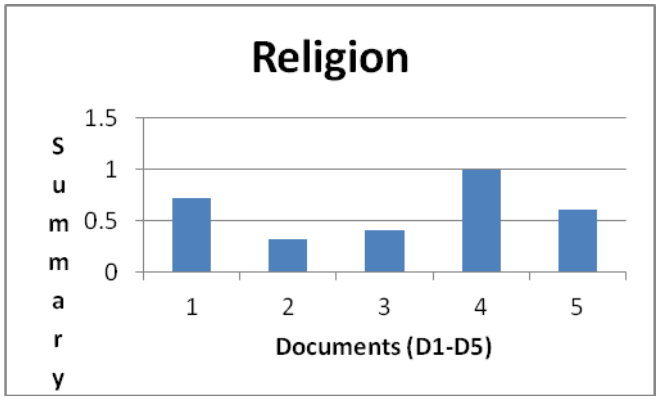


Fig 1.47 comparison of Human (expert8) summary Vs Machine summary2 (Religion)

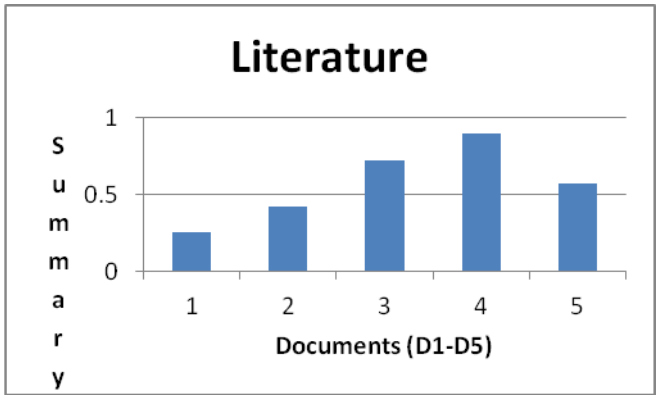


Fig 1.48 comparison of Machine Summary1 Vs Machine summary2 (Literature)

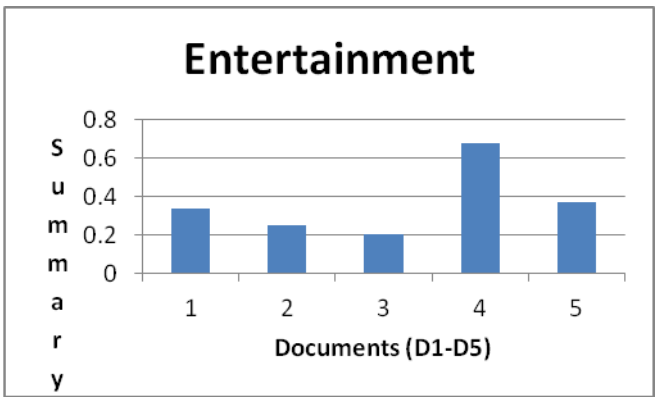


Fig 1.49 comparison of Machine Summary1 Vs Machine summary2 (Entertainment)

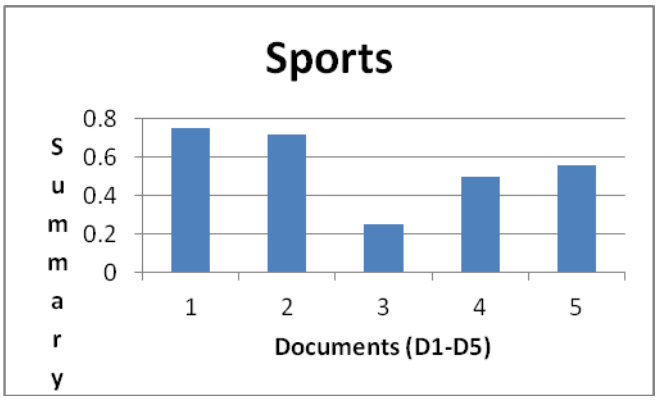


Fig 1.50 comparison of Machine Summary1 Vs Machine summary2 (Sports)

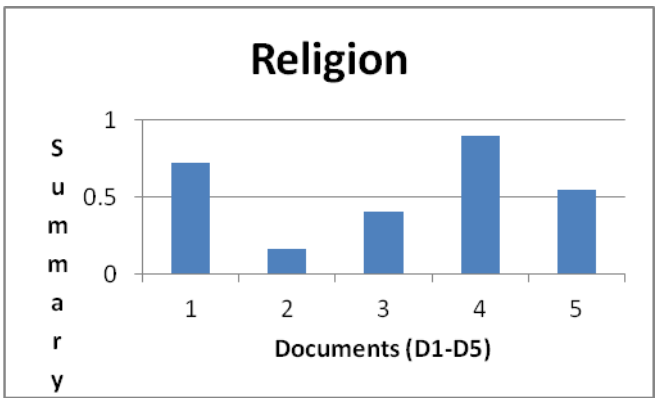


Fig 1.51 comparison of Machine Summary1 Vs Machine summary2 (Religion)

The results show a clear indication of the fact that human summaries are far more realistic to be achieved. The comparison between two machine generated summaries indicates that both approaches are very similar. But human summaries do not have commonality to machine generated summaries. Machine generated summaries given here lack coherence as the summarization is sentence limited, which could be addressed in future work.

Conclusion

The results though not promising, indicate the fact that, summarization can be very effective if the algorithm can incorporate techniques for achieving coherence. Human summaries are far more effective if knowledge about the document being summarized is known in advance. It is evident that human created summary is far superior than the machine generated summary. We can further explore the possibility of introducing better techniques which can match machine summary on par with human summary. Our future work is application of Artificial Neural Network concepts to check if summaries generated by neural networks are superior to the method discussed here.

References

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

[2] Mari-Sanna Paukkeri and Timo Honkela,'Likey: Unsupervised Language-independent Keyphrase Extraction,Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pp 162–165,Uppsala, Sweden, 15-16,2010.

[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, 15-16,2010.

[4] You Ouyang Wenjie 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, 2010.

[5] Su Nam Kim,Ä Olena Medelyan,~ 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 , 2010.

[6] Fumiyo Fukumoto Akina 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, 2010.

[7] Michael J. Paul,ChengXiang Zhai,Roxana Girju,'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, 2010.

[8] You Ouyang Wenjie 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, pages 142–145,Uppsala, Sweden, 15-16, 2010.

[9] Xiaojun Wan, Huiying Li and Jianguo Xiao,'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 , 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, 2010

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

[12] Zhiyuan Liu,Wenyi Huang, Yabin Zheng and Maosong Sun,'Automatic Keyphrase Extraction via Topic Decomposition',Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp 366–376,MIT, Massachusetts, USA, 9-11,2010.

[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.

[15] Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo and Genichiro Kikui,' Opinion Summarization with Integer Linear Programming Formulation for Sentence Extraction and Ordering', Coling 2010: Poster Volume, pages 910–918, Beijing, 2010.

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

[17] Jayashree.R,Srikantamurthy.K,' Text Document Summarization in the Kannada Language using Key word Extraction',Proceedings of the Ist International Conference on Artificial Intelligence,Soft computing and Applications(AIAA)-2011,pp 48-54,Tirunelveli,India,2011.

[18] Jayashree.R,Srikantamurthy.K,Basavaraj S Anami,'Categorized Text Document Summarization in the Kannada Language by Sentence Ranking',Proceedings of the Fourth International Conference on Soft Computing and Pattern Recognition,(SOCPAR-2012),PP 7-12,Brunei,December , published by IEEE,2012.

[19] Marina Litvak, Mark Last,' Graph-Based Keyword Extraction for Single-Document Summarization', Coling 2008: Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, pages 17–24,Manchester, 2008.

[20] Vishal Gupta, Gurupreet Singh Lehal,' Automatic Punjabi Text Extractive Summarization System', Proceedings of COLING 2012: Demonstration Papers, pages 191–198,COLING 2012, Mumbai, December 2012.

[21] Haiqin Zhang, Zheng Chen Wei-ying Ma, Qingsheng Cai,'A Study for Documents Summarization based on Personal Annotation, HLT-NAACL 2003 workshop: Text Summarization Workshop.

Author Biographies

Jayashree.R is an Associate Professor in the Department of Computer Science and Engineering, at PES Institute of Technology, Bangalore. She has over 18 Years of teaching experience. She has published several papers in international conferences and journals. Her research areas of interest are Natural Language Processing and Pattern Recognition.

Dr Srikantamurthy K is a Professor and Head, Department of Computer Science and Engineering, at PES Institute of Technology, Bangalore. He has put in 26 years of service in teaching and 7 years in research. He has published 13+ papers in reputed International Journals; 60+ papers at various International Conferences. He is currently guiding 5 PhD students .He is also a member of various Board of Studies and Board of Examiners for different universities. His research areas include Image Processing ,Document Image analysis, Pattern Recognition, Character Recognition ,Data Mining and Artificial Intelligence.

Dr Basavaraj S Anami is presently working as the Principal, K.L.E Institute of Technology, Hubli, since August 2008. He completed his Bachelor of Engineering in Electrical Stream during November 1981. Then he completed his M.Tech in Computer Science at IIT Madras in March 1986. Later he received his Doctrine (PhD) in Computer Science at University of Mysore in January 2003.He began his academic journey as the Lecturer in Electrical department in BEC, Bagalkot from September 1983 up to December 1985. Then he was promoted as the In-charge Head of Department of Computer Science in the same college and in February 1990 as the Head of Department, Computer Science, up to July 2008.He has over 10 years of experience in research and has 15 international journals, 20 national journals, 10 international conference papers and 40 national conference papers to his credit.