A Survey of Text Summarizers for Indian Languages and Comparison of their Performance

Vishal Gupta
UIET, Panjab University, Chandigarh, India
Email: vishal@pu.ac.in

***Abstract*—Automatic text summarization is technique of compressing the original text into shorter form which will provide same meaning and information as provided by original text. The brief summary produced by summarization system allows readers to quickly and easily understand the content of original documents without having to read each individual document. The overall motive of text summarization is to convey the meaning of text by using less number of words and sentences. Summaries are of two types: Abstractive summaries and Extractive summaries. Extractive summaries involve extracting relevant sentences from the source text in proper order. The relevant sentences are extracted by applying statistical and language dependent features to the input text. On the other hand, abstractive text summaries are made by applying natural language understanding. Human beings usually make summaries in abstractive way. Moreover abstractive summaries can also involve the words or sentences which are not present in the input text. Automatic generation of abstractive summary is more difficult as compared to producing extractive text summary. This paper concentrates on survey and performance analysis of automatic text summarizers for Indian languages.**

***Index Terms*—Indian summarizers, summarizers, text summarization system**

I. Introduction

Automatic text summarization [1] is technique of compressing the original text into shorter form which will provide same meaning and information as provided by original text. The brief summary produced by summarization system allows readers to quickly and easily understand the content of original documents without having to read each individual document. The overall motive of text summarization is to convey the meaning of text by using less number of words and sentences. Text Summaries are of two types: Abstractive summaries and Extractive summaries. Extractive summaries involve extracting relevant sentences from the source text in proper order. The relevant sentences are extracted by applying statistical and language dependent features to the input text. On the other hand, abstractive text summaries are made by applying natural language understanding. Human beings usually make summaries in abstractive way. Moreover abstractive summaries can also involve the words or sentences which are not present in the input text. Automatic generation of abstractive summary is more difficult as compared to producing extractive text summary.

Automatic Text Summary generated by Microsoft Word is type of extractive summary for English language. In some of summarization systems, users can specify percentage of total source text in final summary. Worthiness of lengthy documents can quickly and easily be judged using text summarization. A summary can be labeled as good summary if it is highlighting different topics of input text and it should not have duplicate sentences. For natural language processing, making automatic text summary is largely used application of it.

Abstractive text summarization generates the summary after thoroughly understanding of input text and reconstructing the summary using less number of words and sentences in same manner as human beings usually make the summary. Abstractive text summarization is difficult because as compared to human beings, computers have limited capabilities of language understanding, so alternative methods must be considered. Difficulties of abstractive summary [1] are as: The main difficulty with abstractive summarization is representation. The abilities of automatic systems are limited by the large number of their representations and capability to produce these representation-structures—Abstractive summarizers can not produce summary of the text which their structures cannot represent. Under restricted category it is possible to formulate proper representations, but a general solution is not feasible and is dependent on general-domain semantic representations. It is not possible to build the automatic systems which can fully understand and represent the natural language of human beings.

Extractive text summarization selects the relevant sentences from input text. The relevant sentences are extracted by applying statistical and language dependent features of sentences. In most of cases in the world we prefer to make extractive text summaries due to its ease in generating text summary. Difficulties of extractive text summary [1] [2] are: 1) As compared to average summaries, extractive summaries are normally lengthy because certain sections of text which are not required in summary may also be included in it. 2) In many cases essential information is usually present across different lines, and usually extractive summaries may not collect it unless it is lengthy enough for covering all these lines.

doi:10.4304/jetwi.5.4.361-366

This paper concentrates on survey and performance analysis of automatic text summarizers for various Indian languages.

II. TEXT SUMMARIZERS FOR INDIAN LANGUAGES

Various automatic text summarization systems are commercially or non-commercially available for most of the commonly used natural languages. Most of these text summarization systems are for English and other foreign languages. Moreover, for commercial products the technical documentation is often minimal or even absent. When it comes to Indian languages, automatic text summarization systems are still lacking. Various text summarizers for Indian languages are discussed below:

Islam and Masum (2004) developed corpus oriented text summarization system 'Bhasa' for Bengali language. It is based on scoring the files of corpus in which query words are having highest frequency and then producing the summary of text documents on the basis of query words by applying vector-space-term-weighting. A tokenizer is used for tokenizing the input documents and then ranking of documents is performed with text summarization on these tokenized text documents. Tokenizer is able to determine different terms, abbreviations, tags and boundary of sentences and to denote terms, headings, titles and sentences using markups by semantic and syntactic analysis. Moreover if lines are identified using shallow-linguistic text analysis then some times text summary may have dangling anaphors [3].

Das and Bandyopadhyay (2010) developed Bengali opinion text summarizer based on given topic which can determine the information on sentiments in the input text. Then this information is aggregated for denoting text summary. It applies a model on topic-sentiment for determination and aggregation of sentiments. It is implemented for theme determination at the discourse level. Moreover aggregation is performed by clustering of theme using k-means approach and by applying theme graph representation at relational level which is ultimately applied for selection of relevant sentences in summary by using page rank standard approach. The Precision, Recall and F-Score of this approach is calculated as 72.15%, 67.32% and 69.65% respectively [4].

Sarkar (2012) proposed Bengali text summarization by sentence extraction and has investigated the impact of thematic term feature and position feature on Bengali text summarization. The proposed summarization method is extraction based. It has three major steps: (1) preprocessing (2) sentence ranking (3) summary generation. The preprocessing step includes stop-word removal, stemming and breaking the input document in to a collection of sentences. After an input document is formatted and stemmed, the document is broken into a collection of sentences and the sentences are ranked based on two important features: thematic term and position. The thematic terms are the terms which are related to the main theme of a document and having TF-IDF score above a given threshold. The positional score of a sentence is computed in such a way that the first sentence of a document gets the highest score and the last sentence gets the lowest score. Long sentences are given preference in summary A summary is produced after ranking the sentences based on their scores and selecting K-top ranked sentences, when the value of K is set by the user. To increase the readability of the summary, the sentences in the summary are reordered based on their appearances in the original text [5].

Sarkar (2012) proposed another approach for summarizing Bengali news documents. It describes a system that produces extractive summaries of Bengali news documents. The ultimate objective of produced summaries is defined as helping readers to determine whether they would be interested in reading a particular document. To this end, the summary aims to provide a reader with an idea about the theme of a document without revealing the in-depth detail. The approach presented here has four major steps (1) preprocessing (2) extraction of candidate summary sentences (3) ranking the candidate summary sentences (4) summary generation. The proposed approach defines TF*IDF, position and sentence length feature in more effective way that helps in improving the summarization performance. The experimental results show that this system performs better than the lead baseline and a more sophisticated baseline that uses TF*IDF and position features both [6].

Kumar and Devi (2011) proposed Tamil language summarization system for scoring of sentences in summary using graph theoretic scoring technique. This system uses statistics of frequency of words and a term positional and weight-age calculation by string pattern for scoring of sentences [7].

Kallimani et al. (2010) proposed a text summarizer for Kannada i.e. "AutoSum" a named IR system using Text Summarization of some Regional Language in India. This system processes the input text and then decides which lines are relevant and which lines are not relevant. User interaction in this system is command based interaction. In it, text is summarized on console. The output summary of this system can be produced either in simple text or in hyper text markup language. If hyper text markup language is used in output then relevant lines are highlighted. It begins its summarization task when input text is given by user which is having 03 steps i) Command is given by user on the terminal ii) In further stages, the input moves through the system and summary is produced iii) The resulting text is sent to the terminal after summary is made or the results of summary can be highlighted in the web browser. This system uses adjectives, adverbs and nouns as key terms. The score value of each term in every sentence is determined and summed up to the score of that sentence. Every line is assigned a score based on the key terms in it [8].

Jayashree et al. (2011) proposed a text summarization system for Kannada named "Kannada text Summarizer based on Key terms Extraction". This system takes pre-classified Kannada documents obtained from online web resources and identifies the thematic words from these documents by mixing GSS (Galavotti, Sebastiani, Simi)

coefficients and Inverse-Document-Frequency techniques with Term-Frequency and then apply these extracted keywords for making summary [9].

Jayashree et al. (2012) proposed another pre-classified documents summarizer for Kannada by scoring of sentences which retrieves key terms from Kannada documents, by combining GSS (Galavotti, Sebastiani, Simi) coefficients and Inverse-Document-Frequency techniques with Term Frequency for retrieving key terms and then applies them for summarizing the text. Overall motive of this technique is to give weight-age to every term of a line, the final weight of a line is the addition of weight-age of each term in that line. Finally it selects n sentences based on sentence scores. Database is specially built for this purpose by selecting a document of a given category. Kannada text files are taken from Webdunia. Webdunia is a special web portal in Kannada that is used for latest News, Entertainment News, Sports related news, Jokes and Shopping etc. Summary is produced based on number of sentences in the input given by user. Then summary evaluation is done by comparing the human produced summary with system produced summary. Other motive of this technique is to extract different features after removing the stop words from input text. Moreover for elimination of stop words from input a new approach has been used which identifies structurally similar type of terms in any text document [10].

Patel et al. (2007) proposed a technique to text summarization for English, Hindi, Gujarati and Urdu documents. The algorithm is based on structural and statistical (rather than semantic) factors. The algorithm has been applied on document understanding conference (DUC) data English documents and various newspaper articles for other languages with corresponding stop words list and modified stemmer. To test the language independence of the summaries generated by this summarizer, it has been tested on 70 news articles of Hindi leading dailies, 50 articles of Gujarati literature and 75 new articles of Urdu from BBC web site. In almost every case, it gives degree of representative ness more than 80% [11].

Garain et al. (2006) proposed text summarization of compressed text pictures for Indian language. This system is used to summarize JBIG2 coded text pictures without using optical character recognition. Compressed pictures are decompressed and then sentences and terms are marked. Four features are determined at the level of sentences. These features are (1) Feature1: Length of sentences (2) Feature2: Position of sentences in each paragraph (3) Feature3: Thematic term features (4) Feature4: Title terms. For values of these features, lines are treated as summary lines or non summary lines. Finally this system produces a set of summary sentences. Moreover, within summary sentences are further ranked. In experiments author only considers Indian language text images. The sentence selection efficiency of this approach is 56% calculated against human generated summary [12].

Automatic text summarization software for Hindi [13] text has been commercially developed by CDAC (Centre for development of advance computing) Noida. This system has applied statistical approach, Language based approach along with heuristic approach for developing text summarization system for Hindi. This summarization software includes 1) Features based on Statistics: Term, Pair of Terms, Particular Cue terms, count, determining Value of Threshold, location of sentences and proper location scheme etc. 2) Analyzing language oriented features: Determine noun terms, terms existing together, finding stop-words, terms which are functional in nature. 3) Language oriented Psycho features: Unique or duplicate terms. 4) Feature belonging to Heuristics: sentence belonging to Title, Location, Number of words in a sentence and Table of contents etc. 5) Giving weight-age, ranking of lines etc. [13]

Gupta et al. (2012) proposed of Punjabi text summarizer. It makes extractive summary for Punjabi text by extracting the important lines based on language oriented features and features belonging to statistics of text. Every line of input text is treated as vector of different features like sentence relative length, Punjabi cue terms, Punjabi terms belonging to nouns, terms belonging to common nouns of English and Punjabi, Punjabi named entities, location of lines, Term-Frequency and Inverse-Sentence-Frequency scores for extracting thematic terms, existence of numeric data in lines etc. Duplicate sentences are eliminated in the pre processing phase. Weight-age of sentence-features which are influencing the different lines are calculated by applying regression which is a weight learning method. For each sentence, the score values of all features are calculated and final score values of all sentences are determined using equation of features and weights. Finally Punjabi sentences with top scores are selected in same order as in input text at given CR (compression ratios). In case of Punjabi news articles, Punjabi text summarizer is showing F-measure 97.87%, 95.32 and 94.63% respectively at 10%, 30% and 50% CR (compression ratios) and in case of Punjabi stories, this system shows F-measure 81.78%, 89.32% and 94.21% respectively at 10%, 30% and 50% CR (compression ratios) [14][15][21].

Kallimani et al. (2012) proposed a new technique for summarizing the longer text documents by considering one of the South Indian regional languages (Kannada). It deals with a single document summarization based on statistical approach. The purpose of summary of an article is to facilitate the quick and accurate identification of the topic of the published document. The objective is to save prospective readers' time and effort in finding the useful information in a given huge article. Moreover in case of Kannada summary, the total frequency of terms in system produced summary is more as comparative to summary produced by human and also %age term frequency is more in both the summaries because the size of summary is increased. This is clear that out of 04 lines in the 20 % summary, 75% lines i.e. 03 lines are common, Out of 05 lines in 30% summary, 80% lines i.e. four lines are common and out of 06 lines in the 40 % summary, 83.33% lines i.e. 05 lines are common. It shows that with

increase in percentage of summary size, the number of common lines have also increased [16].

Banu et al. (2007) proposed text summarizer for Tamil documents using technique of semantic graph by identifying Subject Object Predicate from individual lines for making semantic-graph of source text document and its corresponding summary generated by human experts [20].

Banu (2010) proposed another technique for summarizing documents of Tamil by using approach of sub graph for selecting lines from source document treated as text summary or another technique for generating a generic summary of document. In this system, syntax of language neutral, which is the system for representing the natural language lines has been applied for compressing the text documents. It has used syntactic analysis of the source text which makes a analysis of logical form has been used for every line. Triples of subject object predicate are selected from individual lines to generate a semantic graph of source document and its corresponding summary generated by human experts. To triples of SOP Semantic Normalization is used for reducing the frequency of nodes of semantic graph of source document. Classifier has provided training by using leaning technique based on support vector machine learning, for identifying triples of SOP from semantic graph of document which belongs to actual summary. Then this classifier is used to extract automatic summaries from test documents [17].

Keyan (2012) proposed multi-lingual (Tamil and English) multi-document summarization by neural networks. The system involves three steps. In first step, the sentences of the documents are converted into vector form. In the second step weight values are assigned to vector form based on sentence features. Depend on sentence weight value, single document summarization is done. The output of single document summarization is used as an input for multi-document Summarization. Final step is a sentence selection, in which output summary is selected based on the similarity and dissimilarity measures. Sentence similarity and dissimilarity measures are used to compare the sentences. From that, resultant summary is produced. The proposed system can be able to summarize both Tamil and English online news papers. [18]

Islam et al. (2007) proposed text summarizer for Bangla using text extraction based summarization technique and reported average highest score of 8.4 (on 0-10 scale) at 40% compression ratio [19].

III. Performance Comparison in Indian Summarizers

Garain et al. (2006) [12] proposed method for automatic summarization of JBIG2 coded textual images for Bengali text without optical character recognition (OCR) with efficiency of about 56% when judged against summarization generated by human. Islam et al. (2007) [19] proposed text summarizer for Bangla using text extraction based summarization technique and reported average highest score of 8.4 (on 0-10 scale) at 40% compression ratio. Das et al. (2010) [4] proposed topic-Based Bengali Opinion Summarization with Precision of 72.15%, Recall of 67.32% and F-measure of 69.65%. Bengali text summarization by sentence extraction is another Bengali text summarization system developed by kamal sarkar [5] and had investigated the impact of thematic term feature and position feature on Bengali text summarization with Average Unigram based Recall Score 0.4122. Automatic text summarization software for Hindi text [13] had been commercially developed by CDAC (Centre for development of advance computing) Noida. Statistics based technique, language oriented & heuristic technique had been applied for this text summarizer for Hindi. Patel et al. (2007) proposed a language independent approach to multilingual text summarization for English, Hindi, Gujarati and Urdu [11] documents based on structural and statistical (rather than semantic) factors with efficiency of 82%. Regarding Kannada, Text summarization system for Kannada named "Information Retrieval by Text Summarization for an Indian Regional Language" [8] had been proposed in 2010 using keywords extraction by taking nouns, adjectives and adverbs as keywords. Another text summarization system for Kannada named "Document Summarization in Kannada using Keyword Extraction" [9] had been proposed in 2011 using extracted key words from pre-categorized Kannada documents collected from online resources with relevant score of 0.7 for literature, 0.8 for entertainment, 0.8 for astrology and 0.76 for sports documents. Banu et al. (2007) [20] proposed text summarizer for Tamil documents using technique of semantic graph by identifying Subject Object Predicate from individual lines for making semantic-graph of source text document and its corresponding summary generated by human experts. Another Tamil text extraction system for an agglutinative language [7] had been introduced in 2011 by proposing an efficient algorithm for sentence ranking based on a graph theoretic ranking model applied to text summarization task with ROUGE-1 score 0.47. TABLE I shows the comparison of performance of some of existing summarizers for Indian languages [21].

TABLE I.
PERFORMANCE COMPARISON OF EXISTING INDIAN SUMMARIZERS [21]

Summarization systems	Performance comparison of existing summarizers for other Indian Languages	
	Accuracy (In %)	Test used
Punjabi Text Summarization System [14][15] [21]	For Stories: 89.32% (At 30% Compression Ratio) For News Documents: 95.32% (At 30% Compression Ratio)	F-Score
Bengali Summarizer using Textual Images [12]	56%	Efficiency
Bengali Summarizer using Text Extraction [19]	84% (At 40% Compression Ratio)	Efficiency

Topic based Bengali Opinion Summarizer [4]	69.65%	F-Score
Multi Lingual Summarizer for English, Hindi, Gujarati & Urdu [11]	82%	Efficiency
Document Summarizer for Kannada [9]	For Literature: 70% For Entertainment: 80% For Sports: 76%	Efficiency
Summarization from large Kannada documents using a novel approach [10]	At 30% Compression ratio: 80% At 40% Compression Ratio: 83.33%	Efficiency
Tamil text extraction system for an agglutinative language [7]	Score : 0.47	ROUGE-1

IV. CONCLUSIONS

Although various automatic text summarization systems are commercially or non-commercially available for most of the commonly used natural languages for English and other foreign languages, but when it comes to Indian languages, automatic text summarization systems are still lacking. But now days lot of research is going on for Indian regional languages and after comparing the performance of various Indian summarizers for Hindi, Punjabi, Kannada, Tamil, Gujarati and Bengali, we can conclude that they are reasonably performing well over wide range of text dataset including news documents, stories, and documents related to literature, sports and entertainment.

REFERENCE

[1] V. Gupta and G. S. Lehal, "A Survey of Text Summarization Extractive Techniques," *International Journal of Emerging Technologies in Web Intelligence*, vol. 2. , pp. 258-268, 2010.

[2] J. Cheung, "Comparing Abstractive and Extractive Summarization of Evaluative Text : Controversiality and Content Selection," *B. Sc. (Hons.) Thesis,* Department of Computer Science of the Faculty of Science, University of British Columbia, 2008.

[3] T. Islam and S. M. A. Masum, "Bhasa: A Corpus Based Information Retrieval and Summarizer for Bengali Text," *Macquarie University*, Sydney, Australia, 2004.

[4] A. Das and S. Bandyopadhyay, "Topic-Based Bengali Opinion Summarization", *International Conference COILING '10*, Beijing, pp. 232–240, 2010.

[5] K. Sarkar, "Bengali text summarization by sentence extraction," *In Proceedings of International Conference on Business and Information Management(ICBIM-2012)*, NIT Durgapur, pp. 233-245, 2012.

[6] K. Sarkar, "An approach to summarizing Bengali news documents," *In proceedings of the International Conference on Advances in Computing, Communications and Informatics*, ACM, pp. 857-862, 2012.

[7] S. Kumar, V. S. Ram and S. L. Devi, "Text Extraction for an Agglutinative Language," *Proceedings of Journal: Language in India*, pp. 56-59, 2011.

[8] J. S. Kallimani, K.G. Srinivasa and B. R. Eswara, "Information Retrieval by Text Summarization for an Indian Regional Language," *In Proceedings of International Conference on Natural Language Processing and Knowledge Engineering*, pp. 1-4, 2010.

[9] R. Jayashree, K. M. Srikanta and K. Sunny, "Document Summarization in Kannada using Keyword Extraction," *Proceedings of AIAA 2011,CS & IT 03*, pp. 121–127 , 2011.

[10] R. Jayashree, "Categorized Text Document Summarization in the Kannada Language by Sentence Ranking," *Proceedings of 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, pp. 776-781, 2012.

[11] A. Patel, T. Siddiqui and U.S. Tiwary, "A language independent approach to multilingual text summarization," *Proceedings of conference RIAO '07*, Pittsburgh PA, U.S.A., 2007.

[12] U. Garain, A. K. Datta, U Bhattacharya and S.K. Parui, "Summarization of JBIG2 Compressed Indian Textual Images," *Proceeding of 18th International Conference on Pattern Recognition (ICPR'06)*, IEEE, Kolkata, India, 2006.

[13] http://cdacnoida.in/SNLP/digital_library/text_summ.asp

[14] V. Gupta and G. S. Lehal, "Complete Preprocessing Phase of Punjabi Language Text Summarization," *International Conference on Computational Linguistics COLING'12*, IIT Bombay, India, pp. 199-205, 2012.

[15] V. Gupta and G. S. Lehal, "Automatic Punjabi Text Extractive Summarization System," *International Conference on Computational Linguistics COLING '12*, IIT Bombay, India, pp. 191-198, 2012.

[16] J. S. Kallimani, K. G. Srinivasa and B. R. Eswara, "Summarizing News Paper Articles: Experiments with Ontology Based, Customized, Extractive Text Summary and Word Scoring", *Journal of Cybernetics and Information Technologies*, Bulgarian Academy of Sciences, vol. 12, pp. 34-50, 2012.

[17] M. Banu, C. Karthika, P. Sudarmani and T.V. Geetha, "Tamil Document Summarization Using Semantic Graph Method", *International Conference on Computational Intelligence and Multimedia Applications*, IEEE, pp. 128-134, 2007.

[18] M.. K. Keyan and K.G. Srinivasagan, "Multi-Document and Multi-Lingual Summarization using Neural Networks", *Proceedings of International Conference on Recent Trends in Computational Methods, Communication and Controls*, pp. 11-14, 2012.

[19] N. Uddin and S. A. Khan, "A Study on Text Summarization Techniques and Implement Few of Them for Bangla Language", *Proceedings of international conference on Computer and information technology*, IEEE, pp. 1-4, 2007.

[20] M. Banu, C. Karthika, P Sudarmani and T.V. Geetha, "Tamil Document Summarization Using Semantic Graph Method", *Proceedings of International Conference on Computational Intelligence and Multimedia Applications*, pp. 128-134, 2007.

[21] V. Gupta and G.S. Lehal, "Automatic Text Summarization for Punjabi Language," *International Journal of Emerging Technologies in Web Intelligence*, vol. 5, pp. 257-271, 2013.

Dr. Vishal Gupta is Senior Assistant Professor in Computer Science & Engineering department at University Institute of Engineering & Technology, Panjab University Chandigarh. He has done his Ph.D. and M.Tech. in Computer Science & Engineering from Punjabi University Patiala in 2013 and 2005 respectively. He is among University toppers. He secured 82% Marks in M.Tech. Vishal did his BTech. in CSE from S.B.S. College of Engineering & Technology Ferozepur in 2003. He is Young Scientist Award Winner-2013 in Engineering & Technology at Punjab Science Congress. He has written around fifty research papers in international and national journals and conferences. He has developed a number of research projects in field of NLP including topic tracking, keywords extraction, named entity recognition, synonyms detection, automatic question answering and text summarization etc. One of his research paper on Punjabi language text processing was awarded as best research paper by Dr. V. Raja Raman at an International Conference at Panipat. He is also a merit holder in 10^{th} and 12^{th} classes of Punjab School education board.