IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

Automatic Summarization for Hindi Text Documents using Bio-inspired Computing

Vipul Dalal[1], Dr. Latesh Malik[2]

Research Scholar, CSE Department, G.H.Raisoni College of Engineering, Nagpur[1]

Computer Department, Government Engineering College, Nagpur[2]

Abstract: Summarizing given text document automatically using an intelligent algorithm is an important text mining task in the field of data mining. In this paper, we are proposing an approach for automatic summarization of Hindi text documents using bio-inspired computing. The paper mainly focuses on pre-processing, machine learning and summary evaluation phases of summarization process. Employability of bio-inspired computing for summarization gives a new dimension especially in the domain of Hindi text summarization.

Keywords: text summarization, summary evaluation, machine learning, bio-inspired computing.

I. INTRODUCTION

A summary can be defined as a brief description of the given text document which provides salient information about the document. It helps readers of the document decide whether to read the complete document or not. It may help readers who have already read the document understand the document in more appropriate way. In past two decades, automatic text summarization has gained attention of research community and a number of methods and approaches are suggested by the researchers. Basically, Automatic Text Summarization (ATS) is a process of generating a summary of an input document using some intelligent algorithm. Based on kind of input taken by the algorithm, kind of approach used in the algorithm and kind of summary generated by the algorithm, ATS can be classified into various categories.

CLASSIFICATION OF ATS

- Indicative Vs. Informative: An ATS which is based on Indicative approach generates a set of pointers to the important information in the input document instead of generating actual summary document. Whereas an informative based ATS generates summary as a separate document from the given input document. In general, informative based ATS process is preferred over an indicative based ATS process as it gives more clear idea about the document to the readers.
- Query oriented Vs. Generic: A Query oriented ATS tries to generate summary of input document as per the query, which usually consists of keywords or sentences, submitted by the readers. It involves more complex and costly summarization model since different readers may have different perspective for the same input document. Unlike query oriented ATS, a generic ATS process tries to extract all important information from the input document without needing the readers to give any form of query. This type of ATS process is preferred over query based ATS since most of times the readers have no clue about the type of information that the input document may contain.
- Multi-document Vs. Single document: A multi-document ATS process is capable of generating a single summary document from multiple input documents, whereas a single document ATS, as the name suggests, can generate a summary document from only a single input document. A multi-document ATS process, in general, is more complex as compared to a single document ATS as maintaining the summarization model consistent across multiple document is more difficult as compared to a single document.
- Cross language Vs. Single language: A Cross language ATS process has a capacity to generate a summary document that is in a different language then the input document. The language of the summary document is usually chosen by the readers in the most flexible form of cross language ATS process. There are two types of cross language ATS architectures possible. The first one in which the corpus database contains the documents in various different languages such as English, French, Spanish, Japanese, etc and the language of the summary document is fixed say English. The second one in which the corpus database contains all the documents in one specific language, say English and the language of the summary document is as chosen by the readers. It is obvious that the cross language ATS process is more complex as different languages differ in terms of grammar, syntax and semantic dependencies among the phrases which the ATS process must take into consideration.
- Abstractive Vs. Extractive: Abstractive summarization consists of understanding the original text and re-telling it in fewer words. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

original text document. Figure 1 depicts this procedure. Abstracts generated using this method may or may not contain the sentences from the original document. Usually abstraction based methods involve more complexity with respect to understanding, compaction and condensation. The natural language generation is also a complex task. So, in general, abstractive summarization methods are more complex but they are capable of generating summaries that are more like human generated summaries.

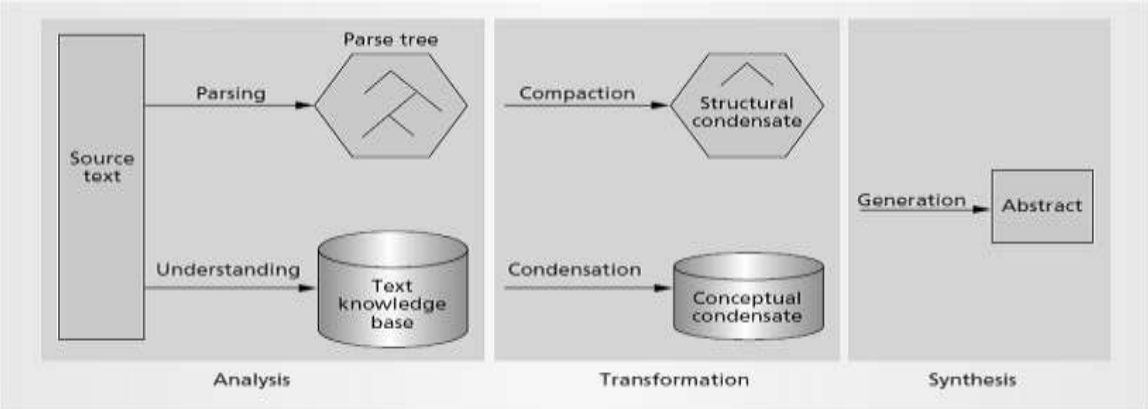

Figure 1. Architecture of an Abstractive ATS system
(Source: "The Challenges of Automatic Text summarization", Udo, Hahn and Inderjeet Mani)

In extractive ATS process, key textual elements such as keywords, clauses or sentences are extracted from the input document using linguistic and statistical analyses.

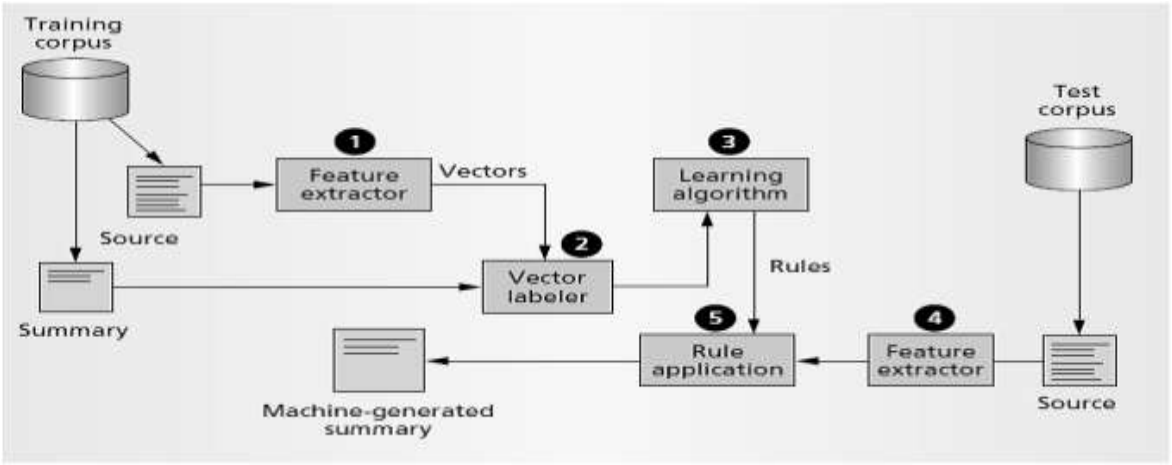

Figure 2. Architecture of an extractive ATS system
(Source: "The Challenges of Automatic Text summarization", Udo, Hahn and Inderjeet Mani)

This extracted text is directly used as a summary document. In extraction based summarization, usually lists of features called feature vectors are extracted from the training corpus and a feature labeler along with summary documents are used to label these feature vectors. A machine learning algorithm is then used to train a classifier to generate rules from these labeled feature vectors. When a document is submitted for summarization, feature vectors are extracted which are labeled using the constructed rules and summary is generated from these labeled feature vectors and corresponding sentences from the input document. Figure 2 describes this procedure.

SEMANTIC GRAPH

Capturing semantic structure of a document is essential for text summarization process [22]. The semantic structure of a document can be captured using semantic graph. One of the possible ways to construct a semantic graph of a document is to use logical form triple subject-verb-object as basic elements. The first step is to perform syntactic analysis of individual sentences to obtain Part-Of-Speech (POS) tag and dependency tag for each word in a sentence. These tags can then be used to extract logical form triple or the semantic structure of the sentence.

Co-reference resolution and semantic normalization are to be done before semantic graph can be constructed. Terms with different surface forms may refer to the same entity. Identifying such terms is referred to as co-reference

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

resolution. Semantic normalization is the process in which each term in the logical form triple is expanded using Wordnet to identify those terms that refer to the same concept. Now the logical form triples having identical or similar terms are merged together to construct the semantic graph. Figure 3 describes the process of semantic graph generation for English text.

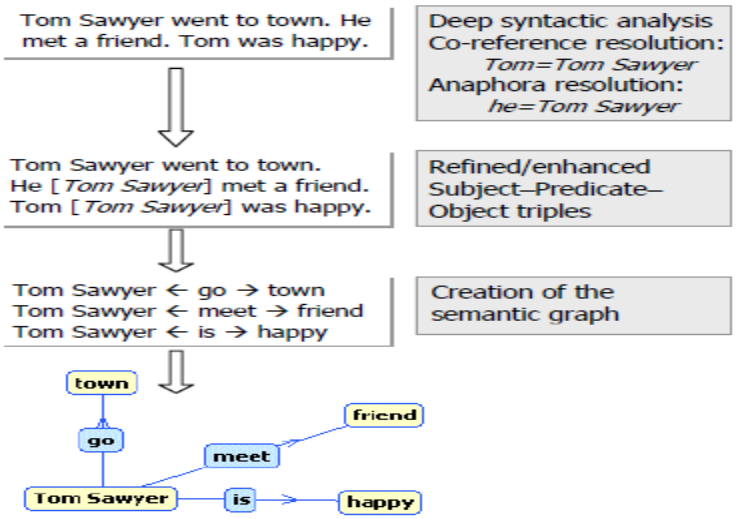

Figure 3. The process of Semantic Graph generation
(Source: "Learning Sub-structures of Document Semantic Graphs for Document Summarization", Jurij Leskovec, Marko Grobelnik, Natasa Milic-Frayling)

II. RELATED WORK

AUTOMATIC TEXT SUMMARIZATION USING BIO-INSPIRED ALGORITHMS

Few efforts for automatic text summarization based on Bio-inspired methods are as follows.

In the area of text summarization, M. S. Binwahlan et al [14] introduced a work for feature selection. They exploited five features regarding to text summarization and the PSO was used to train the system to obtain the weights of each feature. These weights have been employed in their next work [15] to generate the best summary. The results shown that, the proposed PSO method generate summaries which are 43% similar to the manually generated summaries, while MS-Word summaries are 37% similar.

Albaraa Abuobieda M. Ali et al [16] presented a feature selection method using (pseudo) Genetic probabilistic-based Summarization (PGPSum) model for extractive single document summarization. The proposed method, working as features selection mechanism, was used to extract the weights of features from texts. Then, the weights were used to tune features' scores in order to optimize the summarization process. In this way, important sentences were selected for representing the document summary. The results showed that, their PGPSum model outperformed Ms-Word and Copernic summarizers benchmarks by obtaining a similarity ratio closest to human benchmark summary.

AUTOMATIC TEXT SUMMARIZATION FOR DOCUMENTS WRITTEN IN INDIAN LANGUAGES

A few attempts have been made by researchers to propose methods for ATS for documents written in Indian languages such as Hindi, Bengali, etc.

Kamal Sarkar [19] proposed an extraction based approach for Bengali text summarization. His approach has three major steps: 1) pre-processing 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. For sentence ranking he used thematic terms and sentence position as features. These features were combined to generate rank for individual sentences. The sentences with top-k score are extracted to form the summary. For evaluation purpose the system generated summaries were compared with human extracted summaries.

Vishal Gupta and Gurpreet Singh Lehal [21] suggested an approach for pre-processing phase for Punjabi text summarization. Their work mainly concentrates on Punjabi language stop words removal, noun stemming, finding common English-Punjabi noun words, finding Punjabi language proper nouns and identification of cue phrases in a sentence. These operations were performed in the sequence given above so as to generate output which can be helpful for developing NLP tools for Punjabi language.

Chetana Thaokar and Latesh Malik [27] proposed an idea for summarizing Hindi text using sentence extraction method. They used Hindi Wordnet to tag POS of words for checking Subject-Object-Verb (SOV) of the sentence. They

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

also employed genetic algorithm to optimize the generated summary so as to maximize the theme coverage and to minimize redundancy.

III.PROPOSED APPROACH

PRE-PROCESSING

The input document, be it a training document or a document to be summarized, first passes through pre-processing phase. The first step of the pre-processing phase is document parsing.

- Parsing

The parsing process first tokenizes each sentence and for each word it gives root word after stemming, Part-Of-Speech (POS) tag, dependency tag and word position in the dependency tree.

- Feature space

The proposed approach extracts three types of features from the document.

o Document Discourse Structure features: This includes – Word frequency, Sentence position, Sentence-to-sentence similarity, Sentence length, TF-ISF.

o Linguistic features: This includes – SOV tags, POS tags, Dependency tags

o Semantic Graph features: This includes – PageRank, Hub, Authority, Number of in-coming links, Number of out going links and Number of next neighbors. These features are applicable only to the words that are identified as Subject/Object in a sentence since these words are the nodes in the semantic graph. The words that are identified as verbs are the edges in the graph and so these features are not applicable.

The table 3.1 gives summary of number of attributes included in each category. These figures are specified after converting each nominal variable into required number of binary variables.

TABLE 3.1 NUMBER OF FEATURES IN EACH CATEGORY

	Document Discourse Structure features	**Linguistic features**	**Semantic Graph features**	**Total**
Subject/Object	05	38	06	49
Verb	05	38	--	43

Finally when features of subject, object and verb are combined to form feature vector for a single SOV triple then the length of this vector is 49+49+43=141 features. It consists of binary and non-binary values.
In the final stage of pre-processing phase, the features are normalized in the range of 0 to 1 and combined to represent individual SOV triples.

ADAPTATION OF PSO ALGORITHM FOR ATS

At the end of pre-processing phase of documents from the training corpus, the training set consists of feature vectors of SOV triples from the training documents and are labelled either 1 or 0 depending up on if the SOV triple occurred in the sub-graph of corresponding summary document or not. This means that the training set contains two types of SOV triples. One representing all those sentences that should be included in the summary and these triples are labelled 1. The other representing non-summary sentences and are labelled 0. A Machine Learning algorithm can be used to train a classifier using these labelled triples. In the proposed approach the PSO algorithm is used for the same.

PSO BASED DATA CLUSTERING FOR ATS

PSO is well known for its optimization capabilities and has been successfully employed for solving many real world problems as discussed in the previous chapter. It is population based optimization approach in which a collection of agents or particles move in the solution space searching for optimal solution. Movement of each particle to a next position depends up on its own best position explored so far and the global best position explored by any other particle in the swarm. When applied for data clustering, each particle has a set of K cluster centroids (K=2, in this case) and searches for the set of optimal centroids. Intra-cluster distance can be used as fitness function to evaluate fitness of all the particles.

The particle with the smallest intra-cluster distance, that is, the highest fitness value is declared as the global best solution. Initially, all the particles are assigned a random position in the search space and a random velocity value to move around the space. This means that initially each particle randomly selects any two triples from the training set as initial cluster centroids and then iteratively searches for the optimal centroids as per the basic PSO algorithm.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

IV. EVALUATION OF GENERATED SUMMARY

In the proposed approach the generated summary is evaluated based on precision, recall, F1 score and G score measures. To calculate these measures the generated summaries are compared with the human extracted summaries since there is no benchmark evaluation system available for Hindi text summarization as of today.

PRECISION

It is given by the following equation.

$$precision = \frac{no\ of\ summary\ sentences\ extracted\ that\ match\ with\ human\ exctracted\ summary}{total\ number\ of\ sentences\ extracted}$$

With respect to text summarization task, precision represents the probability that the sentence extracted by the system as a summary sentence is actually a summary sentence as per the human extracted summary. It indicates how good the system is at picking up or selecting a sentence as a summary sentence. A precision value of 1 indicates that all the sentences extracted as summary sentences are actually summary sentences. It is also called accuracy of the system

RECALL

It is given by the following equation.

$$recall = \frac{no\ of\ summary\ sentences\ extracted\ that\ match\ with\ human\ exctracted\ summary}{no\ of\ actual\ summary\ sentences\ in\ human\ extracted\ summary}$$

With respect to text summarization task, recall represents the probability that a sentence in the document that is actually a summary sentence will be picked up or selected by the system as a summary sentence. It represents completeness of the system. A recall value of 1 indicates that all the actual summary sentences in the document are selected by the system. It is also called sensitivity of the system.

The problem with precision measure is that it specifies how good the system is at selecting a sentence from the document as a summary sentence but it doesn't specify whether the system is capable of selecting all the actual summary sentences as the summary sentences.

The problem with recall measure is that it specifies what fraction of actual summary sentences in the document are selected by the system, but it doesn't specifies whether the system is wasting efforts by selecting those sentences also that are actually not summary sentences.

F1 SCORE

It is given by the following equation.

$$F1 = 2\frac{precision \cdot recall}{precision + recall}$$

It combines precision and recall measures and is the harmonic mean of these two measures. It is called so because both precision and recall are evenly weighted. F-measure has an intuitive meaning. It tells us how precise the summarizer is (how many summary sentences it selects correctly), as well as how robust it is (it does not miss a significant number of actual summary sentences).

G SCORE

It is given by the following equation.

$$G = \sqrt{precision \cdot recall}$$

It is geometric mean of precision and recall.

V. EXPERIMENTAL SETUP AND DATASET

The proposed approach is implemented entirely using Java platform. A Hindi text document corpus, CLINSS (Cross Language Indian News Story Search), was downloaded from the Internet. This corpus contains Hindi news articles related to politics, events, sports, historical incidents, stories, etc. About 50 documents were selected as training document set from this corpus and manually processed for anaphora resolution and co-reference resolution. To parse the input document, Hindi Dependency Parser developed by Siva Reddy [28] was used. It is claimed by the developer that the parser has accuracy of 72%.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

RESULTS

Table 5.1 below gives the overall metrics obtained for the system.

TABLE 5.1 OVERALL PERFORMANCE METRICS FOR THE SYSTEM

Recall	60
Precision	42.86
F1 score	50.01
G score	50.71

COMPARISON WITH EXISTING WORK

The proposed approach is inspired from [22]. In their work [22], the researchers have used semantic graph based features along with conventional statistical features for summarizing English text. They used SVM to train the classifier. In the proposed work we have used PSO based classifier. In the existing work, the researchers had used varying size training set of 10, 20, 50 and 100 documents to train the classifier. The performance of the classifier for training set of 50 documents is given in table 5.2 with which we can compare performance of the proposed system.

TABLE 5.2 PERFORMANCE OF THE EXISTING WORK

Training set	Precision	Recall	F1-score	G-score
50 documents	25.75	72.81	38.04	43.3

As it can be seen in table 5.1, like existing work, the proposed system has higher recall rate as compared to precision.

VI. CONCLUSION

In this work we have proposed an approach for summarizing Hindi text document using semantic graph and particle swarm optimization algorithm. The concept of semantic graph has been used extensively in document analysis including summarization of English text documents. It captures semantic structure of the document which is one of the important information for generating meaningful summaries. It has never been applied to Hindi text for summarization purpose. Similarly, the concept of particle swarm optimization has also been used successfully in variety of applications including English text summarization. It has proven ability in searching optimal solution, in spite of large dimensionality of the solution space. But its potential was never explored in the domain of summarizing Hindi text.
So, our work gives a new dimension to the field of text summarization especially to the documents written in Indian languages.

REFERENCES

[1] 1. Luhn, H.P., 1958. “The automatic creation of literature abstracts”. IBM J. Res. Develop., 2: 159-165.
[2] 2. P. B. Baxendale, "Machine-made index for technical literature: an experiment," IBM J. Res. Dev., vol. 2, pp. 354-361, 1958.
[3] 3. Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM, 16(2):264-285.
[4] 4. Lin, C.Y. 1999. “Training a selection function for extraction”. Proceedings of the 18th Annual International ACM Conference on Information and Knowledge Management, pp:55-62.
[5] 5. Massih R. Amini, Nicolas Usunier, and Patrick Gallinari, "Automatic Text Summarization Based on Word-Clusters and Ranking Algorithms", ECIR 2005, LNCS 3408, pp. 142–156, (2005).
[6] 6. Rafeeq Al-Hashemi, "Text Summarization Extraction System (TSES) Using Extracted Keywords", International Arab Journal of e-Technology, Vol. 1, No. 4, June, pp. 164-168, (2010).
[7] 7. Jade Goldstein, Jaime Carbonell. “SUMMARIZATION: (1) USING MMR FOR DIVERSITY- BASED RERANKING AND (2) EVALUATING SUMMARIES”. Carnegie Group Inc.'s Tipster III Summarization Project
[8] 8. Aysun Güran, Eren Bekar, Selim Akyokuş “A Comparison of Feature and Semantic-Based Summarization Algorithms or Turkish”. INISTA 2010, International Symposium on Innovations in Intelligent Systems and Applicaitons, 21-24June 2010, Kayseri & Cappadocia,TURKEY.
[9] 9. Ono, K., Sumita, K., and Miike, S. (1994). “Abstract generation based on rhetorical structure extraction.” In Proceedings of Coling '94, pages 344{348, Morristown,
[10] NJ, USA.
[11] 10. Marcu, D. (1998a). “Improving summarization through rhetorical parsing tuning”. In Proceedings of The Sixth Workshop on Very Large Corpora, pages 206-215, pages 206,215, Montreal, Canada.
[12] 11. Giuseppe Carenini and Jackie Chi Kit Cheung, “Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality”.
[13] 12. Pierre-Etienne Genest, Guy Lapalme, “Framework for Abstractive Summarization using Text-to-Text Generation”, Workshop on Monolingual Text-To-Text Generation, pages 64–73,Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 64–73,Portland, Oregon, 24 June 2011. c 2011 Association for Computational Linguistics.

[14] 13. Vipul Dalal, Dr. Latesh Malik.: "A Survey of Extractive & Abstractive Text Summarization", 6th International Conference on Emerging Trends in Engineering & Tecnology (ICETET), 2013
[15] 14. M. S. Binwahlan, Salim, N., & Suanmali, L.: "Swarm based features selection for text summarization", International Journal of Computer Science and Network Security IJCSNS, vol. 9, pp. 175-179, 2009b.
[16] 15. M. S. Binwahlan, Salim, N., & Suanmali, L.: "Swarm Based Text Summarization", Computer Science and Information Technology – Spring Conference, 2009. IACSITSC '09. International Association of, 2009, pp. 145-150.
[17] 16. Albaraa Abuobieda M. Ali, Naomie Salim, Rihab Eltayeb Ahmed, Mohammed Salem Binwahlan, Ladda Sunamali, Ahmed Hamza.: "Pseudo Genetic And Probabilistic-Based Feature Selection Method For Extractive Single Document Summarization", Journal of Theoretical and Applied Information Technology, 15th October 2011. Vol. 32 No.1, ISSN: 1992-8645, E-ISSN: 1817-3195.
[18] 17. Alkesh Patel, Tanveer Siddiqui, U. S. Tiwary.: "A language independent approach to multilingual text summarization", Conference RIAO2007, Pittsburgh PA, U.S.A. May 30-June 1, 2007 - Copyright C.I.D. Paris, France
[19] 18. Naresh Kumar Nagwani, Shrish Verma.: "A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 17– No.2, March 2011.
[20] 19. Kamal Sarkar.: "Bengali Text Summarization By Sentence Extraction"
[21] 20. Upendra Mishra, Chandra Prakash.: MAULIK: "An Effective Stemmer for Hindi Language", International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397, Vol. 4 No. 05 May 2012
[22] 21. Vishal Gupta, Gurpreet Singh Lehal.: "Preprocessing Phase of Punjabi language Text Summarization"
[23] 22. Jurij Leskovec, Natasa Milic-Frayling, Marko Grobelnik.: "Extracting Summary Sentences Based on the Document Semantic Graph, Microsoft Research, Microsoft Corporation
[24] 23. Regina Barzilay, Michael Elhadad.: "Using Lexical Chains for Text Summarization", In Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97). Madrid: ACL, 1997. 10-17.
[25] 24. Kavita Ganesan, ChengXiang Zhai, Jiawei Han.: "Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions".
[26] 25. Eduard Hovy and Chin-Yew Lin.: "Automated Text Summarization in SUMMARIST", In I. Mani and M. Maybury (eds), Advances in Automated Text Summarization. MIT Press.
[27] 26. Udo Hahn, Inderjeet Mani. : "The Challenges of Automatic Text Summarization", IEEE Computer Society Press Los Alamitos, CA, USA, Volume 33 Issue 11, November 2000, Page 29-36 ISSN:0018-9162.
[28] 27. Chetana Thaokar, Latesh Malik, "Test Model for Summarizing Hindi Text using Extraction Method", Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013).
[29] 28. Reddy Siva. Natural Language Processing Tools. December. 2012 URL: http://sivareddy.in/downloads