A Review on Automatic Text Summarization Based on Gujarati Language

Bhavin Jadav
Department of Computer Science and Technology , Uka Tarsadia University Bardoli Mahua Road, Bardoli-394350,Gujarat, India.
bhavinjadav61@gmail.com

Jenish Keriwala
Department of Computer Science and Technology , Uka Tarsadia University Bardoli Mahua Road, Bardoli-394350,Gujarat,India.
keriwalajenish@gmail.com

Vipul Gamit
Department of Computer Science and Technology , Uka Tarsadia University Bardoli Mahua Road, Bardoli-394350,Gujarat,India.
vipul.gamit24@gmail.com

Hardik Vyas
Department of Computer Science and Technology , Uka Tarsadia University Bardoli Mahua Road, Bardoli-394350,Gujarat,India.
hardik.vyas@utu.ac.in

Abstract **— as the problem of information overload has grown, and as the quantity of data has increased day by day. Readers are overloaded with lengthy text documents. All computer users are particularly affected by this predicament. This paper focuses to investigate some of the most reliable and efficient approaches of automatic text summarization for Guajarati language text document. A summarization system consists of reduction of a text document to generate a new form which conveys the key meaning of the contained text. Pre-processing phase and some of methods are used to resolve native language text summarization paradigm.**

***Keywords*— Summarization, Gujarati NLP, Pre-processing, Stop words, Named Entity Recognition.**

I. INTRODUCTION

Computerized summarization had begun 60 years ago by H.P. Luhn's "The Automatic creation of Literature Abstract" in IBM journal of Research Development 2(2) 159-165 1958. [1] Before knowing about Automatic Text Summarization we must know what is Text Summarization? Text Summarization is a process of converting larger text document in short document that contains overall meaning which is done by human. When we talk about Automatic Text Summarization it is nothing but doing same process of summarization with the help of computer program .The product of this process also contains the most important points of the original text [1] that reduces user's time to find key information from the document.

Nowadays, a swarm of information comes from the Internet in a Guajarati textual form as well. Automatic summarization is a reliable and effective way to resolve the information overload problem. There are mainly two approaches to summarize text document Extraction and Abstraction although automatic summarization is a hot topic of research nowadays, only very few software tools are available to the end users and none of them are particularly designed for Gujarati language. The reason for this should be the low quality of automatically produced summaries. In general, creation of a good summary using abstraction method requires a lot of intelligence.

Current need for Automatic Text Summarization: [2]

In today's environment people don't have much time to read entire document in short period of time to get key information there must be some tools to reduce the time spent in manually extracting the main idea from the original text document.

- News article summary
- Email summary
- Short message news on mobile
- Information summary for businessman, government officials, researchers, online search engines to receive the summary of pages found.

Types of Text Summarization [3]**:**

Based on Document:

Text summarization techniques can be classified on the basis of number of text documents available in the text database.

1. **Single Document Summarization:** Automatic Text Summarization applied on single document to produce abstract headline.
2. **Multiple Document Summarizations:** Multiple documents are given to produce gist of the document
   a) A series of news stories of the same event
   b) A set of webpages about some topic or questions

Based on Language:

1. **Single Language Summarization:** Single language document is given to produce summary in such summarization whole document is written in single language only. So we don't need to identify language used in document.
2. **Multiple Language Summarizations:** In such type of summarization document may contain multiple languages. It is an automatic procedure designed to extract the information from multiple text documents written about the same topic. The multi-

document summarization task is much more complex than summarizing a single language document.

a) Need to perform language identification
b) Different methods used to summarize different language.

II. LITARATURE REVIEW

Approaches to Automatic Summarization:

1. **Extraction**

In extraction based text summarization important text segments of the original text are identified and presented as they are [3].

Advantages of Extraction:

-Quicker than abstraction approaches. Take less time to summarize a document.

-No depth knowledge of language is required [3].

Disadvantages of Extraction [3]**:**

-Inconsistencies

-Anaphoric reference may break

-Relationship between sentences is not managed.

2. **Abstraction**

In Abstraction based text summarization original text is interpreted and is written in a condensed form so that the resulting summary contains the essence of the original text [4]..

Pre-processing phases of Text Summarization:

Pre-processing phase of the summarization Process is employed with the aim of decreasing the number of words for the further summarization processes. [15]

Pre-processing phase includes words identification, sentences identification, stop words elimination, language stemmer for nouns and proper names, and elimination of duplicate sentences or words.

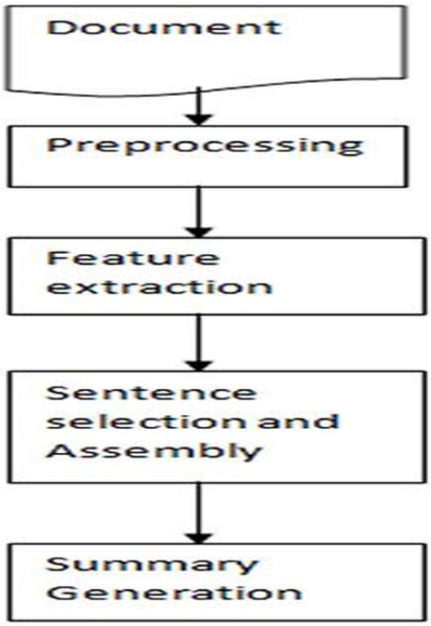

Figure 1: pre-processing phase of text summarization. [7]

Stop word Elimination:

Stop words are the words which appear frequently in document but provide less meaning in identifying the important content of the document such as "a□, "an□, "the□, etc. [7]

Content Words = Total word – Stop words [6]

Given below table contains some stop words in Hindi, English, and Gujarati Languages.

ENGLISH	HINDI	GUJARATI
An	यह	તો
As	और	હતી
Be	होता	છે
By	तो	હતા
Com	हुआ	પણ
Else	बाद	સાથે
Etc.	लिए	શકાય

Table 1: Sample List of sample stop words

In case of English language, the stop words list is directly available but for other languages we have to make a list of stop words. For that we are finding the most frequent words that are occurring throughout the test sets. These words are then analysed by language expert to make a final list of stop words. We have prepared list of 275 stop words for Hindi, 258 stop words for Gujarati and 398 stop words for Urdu initially and these lists are updated as and when required.[16]

Lexical Analysis:

Process of converting sequence of character in the sequence of token. A token is a string of characters, categorized according to the rules as symbols (e.g. IDENTIFIER, NUMBER, COMMA, etc.). Objective of lexical analysis process is the identification of the words in a text.

TOKENS: Money, Number, Time and Date, Gender, etc. [8]

Example:

સચીન વિશ્વ નો મહાન ખેલાડી છે.સચીને પોતાની કારકિર્દી મા ઘણી ખ્યાતિઓ પોતાના નામ કરી છે.હાલમા સચીન સંસદનો કાયૅરત સદસ્ય છે.

Generate token:

<GEN>સચીન</GEN> વિશ્વ નો મહાન ખેલાડી છે.<GEN>સચીને</GEN> પોતાની કારકિર્દી મા ઘણી ખ્યાતિઓ પોતાના નામ કરી છે.હાલમા<GEN>સચીન</GEN> સંસદનો કાયૅરત સદસ્ય છે.

Output:

સચીન વિશ્વ નો મહાન ખેલાડી છે.તેણે પોતાની કારકિર્દી મા ઘણી ખ્યાતિઓ પોતાના નામ કરી છે.હાલમા તે સંસદનો કાર્યરત સદસ્ય છે.

Stemmer for noun and proper nouns:

Stemming is a technique for converting derived terms into corresponding root or stem words. The major task of a stemmer is to find root words that are not in original form of and also absent in the language specific dictionary. [9]

$\{stem_1+suffix_1, stem_2+suffix_2, \ldots, stem_L+suffix_L\}$

પાણીમાં={પ + ાણીમાં, પા + ણીમાં, પાણ + ીમાં, પાણી + માં, પાણીમ + ાં, પાણીમા + ં, પાણીમાં + NULL}

Figure 3: Example of stemming in Gujarati language [10]

Steps to reach root word for Hindi language: [17]

Step I: Enter Input Text in Hindi

Step II: Segment this text into words and search each Hindi word in Hindi Word-Net [18].If that word is found in Hindi Word-Net and is tagged as noun then word is already in root form. Else if word is not found in Hindi Word-Net then go to Step III for performing its stemming

Step III: If Suffix of that word matches with any of suffixes: or or or or or or or Then eliminate this suffix from end of that word and this stemmed word is again searched in Hindi Word-Net for noun possibility. If word is tagged as Noun then that word is returned as result. Otherwise go to Step IV.

Step IV: If Suffix of that word matches with any of suffixes: or then eliminate this suffix from end of that word and add at end of stemmed word. Search this resulting word in Hindi Word-Net for noun possibility. If word is tagged as Noun Then that word is returned as result. Otherwise go to Step-V.

Step V: If Suffix of that word matches with any of suffixes: or or or or or Then eliminate this suffix from end of that word and add at end of stemmed word. Search this resulting word in Hindi Word-Net for noun possibility. If word is tagged as Noun Then that word is returned as result. Otherwise go to Step VI.

Step VI: If Suffix of that word matches with then eliminate this suffix from end of that word and add at end of stemmed word. Search this resulting word in Hindi Word-Net for noun possibility. If word is tagged as Noun Then that word is returned as result. Otherwise go to Step VII.

Step VII: The word is not Hindi Noun.

Procedure Input: “Medicines”, “Sparrows”, “Girls”, “Cats”

Procedure Output: “Medicine”, “Sparrow”, “Girl”, “Cat”

Suffix	Replacement	Count	Example	Category
ી(*ee*)	વું (*vun*)	392	પીગળી (*peegalee, melting*) - પીગળવું (*pigalavu, to melt*)	verb
ના (*naa*)	-	358	સલાહકારના (*salaahakaaranaa, advisor's*) - સલાહકાર (*salaahakaar, advisor*)	noun
ની (*nee*)	-	347	સલાહકારની (*salaahakaaranee, advisor's*) - સલાહકાર (*salaahakaar, advisor*)	noun
ને (*ne*)	-	318	સલાહકારને (*salaahakaarane, to advisor*) - સલાહકાર (*salaahakaar, advisor*)	noun
ા(*aa*)	્. (*un*)	278	બતાવવા (*bataavavaa, for showing*) - બતાવવું (*bataavavun, to show*)	verb
ા(*aa*)	્ (*un*)	275	જોડાયેલા (*jodaayelaa, connected(p)*) - જોડાયેલું (*jodaayelu, connected(s)*)	adj
માં (*maan*)	-	273	વિચારમાં (*vichaarmaan, in thought*) - વિચાર (*vichaar, thought*)	noun
ે(*e*)	વું (*vun*)	269	જાળવે (*jaalave, preserve*) - જાળવવું (*jaalavavun, preserve*)	verb
તા (*taa*)	વું (*vun*)	258	બતાવતા (*bataavataa, showing*) - બતાવવું (*bataavavun, to show*)	verb
નો (*no*)	-	258	કટોકટીનો (*katokateeno, of urgency*) - કટોકટી (*katokatee, urgent*)	noun

Table 2: Example of stemming in Gujarati language [14]

Named Entity Recognition:

Named Entity Recognition (NER) is used to locate and classify predetermined classes such as the names of persons, organizations, locations, concepts etc. Various rules have been developed like prefix rule, suffix rule, proper name rule, middle name rule and last name rule.

ਪੁਰਾ (purā)	ਪੁਰੀ (purī)	ਜੀਤ (jīt)
ਮੀਤ (mīt)	ਜੋਤ (jōt)	ਦੀਪ (Dīp)

Figure 4: Example of last name suffix in Punjabi language [11]

ਕੁਮਾਰ (Kumār)	ਲਾਲ (lāl)
ਕੌਰ (kaur)	ਸਿੰਘ (siṅgh)
ਕੁਮਾਰੀ (kumārī)	ਕੁਮਾਰੀ (kumārī)

Figure 5: Example of first name suffix in Punjabi language [11]

ISSUE: above examples shows the NER in Punjabi language by identifying prefix and postfix of person name but only for the Punjabi community names. But in Gujarati language document it is not necessary that all person names will have same prefix and postfix same problem is for any natural language.

III. METHODS

Abstractive summarization techniques are broadly classified into two categories: Structured based approach and Semantic based approach.

1. **Structured Based Approach**

Structured based approach encodes most important information from the document(s) through cognitive schemas [19] such as templates, extraction rules and other structures such as tree, ontology, lead and body phrase structure. Different methods used this approach are discussed as follows.

1.1 Tree based method

-Sentence generation
-Determine common phrases
-Order them
-Language generator (FUF/SURGE)

This technique uses a dependency tree to represent the text/contents of a document. Different algorithms are used for content selection for summary e.g. theme intersection algorithm or algorithm that uses local alignment across pair of parsed sentences. The technique uses either a language generator or an algorithm for generation of summary. Related literature using this method is as follows. [18]

1.2.Template based method

-Linguistic patterns (study of grammar)
-Multi document summarizer GISTEXTER use output of CICERO extraction system to generate final summary

This technique uses a template to represent a whole document. Linguistic patterns or extraction rules are matched to identify text snippets that will be mapped into template slots. These text snippets are indicators of the summary content. Related literature using this method is as follows. The approach proposed in [20] presents a multidocument summarization system, GISTEXTER, which produces abstract summaries of multiple newswire/newspaper documents relying on the output of the CICERO Information extraction system.

2. **Semantic Based Approach**

In Semantic based method, semantic representation of document is used to feed into natural language generation system. This method focuses on identifying noun phrases and verb phrases by processing linguistic data [21]. Different methods using this approach are discussed here

2.1 Multimodal semantic models

- Involve in text and image multimodal document
- Knowledge representation is based on concept
- It will check the density matrix
- Generate important concepts expressed as a sentence
- Adv. Produce abstract summary
- Dis. manually evaluated by human

In this method, a semantic model, which captures concepts and relationship among concepts, is built to represent the contents (text and images) of multimodal documents. The important concepts are rated based on some measure and finally the selected concepts are expressed as sentences to form summary. [18]

2.2 Semantic Graph Based Method

This method aims to summarize a document by creating a semantic graph called Rich Semantic Graph (RSG) for the original document, reducing the generated semantic graph, and then generating the final abstractive summary from the reduced semantic graph. The abstractive approach proposed by [22] consists of three phases as shown in figure 1.

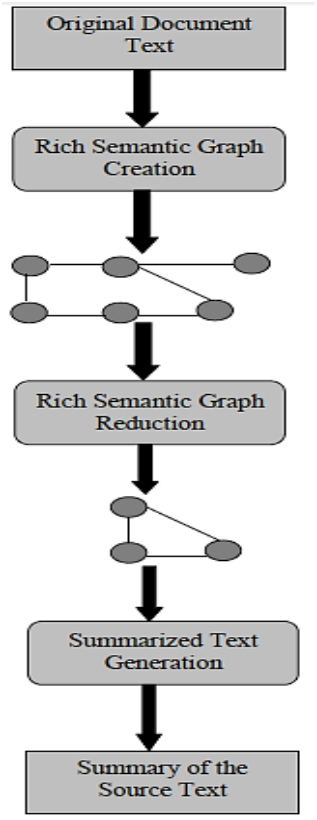

Figure 6: Semantic Graph Reduction for Abstractive Text Summarization [18]

IV. CONCLUSION

From above literature review we conclude that there is still a long trail to walk in the field of automatic text summarization in Gujarati language. Summarizing natural languages like Hindi, Gujarati etc. is more challenging task in compare to English language. To perform automatic summarization in natural languages we require dipper or linguistic knowledge of particular language. Abstractive summarization is preferred to get fruitful output in natural languages like Gujarati but there are some issues in pre-processing phase with the aim of decreasing the number of words.to identify stop word we can create dictionary manually because there is no stop word dictionary is available for Gujarati language. In NER to identify location from dictionary but to identify person name is challenging task it is not necessary that all names will have some common prefix or postfix words. Output of Pre-processing phase will be used by abstractive summarization methods like tree based method, multimodal based method, template based methods or graph based method to perform further summarization process to get final summary.

V. REFERENCES

[1] O.M. Foong, A. Oxley and S. Sulaiman,"Challenges and Trends of Automatic Text Summarization," ISSN: 0976–5972, vol.1, Issue 1, pp. 34-39, 2010.

[2] M. Haque*, et al., "Literature Review of* Automatic Multiple Documents Text Summarization," *International Journal of Innovation and Applied Studies, vol. 3, pp.* 121-129, 2013.

[3] Karel Jezek and Josef Steinbergerm ," Automatic Text Summarization," The state of art 2007 and new challenges.

[4] R.V.V Murali Krishna and Ch. Styananda Reddy, "A Sentence scoring method for extractive text summarization based on natural language queries ," IJCSI,vol.9,Issue 3,No 1,May 2012.

[5] Gleb Sizov , "Extraction-Based Automatic Summarization", Norwegian University of Science and Technology, June 2010.

[6] Aqil Burney, Badar Sami, Nadeem Mahmood and Zain Abbas ," Urdu Text Summarizer using Sentence Weight Algorithm for Word Processors ",IJCA (0975-8887) volumn 46-No.19 ,May 2012.

[7] Anita.R.Kulkarni and Dr Mrs. S.S.Apte, "An automatic Text Summarization using feature terms for relevance measure",IOSR-JCE , e-ISSN*:* 2278-0661, p- ISSN: 2278-8727Volume 9*,* Issue *3* (Mar. - *Apr.* 2013), PP 62 -66.

[8] R. Barzilay and M. Elhadad, "Using Lexical Chain for Text Summarization", 84105 Israel.

[9] Vishal Gupta, "Hindi Rule Based Stemmer for Nouns," International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 1, January 2014

[10] Kartik Suba, Dipti Jiandani , and Pushpak Bhattacharya," Hybrid Inflectional Stemmer and Rule-based Derivational Stemmer for Gujarati",Department of Computer Engineering Dharmsinh Desai University.

[11] Vishal Gupta and Gurpreetsingh Lehal,"Named Entity Recognition for Punjabi Language Text Summarization", Department of Computer Science Punjab University,Patiala,India.

[12] http://en.wikipedia.org/wiki/Automatic_summarization ,Reviewed on 5th October 2014

[13] R. Barzilay and M. Elhadad, "Using Lexical Chain for Text Summarization", 84105 Israel.

[14] Niraj Aswani, Robert Gaizauskas, "Developing Morphological Analysers for South Asian Languages: Experimenting with the Hindi and Gujarati Languages", EU funded MUSING project (IST-2004-027097).

[15] Vishal Gupta, "Processing Phase of Summarizer for Multiple News Single Punjabi Documents", International Journal of Engineering Trends and Technology (IJETT) – Volume 6 Number 7- Dec 2013.

[16] Alkesh Patel, Tanveer Siddiqqui, U. S. Tiwari , " A language independent approach to multilingual text summarization", Indian Institute of Technology, Allahabad.

[17] http://www.cfilt.iitb.ac.in/wordnet/webhwn, Reviewed on 26th November 2014

[18] Atif Khan, Naomia Salim, "A Review on Abstractive Summarization Methods", Journal of Theoretical and Applied Information Technology, 10th January 2014. Vol. 59 No.1

[19] P.E. Genest and G. Lapalme, "Framework for abstractive summarization using textto- text generation," in Proceedings of the Workshop on Monolingual Text-To-Text Generation, 2011, pp. 64-73.

[20] S. M. Harabagiu and F. Lacatusu, "Generating single and multi-document summaries with gistexter," in Document Understanding Conferences, 2002.

[21] H. Saggion and G. Lapalme, "Generating indicative-informative summaries with sumUM," Computational Linguistics, vol. 28, pp. 497-526, 2002.

[22] I. F. Moawad and M. Aref, "Semantic graph reduction approaches for abstractiveText Summarization," in Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on, 2012, pp. 132-138