Pre-Processing Phase of Text Summarization Based on Gujarati Language

Ashish B. Tikarya, Kothari Mayur, Pinkeh H. Patel

Abstract— **A text summarization is a technique for the text that is produced from one or more texts, that contain a significant portion of the information in the original text(s), and that is no longer than half of the original text. Text summarization technique advantage is no need to read entire document. This paper presents overview of text summarization, methods of text summarization and helps to know what the problems in summarization. Authors focused on all issues regarding all phases of text summarization preprocessing in Gujarati language.**

Index Terms— **stemming, stop-word, elimination of duplicate sentence**

I. INTRODUCTION

In today's environment people have no enough time to read entire document or paper which sometimes very large and take lots of time to read in quick period of time. So it is necessary to discover some techniques which can provide us such functionalities to cover major parts from the document in summarized. Text Summarization is the process of automatically creating a compressed version of the given text. This compressed version is called summary. There are two types of text summarization methods namely 1) Extractive method consists of selecting important sentences, paragraphs from the original document and concatenating them into shorter form. It selects the best-scoring sentences from the original document based on a set of extraction criteria and 2) Abstractive method consists of understanding the original text and re-telling it in fewer words. It uses linguistic methods to examine and interpret the text and find the new concepts. This is done by generating a new shorter text that conveys the most important information from the original text document [1].

Several types of application available for text summarization like Competitive Intelligence, Extraction Transformation Loading, Human resource management, Customer Relationship Management (CRM), Market Analysis (MA), Text Mining Applications in Technology watch, Text Mining Applications in Natural Language Processing and Multilingual Aspects, Questioning in Natural Language, Multilingual Applications of Natural Language Processing[6].

Text summarization process can be divided into three phases:

- Pre Processing phase is structured representation of the original text. Various features influencing the relevance of sentences are calculated.
- In processing phase, final score of each sentence is determined using feature weight equation.
- In final phase, Top ranked sentences are selected for final summary.

II. PRE-PROCESSING PHASE

Pre-processing is structured representation of the original inputted text. The importance of pre-processing is used in almost every developed system related with text processing and natural language processing. Pre-processing phase includes words identification, sentences identification, stop words elimination, language stemmer for nouns and proper names, allowing input in proper format and elimination of duplicate sentences or words. Pre-processing phase reduce size of text. Author will more focus on pre-processing phase in following sections.

III. LITERATURE REVIEW

Text Summarization is shortening the text into shorter form and holding its information and meaning. The goal of automatic text summarization is to present most significant contents from information source to the user in a shorter version.

Several types of technologies are available for Text Summarization.

1) Information Extraction 2) Topic tracking 3) Summarization 4) Categorization 5) Clustering 6) Concept Linkage 7) Question Answering 8) Information visualization [6]. Current trend in text summarization is Online Product Review, Bio-Medicine, Education, Emails and Blogs [7].

IV. PROBLEMS

Extractive Method

- Extracted sentences usually longer than average.
- Due to this part of the segments that are not essential for summary and it also time consuming space.
- Importance or relevant information is usually spread across sentences, so extractive summary cannot catch

Manuscript received July 23, 2014.
Ashish B. Tikarya, SRIMCA, Uka Tarsadia University, Surat, India, 08347304846, (e-mail: 201204100110024@srimca.edu.in).
Mayur Kothari, SRIMCA, Uka Tarsadia University, Surat, India, 08980914049, (e-mail: 201204100110171@srimca.edu.in).
Pinkeh H. Patel, SRIMCA, Uka Tarsadia University, Surat, India, 09601851159, (e-mail: pinkesh.patel@utu.ac.in).

this important information.
- Conflicting information may not be presented accurately.
- Sentences often contain pronounces which lose their referents or actual meaning when extracted out [1].

Abstractive Method
- An Abstractive Summarization method consists of understanding the original text and re-telling it in fewer words.
- It uses linguistic methods to examine and interpret the text and find the new concepts.
- This done by generating a new shorter text that conveys the most important information from the original text document [11].
- Representation problem for sentences with grammar pronouns.
- A system capability has constraint of their representation structure.
- Therefore, system cannot summarize their representation and also cannot capture sentences.
- In limited domains, it may be feasible to devise appropriate structures, but a general-purpose solution depends on domain specific analysis [1].

Example: News blaster is a good example of a text summarizer, that helps users find the news that is of the most interest to them.

V. ISSUES OF TEXT SUMMARIZATION

Solved Issue
- The importance of sentences is decided based on statistical and linguistic of sentences. Sentences select based on statistical and linguistic approach of Text Summarization.
- Selection of important sentences from the original text.
- Effective summary in less time and with least redundancy. In pre-processing phase of Text Summarization remove duplicate word and duplicate sentences then generate summery so least redundancy is not possible.
- Author calculated the rank of the sentences by summing up these scores. The top n ranked sentences were picked up to be included in summary [3].In processing give rank to all sentences and summery in selected top n ranked sentences.

Unsolved Issues
- Improve system by adding sentence simplification technique for producing summary [2]. In a text hyphen, punctuation mark, stop word are available so, sentences not simple. Difficult to understand original meaning of text.
- The biggest challenge for text summarization is to summarize content from a number of textual and semi structured sources, including databases and web pages, in the right way (language, format, size, time) for a specific user [1].

VI METHODOLOGY

Pre-Processing phase of Gujarati Text Summarization

i. Lexical Analysis

The main objective of lexical analysis process is the identification of the words in a text. Usually, the following cases are considered: digits, hyphens, and punctuation marks. In digits, if texts like ત્રણ so author change it in digit 3.In hyphens, if texts like જાત-પાત so author remove hyphen જાત, પાત. All lines in remove punctuation marks because understand meaning of word.

ii. Stop Word Elimination

Gujarati language Stop words are frequently occurring words in Gujarati text. We have to eliminate these words from original texts; otherwise, sentences containing them can get influence unnecessarily. Author has made a list of Gujarati language stop words by creating a frequency list from a Gujarati corpus. Analysis of Gujarati corpus taken from popular Gujarati newspapers has been done. Some commonly occurring stop words are
છે, કે, તો, પણ, શકાય, હતી, હતું, હતા, સાથે, etc.

Steps to Stop Word Elimination

Step 1: If current word is stop word then delete all the occurrences of it from current sentence
Step 2: If current word is not present in dictionary, noun morph, common nouns list, proper nouns list then apply noun and proper noun Stemmer for the possibility of noun or proper noun.
Step 3: Delete redundant (duplicate) words from input text, to prevent them occurring in final summary and produce output of preprocessing phase [8].

iii. Stemmer For Nouns 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.[13]It is not necessary that stem terms should be similar to root of that term [13].
For example, a typical Guajarati stemmer can convert different variants ભાજપે, ભાજપનો, ભાજપને, ભાજપના like into root term of ભાજપ.

Procedure of Stemming For Gujarati Noun and Proper-Nouns

Step 1: Enter Input Text in Gujarati
Step 2: Segment this text into words and search each Guajarati word in Guajarati Word-Net.
Step 3: Obtain the stem or radix of that word which are not found in Word-Net dictionary.
Step 4: If find present in dictionary, then that is genuine word, otherwise it may proper name or some invalid word.
Step 5: Once again check the stem or word in Guajarati Word-Net dictionary.

Step 6: If word found in the dictionary checked that stemmed word is noun or not.
Step 7: For depth analysis of corpus was made and various possible noun suffixes were identified.
Step 8: if found suffix then eliminate the suffix from end of that word.

Word	Stem	Suffix
શહેરી	શહેર	ી
વિસ્તારોમાં	વિસ્તાર	ોમાં
ભાજપનો	ભાજપ	નો
સફાયો	સફ	ાયો
દેશને	દેશ	ને
બુટાસિંહને	બુટાસિંહ	ને
અદાલતને	અદાલત	ને
અસીલોએ	અસીલ	ોએ
વકીલોની	વકીલ	ોની

Fig-1 Example of Suffix

iv. Allowing input restriction of input text

Steps for Input Restriction

Step 1: If Uploaded input file is in .txt format then calculate input character length otherwise display the error message.

Step 2: If input character length> 1, 00,000 characters then display error message.

Step 3: From the input text, calculate length of characters, Punctuation mark characters, numeric characters, English characters and other characters.

If characters length is less than equal to Punctuation character length or numeric characters length or Gujarati Characters length or other characters length then display error. Else if other language characters length is greater than equal to 10% of total input characters length then display error message.

v. Elimination of duplicate sentences

Two Techniques are used:

1) Delete the duplicate sentences in the output summary.
2) Delete sentence in the input itself.

It is desirable to delete the duplicate sentences from input because numbers of input sentences are reduced and processing phase takes less time. Duplicate sentences are deleted from input by searching the current sentence in to the sentence list which is initially empty. If current sentence is found in sentence list then that sentence is set to null otherwise it is added to the sentence list being the unique sentence.

VII. CONCLUSIONS

In this Paper, Author have discuss the phases of pre-processing of Text Summarization for reduce size of text. In pre-processing phase remove hyphen, remove stop word, remove duplicate word, normalize stemming, and remove duplicate sentences. Some unsolved issues of text summarization define in this paper. Author can solve all unsolved issue using all phases of pre-processing phase that define in this paper.

VIII. FUTURE ENHANCEMENT

In entire semester duration author focus on Text Summarization and author comes at the end that there are various unsolved issues in it. Author want to finally apply Text Summarization on Gujarati language and on the basis of entire semester research author comes at the end that only some phases of Pre Processing Phase apply on it so that its result is not as required in final summary at the end. So in future author can apply all phases of Pre Processing in Text Summarization. So author can get better result in final summary. There are various phases of Pre Processing namely:

- Lexical analysis
- Stop word elimination
- Stemmer for nouns and proper nouns
- Normalization of nouns in noun morph
- Allowing input restrictions to input text
- Elimination of duplicate sentences

REFERENCES

[1] Gupta V., Lehal G.S., "A Survey of Text Summarization Extractive Techniques," Journal of Emerging Technologies in Web Intelligence, Vol. 2, PP. 258-268 No 3, August 2010.
[2] Anjali R. Deshpande, Lobo L. M. R. J., "Text Summarization using Clustering Technique," International Journal of Engineering Trends and Technology (IJETT) – Vol. 4, PP. 3348-3351, Issue 8 August 2013.
[3]Anita.R.Kulkarni, Dr Mrs. S.S.Apte, "An automatic Text Summarization using featureterms for relevance measure", IOSR Journal of Computer Engineering (IOSR-JCE)Vol. 9, PP. 62-66, Issue 3 March – April 2013.
[4] Alok Ranjan Pal, Projjwal Kumar Maiti, and Diganta saha, "An Approach To Automatic Text Summarization Using Simplified Lesk Algorithm And Wordnet," International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.3, PP. 15-22, September 2013.
[5]Mrs.Pimpalshende A. N., "Overview of Text Summarization Extractive Techniques,"International Journal of Engineering and Computer Science Vol. 2, PP. 1205-1214, Issue 4 April 2013.
[6] Gupta V., Lehal G.S., "A Survey of Text Mining Techniques andApplications,"JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, PP. 60-76 NO. 1, AUGUST 2009.
[7] Oi Mean Foong, Alan Oxley and Suziah Sulaiman, "Challenges and Trends of Automatic Text Summarization,"International Journal of Information and Telecommunication Technology IJITT Vol. 1, PP. 34-39 Issue 1 2010.
[8] Gupta V., Lehal G.S., "Complete Pre Processing phase of Punjabi Text Extractive Summarization System", Demonstration Papers, PP. 199–206, COLING 2012, Mumbai, December 2012.
[9] R.V.V Murali Krishna and Ch. Satyananda Reddy, "A sentence scoring method for extractive text summarization based on Natural language queries,"IJCSI International
Journal of Computer Science Issues, Vol. 9, PP.
259-262, Issue 3, No 1, May 2012.
[10] Long Duong, Paul Cook, Steven Bird and Pavel Pecina, "Simpler unsupervised POS tagging with bilingual projections,"Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, PP. 634–639 August 4-9 2013.
[11] Khosrow Kaikhah, "Automatic Text Summarization with Neural Networks,"in Proceedings of second international Conference on intelligent systems, IEEE, PP. 40-44, Texas, USA, June 2004.
[12] PadmaPriya, G. and K. Duraiswamy,"An Approach For Text Summarization Using Deep Learning Algorithm," Journal of Computer Science 10 PP. 1-9, 2014.
[13] Vishal Gupta, "Hindi Rule Based Stemmer for Nouns," International Journal of Advanced Research in Computer Science and Software

EngineeringVol. 4, PP. 62-65 Issue 1, January 2014.
[14] Hovy, E.H., "Automated Text Summarization," In R. Mitkov (ed), The Oxford Handbook of Computational Linguistics, pp. 583–598, 2005.
[15]MojoJolo, "Automatic Text Summarization," Internet:http://en.wikipedia.org/wiki/Automatic_summarization 10 February 2014 [15 February 2014].
[16] K. Nandhini and S. R. Balasundaram, "Use of Genetic Algorithm for Cohesive SummaryExtraction to Assist Reading Difficulties,"Applied Computational Intelligence and Soft Computing, PP.1-11,2013.

Mr. Ashish B. Tikarya is pursuing in master of computer application in Uka Tarsadia university, Dist.:-surat, Gujarat. Area of interest – Text-summarization.

Mr. Mayur Kothari is pursuing in master of computer application in Uka Tarsadia university, Dist.:-surat, Gujarat. Area of interest – Text-summarization

Mr. Pinkeh H. Patel has done master of computer application from sardar patel university, Vidyanager, Gujarat. He has interest in area of Text-summarization and Information retrieval. He has 2 year of industry and 3 year of teaching experience in under graduates and post graduates.