Ruled Based Text Summarizer for History Documents

Anjusha Pimpalshende

Assoc. Prof, CSE, MLR Institute of Technolgy,Hyderabad

Dr A.R.Mahajan

Prof,HOD ,ITGovernment polytechnic,Nagpur,

Abstract: A summary is the concise text conveys the most important information from the source document. Summary of the text can be generated from a single document or from multiple documents. In digital era lot of information available on web. To search the required information search engine takes more time and gives excess data or information ,Automatic text summarization resolves this problem. In our rule based summarizer system we are generating summary by extracting most important sentences from the text document using some features like named entity, numerical data, title word .This system is mostly useful in history text document. History text document is so lengthy and boring to Read, by this summarizer we can generate important sentences so that user can read it in short time with better understanding.
Keywords : summarizer, extractive, abstractive

I. INTRODUCTION

There is need of efficient Automatic text summarization as Internet provides the access to a very large amount of data in a particular language. People in today's world go to movies, various tourist places on the basis of reviews they get online. This type of text summarization tool helps them in making decisions in a lesser duration. This paper targets the problem of information overload and proposes a system for extractive text summarization which compresses the input document but not losing the important content in the document. As access to data from anywhere has increased so the demand for an automatic text summarization has also increased. Automatic summarization system reduces a text document into shorter set of texts or paragraph that conveys the actual semantics of the text and also should not lose the main meaning described in the text.

Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important content of the original document.The main idea of summarization is to find a representative subset of the data, which contains the information of the entire set. Summarization technologies are used in a large number of sectors in industry today. An example of the use of summarization technology is search engines such as Google.

II. TEXT SUMMARIZATION APPROACHES

There are two approaches for text summarization
1.Extractive Summarization.
2.Abstractive Summarization.

Extractive Summarization:

An extractive summarization method only decides for each sentence whether or not it will be included in the summary. In extractive summarization system different weights are assigned to each sentence of the document on which sentence is selected to get added for the summary. Weights can be assigned to the sentences according to the position of the sentence in the document i.e. sentences in the beginning and at the end are assigned more weight as they are supposed to contain more valuable information. Weights can also be assigned according to the type of information they contain. For example if a sentence contain name of person, date of the event occurred then more weight is given to that sentence than those which do not contain any Named Entity.

Abstractive Summarization:

Abstractive summarization process consists of "understanding" the original text and "retelling" it in fewer words. In abstractive summarization semantic analysis of the document(s) is done on basis id which summary of the document is generated. In this type summarization interpretation of each of the sentence is done and may be represented in the different style from the original one. In both extractive and abstractive summarization technique rule basedapproach can be used in which various handcrafted rules are to be created on the basis of which summary of the text document can be generated

III . LITERATURE SURVEY

1. *Mandeep Kaur and Jagroop Singh* "**A survey on different Text Summarized techniques and deadwood is eliminated andremove from the summary".** In this paper, an author proposes a system for detection and removal of five different features forthe assignment of weight to the sentences. In the next step the highest scoring sentences are selected to form the summary. Inthe last steps the Deadwood in summaries for Punjabi language. Deadwood means word or phrase that can be omitted withoutloss in meaning. Removing it shortens and clarifies the summary. Proposed system works in two phases which are semanticanalysis and Adjective Removal Rule.
2. *Visual Gupta and Gurpreet Singh Lehal* **"Automatic Punjabi Text Extractive Summarization system".** In this paper author describe the Punjabi text extractive system which consist of two phases 1) Pre Processing 2) Processing. In this paperterm pre processing is defined as the phase which identify the word boundary, sentence boundary, Punjabi stop wordselimination etc. and the processing phase sentence features are calculated and a weight is assigned to each sentence on thereference of which unwanted sentences are eliminated from the input text. It is described that the author tested the proposedsystem over fifty Punjabi news documents (with 6185 sentences and 72689 words) from Punjabi Ajit news paper and fiftyPunjabi stories (with 17538 sentences and178400 words). Accuracy of the system is varies from 81% to 92 %.[2]
3. *Ng Choon-Ching & Ali Selamat* **"Text Summarization Review"** In this paper author describe an existing need for text summarizers that small devices like PDA has emerged the development of text summarization of web pages. Authors have identified problems for text summarization in several areas such as dynamic content of web pages, diverse summary definition of text, and different benchmark of evaluation measurements. Besides, authors also found advantages of certain methods that increased the accuracy of web page classification. In the future work, author plan to investigate machine learning techniques to incorporate additional features for the improvement of text summarization quality. The additional features authors are currently considering include linguistic features such as discourse structure, lexical chains, semantic features such as name entities, time, location information etc. [3]
4*.4. Josef Steinberger, Karel Jezek* **"Using Latent Semantic Analysis in Text Summarization and Summary Evaluation"** This paper deals with using latent semantic analysis in text summarization. In this paper author describe a generic text summarizationmethod which uses the latent semantic analysis technique to identify semantically important sentences. The proposed methodhas been further improved. Then author propose two new evaluation methods based on LSA, which measure content similarity between an original document and its summary. In the evaluation part author compare seven summarizers by a classical content- based evaluator and by the two new LSA evaluators. Author also studies an influence of summary length on its quality from theangle of the three mentioned evaluation methods. [4]
5. *Vishal Gupta and Gurpreet Singh Lehal***,"**
Preprocessing phase of Punjabi text Extractive Techniques." In this paper author describe the Punjabi text extractive system which consist of two phases 1) Pre Processing 2) Processing. In this paper term pre processing is defined as the phase which identify the word boundary , sentence boundary, Punjabi stop words limination etc.and the processing phase sentence features are calculated and a weight is assigned to each sentence on the reference of whichunwanted sentences are eliminated from the input text. It is described that the author tested the proposed system over fiftyPunjabi news documents. [5]

IV. TEXT SUMMARIZATION BASIC CONCEPTS

- Coherence: A summary is said to be coherent if all its sentences or text units form an integrated whole and the sequence of ideas progressed logically.
- Compression Rate: It is a ratio of summary length to source length expressing the degree of summarization required
- Compaction of text: It is a process of removing less salient phrases or words from sentences.

V. METHODOLOGY USED IN RULE BASED SUMMARIZER

Summarization system consists of three major steps.
1)Pre-processing: Is structured representation
of the original documents.
Sentence segmentation: Decompose sentences along with its word count.
Ex. There are events one never forgets, anyone who was alive in 1963 remembers where he was when President Kennedy was assassinated. And anyone who was at Candlestick Park Tuesday night for Game 3 of the World Series will never forget. I was in a trailer just outside the stadium, about to watch the telecast when the earthquake hit

Example of sentence segmentation.
S1 : There are events one never forgets, anyone who was alive in 1963 remembers where he was when President Kennedy was assassinated.

S2 : And anyone who was at Candlestick Park Tuesday night for Game 3 of the World Series will never forget.

S3 : I was in a trailer just outside the stadium, about to watch the telecast when the earthquake hit.
Tokenization: Process of splitting of sentences into words by identifying Full stops (.).
Stop word removal: Remove common words with no semantics in English a, an , and in
List of stop words.

a	again	although	anyone
because	before	below	between
came	causes	com	considering
do	despite	different	doesn't
each	else	et	everybody
first	follows	formerly	from
getting	go	gone	greetings
have	hence	hereupon	his
if	indeed	instead	it'd
mainly	me	might	mostly
name	need	next	needs
ones	otherwise	outside	ok

Stemming: Stemming is used to check similarity feature. In this obtain the stem or root of the word or obtain similar words. e.g. walk, walking and walked are counted as same and derived from a stem word walk
2) Processing: Check each sentence for all feature is there or not Increase the weight by one for existence of one feature.

3)Sentence Generation: Sort the sentences as per weight and extract the sentences in summary as per the compression ratio.

VI. ALGORITHM:

Input :English document
Output : summarized text depending on compression ratio.
Step1: Select the document D
//------ *Pre-processing------//*
Step2. Apply sentence segmentation.
It is the process of decomposing the given text document into its constituent sentences along with its wordcount . In English, sentence is segmented by identifying the boundary of sentence which ends with full Stop (.)
Step 3: Apply Tokenization.
It is the process of splitting the sentences into words by identifying the spaces, comma and special symbols between the words. So list of sentences and words are maintained for further processing.
Step 4. Remove stop words.
To effectively use word feature score we need to only consider the words in the document which have importance.
Stop words are common words that carry less important meaning than keywords .
These words should be eliminated otherwise sentence containing them can influence summary generated.
d= [s1, s2,...., sn] , *si= [w1 ,w2 ,...., wn] , set of words.*
Step 5: Stemming
In Stemming process, the suffixes are ignored and removed from words to get the common origin. Syntactically similar words, such as plurals, verbal variations, etc. are considered similar. Suffix striping algorithm is used for stemming.

e.g. walk, walking and walked are counted as same and derived from a stem word walk.

Step 6: Sentence Extraction

Assign weight to each sentence depends on the feature one of the feature is named entity to check name entity we stored 2000 entries in a tables

Name entity:

1.Table contain person names
2.Table contain location names
3.Table contain city names
4.Table contain state and country names.

Features	TF	SL	SP	ND	TW	TOTAL
S1	0.83	0.47	0.36	0.94	0.52	3.12
S2	0.61	0.45	0.64	0.58	0.33	2.61
S3	0.73	0.61	0.81	0.57	0.63	3.35
…	…	…	…	…	…	…
S10	0.89	0.71	0.44	0.89	0.65	3.58

Example of feature score

VI RESULTS

Summary is generated as per the compression ratio.

Ex.

There are events one never forgets, anyone who was alive in 1963 remembers

where he was when President Kennedy was assassinated. And anyone who was at Candlestick Park Tuesday night for Game 3 of the World Series will never forget. I was in a trailer just outside the stadium, about to watch the telecast when the earthquake hit. It was shortly after the network went on the air. My immediate reaction was that a jet was flying very low overhead, but soon I knew what was happening. To me, the earthquake was not as bad as the Whittier quake in October, 1987, mainly because I was awake this time. I opened the door to the trailer and looked out. The stadium was shaking and the special concrete joint -- one of several strategically spaced aroud the top of the stadium to prevent earthquake damage -- was doing was it was designed to do. It was opening and closing as if the place was made of cardboard instead of concrete. A few people were running out of the stadium, but there did not seem to be great alarm, because the quake did not last long. I walked into the stadium to interview fans. And again, I sensed little panic. Most people stayed in their seats, waiting to hear whether the game would be postponed. And later, when fans started leaving en masse, the reaction of those I talked to varied. Those who were sitting in the upper deck seemed levels

Generated Summary:

> Experience of a reporter at Candlestick Park for the World Series the night of the 1989 San Francisco quake: I was sitting in a trailer outside the stadium when the earthquake hit. I looked out and the stadium was shaking. A special concrete joint on top of the stadium was opening and closing like cardboard. A few people were running out. I walked into the stadium to interview fans and sensed little panic. Most people stayed, waiting to hear whether the game would be postponed. When leaving, those in the upper deck seemed more shaken that those in lower levels.

VII. EVALUATION OF SUMMARY

Automatic generated summaries can be evaluated using following parameters.

- Precision: It evaluates correctness for the sentences in the summary.

p = Retrieved sentences Ω Relevant sentences/Retrieved sentences

- Recall: It evaluates proportion of relevance included in the summary.

R= Retrieved sentences Ω Relevant sentences/ Relevant sentences

- F1 score: The F1 score is a standard way to mix the two numbers in a single score:

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{Precision} + \text{recall}}$$

IX. CONCLUSION

In automatic text summarization, there are several techniques which used for selecting important sentences. The features were used to determine these sentences that should be selected in the final summary. The feature is an important component in the summary process. In this work the important feature is named entity and numerical data,based on theses main features sentences are extracted. This summarizer is basically used in historical documents . History documents are very lengthy to read using this we can read important content in short time.

REFERENCES

[1] H. P. Luhn, "*The automatic creation of literature abstracts,*" IBM Journal of Research and Development, vol. 2, no. 2, 1958.

[2] Madhavi K. Ganapathiraju, "*Overview of summarization methods*" Self- paced lab in Information Retrieval, 2002

[3] U. Hahn and I. Mani, "*The challenges of automatic summarization,*" IEEE Computer, vol. 33(11) pp. 29–36, November 2000

[4] R. C. Eberhart, and Y. Shi, "*Particle swarm optimization: Developments, applications and resources*". Proceedings of the 2001 Congress on Evolutionary Computation. 27-30 May 2001.Seoul, Korea: IEEE, PP. 81-86.

[5] K. S. Jones, "*Automatic summarizing: The state of the art,*" Inf. Process.Manage., vol. 43, no. 6, pp. 1449– 1481, 2007.

[6] Mohammed Salem Binwahlan1 "*Swarm Based Text Summarization*" 2009 International Association of Computer Science and Information Technology - Spring Conference

[7] Mohamed Abdel Fattah and Fuji Ren, "*Automatic Text Summarization*", In Proceedings of World Academy of Science, Engineering and Technology, Vol. 27, pp192-195, 2008 .

[8] Vishal Gupta and Gurpreet Singh Lehal, "*A Survey of Text Summarization Extractive Techniques*", Journal of Emerging Technologies in Web Intelligence, Vol. 2, No. 3, pp258-268, 2010

[9] ALGULIEV, R. M., and R. M. ALYGULIEV.. "*Automatic text documents summarization through* sentences clustering". Journal of Automation and Information Sciences, **40**(9): 53–63. 2008