Semantic Analysis Based Text Summarization

Harsh Desai
Student, Department of Information Technology
M.C.T Rajiv Gandhi Institute of Technology.
Mumbai, India
harsh.desai07@gmail.com

Hamza Moiyadi
Student, Department of Information Technology
M.C.T Rajiv Gandhi Institute of Technology.
Mumbai, India
hamza.moiyadi@gmail.com

Dhairya Pawar
Student, Department of Information Technology
M.C.T Rajiv Gandhi Institute of Technology.
Mumbai, India
pawar.dhairya@yahoo.co.in

Geet Agrawal
Student, Department of Information Technology
M.C.T Rajiv Gandhi Institute of Technology.
Mumbai, India
geetagrawal1995@gmail.com

Prof. Nilesh Patil
Professor, Department of Information Technology
M.C.T Rajiv Gandhi Institute of Technology.
Mumbai, India
nilesh.patil@mctrgit.ac.in

Abstract— Automatic summarization has become an important part in the study of natural language processing since the advent of the 21st century, since a majority of the data online is textual. Summarization of text will lead to a reduction of data while maintaining the context of it. Having such summarization activity being done automatically also helps in reducing human effort. Summarization is the process of generation of the summary of input text by extracting the representative sentences from it. In this project, we present a novel technique for generating the summarization of domain specific text by using Semantic Analysis for text summarization, which is a subset of Natural Language Processing.

Keywords:*latent semantic analysis, natural language processing, python, summarization*

*****

I. INTRODUCTION

Text summarization (or automatic summarization) is the construction of a reduced version of a text by a computer program. The product of this procedure still contains the most important points of the original text and is generally referred a summary. There are two approaches to text summarization: extraction and abstraction. Extraction techniques simply copy information considered to be most important by the system to the summary, while abstraction techniques include interpreting sections of the input document. In general, abstraction can produce summaries that are more accurate than extraction, but these programs are considered much harder to develop. Both techniques use natural language processing and/or statistical methods for generating summaries. And, the classical approaches to text summarization proposed by Luhn et al have established the basis for the discipline of text summarization techniques. Text summarization is increasingly being used in the commercial sector, in areas of telecommunications, data mining, information retrieval, and in word processing with high probability rates of success. In addition to its wide range of applications in the commercial sector, emerging areas of text summarization include, multimedia and multi-document summarization; however, there has been less work performed in meeting summarization.

The goal of this report is to capture the product evaluation process in 4 distinct phases:

1) Preparation
2) Criteria establishment
3) Characterization, and
4) Testing

First and foremost, the preparation phase consists of requirement analysis and product research that identify three feasible products (text summarization tools). In the criteria establishment phase, evaluation criteria are established for the two sub-criteria (characteristic and testing). While the characterization phase comprises of the data collection for the criteria defined. Followed by the evaluation experiment (or testing) performed on the established testing criteria, as the final phase of the evaluation process. Furthermore, the discussion section discloses the results of the experiment and any follow-up work to be carried out.

II. LITERATURE SURVEY

Rasim et al proposed a system for automatic summarization using the extractive methodology using an evolutionary algorithm. In their study, they proposed an unsupervised document summarization method that creates the summary by clustering and extracting sentences from the original document[5]. On the other hand, MandarMitra et al, from the department of computer science, in Cornell University proposed a similar system for text summarization but instead of using the sentence extraction method proposed before, they use another method based on paragraph extraction. In their study they used text traversal & text relation maps to generate summaries[3]. In 2014, M. S. Patil et al, suggested a summarization system based on several extractive text summarization approaches, and on the Support-Vector-Machine(SVM). This system tries to improve the performance and quality of the summary generated by the clustering technique by cascading it with SVM[6]. Anne Hendrik Buist et al, deliberated the disclosure of audio-visual meeting recordings is a new challenging domain studied by several large scale research projects in Europe and the US. Automatic meeting summarization is one of the functionalities studied. They published a report on the results of a feasibility study on a subtask, namely the summarization of meeting transcripts. The authors concluded that the system produces fairly readable summaries, and identified the bottleneck of the system to be the lack of structure in meetings, and related to this the absence of good features[8]. Josef Steinberger et al, described a generic text summarization method which used the latent semantic analysis technique to identify semantically important sentences and suggested two new evaluation methods based on LSA, which measure content resemblance between an original document and its summary[1]. Jen-Yuan Yeh et al, used a trainable summarizer for summarization. A trainable summarizer considers several features such as position, positive keyword, negative keyword, centrality, and the resemblance to the title, to generate Summaries. They also proposed a second approach which used latent semantic analysis (LSA) to derive the semantic matrix of a document and used semantic sentence representation to construct a semantic text relationship map[11]. Ronan Collobert et al, attempted to define a unified architecture for Natural Language Processing which learns features that are relevant to the tasks at hand given very limitedprior knowledge. These tasks include Part-Of-Speech Tagging (POS), Chunking, Named Entity Recognition (NER), Semantic Role Labeling (SRL), Language Models and Semantically Related Words (“Synonyms”) [9]. Dipanjan Das et al, explored few approaches in the areas of single and multiple document summarization and gave special emphasis to empirical methods and extractive techniques[4]. Recently, Hovy and Lin devised a multilingual automatic summarization system called SUMMARIST which summarizes text documents using Information Retrieval & statistical techniques, but at the time of writing this review, not all the modules of SUMMARIST were performing optimally[10]. In 2016, Dr.A.Jaya et al, studied the various techniques available for abstractive summarization and put forward the fact that very little work is available in abstractive summary field of Indian languages. They also described the various works currently available in Indian languages. [2]

III. COMPARISON TABLE

Paper Title	Authors	Technology Used	Remarks	Extractive/ Abstractive
Evolutionary Algorithm for Extractive Text Summarization	RasimAlguliev, RamizAliguliyew	Sentence Based Extractive Document summarization	Uses the usual extractive method of sentence extraction with an algorithm that moulds itself to every document to give the best summary possible	Extractive
Automatic Text Summarization By Paragraph Extraction	MandarMitra, Amit Singhal, Chris Buckley	Paragraph Extraction	Expands on the sentence extraction technique by implementing a more generalised technique	Extractive
A Hybrid Approach for Extractive Document Summarization Using Machine Learning and Clustering Technique	M. S. Patil, M. S. Bewoor, S. H. Patil	Machine Learning and Clustering Technique	Implements a machine learning algorithm to the summarizing system which trains the system everytime a document is given to it so that the summary is better each time	Extractive

Automatic Summarization of Meeting Data: A Feasibility Study	Anne Hendrik Buist, Wessel Kraaij and Stephan Raaijmakers	Maximum Entropy based extractive summarization	Provides a novel way of summarizing documents which are a record of meetings.	Extractive
Using Latent Semantic Analysis in Text Summarization and Summary Evaluation	Josef Steinberger, Karel Ježek	Latent Semantic Analysis	In-depth paper on semantic analysis for text summarization which also proposes evaluation methods for summary accuracy	Abstractive
Text summarization using a trainable summarizer and latent semantic analysis	Jen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang, I-HengMeng	Latent Semantic Analysis + Text Relationship Mapping	Adds T.R.M to an existing LSA text summarizer to improve the accuracy with minimal training	Abstractive
A Survey on Automatic Text Summarization	Dipanjan Das, Andre F.T. Martins	-	Looks at extractive and abstractive summaries and evaluates both.	-
A Study on Abstractive Summarization Techniques in Indian Languages	Sunitha C., Dr. A. Jaya, Amal Ganesh	Semantic Graph	Studies on summaries based on indian languages are very few, and this paper is highly informative for the same	Abstractive
Automated Text Summarization And the SUMMARIST System	Edward Hovy, Chin-Yew Lin		So far one of the most successful extractive summarizers, with support for 5 languages and available for students to study	Extractive

IV. PROPOSED SYSTEM

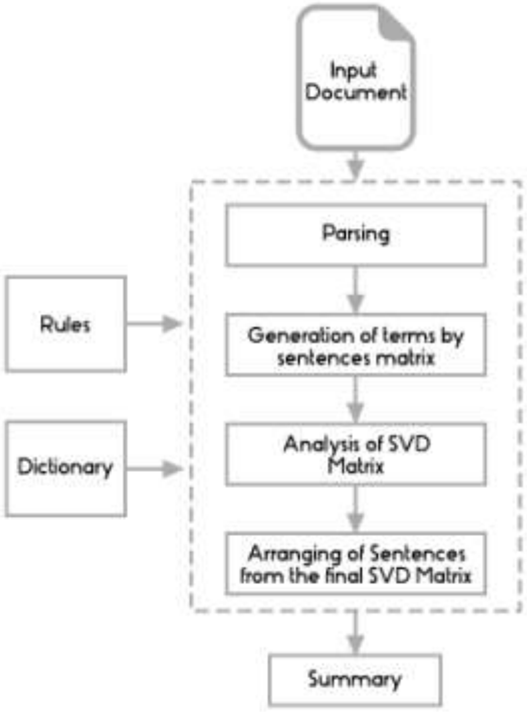
Input Document
Parsing
Rules
Generation of terms by sentences matrix
Dictionary
Analysis of SVD Matrix
Arranging of Sentences from the final SVD Matrix
Summary

Fig.1: Overview of Text Summarization using LSA

The above system uses Latent Semantic Analysis [1] to summarize documents from the user. The user inputs a document to the summarizer (denoted by dashed box) which has classes derived from the NLP libraries implemented on it. These classes are a collection of semantic rules (which allows the system to group the content using world knowledge) and dictionaries, which help in the semantic analysis and SVD phases in the summarizer. The input document is first parsed or pre-processed, wherein there unneeded words such as ‘stop words’ which are simply small words, like “the”, “and”, “a”, which do not contribute meaning to the text summary are removed. Generation of a Singular Value Decomposition (SVD) matrix, which is a m x n matrix, where m is the total number of terms in the original text and n is the number of sentences in the original text is the second stage. The SVD Analysis stage derives the latent semantic structure from the document represented by matrix A. In the summarization process, the sentences generated from the SVD Analysis stage are arranged by the system by semantically placing them in a way that the summary incorporates all the concepts of the original text. The final summary is then given back to the user.

V. IMPLEMENTATION

Inspired by the latent semantic indexing, we applied the singular value decomposition (SVD) to generic text summarization. The process starts with the creation of a term by sentences matrix $A = [A_1,A_2,\ldots,A_n]$ with each columnvectorA_i representing the weighted term-frequency vector of sentence i in the document under consideration. If thereare a total of m terms and n sentences in the document, then we will have an $m \times n$ matrix A for thedocument. Sinceevery word does not normally appear in each sentence, thematrix A is usually sparse. Given a $m \times n$ matrix A, where without loss of generalitym $\geq$ n, the SVD of A is defined as:

$$A=U\sum V^T$$

Where $U = [u_{ij}\ \ ]$ is an $m \times n$ column-orthonormal matrix whose columns are called left singular vectors; $\sum = \mathrm{diag}(\sigma_1,\sigma_2,\ldots,\sigma_n\ )$ is an $n \times n$ diagonal matrix whose diagonal elements are non-negative singular values sorted in descending order, and $V = [v_{ij}\ \ ]$ is an $n \times n$orthonormalmatrix whose columns are called right singular vectors. Ifrank(A) = r, then $\sum$satisfies

$$\sigma_1 \geq \sigma_2 \ldots \geq \sigma_r \geq \sigma_{(r+1)} = \cdots = \sigma_n = 0$$

The interpretation of applying the SVD to the terms by sentences matrix A can be made from two different view-points. From transformation point of view, the SVD derives a mapping between the m-dimensional space spanned by the weighted term-frequency vectors and the r-dimensional singular vector space with all of its axes linearly-independent. This mapping projects each column vector i in matrix A,which represents the weighted term-frequency vector of sentence i, to column vector $\psi_i = 〖[v_{i1},v_{i2},\ldots,v_{ir}\ ]〗^T$ of matrixV^T , and maps each row vector j in matrix A, which tellsthe occurrence count of the term j in each of the documents,to row vector $\varphi_j=[u_{j1},u_{j2},\ldots,u_{jr}]$ of matrix U. Here eachelementv_{ix}ofψ_i, u_{jy} of $\varphi_j$is called the index with the $〖x'〗^{th}$, $〖y'〗^{th}$ singular vectors, respectively.From semantic point of view, the SVD derives the latentsemantic structure from the document represented by matrix A. This operation reflects a breakdown of the original document into r linearly-independent base vectors orconcepts. Each term and sentence from the document is jointly indexed by these base vectors/concepts. A unique SVD feature which is lacking in conventional IR technologies is that the SVD is capable of capturing and modeling interrelationships among terms so that it can semantically cluster terms and sentences. Consider the words doctor, physician, hospital, medicine, and nurse. The words doctor and physician are synonyms, and hospital, medicine, nurse are the closely related concepts. The two synonyms doctor and physician generally appear in similar contexts that share many related words such as hospital, medicine, nurse, etc. Because of these similar patterns of word combinations, the words doctor and physician will be mapped near to each other in the r-dimensional singular vector space. Furthermore, as demonstrated in [10], if a word combination pattern is salient and recurring in a document, this pattern will be captured and represented by one of the singular vectors. The magnitude of the corresponding singular value indicates the importance degree of this pattern within the document. Any sentences containing this word combination pattern will be projected along this singular vector, and the SSsentence that best represents this pattern will have the largest index value with this vector. As each particular word combination pattern describes a certain topic/concept in the document, the facts described above naturally lead to the hypothesis that each singular vector represents a salient topic/concept of the document, and the magnitude of its corresponding singular value represents the degree of importance of the salient topic/concept. Based on the above discussion, we propose the following SVD-based document summarization method.

1. Decompose the document D into individual sentences,and use these sentences to form the candidate sentenceset S, and set k = 1.

2. Construct the terms by sentences matrix A for the document D.

3. Perform the SVD on A to obtain the singular value matrix$\sum$, and the right singular vector matrix V^T .Inthe singular vector space, each sentence i is representedby the column vector $\psi_i = 〖[v_{i1},v_{i2},\ldots,v_{ir}]〗^T$of V^T.

4. Select the k'th right singular vector from matrix V^T.

5. Select the sentence which has the largest index value with the k'th right singular vector, and include it in the summary.

6. If k reaches the predened number, terminate the operation; otherwise, increment k by one, and go to Step 4.

VI. RESULT

Since the inception of our project Latent Semantic Analysis Summariser we have carried out different phases of the Software Development Lifecycle. We have clearly defined the purpose of the project and also the scope determining the goals and milestones in project lifecycle. The benefits and limitations of the project are well listed in this synopsis. The aims and objectives of LSA Summarizer are achievable within the next phases of lifecycle.

Original Text: "Russia has lost its bid to become a member of the UNs human rights council, in a defeat that reflects the diplomatic cost of its war in Syria.

Russia was beaten on Friday by Hungary and Croatia in the competition for two seats on the council allotted to eastern European states. It was the first time one of the permanent

five members of the security council had failed to get elected to the HRC since its formation a decade ago, and followed a campaign by human rights groups opposing Russian membership because of its role in the bombing of Syrian cities, eastern Aleppo in particular.
They bomb a hospital one day, they run for the Human Rights Council the next. And they wonder why they missed the cut, a western diplomat said.
VitalyChurkin, the Russian envoy, shrugged off the rebuff, saying the countries who beat Russia are not as exposed to the winds of international diplomacy.
Russia is quite exposed, Churkin said.
Human rights groups also campaigned against Saudi Arabia for the high civilian death toll of its bombing campaign in Yemen, but the kingdom won one of the four seats reserved for the Asia-Pacific region.
The 193-member general assembly on Friday elected 14 members to the 47-nation council, the UNs main body charged with promoting and protecting human rights.
Brazil, China, Cuba, Egypt, Iraq, Japan, Rwanda, South Africa, Tunisia, the UK and the US were also elected to the council.
At the same time as Russia suffered its diplomatic setback, the Kremlin announced that president Vladimir Putin had turned down the Russian militarys demand to resume bombing of Aleppo, to keep open humanitarian corridors for rebels and civilians to leave the city.
In rejecting Russias bid for re-election to the Human Rights Council, UN member states have sent a strong message to the Kremlin about its support for a regime that has perpetrated so much atrocity in Syria. It also shows how important it is to have competitive slates in UN elections, Louis Charbonneau, UN director at Human Rights Watch, said.
Countries should have a chance to reject those whose candidacies are so severely compromised, as they did today. We have already said that Saudi Arabia, which was re-elected without competition, doesnt belong on the council in light of its indiscriminate attacks on civilians in Yemen. Well be keeping all members rights records under the microscope while theyre on the council. Next year, UN member states should make sure that all regional groups have real competition so no one is guaranteed victory, he said.
Russia currently holds the presidency of the UN security council but has alienated many UN member states by its support for the Syrian regimes airstrikes against rebel-held cities, and by its verbal attacks on UN officials who had criticised the airstrikes. On Thursday, Churkin shrugged off the findings of a UN investigation that the Syrian regime had used chemical weapons, saying the regime itself should have its own enquiry.
Russia deserves this defeat, but it will only increase Moscows contempt for the UN, said Richard Gowan, a UN expert at the European Council for Foreign Relations. The Russians only really care about the security council anyway, and they may well respond by stirring up more trouble there over Syria or other crises.
In 2001, the US was voted off the HRCs predecessor, the UN Human Rights Commission, in a gesture of disapproval over the George Bush administrations unilateralist leanings.
”

Summarized text: “It was the first time one of the permanent five members of the security council had failed to get elected to the HRC since its formation a decade ago, and followed a campaign by human rights groups opposing Russian membership because of its role in the bombing of Syrian cities, eastern Aleppo in particular.

Human rights groups also campaigned against Saudi Arabia for the high civilian death toll of its bombing campaign in Yemen, but the kingdom won one of the four seats reserved for the Asia-Pacific region.

At the same time as Russia suffered its diplomatic setback, the Kremlin announced that president Vladimir Putin had turned down the Russian militarys demand to resume bombing of Aleppo, to keep open humanitarian corridors for rebels and civilians to leave the city.

In rejecting Russias bid for re-election to the Human Rights Council, UN member states have sent a strong message to the Kremlin about its support for a regime that has perpetrated so much atrocity in Syria.

Russia currently holds the presidency of the UN security council but has alienated many UN member states by its support for the Syrian regimes airstrikes against rebel-held cities, and by its verbal attacks on UN officials who had criticised the airstrikes.**”**

VII. CONCLUSION

Text summarization is one of the major problems in the field of Natural Language Processing, and yet it is even after years of research and implementations, fraught with complications. However, there have been some major breakthroughs in the past, such as Columbia University’s Multigen (1999) and Copy and Paste (1999)[11], and USC’s ISI Summarist[9]. Many different methods were used to arrive at the final summary, whether that summary was abstractive or extractive. Methods such as Deep Understanding, Sentence Extraction, Paragraph Extraction, Machine Learning, and even some which employ all these methods along with Traditional NLP Techniques(Semantic Analysis, etc.). As such, keeping these accomplishments in mind, there is still ample amount of research left in the domain of Text Summarization, as a meaningful summary is still difficult to attain in all domains and langauges.

REFERENCES

[1] Josef Steinberger, Karel Ježek, “Using latent Semantic analysis In Text Summarization and Summary Evaluation”, Department of Computer Science and Engineering, UniverzitníCZ-306 14 Plzeň

[2] Sumitha C., Dr. A. Jaya, Amal Ganesh, “A study on Abstract Summarization Techniques in Indian Languages”, Elsevier Proceeding of Computer Science, No. 87, pp.25-31, 2016.

[3] MandarMitra, Amit Singhal, Chris Buckley, “Automatic Text Summarization by Paragraph Extraction”, Department of Computer Science Cornell University, AT&T Labs Research.

[4] Dipanjan Das, Andre F.T. Martins, “A Survey on Automatic Text Summarization”, Language Technologies Institute, Carnegie Mellon University, November 2007.

[5] RasimAlguliev, RamizAliguliyev, “Evolutionary Algorithm for Extractive Text Summarization.” Intelligent Information Management, 1, pp. 128-138, November 2009.

[6] M. S. Patil, M. S. Bewoor, S. H. Patil “A Hybrid Approach for Extractive Document Summarization Using Machine Learning and Clustering Technique.” International Journal of Computer Science and Information Technologies, Vol. 5, Issue No. 2, ISSN: 0975-9646, pp.1584-1586, 2014.

[7] Michael Ji, “Text Summarization Tool Evaluation: A Feasibility Study for Generating Meeting Summaries.” CPSC503 Final Report, Department of Computer Science, University of Calgary.

[8] Anne Hendrik Buist, Wessel Kraaij and Stephan Raaijmakers, “Automatic Summarization of Meeting Data: A Feasibility Study.”

[9] Ronan Collobertcollober, Jason Weston, “A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning.”

[10] Edward Hovy, Chin-Yew Lin, “Automated Text Summarization And The SUMMARIST System”, Information Sciences Institute of the University of Southern California

[11] Jen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang, I-HengMeng, “Text summarization using a trainable summarizer andlatent semantic analysis”, Elsevier Proceeding of Information processing and management, No. 41, pp 75-95, 2016.