NLP Based Text Summarization Using Semantic Analysis

Hamza Shabbir Moiyadi[1], Harsh Desai[2], Dhairya Pawar[3], Geet Agrawal[4], Nilesh M.Patil[5]

[1,2,3,4]Student, MCT's Rajiv Gandhi Institute of Technology, Mumbai, India
[5]Assistant Professor,MCT's Rajiv Gandhi Institute of Technology, Mumbai, India

Abstract— *Due to an exponential growth in the generation of textual data, the need for tools and mechanisms for automatic summarization of documents has become very critical. Text documents are vital to any organization's day-to-day working and as such, long documents often hamper trivial work. Therefore, an automatic summarizer is vital towards reducing human effort. Text summarization is an important activity in the analysis of a high volume text documents and is currently a major research topic in Natural Language Processing. It 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— NLP, Text summarization.

I. INTRODUCTION

Text summarization (or automatic summarization) is the creation of a shortened 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 to as an abstract or a summary. Broadly, one distinguishes two approaches to text summarization: extraction and abstraction. Extraction techniques merely copy information deemed to be most important by the system to the summary, while abstraction involves paraphrasing sections of the source document. In general, abstraction can produce summaries that are more condensed than extraction, but these programs are considered much harder to develop. Both techniques exploit the use of 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. The applicability of text summarization is increasingly being exploited 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 applicability 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. Therefore, as for our initial basis for the Alan project – robotic partner for agile software engineering team - our goal is to extend this applicability to the meeting domains to produce high-quality meeting summaries. To accomplish our task in hand requires a text summarization tool. But, rather than developing our own tool, a feasibility study was instigated to determine the success of making use of third party software. This in turn required a product evaluation to be carried out.

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 REVIEW

Rasimet 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 HendrikBuist 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 inmeetings, 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].The goal of the report published by Michael Ji [7] was 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 consisted of requirement analysis and product research that identified three feasible products (text summarization tools). In the criteria establishment phase, evaluation criteria were established for the two sub-criteria (characteristic and testing). While the characterization phase comprised of the data collection for the criteria defined. It was followed by the evaluation experiment (or testing) performed on the established testing criteria, as the final phase of the evaluation process.Table 1 below gives the comparison of various researches done for text summarization.

Table.1: 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, AmitSinghal, Chris Buckley	Paragraph Extraction	Expands on the sentence extraction technique by implementing a more generalised technique	Extractive
A Hybrid Approach for Extractive Document	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	Extractive

Summarization Using Machine Learning and Clustering Technique			everytime a document is given to it so that the summary is better each time	
Automatic Summarization of Meeting Data: A Feasibility Study	Anne HendrikBuist, 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, KarelJežek	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-RenKe, 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

III. DISCUSSION

As per our research, it is quite evident that extractive based summarizing implementations have had a greater deal of success than abstractive based. However, even though the implementations within the bounds of the domains to which the studies have been restricted have been successful, they are still not as accurate as would be expected to a normal user of that system. As far as the research on abstractive summarization is considered, successful implementations are a rarity, though the research conducted on it, at least theoretically, proves that if a successful implementation is attained, the summary generated will make more sense than the summary from an extraction based summary.

IV. PROPOSED SYSTEM

The proposed system as shown in figure 1 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 aid in the semantic analysis and SVD phases in the summarizer. The input document is first parsed or pre-processed, wherein there is a removal of unneeded words such as ‘stop words’ which are simply small function words, like “the”, “and”, “a”, which do not contribute meaning to the text summary. The next stage is the 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. The SVD Analysis stage derives the latent semantic structure from the document represented by matrix A. Finally in the summarization process, the system arranges the sentences generated from the SVD Analysis stage by semantically placing them in a way that the summary encompasses all the concepts of the original text. The final summary is then given back to the user.

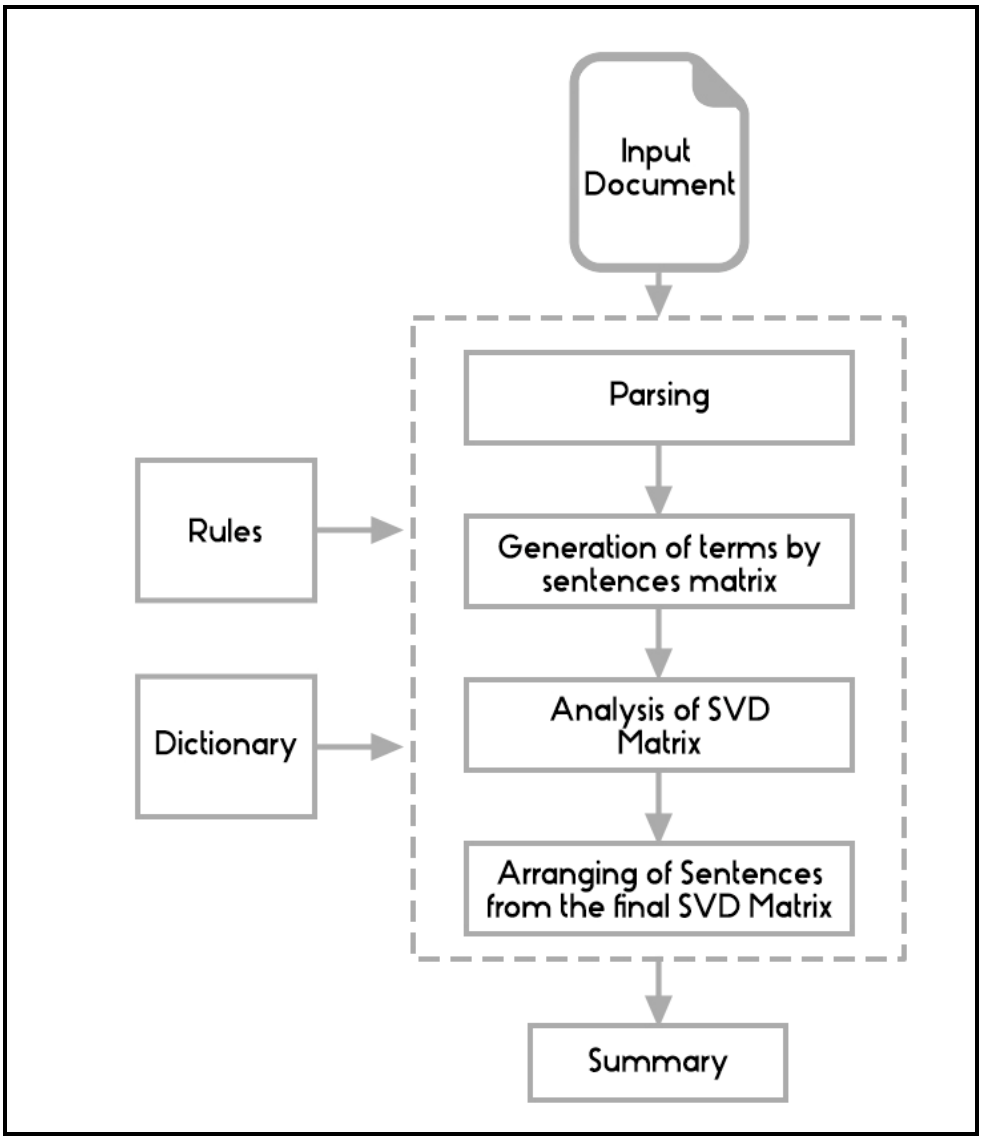

Fig.1: Proposed System

V. IMPLEMENTATION

The below given is the code for implementation of Latent Semantic Analysis (LSA) using Python library.

```
//Implementataion of LSA in Python

# coding: utf-8
importnumpy as np
frombaseclass import BaseSummarizer
fromscipy.sparse.linalg import svds
from warnings import warn
classBaseLsaSummarizer(BaseSummarizer):
    """
    This is an abstract base class for summarizers using the LSA method.
    """

    @classmethod
def _svd(cls, matrix, num_concepts=5):
        """
        Perform singular value decomposition for dimensionality reduction of the input matrix.
        """
u, s, v = svds(matrix, k=num_concepts)
return u, s, v
```

```
    @classmethod
def _validate_num_topics(cls, topics, sentences):
        # Determine the number of "linearly independent"
sentences
        # This gives us an estimate for the rank of the matrix
for which we will compute SVD
sentences_set = set([frozenset(sentence.split(' ')) for
sentence in sentences])
est_matrix_rank = len(sentences_set)

ifest_matrix_rank<= 1:
raiseSvdRankException('The sentence matrix does not
have sufficient rank to compute SVD')

if topics >est_matrix_rank - 1:
warn(
            'The parameter "topics" must be <=
rank(sentence_matrix) - 1 to avoid rank '
            'deficiency in the SVD computation. The
number of topics has been adjusted '
            'to equal rank(sentence_matrix) - 1 but this
could result in a poor summary.',
            Warning
        )
topics = est_matrix_rank - 1

return topics
classSvdRankException(Exception):
pass
classLsaSteinberger(BaseLsaSummarizer):

def summarize(self, text, topics=4, length=5,
binary_matrix=True, topic_sigma_threshold=0.5):
        """
        Implements the method of latent semantic analysis
described by Steinberger and Jezek in the paper:
        J. Steinberger and K. Jezek (2004). Using latent
semantic analysis in text summarization and summary
evaluation.
        Proc. ISIM '04, pp. 93–100.
:param text: a string of text to be summarized, path to a
text file, or URL starting with http
:param topics: the number of topics/concepts covered in
the input text (defines the degree of
dimensionality reduction in the SVD step)
:param length: the length of the output summary; either a
number of sentences (e.g. 5) or a percentage
of the original document (e.g. 0.5)
:parambinary_matrix: boolean value indicating whether
the matrix of word counts should be binary
        (True by default)
:paramtopic_sigma_threshold: filters out topics/concepts
with a singular value less than this
percentage of the largest singular value (must be between
0 and 1, 0.5 by default)
:return: list of sentences for the summary
        """

text = self._parse_input(text)

sentences, unprocessed_sentences =
self._tokenizer.tokenize_sentences(text)

length = self._parse_summary_length(length,
len(sentences))
if length == len(sentences):
returnunprocessed_sentences

topics = self._validate_num_topics(topics, sentences)

        # Generate a matrix of terms that appear in each
sentence
weighting = 'binary' if binary_matrix else 'frequency'
sentence_matrix = self._compute_matrix(sentences,
weighting=weighting)
sentence_matrix = sentence_matrix.transpose()

        # Filter out negatives in the sparse matrix (need to do
this on Vt for LSA method):
sentence_matrix =
sentence_matrix.multiply(sentence_matrix> 0)

s, u, v = self._svd(sentence_matrix,
num_concepts=topics)

        # Only consider topics/concepts whose singular
values are half of the largest singular value
if 1 <= topic_sigma_threshold< 0:
raiseValueError('Parameter topic_sigma_threshold must
take a value between 0 and 1')
sigma_threshold = max(u) * topic_sigma_threshold
u[u <sigma_threshold] = 0  # Set all other singular values
to zero
        # Build a "length vector" containing the length (i.e.
saliency) of each sentence
saliency_vec = np.dot(np.square(u), np.square(v))

top_sentences = saliency_vec.argsort()[-length:][::-1]
        # Return the sentences in the order in which they
appear in the document
top_sentences.sort()

return [unprocessed_sentences[i] for i in top_sentences]
```

User End Script for Summarizing txt file

```
# coding=utf-8
frompytldr.summarize.lsa import LsaSteinberger

if __name__ == "__main__":
        demo = open('demo.txt', 'r')
        txt = demo.read()


lsa_s = LsaSteinberger()


print '\n\nLSA Steinberger:\n'
summary = lsa_s.summarize(txt, length=0.5,
binary_matrix=True, topics=5,
topic_sigma_threshold=0.8)
for sentence in summary:
        print sentence
```

VI. RESULTS

In this section, we show the result of summarization of the text document using the Latent Semantic Analysis Summarizer in Python.

Original Text

In a no-holds-barred email to the board seen by the BBC, Cyrus Mistry says he had become a "lame duck" chairman and alleges constant interference, including being asked to sign off on deals he knew little about.
He also warned the company risks huge writedowns across the business.
Tata said it currently had no response to the allegations.
The Bombay Stock Exchange has sought clarification from Tata on the contents of Mr Mistry's letter.
Tata Sons, the holding company of Tata Group, unexpectedly replaced Mr Mistry with his predecessor Ratan Tata on Monday, giving no explanation or details about its decision.
But analysts say there was a clash over strategy, with the Tata family unhappy at Mr Mistry's policy of looking to sell off parts of the business - including Tata's European steel business - rather than holding on to assets and extending the firm's global reach.
Whatever the reasons, Mr Mistry has come out fighting. In his blistering five-page attack, he wrote that the board had "not covered itself with glory" and that the nature of his dismissal had done "immeasurable harm" to both his own reputation and that of the firm.
And he said that when he moved from being a non-executive director to chairman in 2012, he did "not have a clear grasp of the gravity" of problems he had inherited.
While saying that he did not want to "air a laundry list", Mr Mistry went on to unleash a brutal assessment of many aspects of the business, warning the firm may face 1.18 trillion rupees ($18bn) in writedowns because because of five unprofitable businesses he inherited.
Issues he raised included:
Huge debts from many of its foreign investments including hotels, its chemicals business in the UK and Kenya, and steel operations in Europe.
A telecoms business that is "continuously haemorrhaging" money as well as facing a fine of at least $1bn
Tata Power struggling because of underestimating coal prices, and getting into clashes with local landowners
Mr Mistry said there was no sign of profitability on the Tata Nano project - which had been launched as the world's cheapest car - and criticised a failure to face up to the reality of its consistently losing money.
"Any turnaround strategy for the company requires to shut it down. Emotional reasons alone have kept us away from that crucial decision," he said.
Tata's foray into the aviation sector was also criticised, with Mr Mistry suggesting he signed up to joint ventures under pressure from the former chairman.
He claimed he was asked by Ratan Tata to sign off quickly on a tie-up with Malaysia's Air Asia to create Air Asia India and that "my pushback was hard but futile".
And he wrote that Tata's 51% stake in Vistara - a venture between Tata and Singapore Airlines - was also foisted upon on him "without the benefit of time and experience to fully evaluate the proposal".
Cyrus Mistry had been hand-picked as a successor to Ratan Tata as the second chairman from outside the Tata family and with high hopes that he would be the right man to steer the company.
He was the sixth chairman in Tata's 148-year history and the first chairman in nearly 80 years to come from outside the Tata family.
But Mr Mistry did not come into the job cold. His family has been a major Tata investor since the 1930s and controls companies holding 18% of Tata Sons.
And he knows the family well, not least because of his sister's marriage to Ratan Tata's half-brother, Noel.

Summarized Text

In a no-holds-barred email to the board seen by the BBC, Cyrus Mistry says he had become a "lame duck" chairman and alleges constant interference, including being asked to sign off on deals he knew little about.
Tata Sons, the holding company of Tata Group, unexpectedly replaced Mr Mistry with his predecessor Ratan Tata on Monday, giving no explanation or details about its decision.
But analysts say there was a clash over strategy, with the Tata family unhappy at Mr Mistry's policy of looking to sell off parts of the business - including Tata's European

steel business - rather than holding on to assets and extending the firm's global reach.

While saying that he did not want to "air a laundry list", Mr Mistry went on to unleash a brutal assessment of many aspects of the business, warning the firm may face 1.18 trillion rupees ($18bn) in writedowns because of five unprofitable businesses he inherited.

Mr Mistry said there was no sign of profitability on the Tata Nano project - which had been launched as the world's cheapest car - and criticised a failure to face up to the reality of its consistently losing money.

Cyrus Mistry had been hand-picked as a successor to Ratan Tata as the second chairman from outside the Tata family and with high hopes that he would be the right man to steer the company.

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), and USC's ISI Summarist. 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 languages.

REFERENCES

[1] Josef Steinberger, KarelJežek, "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, AmitSinghal, 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 HendrikBuist, 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-RenKe, 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.