IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

In-a-Nutshell Document Summarizer

Mrunmayi Anchawale[1], Shravani Joshi[2], Rajat Shenoi[3], Shreya Bamne[4]

Student, Information Technology Department, Vidyalankar Institute of Technology, Mumbai, India[1,2,3]

Lecturer, Information Technology Department, Vidyalankar Institute of Technology, Mumbai, India[4]

Abstract: Automatic data summarization is part of machine learning and text mining, in which source text is condensed into a shorter version preserving its information content and overall meaning. First developed as a labour-intensive manual discipline in the 1980s, text mining has become ever more efficient as computing power has increased. In-A-Nutshell is an attempt to create a robust automated text summarization system, based on sentence scoring.

Keywords: text mining, summarization, NLP, extraction, abstraction, cue phrases, sentence generation.

I. INTRODUCTION

A. The Concept

Text summarization is the process of searching through countless pages of plain-language digitized text to find useful information that's been hiding in plain sight. It is more about finding unseen connections and patterns in plain-language narratives.

Several existing systems, including some Web browsers, claim to perform summarization. However, an analysis of their output shows that their summaries are simply portions of the text, produced verbatim. While there is nothing wrong with such extracts, the word 'summary' usually connotes something more, involving the fusion of various 'concepts of the text' into a 'smaller number of concepts'. Many methods have emerged over time for generation of summaries.

An extractive method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences.

An abstractive method consists of understanding the original text and re-telling it in fewer words. Abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what humans might generate.

In addition to extracts and abstracts, summaries may differ in several other ways. Some of the major types of summary that have been identified include indicative (keywords indicating topics) vs. informative (content laden); generic (author's perspective) vs. query-oriented (user-specific); background vs. just-the-news; single document vs. multi-document; neutral vs. evaluative.

This process can be used in many applications such as information retrieval, intelligence gathering, information extraction, text mining, and indexing [5][7][14]. The texts that are mined could be newspaper or website articles, research papers, blog entries, patent applications;

A summary should meet two conditions: maintain a wide coverage of the document topics and keep low redundancy at the same time [7][14]. A good generic summary should contain the major topics of the document and minimize redundancy. A full understanding of the major dimensions of variation, and the types of reasoning required to produce each of them, is still a matter of investigation. This makes the study of automated text summarization an exciting area to work in.

B. Overview of In-A-Nutshell

As the problem of information overload has grown with a large volume of text documents, presenting the user with a summary of each document greatly facilitates the task of allowing the user to read less data but still receive the most important information.

In-A-Nutshell is a java web application that allows users to upload files of the type .doc, .docx, .txt, .pdf containing arbitrary English input text and provides both extracted summary as well as abstracted summary after processing the input.

Our application is a score based summarization technique which considers various factors for scoring a sentence like similarity of words, position relevancy, named entity recognition and cue phrases, along with some parts of Natural Language Processing (NLP).

The main objective of In-A-Nutshell is to:

1. Reduce the human effort and time required for the generation of summary
2. Allow users to obtain a quick overview of a given document

Any summary must consists of all the necessary details of the parent document and the length of summary must be less than the original document. In the previous methodology used for this particular task, it was felt that few of the important sentences were excluded from the summary due to the fact that their frequency does not satisfy the threshold value of sentence score because of usage of different phrases used to represent the same fact. The proposed technique will remove this problem up to a certain extent by considering the semantic similarity between sentences.

In-A-Nutshell makes use of following JAVA libraries:

Apache Lucene Core - Apache Lucene is a high-performance, full-featured text search engine library

DOI 10.17148/IJARCCE.2016.53195

written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform. It has been used here for the elimination of stop words, which is a part of our application's pre-processing module.

Apache OpenNLP - The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text. It supports the most common NLP tasks, such as tokenization, sentence detection, part-of-speech tagging, named entity extraction, chunking, parsing, and co-reference resolution. These tasks are usually required to build more advanced text processing services.

II. THE NEED OF AUTOMATIC SUMMARIZER

Businesses use data and text mining to analyse customer and competitor data to improve competitiveness; the pharmaceutical industry mines patents and research articles to improve drug discovery; within academic research, mining and analytics of large datasets are delivering efficiencies and new knowledge in areas as diverse as biological science, particle physics and media and communications. Economic, academic and social activities generate ever increasing quantities of data.

Businesses collect trillions of bytes of information on customer transactions, suppliers, internal operations and indeed competitors; the global research community generates over 1.5 million new scholarly articles per annum; and social networking sites such as Facebook and twitter enable users to share over 1.3 billion pieces of information/content per day. According to the McKinsey Global Institute's (MGI) 'Big Data' report 6, the generation of information and data has become a 'torrent', pouring into all sectors of the global economy and is predicted to increase at a rate of 40% annually. (Mar 14, 2012).

Exploitation of this vast data and information resource can generate significant economic benefits, says the report, including enhancements in productivity and competitiveness, as well as generating additional value for consumers.

III. STRUCTURE OF IN-A-NUTSHELL

The design of this application is based on the following modules:

A. User Interface
B. Pre-processor
C. Sentence connectivity calculator
D. Sentence Scorer
E. Extractor
F. Abstractor

Each module employs several different, complementary, methods.

A. User Interface

The user interface of our application is divided into two sections, one for accepting input and the other for displaying the results.

Input section consists of:

1) Upload button for allowing user to upload a file.
2) List of file types supported for upload.
3) Summarize button to provide the user with extracted as well as abstracted summary.
4) Statistics button to provide the user with additional information such as number of words in original text, number of words in summary, time taken for the application to generate summary and time taken to generate same summary manually.

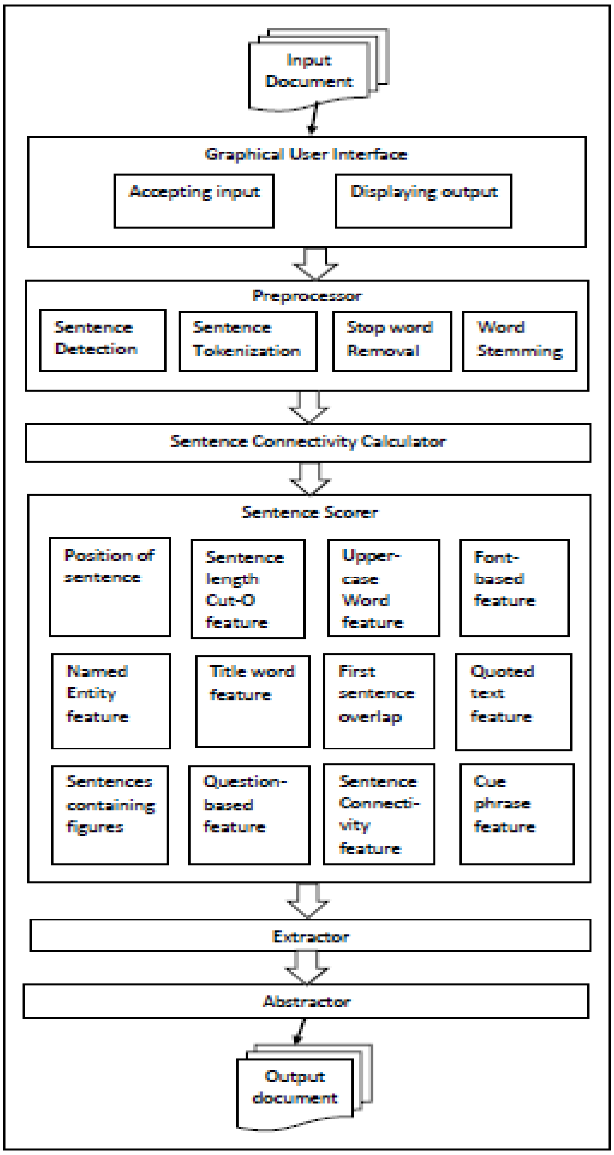


Fig. 1 Overall Methodology of In-A-Nutshell Document Summarizer

B. Pre-processor

The foundation of a summary is its sentences. Before generating these sentences, it is necessary to select the correct sentences from input document and place them in a desired format, by following a set of pre-processing steps, namely, sentence detection, tokenization, stop words removal and word stemming.

1) Sentence Detection:

This step deals with classification of document into sentences. While doing this, the fact that only a full-stop

DOI 10.17148/IJARCCE.2016.53195

does not indicate end of sentence is also considered. The method used for detection of sentences is “the longest white space trimmed character sequence between two punctuation marks.”

2) *Sentence Tokenization:*
This step deals with dividing the sentences into collection of unique tokens. This is done since text processing components like part-of-speech taggers, parsers, stemmers and so on, work with tokenized text.

3) Stop Word Removal:
Sometimes, some extremely common words which do not contribute to help select sentences which are to be kept in summary and are of little value are excluded from the document entirely. These words are called *stop words*. Some of the common English stop words include ‘a’, ‘the’, ‘is’, ‘from’, ‘he’, ‘will’ etc. The general strategy for determining a stop list is to sort the terms by *collection frequency* (the total number of times each term appears in the document), and then to take the most frequent terms.

4) Word Stemming:
Stemming in linguistics refers to the process of obtaining root form of a word. The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance:

am, are, is ⇒ be

car, cars, car's, cars' ⇒ car

Stemming is done in order to ensure that two words, which have the same root word, get the same score in the Sentence Scorer module.

C. Sentence connectivity calculator
In this module, the similarity (interconnectivity) of each sentence with every other sentence of the document is computed. This is known as *Sentence-to-Sentence Cohesion.*

Interconnectivity function:
This function receives two sentences, and returns a score for the intersection between them. We just split each sentence into words/tokens, count how many common tokens we have, and then we normalize the result with the average length of the two sentences. f (s1, s2) = |{w | w in s1 and w in s2}| / ((|s1| + |s2|) / 2)

In the first step we split the text into sentences, and store the intersection value between each two sentences in a matrix (two-dimensional array). So values[0][2] will hold the intersection score between sentence #1 and sentence #3. In the second step we calculate an individual score for each sentence and store it in a key-value dictionary, where the sentence itself is the key and the value is the total score. We do that just by summing up all its intersections with the other sentences in the text (not including itself).

D. Sentence Scorer
In this module, features influencing the relevance of sentences are decided and then scores are assigned to these features. Final score of each sentence is determined by adding scores from each feature. Top ranked sentences are selected for final summary.

The scoring mechanism of In-A-Nutshell awards certain points to a pre-processed sentence based on following features:

1) Position-of-sentence feature:
This method exploits the fact that in some genres, certain sentence positions tend to carry more topic material than others [5][7]. Optimal Position Policy (OPP) is defined as a list that indicates in what ordinal positions in the text, high-topic bearing sentences occur. This work, described in [22], is the first systematic study and evaluation of the Position method reported.

For the Ziff-Davis corpus (13,000 newspaper articles announcing computer products) research has found that the OPP is [T1, P2S1, P3S1, P4S1, P1S1, P2S2, {P3S2, P4S2, P5S1, P1S2}, P6S1,…] i.e., the title (T1) is the most likely to bear topics, followed by the first sentence of paragraph 2, the first sentence of paragraph 3, etc. In contrast, for the Wall Street Journal the OPP is [T1, P1S1, P1S2, ...]

Generalizing this approach to all topics, it is found that sentences occurring in initial and final position of entire document(i.e. the first and last sentence) as well as first sentence of individual paragraphs have a higher probability of being relevant, and hence they obtain a higher score.

2) Sentence Length Cut-O Feature:
Sentences containing less than a pre-specified number of words are not included in the summary

3) Upper-case word feature:
Sentences containing acronyms are given a higher score.

4) Font based feature:
Sentences containing words appearing in upper case, bold, italics or underlined fonts are usually more important, and hence are given higher score.

5) Sentence Connectivity Score
This feature calculates total connectivity score of a sentence, which is the sum of the relative connectivity scores, obtained from Sentence Connectivity Calculator, mentioned above.

6) Named Entity feature:
Sentences that contain Proper nouns, Names of people, Places and Dates are considered as important and are given a higher score.

7) Sentence with figures feature:
Writing a scholarly manuscript often requires the use of numbers to express important information, particularly in the science field. Also, news articles and articles related to stock market are full of numbers. Our application considers figures as an important parameter to score the sentences.

8) Title word feature:
If the user input document comes with an already available title, then the sentences in the document which contain words that appear in the title are also indicative of the theme of the document. These sentences have higher chances of inclusion in summary.

DOI 10.17148/IJARCCE.2016.53195

9) *Cue-Phrase feature:*
Phrases such as 'in summary', 'in conclusion', and superlatives such as 'the best', 'the most important' can be good indicators of important content in a document [6][21]. Cue phrases are generally genre dependent.

For example, 'rise' and 'theses' would be present in documents related to stock market and scientific literature respectively. After careful research on a number of documents of various genres, we have successfully developed a list of 510 generalised cue phrases applicable to any document of any genre. Our application uses this research as a foundation for scoring of sentences, which includes cue phrases like 'this paper', 'this article', 'the fact', 'outline', 'proof' to name a few.

In-A-Nutshell divides the cue phrases into three categories based on their importance.

Category 1: It includes cue phrases that are best indicators of important content which simply must appear in the summary. Sentences containing cue phrases from this category are given the highest score. Examples include 'as a result', 'defined', 'important' etc.

Category 2: It includes cue phrases whose roles differ according to context i.e. cue phrases that appear to be very important in one context but not so important in another context. Sentences containing cue phrases from this category are given a medium score. Examples include 'recently', 'although', 'classify' etc.

Category 3: It includes cue phrases that might give a slight clue about the overall topic of any user document. Sentences containing cue phrases from this category are given the least score. Examples include 'likewise', 'relate' etc.

Our application also considers the position dependency of a cue phrase and scores it accordingly. For example, the sentence "Starters would be served *first*" is not in the same importance class as "The *first* person to go on moon was Neil Armstrong"

For more accuracy, our application assumes that a sentence containing *many* cue phrases from category 3 are of the same importance as that of a sentence containing *only one* cue phrase of category 2, and hence they are scored equally. This is the NP-IP condition (i.e. the number of cue phrases in a sentence as well as their importance is considered)

10) *Quoted text feature:*
The sentences having quotes are also given higher score. If an entire sentence is in quotes and it contains words such as 'I', 'you', 'we' etc, our application regards this as a conversation sentence and does not give a higher score to it.

11) *Question based feature:*
If a sentence is interrogative, then this sentence and its next one are given a high score. This is based on the fact that the next sentence of an interrogative sentence might contain answer to the question asked in previous sentence. For example, "What is a database? A *database* is a collection of information that is organized so that it can easily be accessed, managed, and updated." However, rhetorical questions are not given any weightage.

12) *First-sentence overlap feature:*
As discussed earlier, the first sentence of the document has a higher probability of being relevant for the summary. This feature checks how much similar (closer in meaning) is each sentence to the first sentence.

E. Extractor
The scored sentences are given to the Extractor module. This module picks out the sentences in descending order of their scores. The user is given the choice of selecting what percent of summary he/she wants (the default being 40%)

F. Abstractor
In-A-Nutshell uses Markov chains to generate new sentences from existing ones.

There are many real-world scenarios where it's useful for a program to create new sentences. For example, *Google Translate* analyzes a sentence in a foreign language, and generates a new sentence in English with the same meaning. *Siri* listens to questions, and generates new sentences that answer those questions. When programs generate sentences, they usually follow a simple trick. First, they analyze lots of existing sentences that are similar to what they want to generate, and record which words and phrases occur frequently. Then, they randomly choose phrases that occur, and rearrange them in a way that makes sense. Markov chains are the simplest way to generate sentences that almost make sense, but really don't. They are based on figuring out the likelihood of a word following another word by looking at existing bodies of text (for example, *Wikipedia*).

Then, to generate sentences you choose a starting word and based on a random variable as well as the probabilities that you've found by looking at existing text, you choose a word following that starting word and repeat.

In the 1948 landmark paper 'A Mathematical Theory of Communication', Claude Shannon founded the field of information theory and revolutionized the telecommunications industry, laying the groundwork for today's Information Age. In this paper, Shannon proposed using a *Markov chain* to create a statistical model of the sequences of letters in a piece of English text. Markov chains are now widely used in speech recognition, handwriting recognition, information retrieval, data compression, and spam filtering.

The basic steps of creating Markov chains are:

1. Select a random starting word to start a new sentence.
2. From all the words that ever follow that word in the input sequence, choose one. Add that word to the end of our new sentence.
3. Continue selecting randomly from the words that can possibly follow the current last word of our sentence until either there are no possible choices or we have made a sentence as long as desired.

DOI 10.17148/IJARCCE.2016.53195

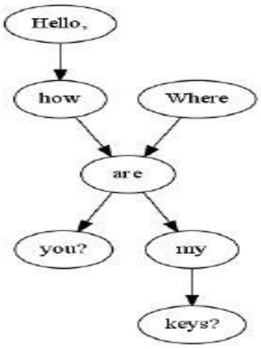


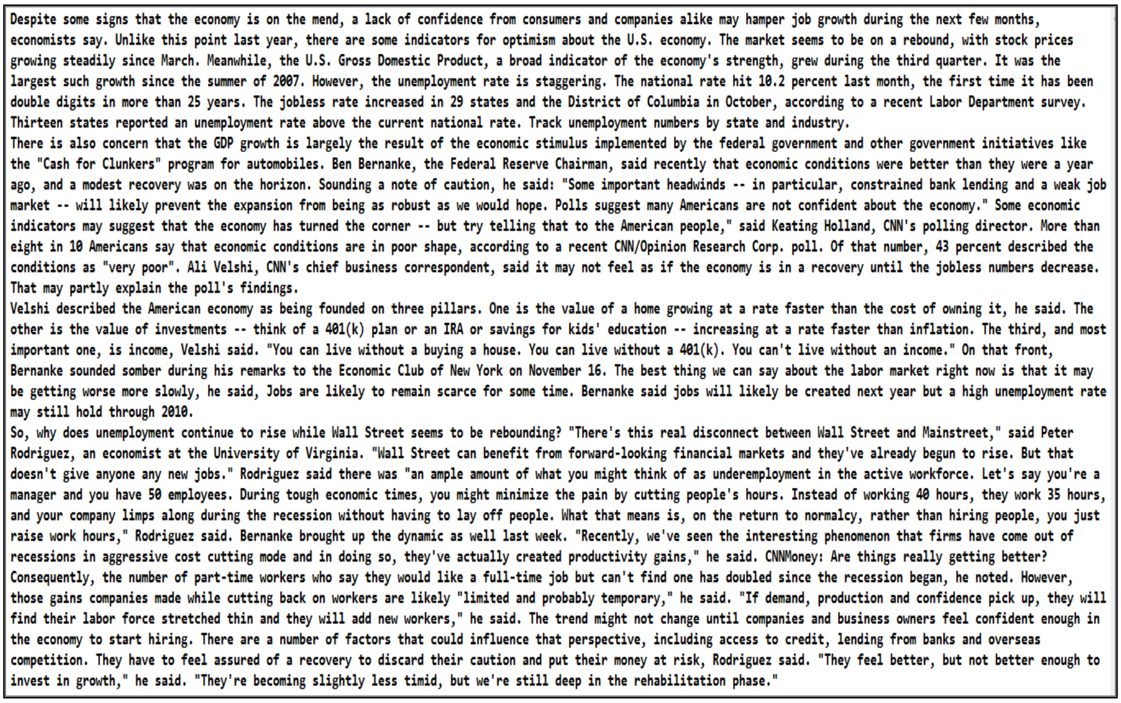
Fig. 2 A graphical representation of the Markov possibilities for "Hello, how are you?" and "Where are my keys?"

E.g. consider two sentences "Hello, how are you" and "where are my keys". If we convert these sentences into a graph showing the possible results, we would get Figure 2

In this graph, each arrow represents a choice we can take based on the last word we added to our sentence, continuing until there are no valid paths to take. Looking at the graph, there are four possible outputs if we start our chain with either "Hello," or "Where":

- Hello, how are you?
- Hello, how are my keys?
- Where are you?
- Where are my keys?

IV.RESULTS



Despite some signs that the economy is on the mend, a lack of confidence from consumers and companies alike may hamper job growth during the next few months, economists say. Unlike this point last year, there are some indicators for optimism about the U.S. economy. The market seems to be on a rebound, with stock prices growing steadily since March. Meanwhile, the U.S. Gross Domestic Product, a broad indicator of the economy's strength, grew during the third quarter. It was the largest such growth since the summer of 2007. However, the unemployment rate is staggering. The national rate hit 10.2 percent last month, the first time it has been double digits in more than 25 years. The jobless rate increased in 29 states and the District of Columbia in October, according to a recent Labor Department survey. Thirteen states reported an unemployment rate above the current national rate. Track unemployment numbers by state and industry.

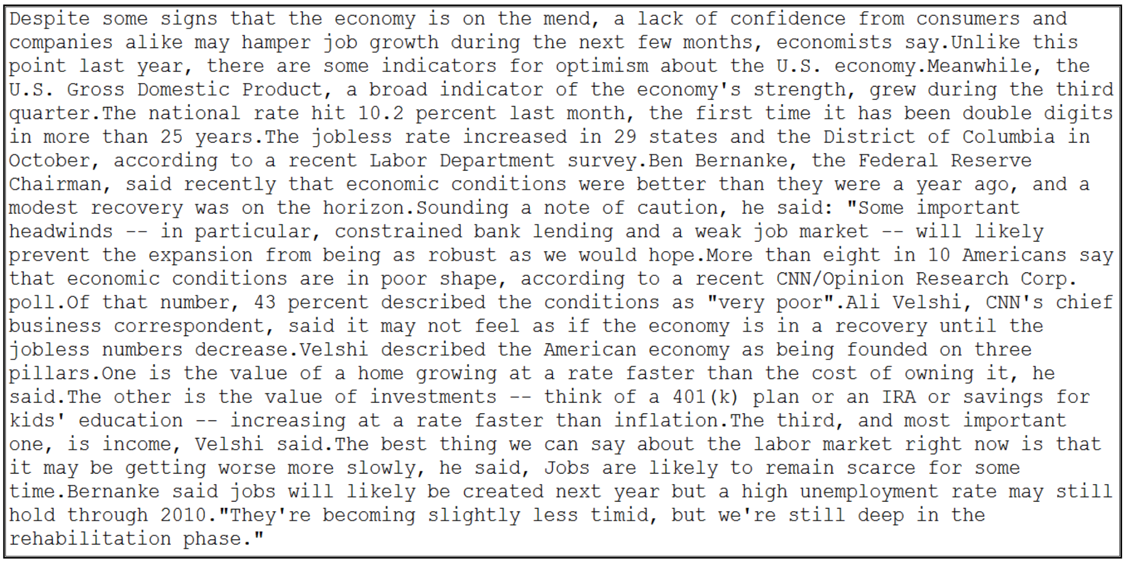
There is also concern that the GDP growth is largely the result of the economic stimulus implemented by the federal government and other government initiatives like the "Cash for Clunkers" program for automobiles. Ben Bernanke, the Federal Reserve Chairman, said recently that economic conditions were better than they were a year ago, and a modest recovery was on the horizon. Sounding a note of caution, he said: "Some important headwinds -- in particular, constrained bank lending and a weak job market -- will likely prevent the expansion from being as robust as we would hope. Polls suggest many Americans are not confident about the economy." Some economic indicators may suggest that the economy has turned the corner -- but try telling that to the American people," said Keating Holland, CNN's polling director. More than eight in 10 Americans say that economic conditions are in poor shape, according to a recent CNN/Opinion Research Corp. poll. Of that number, 43 percent described the conditions as "very poor". Ali Velshi, CNN's chief business correspondent, said it may not feel as if the economy is in a recovery until the jobless numbers decrease. That may partly explain the poll's findings.

Velshi described the American economy as being founded on three pillars. One is the value of a home growing at a rate faster than the cost of owning it, he said. The other is the value of investments -- think of a 401(k) plan or an IRA or savings for kids' education -- increasing at a rate faster than inflation. The third, and most important one, is income, Velshi said. "You can live without a buying a house. You can live without a 401(k). You can't live without an income." On that front, Bernanke sounded somber during his remarks to the Economic Club of New York on November 16. The best thing we can say about the labor market right now is that it may be getting worse more slowly, he said, Jobs are likely to remain scarce for some time. Bernanke said jobs will likely be created next year but a high unemployment rate may still hold through 2010.

So, why does unemployment continue to rise while Wall Street seems to be rebounding? "There's this real disconnect between Wall Street and Mainstreet," said Peter Rodriguez, an economist at the University of Virginia. "Wall Street can benefit from forward-looking financial markets and they've already begun to rise. But that doesn't give anyone any new jobs." Rodriguez said there was "an ample amount of what you might think of as underemployment in the active workforce. Let's say you're a manager and you have 50 employees. During tough economic times, you might minimize the pain by cutting people's hours. Instead of working 40 hours, they work 35 hours, and your company limps along during the recession without having to lay off people. What that means is, on the return to normalcy, rather than hiring people, you just raise work hours," Rodriguez said. Bernanke brought up the dynamic as well last week. "Recently, we've seen the interesting phenomenon that firms have come out of recessions in aggressive cost cutting mode and in doing so, they've actually created productivity gains," he said. CNNMoney: Are things really getting better?

Consequently, the number of part-time workers who say they would like a full-time job but can't find one has doubled since the recession began, he noted. However, those gains companies made while cutting back on workers are likely "limited and probably temporary," he said. "If demand, production and confidence pick up, they will find their labor force stretched thin and they will add new workers," he said. The trend might not change until companies and business owners feel confident enough in the economy to start hiring. There are a number of factors that could influence that perspective, including access to credit, lending from banks and overseas competition. They have to feel assured of a recovery to discard their caution and put their money at risk, Rodriguez said. "They feel better, but not better enough to invest in growth," he said. "They're becoming slightly less timid, but we're still deep in the rehabilitation phase."

Fig 3. Original document



Despite some signs that the economy is on the mend, a lack of confidence from consumers and companies alike may hamper job growth during the next few months, economists say.Unlike this point last year, there are some indicators for optimism about the U.S. economy.Meanwhile, the U.S. Gross Domestic Product, a broad indicator of the economy's strength, grew during the third quarter.The national rate hit 10.2 percent last month, the first time it has been double digits in more than 25 years.The jobless rate increased in 29 states and the District of Columbia in October, according to a recent Labor Department survey.Ben Bernanke, the Federal Reserve Chairman, said recently that economic conditions were better than they were a year ago, and a modest recovery was on the horizon.Sounding a note of caution, he said: "Some important headwinds -- in particular, constrained bank lending and a weak job market -- will likely prevent the expansion from being as robust as we would hope.More than eight in 10 Americans say that economic conditions are in poor shape, according to a recent CNN/Opinion Research Corp. poll.Of that number, 43 percent described the conditions as "very poor".Ali Velshi, CNN's chief business correspondent, said it may not feel as if the economy is in a recovery until the jobless numbers decrease.Velshi described the American economy as being founded on three pillars.One is the value of a home growing at a rate faster than the cost of owning it, he said.The other is the value of investments -- think of a 401(k) plan or an IRA or savings for kids' education -- increasing at a rate faster than inflation.The third, and most important one, is income, Velshi said.The best thing we can say about the labor market right now is that it may be getting worse more slowly, he said, Jobs are likely to remain scarce for some time.Bernanke said jobs will likely be created next year but a high unemployment rate may still hold through 2010."They're becoming slightly less timid, but we're still deep in the rehabilitation phase."

Fig 4. Summary

DOI 10.17148/IJARCCE.2016.53195

V. CONCLUSION AND FUTURE SCOPE

In this paper a single document sentence scoring based text summarization algorithm is introduced. The result shown by this technique is found to be more efficient than the previously used technique which considers the frequency of text only. Semantic similarity is also used in this algorithm. The proposed algorithm is implemented using java platform and is verified over the standard text mining corpus. The discovered results are interesting and gist of the summarized document is also preserved. The future direction for the proposed work is to apply the similar concept in multi-document summarization. We are also looking forward to extending our system to facilitate search engine optimization (selection of precise and relevant web pages or documents) based on a user query.

Automated summarization is an old topic (work on it dates from the 1950's) and a new topic as well. It is so difficult that an interesting headway can be made for many years to come. We are excited about the possibilities offered by the combination of semantic and statistical techniques in what is, quite possibly, the most complex task of all NLP.

ACKNOWLEDGMENT

The authors would like to thank all reviewers who have provided constructive feedback on this paper.

REFERENCES

[1] Gobinda G. Chowdhury, "Natural Language Processing", Annual Review of Information Science and Technology, Vol: 37, pp: 51–89, 2003.

[2] Inderjeet Mani, "Recent Developments in Text Summarization", In Proceedings of the tenth international conference on Information and knowledge management, ACM Press, pp: 529 - 531, 2001.

[3] Yan Liu, Sheng-hua Zhong, Wen-jie Li, "Query-oriented Unsupervised Multi-document Summarization via Deep Learning", Under review in Journal of Neural Networks (NN).

[4] M. S. Binwahlan, N. Salim, L. Suanmali, "Intelligent Model for Automatic Text Summarization", Information Technology Journal, pp: 1249-1255, 2009.

[5] H. Luhn, "The automatic creation of literature abstracts", IBM Journal of Research and Development, Vol: 2, Number: 2, pp: 159-165, 1958.

[6] H. Edmundson, "New methods in automatic extracting", Journal of the Association for Computing Machinery, Vol: 16, No. 2, pp: 264-285, 1969.

[7] I. Mani, M. Maybury, "Advances in Automatic Text Summarization", MIT Press, 1999.

[8] Michel Gagnon, Lyne Da Sylva, "Text Summarization by Sentence Extraction and Syntactic Pruning", In Proceedings of Computational Linguistics in the North East, 2005.

[9] Kevin Knight, Daniel Marcu, "Summarization beyond sentence extraction:A probabilistic approach to sentence compression", In Artificial Intelligence, Vol: 139, Issue 1, pp: 91–107, 2002.

[10] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz, "Multi-document summarization by sentence extraction", ANLP/NAACL Workshop, pp: 40–48, 2000.

[11] Naresh Kumar Nagwani, Dr. Shrish Verma, "A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 17– No.2, March 2011.

[12] Dragomir R. Rade and Weiguo Fan and Zhu Zhang, "WebInEssence: A Personalized Web-Based Multi Document Summarization and Recommendation System"

[13] Goldstein J., Kantrowitz M., MittalV., Carbonell J.:Summarizing Text Documents: Sentence Selection and Evaluation Metrics. Proceedings of the 22th ACM SIGIR, 121-127, (1999).

[14] I. Mani, Automatic Summarization, John Benjamins Publishing Co. (2001) 1-22.

[15] Mitra M., Singhal A., Buckley C.: Automatic Text Summarization by Paragraph Extraction. Proceedings of theACL'97/EACL'97Workshop on Intelligent Scalable Text Summarization, pp. 31–36 (1997).

[16] Rafeeq Al-Hashemi, "Text Summarization Extraction System (TSES) Using Extracted Keywords", International Arab Journal of e-Technology, Vol. 1, No. 4, June, pp. 164-168, (2010).

[17] Wooncheol Jung, Youngjoong Ko, and Jungyun Seo, "Automatic Text Summarization Using Two-Step Sentence Extraction", AIRS 2004, LNCS 3411, pp. 71 – 81, (2005).

[18] Yulia Ledeneva, Alexander Gelbukh, and René Arnulfo García Hernández, "Terms Derived from Frequent Sequences for Extractive Text Summarization", CICLing 2008, LNCS 4919, pp. 593–604, (2008).

[19] ArchanaAB,Sunitha 2013 An overview on document summarization technique. International Journal on Advanced Computer Theory and Engineering (IJACTE), Volume-1, Issue-2.

[20] A.P. Siva Kumar, Dr. P. Premchand, Dr. A Govardhan 2011 Query-Based Summarizer Based on Similarity of Sentences and Word Frequency. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.3.

[21] Baxendale, P.B. October 1958. Machine-made index for technical literature—an experiment. IBM Journal (354–361).

[22] Lin, C.Y. and E.H. Hovy. 1997. Identifying Topics by Position. Proceedings of the Applied Natural Language Processing Conference, Washington, DC.

DOI 10.17148/IJARCCE.2016.53195