International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Automatic Text Summarization: A Detailed Study

Nimisha Dheer[1], Chetan Kumar[2]

[1]MTech Scholar, Department of Computer Science Engineering, Kautilya Institute of Technology & Engineering, Jaipur, India

[2]Associate professor, Department of Computer Science Engineering, Kautilya Institute of Technology & Engineering, Jaipur, India

Abstract: ***During this paper we've got studied the conception and trends within the field of text report. Text report is compression the supply text into a shorter version conserving its info content and overall that means. It's terribly tough for people to manually summarize large documents of text. Text report ways will be classified into extractive and abstractive report. Associate extractive report methodology consists of choosing vital sentences, paragraphs etc. from the first document and concatenating them into shorter kind. The importance of sentences is set supported applied mathematics and linguistic options of sentences. An abstractive summarization methodology consists of understanding the first text and re-telling it in fewer words. It uses linguistic ways to look at and interpret the text then to seek out the new ideas and expressions to best describe it by generating a brand new shorter text that conveys the most necessary info from the first text document.***

Keywords: Abstractive Summary, Extractive Summary, Summary, Text Summarization

1. Introduction

Text account has become a very important and timely tool for aiding and interpreting text data in today's aggressive modern era. It's terribly troublesome for human beings in general to manually summarize large documents of text. There's an abundance of text material obtainable on the net. However, sometimes the net provides a lot of data than is required. Therefore, a twofold downside is encountered: checking out relevant documents through an amazing variety of documents obtainable, and fascinating an outsized amount of relevant data. The goal of automatic text account is compression the supply text into a shorter version protective its data content and overall that means. An outline may be employed in an indicative approach as a pointer to some components of the first document, or in informative thanks to cover all relevant data of the text. In each case the foremost necessary advantage of employing an outline is its reduced reading time. An honest outline system ought to replicate the various topics of the document whereas keeping redundancy to a minimum. Summarization tools may additionally seek for headings and different markers of subtopics so as to spot the key points of a document. Microsoft Word's AutoSummarize operate may be an easy example of text account.

Text account ways may be classified into extractive and theoretical summarization. An extractive account methodology consists of choosing vital sentences, paragraphs etc. The importance of sentences is set supported applied math and linguistic options of sentences.

A theoretical account tries to develop an understanding of the most ideas in a very document and then categorical those ideas in clear language. It uses linguistic ways to look at and interpret the text then to seek out the new ideas and expressions to best describe it by generating a replacement shorter text that conveys the foremost vital data from the first text document. This paper focuses on extractive text account ways.

Extractive summaries are developed by extracting key text segments (sentences or passages) from the text, based mostly on applied math analysis of individual or mixed surface level options like word/phrase frequency, location or cue words to find the sentences to be extracted. The —most important" content is treated because the most frequent content. Such an approach so avoids any efforts on deep text understanding. They're conceptually easy, simple to implement. Extractive text account method may be divided into 2 steps:

1) Preprocessing step and
2) Processing step.

Preprocessing is structured illustration of the first text. it always includes:

a) Sentences boundary with presence of dot at the tip of sentence.
b) Stop-Word Elimination—Common words with no semantics and that don't combination relevant data to the task are eliminated.
c) Stemming—the purpose of stemming is to get the stem or base of every word that emphasize its semantics.

In process step, options influencing the connation of sentences are set and calculated then weights are assigned to those options using weight learning methodology. Final score of every sentence is set using Feature-weight equation. Prime hierarchal sentences are elite for final outline.

Problems with the extractive outline are:

1) Extracted sentences sometimes tend to be longer than average. Owing to this, elements of the segments that aren't essential for outline conjointly get enclosed, intense area.
2) Necessary or relevant data is typically unfolded across sentences, and extractive summaries cannot capture this (unless the outline is long enough to carry all those sentences).
3) Conflicting data might not be given accurately.
4) Pure extraction typically results in issues in overall coherence of the summary—a frequent issue considerations —dangling" anaphora. Sentences typically contain pronouns that lose their referents once extracted out of context. Worse yet, sewing along de-contextualized extracts could result in a deceptive interpretation of anaphors (resulting in an inaccurate

illustration of supply data, i.e., low fidelity). Similar problems exist with temporal expressions. These issues become a lot of severe within the multi-document case, since extracts area unit drawn from totally different sources. A general approach to addressing these problems involves post-processing extracts, as an example, exchange pronouns with their antecedents, exchange relative temporal expression with actual dates, etc. [1]

2. Importance and Relevance of the Study

According to paper ―Using Latent Semantic Analysis in Text Summarization and Summary Evaluation Josef Steinberger, and Karel Ježek ".

Generic text summarization may be a field that has seen increasing attention from the information science community. The particular immense quantity of electronic data has got to be reduced to modify the users to handle this data a lot of effectively. We have a tendency to mention here categories of summarization ways and a technique supported LSA that has been recently printed. We've got any changed and improved this technique. One foremost contentious elements of the outline analysis is its analysis method. Next a part of the article deals with potentialities of outline analysis. We have a tendency to propose there 2 new analysis ways supported LSA, that live a content similarity between an imaginative document and its outline. At the tip of the paper we have a tendency to gift analysis results and any analysis directions.

Yihong Gong and Xin Liu have printed the concept of using LSA in text summarization in 2002 [1]. They, impressed by the latent linguistics compartmentalization, applied the singular value decomposition (SVD) to generic text summarization show in Figure1. The method starts with creation of a term by sentences matrix A = [A1, A2, …, An] with every column vector Ai, representing the weighted term-frequency vector of sentence I within the document into account. If there are unit a complete of m terms and n sentences within the document, then we'll have an m × n matrix A for the document. Since each word doesn't unremarkably seem in every sentence, the matrix A is thin.

Given an m × n matrix A, wherever while not loss of generality $m \geq n$, the SVD of A is outlined as:

$$A = U\Sigma V^T, \quad (1)$$

where $U = [u_{ij}]$ is an m × n column orthonormal matrix whose columns are referred to as left singular vectors; $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ is an n × n diagonal matrix, whose diagonal components are non-negative singular values sorted in descending order, and $V = [v_{ij}]$ is an n × n orthonormal matrix, whose columns are referred to as right singular vectors. If rank (A) = r, then Σ satisfies:

$$\sigma_1 \geq \sigma_2 \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0 .$$

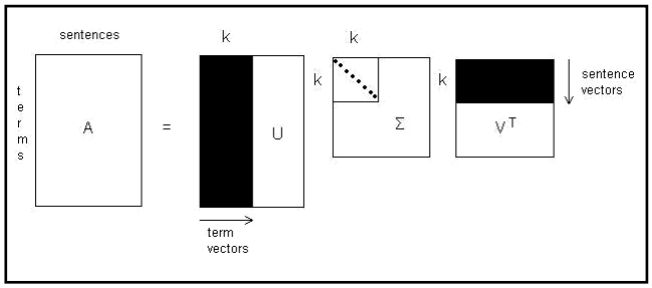

Figure 1: Singular Value Decomposition

The interpretation of applying the SVD to the terms by sentences matrix A will be made up of 2 totally different viewpoints. From transformation purpose of read, the SVD derives a mapping between the m-dimensional space spawned by the weighted term-frequency vectors and also the r-dimensional singular vector space. From linguistics purpose of read, the SVD derives the latent linguistics structure from the document drawn by matrix A. This operation reflects a breakdown of the initial document into r linearly-independent base vectors or ideas. Every term and sentence from the document is together indexed by these base vectors/concepts. A singular SVD feature is that it's capable of capturing and modeling interrelationships among terms in order that it will semantically cluster terms and sentences. Further-more, as incontestable in [5], if a word combination pattern is salient and continual in document, this pattern are captured and drawn by one in every of the singular vectors. The magnitude of the corresponding singular price indicates the importance degree of this pattern at intervals the document. Any sentences containing this word combination pattern are projected on this singular vector, and also the sentence that best represents this pattern can have the biggest index price with this vector. As every explicit word combination pattern describes a particular topic/concept within the document, the facts delineate on top of naturally because the hypothesis that every singular vector represents a salient topic/concept of the document, and also the magnitude of its corresponding singular price represents the degree of importance of the salient topic/concept.

Based on the on top of discussion, authors proposed a summarization technique that uses the matrix Green Mountain State. This matrix describes an importance degree of every topic in each sentence. The summarization method chooses the foremost informative sentence for every topic. It means the k^{th} sentence we elect has the biggest index price in k^{th} right singular vector in matrix V^T. [2]

Another paper is ―Automated Text Summarization in SUMMARIST, Eduard Hovy and Chin-Yew Lin ,Information Sciences Institute of the University of Southern California"

The task of a text summarizer is to provide an abstract of any document (or set of documents) submitted thereto. the amount of sophistication of a abstract will vary from an easy list of isolated keywords that indicate the main content of the document(s), through an inventory of independent single sentences that along categorical the main content, to a coherent, totally planned and generated text that compresses the document(s). The lot of subtle an abstract, the lot of effort it typically takes to provide. Many existing systems,

together with some internet browsers, claim to perform summarization. However, a cursory analysis of their output shows that their summaries area unit merely parts of the text, made verbatim. whereas there's nothing wrong with such extracts, per se, the word ‗summary' typically connotes one thing a lot of, involving the fusion of various ideas of the text into a smaller range of ideas, to create an abstract. we tend to outline extracts as consisting completely of parts extracted verbatim from the initial (they is also single words or whole passages) and abstracts as consisting of novel phrasings describing the content of the initial (which would possibly be paraphrases or totally synthesized text). Generally, manufacturing abstract needs stages of topic fusion and text generation not required for extracts.

In addition to extracts and abstracts, summaries might dissent in many alternative ways that. A number of the main kinds of outline that have been known embody indicative (Keywords indicating topics) v/s informative (content-Laden); generic (author's perspective) v/s query-oriented (user-specific); background vs. just-the-news; single-document vs. multi-document; neutral vs. evaluative. A full understanding of the main dimensions of variation, and therefore the kinds of reasoning needed to provide every of them, continues to be a matter of investigation. This makes the study of machine-driven text summarization associate exciting space within which to.

To produce abstract-type summaries, the core method may be a step of interpretation. During this step, 2 or a lot of topics are united along to create a 3rd, a lot of general, one. (We outline topic as a specific subject that we tend to pen or discuss.). This step should occur within the middle of the summarization procedure (show in figure 2): initial, associate initial stage of topic identification and extraction is needed to seek out the central topics within the input text; finally, to provide the outline, a terminal stage of sentence generation is required. So SUMMARIST relies on the subsequent ‗equation':

Summarization = Topic Extraction + Interpretation+ Generation

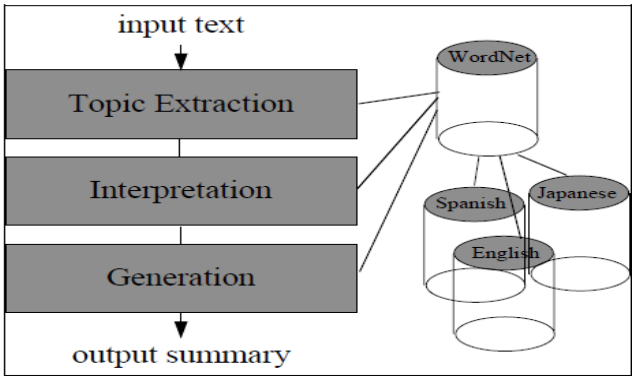

Figure 2: Architecture of SUMMARIST

This breakdown is driven as follows:
1. Extraction: The goal is to filter the input to retain only the foremost vital, central, topics. For generality we assume that a text will have several (sub)-topics, which the subject extraction method may be parameterized in at least two ways: initial, to incorporate a lot of or fewer topics to produce longer or shorter summaries, and second, to incorporate solely topics relating to the user's expressed interests. Typically, topic identification may be achieved using numerous complementary techniques, together with those supported conventional text structure, cue words, high-frequency indicator phrases, and discourse structure.

2. Interpretation: Once the specified central topics are known, they will merely be output, to form an extract. In human summaries, however, a method of interpretation is typically performed to win any compaction. In one study, counted what percentage clauses had to be extracted from a text so as to totally contain all the fabric enclosed in a very human abstract of that text.

3. Generation: The goal is to develop the extracted and united material into a coherent, densely phrased, new text. If this stage is skipped, the output may be a verbatim quotation of some portion(s) of the input, and isn't likely to be high-quality text (although this may be sufficient for the application). [3]

Another paper is ―An Automatic Text Summarization Using Lexica Cohesion And Correlation Of Sentences"
A.R.Kulkarni Computer Science & Engineering Department, Walchand Institute of Technology, Sholapur**.**

Text report is that the method of manufacturing a condensed version of original document. This condensed version should have vital content of the initial document. Analysis is being done since a few years to get coherent and indicative summaries using completely different techniques. Per (Jones, 1993) the text report is represented as 2 step method
i) Building a supply illustration from the initial document.
ii) Generating outline from the supply illustration.

Text report will be generally classified into 2 types: Single document report and multi-document report. This paper focuses on single document report that generates outline of single document. The text report will be categorized into extractive and theoretic supported the character of text illustration within the outline.

Many ways are planned until currently on generating a coherent outline. The sooner ways used solely applied math ways that targeted on term frequency [4] for selecting vital sentences. These ways weren't found to be economical because it failed to take into account all the contexts of the word or determine semantically connected terms called cohesion.

Currently a day's text report is taken into account as a language process task. Lexical chains a simplest kind of lexical cohesion was introduced by Morns & Hirst[5]. But it absolutely was found that every one doable senses of the word weren't taken under consideration. .

Berzilay & Elhada [6] gave an improved formula that constructs all doable interpretations of the supply text using lexical chains. It's an economical methodology for text report as lexical chains determine and capture vital ideas of the document while not going into deep linguistics analyses.

Lexical chains are created using some mental object that contains nouns and its numerous associations.

Our formula is predicated on the tactic used higher than. We've used Word net to get domain-specific extractive outline using Lexical chains for the nouns within the document. The formula segments the given content into sentences & then into tokens. These tokens are labeled using POS tagger. The Nouns are hand-picked & for every noun within the section, we tend to take into account its sense using Word net. Then we tend to conceive to merge these senses into all of the present chains altogether doable ways that, thence building each doable interpretation of the section. Next merge chains between segments that contain a word within the same sense in common. [6]

Another paper is ―COMPENDIUM: A text summarization system for generating abstracts of research papers, By Dr. Elena Lloret,Prof. Dr. M Teresa Romá-Ferri".

This article analyzes the appropriateness of a text report system, COMPENDIUM, for generating abstracts of medicine papers. 2 approaches are suggested: Associate in Nursing extractive (COMPENDIUME), that solely selects and extracts the foremost relevant sentences of the documents, an abstractive-oriented one (COMPENDIUME–A), so facing conjointly the challenge of theoretical report. This novel strategy combines extractive info, with some items of knowledge of the article that are antecedently compressed or amalgamated. Specifically, during this article, they need to study: I whether or not COMPENDIUM produces sensible summaries within the medicine domain; ii) that report approach is a lot of suitable; and iii) the opinion of real users towards automatic summaries. Therefore, 2 forms of analysis were performed: quantitative and qualitative, for evaluating each the knowledge contained within the summaries, additionally because the user satisfaction. Results show that extractive and abstractive-oriented summaries perform equally as so much because the info they contain, thus each approaches are able to keep the relevant info of the supply documents, however the latter is a lot of applicable from an individual's perspective, once a user satisfaction assessment is dole out. This conjointly confirms the suitableness of their advised approach for generating summaries following an abstractive-oriented paradigm. [7]

Another paper is ―Extractive Text Summarization, By Namita Mittal, Basant Agarwal, Himanshu Mantri, Rahul Kumar Goyal and Manoj Kumar Jain"

Text summarization also helps in reducing the length of a text whereas conserving its info content. In this paper, a text report approach is projected based mostly on removal of redundant sentences. Initially, every sentence from original text (input) is scored supported what quantity redundant the sentence is and at what extent that sentence is in a position to hide different sentences by itself. This approach is best effective on the documents that square measure extremely redundant and contain repetitive opinions regarding a topic. The report takes places in 2 stages whereby the input of a stage is the output of previous stage and once every stage the output outline is a smaller amount redundant than the previous one.[8]

Another paper is ―AUTOMATIC TEXTS SUMMARIZATION: CURRENT STATE OF THE ART By Nabil ALAMI, Mohammed MEKNASSI ―

To facilitate the task of reading and looking info, it became necessary to realize a manner to scale back the size of documents while not affecting the content. The answer is in Automatic text account system, it allows, from an input text to result another smaller and additional condensed while not losing relevant data and that meaning sent by the original text. The analysis works carried out on this space have observed lately strong progress particularly in English language. However, researches in Arabic text account square are quite few and it still in their starting. During this paper they expose a literature review of recent techniques and works on automatic text summary.

Arabic as a vital language within the world has not been studied enough, and also the numbers of researches still few in Arabic natural language process .That is as a result of the advanced nature of Arabic language. Some of those reasons are, initial the totally different ways in which that bound mixtures of characters will be written. Second, the big selection of derivations and inflection of purposeful words makes the task of morphology analysis terribly advanced and complex. Third, Arabic words are usually ambiguous as a result of the tri-literal root system.[9]

3. Conclusion

The document summarization problem is a very important problem due to its impact on the information retrieval methods as well as on the efficiency of the decision making processes, and particularly in the age of Big Data Analysis. Though a good kind of text summarization techniques and algorithms are developed there's a requirement for developing new approaches to supply precise and reliable document summaries that may tolerate variations in document characteristics.

4. Acknowledgement

I take this opportunity to express a deep sense of gratitude towards my guide Mr. Chetan Kumar. I am thankful to all faculties especially Mr. Gopal Kumar, Assistant professor of Kautilya Institute of Technology and Engineering who taught the fundamental essential to undertake such a synopsis. Without their valuable guidance it would have been extremely difficult to grasp and visualize the synopsis.

References

[1] Vishal Gupta, "A Survey of Text Summarization Extractive Techniques", JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010.
[2] Josef Steinberger, "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation", Department of Computer Science and Engineering, Univerzitní 22, CZ-306 14 Plzeň
[3] Eduard Hovy and Chin-Yew Lin, "Automated Text

Summarization in SUMMARIST", Information Sciences Institute of the University of Southern California 4676 Admiralty Way Marina del Rey, CA 90292-6695, U.S.A

[4] Canasai Kruengkari and Chuleer at Jaruskulchai, "Generic Text Summarization Using Local and Global Properties of Sentences", Proceedings of the IEEE/WIC international Conference on Web Intelligence (WI‘03),2003.

[5] Morris, J. and G. Hirst –Lexical cohesion computed by thesaurus relations as an indicator of the structure of the text”. In Computational Linguistics, 18(1):pp21-45. 1991.

[6] Barzilay, Regina and Michael Elhadad –Using Lexical Chains for Text Summarization. in Proceedings of the Intelligent Scalable Text Summarization” Workshop.(ISTS‘97), ACL Madrid, 1997.

[7] Dr. Elena Lloret,Prof. Dr. M Teresa Romá-Ferri, COMPENDIUM: –A text summarization system for generating abstracts of research papers”, November 2013

[8] Namita Mittal, Basant Agarwal, Himanshu Mantri, Rahul Kumar Goyal and Manoj Kumar Jain, "Extractive Text Summarization", INPRESSCO ,2014

[9] Nabil ALAMI, Mohammed MEKNASSI , "AUTOMATIC TEXTS SUMMARIZATION: CURRENT STATE OF THE ART" , Journal of Asian Scientific Research, 2015