International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

# A Survey on Query based Automatic Text Summarization

**Payolina Nanda*[1], Dr.Ajit Kumar Nayak[2]**
[1]Scholar, Department of computer science and information technology, ITER, SOA University, Bhubaneswar, Odisha, India
[2]Head of the department, Department of computer science and information technology, ITER, SOA University, Bhubaneswar, Odisha, India

## ABSTRACT

Text summarization is an important problem in natural language processing (NLP). The process in which collection of crucial information takes place from an original document and representing its information in the form of a summary is known as Automatic Text Summarization [17]. We know the history has been an evidence where it is a tasking job for a human being to synopsize a bulk document and a time consuming job to create a summary from the document by considering the key points and the essence of the document. There are two genres of text summarization and it has been categorized as extractive method and abstractive method. Here in our study we will be mainly focusing on extractive text summarization based on a query defined by the user. The maximum inquiring problem in text summarization is to produce a brief text which is elucidative depending on the query given by the user. The problem here for query based text summarization has been plenteously researched and many techniques have been designed for its elucidation. But we need a path landing solution which will provide informative summary without containing any redundancy and ambiguity and which will produce a fluent, well-organised summary for a given query. An inspection which has been carried out here for query-based summarization approach with their accession for single and multi-document summarization, primarily basing on knowledge forms and machine level learning routines. Other than this there are different methods for choosing the highly correlated sentence from the source document with respect to a given query.

**Keywords:** Automatic text summarization, Query based extractive text summarization, Single and multi-document, graph based approach, machine learning approach, sentence coring method.

## I. INTRODUCTION

In natural language processing (NLP) automatic text summarization is a one of the major problem which shows that how a computer can understand, analyse and derive meaning from human language. Extraction of information contained in single document or multiple document is a very time consuming and difficult job for human being. So automatic text summarization can be a key solution for this problem. The goal is to reduce the information of the original large document into shorter version preserving the content and the overall meaning. This process involves the collection of crucial data from the original document and representing the document's record in the form of a compacted text.

Text summarization method has been partly divided into two types, Extractive text summarization and abstractive text summarization [20]. The way in which collection of crucial sentences and passages from the main document and concatenate them into brief explainable way is known as extractive text summarization. The sentences which are considered as important are chiefly basing upon analytical and syntactical features of the sentence. Whereas, an abstractive summarization is a process of grasping the key concept from the source document and represents the content in own natural language. Linguistic method is used in this process to scrutinize the text. Next step is to adapt that particular texts and discover the new meaning or the concept. Explanations here are provided in a best way by generating a blunt face of the text. This survey paper focuses on the hub for discussing of extractive text summarization methods.

To provide a significant extract based on the end user defined question is the most challenging problem for the arena of extractive text summarization. The goal is to design a question answering system which will provide a fluent, well efficient and organised summary for the given query. In the process of query based text summarization different summarization method and summarizer have been attempted to produce the summary. Here we will differentiate the summarization method for query based summarization and will find out the best summarizer that is used to produce the required summary.

## II. LITERATURE SURVEY

*Mariana damova and Ivan koychev, "Query-based Summarization: A survey*", in this paper the author has explained an outline of present-day scenario of query focused extractive text summarization methods and different entrance for single and multi-document text summarization. Knowledge-based and machine learning method has been taken in consideration for finding out the most appropriate sentences or phrases from the document regarding the given query. This paper is galvanized by following the essence of constructing e-book more knowledgeable, in particular designing the system for the acknowledgement of end users question in reduced time. [8]

*Wauter Bosma, "Query-based Summarization using Rhetorical Structure Theory*", in this paper the author has taken an effort to inform everyone, how the existing question answering system, aims at turning up for the needed answers to queries. These all are developed by making the use of summarization procedures to quote for more than just an answer which is presents in the document. Here a graph search algorithm is used to search for the relevant sentences. The graphical symbolization of the document, is where the output consists of an extensive answer, which is not only the answer of the question but also provides the user a chance to examine the perfection of the answer, this process is known as Rhetorical Structure Theory (RST). [1]

*John M. Conroy, Judith D. Schlesinger and Jade Goldstein Stewart, "CLASSY Query-based Multi-Document Summarization"*, this paper is posted on HMM (Hidden Markov Model) which is comprising of scripts from a document and making the use of a pivoted QR algorithm for the generation of a multi-document summary. The features used by HMM model have modified by the authors and linguistic capabilities have added to improve the summaries that is generated. The summarization system here is known as CLASSY (Clustering, Linguistic and Statistics for Summarization yield) which pre-processes each of the document by utilising word and phrase elimination techniques. The research in this paper prioritizes the process of query word generation from theme explanation along with the advanced experiments using named entity extraction. Therefore a break down study is presented using both Rough and pyramid scoring evaluation. [2]

*R.V.V. Murali Krishna and Ch. Satyananda Reddy, "A Sentence scoring method for extractive text summarization based on Natural language queries",* an apt here has been structured on how relevant it is for the use of traditional stoplists in the case of sentence scoring, an advanced method is considerably focuses on POS (part of speech tagging).

Combining the respective designs with semantic analysis has been outfitted with excellent outcomes which when is provided to extract the link between natural language queries and text in a document. [15]

*Ahmed A. Mohamed, "Improving Query-Based Summarization Using Document Graphs",* in this paper the author has extensively studied the number of approaches proposed in the literature for query based summarization and inspect for new procedures that resolves this issue and find out a new fix by using document graphs. Summaries here are based upon establishing the foremost object that is existing and concepts and their alliance in the text document. Here three different summarizer are carried out and comparison between the summarizers is performed to give the best summary for query based text summarization process. [16]

## III. QUERY-BASED SUMMARIZATION

Summarization can be divided into two methods:

- ✓ **Abstractive**: Abstractive automatic text summarization includes rewriting the text present in the original document into fewer words using own natural language.
- ✓ **Extractive**: Extractive automatic text summarization includes the collection of essential sentences, phrases and paragraphs from the source document and represents the information in the summary form. Each sentence in extractive text summarization is almost copied from the original document.

**Multi document summarization:** It includes the greater allowable information from a document set.
**Single document summarization:** Only a single document is used here.

Multi document summaries and single document summaries should deal with three main problems

- Coping and recollecting with redundancy;
- Identification of foremost differences among document set;
- Providing the summary relations;

**Query based summarization:** A query based summarization is a tailored system that suits the user defining information needs.
**Generic summarization**: Generic summarization focuses the scripter's communicative intent as guided by the source document.

**Summary construction method:**

- Abstractive summaries generates text from the main part of the document.
- Extractive summaries pinpoints the crucial part of the text and use them in the output summary as they are.

**Indicative summary:** It points to the information in the document, which helps to decide whether the document should be read or not by the user.
**Informative summary:** It provides all the related information to represents the aspect of the original document.

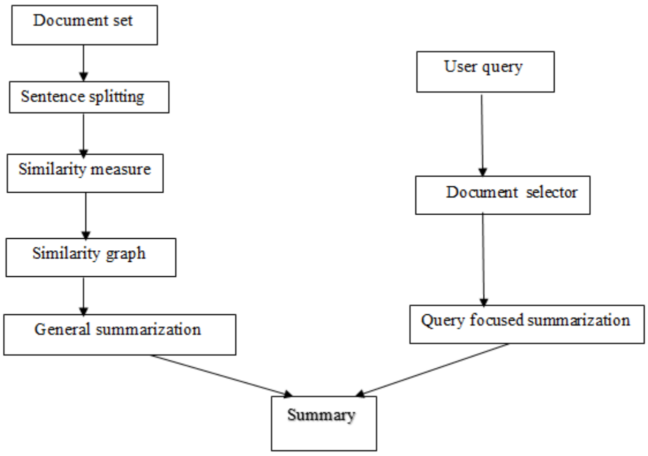

**Figure 1**. (Overall architecture of query based summarization method)

Query based techniques gives consideration to user preferences which can be formulated as a query. In query based summarization the summarization process will not consists of any sentences that are not under consideration in the original document. The question that is given by the individual is known as a query. The answer is always recognized by a question answering engine in the document. For query based summarization technique a pointer to an answer can be used as a framework. For an appropriate answer, a candidate sentence can only be involved to develop the extract if the link in between the candidate sentence and the answer sentence is well-known. This relation includes the statistical measure of text similarity. [8]

## IV. QUERY-BASED SUMMARIZATION APPROACHES

### 4.1. Sentence scoring approach

Short summaries contained in a documents are used by presently available web search engines to bring into view their results. Google creates document summaries using query based technique. Query words appearing in the documents are output together with some of their context. For the improvement of the performance of web search engine, the following system of summarization is proposed. The length of the summary provided in search result is increased first. The user would need to scroll too much if the longer summaries are displayed under the corresponding title. The title of each of the link are catalogued and the document's summary is shown in a parted frame when the cursor of the mouse is moved by the user on a specific link to avoid such problem. In this approach the summaries are limited to their size of the area of the screen that can be displayed without scrolling. This system achieves natural language processing methods for the purpose of summarization as well as widely used term-frequency statistics. The summarization algorithm is given in the following sections. [21]

Depending on the number of frequency counts of the terms that is (words or phrases), the sentences contained in the original source documents are scored in query focused summarization method. [21]. Highest scores are given to those sentences which includes the query phrases rather than the sentences which involves the single query word. Sentences having highest scores are involved in the output summary along with their structural framework. Some portion may be collected from various sections and subsections. Those collected sections are the composition of the output summary. The structural context of those sentences that is displayed which rely upon the frame size of the summary that is fixed to the screen size which can be viewed without scrolling. In the algorithm of sentence extraction, whenever a selected sentences that are to be included in output result, some of the headings are also collected in that context.

The query based sentence extraction algorithm is given below:

1. According to their score all the sentences will be ranked.
2. Main document title will be added to the summary.
3. The first heading will be added to the summary.
4. While (the summary size limit is not exceeded)
5. The next highest scored sentence will be added.
6. The structural context of the sentences will be added. (if any and not already included in the summary)
7. The highest level of heading will be 1added to the extracted text. ( calling this heading h)
8. The heading will be added before h in same level.
9. The steps 7, 8 and 9 will be repeated for the next highest level headings.
10. End while

A framework called GATE is used here. There are various benefits of this framework because it includes more often use of natural language processes like POS (part of speech tagging). In this GATE framework the Google API provided as a GATE plug-in is used for query processes and build the document corpus that contains the search result returned by Google for the query. GATE framework provides English tokenizer, sentence splitter and POS tagger that are to be used to divide the text into separate individual sentences and tokens. For finding out noun phrases and verb phrases GATE plug-in is used.

Based on an inclusive task based calculation, the system will be evaluated in forthcoming work. The analysis can further be continued in different direction. First of all the sentence scoring techniques along with query based method can be bettered and in second, the system can be improved by the consideration of various category of search tasks for example, searching for the background information about the topic.

### 4.2. Machine learning approach

In this approach to producing the summary, data recovery methods are united with the summarization methods [3]. The scoring of the sentences are based on some features from all sentences, the sentences that are given maximum scores, based on their features they are collected for final scoring of sentences and then evaluated by the use of a weighted linear combination of particular component value. The sentences with highest scores are chosen for the output summary until the length of the resulted summary reaches the desired limit.
The formula that is used to measure the scoring of sentences based on the distribution of the constituents words can be described in following two part:

**a.** Rank the document depending on the query
**b.** The specific idea of significance of a document.

For the ranking process the above mechanism allows the addition of demonstration of query independent framework. Different other methods have been used to generate query based summaries such as "SVM" [29], a machine learning approach, "LARS" (Least angle regression)[30], "Sumbasic" which acquires acceptable results by including only one characteristic that is the frequency of words in document clusters [31]. Here a feature analysis is provided by implementing LARS. FastSum [7] can depend on a least possible component set determined by LARS. 1250 news documents can be processed within 1 minute. When different summarization system are compared scalability is normally not considered.

Here fastSum is focuses on the selection of minimum set of characters which are less costly from others. Fastsum is based on word-frequency characters of the clusters, splitting of documents, filtering the main candidate sentence and calculating the word frequency in the description of the topic contained in the document and the title of the topic. A topic title and description of the topic are presented in the ranked topics. The topic contains a topic title and topic description. A lists of key word phrases that contains query words, describes the topic. Word-based features and sentence-based features are computed depending on the characters that contains the sentence position, length and probability of words for different domain. Fastsum mechanism can dependd on a minimum set of feature that leads to speedy processing of documents.

### 4.3. Approaches using linguistic

This approach is based on HMM (Hidden Markov Model) for selection of sentences from a set of document and an algorithm based on a system of question and answering mechanism to generate summary of multiple document [2]. A technique known as CLASSY (clustering, linguistic and statistics for summarization yield) is developed for use of linguistic methods and phrase elimination methods. Techniques, in this technique the sentences are processed with a part of speech tagger (POS). Here some full process elimination process is used along with phrase elimination process. [26]

The following eliminations were made:

- ✓ Ground clauses
- ✓ Restricted relative-clause appositives
- ✓ Intra sentential attribution
- ✓ Lead adverbs

The above elimination of phrases have been proven beneficial for applying on the full document, prior to summarization. Here the patterns are developed using "shallow parsing". For the detection of topic description and obtaining the capacity of question answering mechanism, there are some cues such as title paragraphs and different other phrases and paragraphs are pre-owned. An entity identifier which is a pre-processing step ran all over the document set and generate an entity list for different classification of objects like date, person, organization, location. It evaluates each topic description by searching for the keyword. For scoring the particular individual sentences HMM model is pre-owned after the generation of query terms and linguistic processes and to categorize them into non-summary and summary sentences. Here two approaches are used by HMM model. First one is associated with the token in each of the sentences, where a token is made up of a white space string in a document set. These tokens are identified using the log statistic [27] and used in the process of summarization according to Lin and Hovy [28]. This characteristics was arranged component wise to be mean and variance as zero and one respectively. In second approach the observations used by HMM is log. The query tokens are generated using topic description. For each of the document the observations is assign to be mean and variance as zero and one respectively. Hidden Markov Model is found to be more beneficial from the observations of query terms on given data.

A multi-document summarizer which make the use of query apprehension for the analysis of the user's profile that is given and narrates the topic for document clusters before the creation of a summary. This approach is based on basic elements, such as a head modifier relation and the representation of content included in the given document which is produced using a parser to generate a parse tree which is a set of "cutting rules" for the extraction of the genuine elements from the parsed tree [5]. Based on their basic valid elements the scores are assigned and then some techniques such as standard filtering and removal of redundancy mechanisms are used before the summary creation, in which the output contains the foremost sentences until the desired sentence level is reached.

### 4.4. Document graph approach

The document graph approach represents the method of an extractive multi document summarization, in which the document sets are represented as a graph [16]. The document graph is generated from a simple plain text documents by tokenizing then parse [11] it into NPs (noun phrase extraction method). The relations are generated by following some NPs heuristic rules.

For the search of the candidate sentence that is to be included in the output summary, to guide the summarizer for the searching of important sentences, from the source documents a graph is constructed which is known as a centric graph. Centric graph is an extremely leafy graph and contains a large number of leafy nodes. A leafy node represents an entity or concept node which is linked to other nodes in a centric graph. The foremost nodes in the centric graph are the leafy nodes and assists in the formation of summary of best features. This contain the core concept or entity which the document is converging. The relations in the centric graph consists of two parts, left side and right side and divided following a relation ("realted_to" or 'isa'). For the generation of the centric graph every relation's weights in all of the source graph has to be evaluated.

The algorithm of query based summarization method using document graph method is given as follows [9]:

- **a.** By tokenising, by following some heuristic rules and by parsing the document graph will be constructed.
- **b.** Centric graph will be constructed.
- **c.** The concepts presents in the query will be compared with the centric graph of the document.
- **d.** The document graph and the query graph will generate and then the similarity between each of the sentence and query will be calculated.
- **e.** The output summary will be generated by following the order of the best sentences in the input source document.
- **f.** By considering the graph of the selected sentences to the query, a query modification technique will be used here.

Here 3 types of summarizers are taken as consideration here such as Q summarizer, QInc summarizer and DGS summarizer. Q summarizer produces the best result among these. [16]

The technique used in [1] demonstrates the answers to the queries can be bettered by the extraction of a bit more information or data contained in the topic using summarization methods for the extraction of query based single document. To develop a graphical representation of the document, RST (Rhetorical Structure Theory) is used here. In a weighted graph each node represents a sentence where the weight of an edge represents the distance in between the two sentences. If a sentence is appropriate to answer to the query then the second sentence is calculated for the relevance depending on the weight of the path between two sentences. The sentence relations are described in a discource graph, then for the extraction of most relevant sentences from the summary of the graph an algorithm for graph search is considered here. Sentences with cheapest path from entry point are selected. RST itself applies to multimodal document without any extensive modifications, but it has to be further explored and developed for the generation of multimedia response.

## V. CONCLUSION

For the generation of a descriptive summary depends on a query that is entered by an end user is the biggest problem facing thing in the process of query based text summarization. This problem has been studied by many researchers and for its solution different techniques have been considered. According to different approaches for query-based text summarization methods, there are different summarization techniques and summarizers are considered to produce the most relevant summaries for single and multi-documents and user given queries. All the described systems are participated in DUC competition. According to DUC evaluation methods, the above proposed system are ranked based on their performance. Among them some scored highly top score such as CLASSY, FastSum and Q-summarizer and these systems produces better summaries from others.

Here in this survey we have presented an analysis of various query based summarization techniques that are performed in various applications. In our future work our aim is to consider various methods of query based summarization process and to find out the best method that will provide better summaries and prove to be the best technique.

## VI. REFERENCES

[1]. Wauter Bosma, "Query based Summarization using Rhetorical Structure Theory", 15th meeting of CLIN, 2005.

[2]. John M.Conoroy, Judith D.Schlesinger and Jade Goldstein Stewart, "CLASSY Query-Based Multi Document Summarization", DUC 05 conference, Boston, USA, 2005.

[3]. Jagadeesh J, Prasad pingali and Vasudeva Varma, "Capturing Sentences prior for Query-

based Multi-Document Summarization", RIAO, 2007.

[4]. Ahmed A.Mohamed and Sanguthevar Rajasekaran, "Query-Based Summarization Based on Document Graphs", IEEE International Symposium on signal processing and information technology, Canada, pp. 408-410, 2006.

[5]. Koychev I, Nikolov R and Dicheva D., "SmartBook: The New Generation e-book", Booksonline 09 workshop conjuction with ECDL, October 2, 2009.

[6]. Liang Zhou, Chin-yew and Edward Hovy, "Summarizing Answers for complicated questions", 5th international conference on language resources and evaluation (LREC), Genoa, Italy, 2006.

[7]. Frank schilder and Ravi kumar Kondadadi, "FastSum: Fast and accurate query-based multi-document summarization", 46th Computational linguistics, Columbus, Ohio, 2006.

[8]. Mariana damova and Ivan Koychev, "Query-Based Summarization: A survey", mathematics and informatics, Sofia, Bulgaria, 29 november, 2014.

[9]. A.A.Mohamed, "Generating User Focused Content Based Summaries for multi-Documents Using Document Graphs", 5th IEEE symposium on signal processing and information technology (ISSPIT), pp.675-679, 2005.

[10]. E.J.Santos, A.A.Mohamed and Q.Zhao, "Automatic evaluation of summaries using document graphs", ACL-04 workshop, Barcelona, Spain, pp-66-73, 2004.

[11]. D-Stetor and D.Temperly, "Parsing English with a Link Grammar", 3rd international workshop on parsing technology, pp.277-292, 1993.

[12]. E.J.Santos, H.Haguyen and S.M.Brown, "KAVANAH: An Active User Interface Information Retrieval Technology", Maebashi, Japan, 2001.

[13]. C.Y.Lin, "ROUGE: A package for Automatic Evaluation of Summaries", workshop on Text Summarization Braches out, Document understanding Conference, Barcelona, Spain, 2004.

[14]. Jagadeesh J, Prasad Pingali and Vasudeva varma, "Sentence extraction based single document summarization", Language technology research centre, Hyderbad, India, 2007.

[15]. R.V.V Murali Krishna and Ch.Satyananda Reddy, "A Sentence Scoring Method for Extractive Text Summarization based on Natural Language Queries", IJCSI (International journal of computer science issues, volume 9, issue 3, may 2012.

[16]. Ahmed.A.Mohamed, "Improving Query-Based Summarization Using Document Graph", ISSPIT (IEEE symposium of signal processing and information technology), August 2006.

[17]. Surajit Karmakar, Tanvi Lad and Hiten Chothani, "A Review paper on Extractive Text Summarization", International Research Journal of Computer Science (IRJCS), issue 1, volume 2, January 2015.

[18]. Raj Bardhan Oak, "Extractive Techniques for Automatic Document Summarization: A survey", International journal of innovative research in computer and communication engineering, volume 4, issue 3, March 2016.

[19]. Deepali K Gaikwad and C.Namrata mahender, "A Review Paper on Text Summarization", (IJARCCE) International Journal of Advanced Research in Computer aand Communication Engineering, volume 5, issue 3, March 2016.

[20]. Vishal Gupta and Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques", Journal of Emerging Technologies in web intelligence, volume 2, no.3, August 2010.

[21]. F. Canan Pembe and Tunga Gungor, "Automated query-based and Structure preserving text summarization on web

documents", International symposium on innovations in intelligent systems and Application, Istanbul, june 2007.

[22]. Richa Sharma and Prachi sharma, "A survey on extractive text summarization", International journal of advanced research in computer science and software engineering, volume 6, issue 4, April 2016.

[23]. S.Mohamed Saleem, R.Krithiga, S.K Ram, and S.Celin Sindhya, "STUDY ON TEXT SUMMARIZATION USING EXTRACTIVE METHODS", International journal of science engineering and technology research (IISETR), Volume 4, Issue 4, may 2015.

[24]. Sitanath Biswas, Sweta Acharya and Sujata Dash, "Automatic text summarization for Odiya language", International journal of computer applications, volume 132, December 2015.

[25]. Y.Chali, "Generic and query-based text summarization using lexical cohesion", Advances in Artificial intelligence: 15th conference of the Canadian society for computational studies of intelligence, pp 293-302, 2002.

[26]. D.M Dunlavy, J.M Conroy, J.D Schlesinger, S.A Goodman, M.E Okurowski and H.Van Halteren, "Performance of a three stage system for multi-document summarization", in DUC 03 Conference proceedings, 2003.

[27]. T.Dunning, "Accurate methods for statistics of surprise and coincidence", computational linguistic, pp. 61-74, 1993.

[28]. chin-yew Lin and Edward Hovy, "The automated acquisition of topic signatures for text summarization, in 18th conference on computational linguistics, pp. 495-501, 2000.

[29]. Li, Y. Ougyang, W.Wang and B.sun, "multi-document summarization using support vector regression", in proceedings of DUC, 2007.

[30]. Efron, T. Hastic, J.M. Johnstone and R.Tibshirani, "Least angle regression", Annals of statistics, pp-407-499, 2004.

[31]. Wenkova and L.Vanderwende, "The impact of frequency on summarization", in MSR-TR, 2005.