Automatic Text Summarization: The Current State of the art

S. SUNEETHA

Research Scholar,

Department of Computer Science and Engineering,

JNTU, Hyderabad

Abstract: The increasing availability of online information has necessitated intensive research in the area of automatic text summarization within the Natural Language Processing (NLP) community. Over the past half a century, the problem has been addressed from many different perspectives, in varying domains and using various paradigms. This literature review intends to observe the current state of the art in text categorization approaches both in the areas of single-document and multiple document summarizations, giving special emphasis to empirical methods and extractive techniques. Some promising approaches that concentrate on specific details of the summarization problem are also discussed. Special attention is devoted to automatic evaluation of summarization systems, as future research on summarization is strongly dependent on progress in this area.

Keywords: Text summarization, natural language processing, taxonomy; Automatic Abstracting, semantic relationship significance, extraction.

Introduction:

The subfield of summarization has been investigated by the NLP community for nearly the last half century. Radev et al [4] define a summary as a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that". This simple definition captures three important aspects that characterize research on automatic summarization:

- Summaries may be produced from a single document or multiple documents,
- Summaries should preserve important information,
- Summaries should be short.

Even if we agree unanimously on these points, it seems from the literature that any attempt to provide a more elaborate definition for the task would result in disagreement within the community. In fact, many approaches differ on the manner of their problem formulations. We start by introducing some common terms in the summarization dialect: extraction is the procedure of identifying important sections of the text and producing them verbatim; abstraction aims to produce important material in a new way; fusion combines extracted parts coherently; and compression aims to throw out unimportant sections of the text [4].

Earliest instances of research on summarizing scientific documents proposed paradigms for extracting salient sentences from text using features like word and phrase frequency [5], position in the text [6] and key phrases [7]. Various work published since then has concentrated on other domains, mostly on newswire data. Many approaches addressed the problem by building systems depending of the type of the required summary. While extractive summarization is mainly concerned with what the summary content should be, usually relying solely on extraction of sentences, abstractive summarization puts strong emphasis on the form, aiming to produce a grammatical summary; this usually requires advanced language generation techniques. In a paradigm more tuned to information retrieval (IR), one can also consider topic-driven summarization, that assumes that the summary content depends on the preference of the user and can be accessed via a query, making the final summary focused on a particular topic. A crucial issue that will certainly drive future research on summarization is evaluation. During the last fifteen years, many system evaluation competitions like TREC[1] DUC[2] and MUC[3] have created sets of training material and have established baselines for performance levels. However, a universal strategy to evaluate summarization systems is still absent.

Taxonomy of Automatic Text Summarization Techniques:

The Automatic Text Summarizations can be classified into single document text summarization and multi document text summarization(Fig 1).

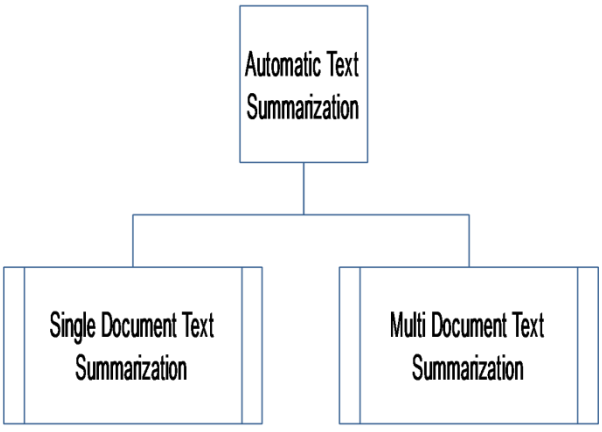

Fig 1: Automatic Text Summarization models

Single-Document Summarization models

Usually, the flow of information in a given document is not uniform, which means that some parts are more important than others. The major challenge in summarization lies in distinguishing the more informative parts of a document from the less ones. Though there have been instances of research describing the automatic creation of abstracts, most work presented in the literature relies on verbatim extraction of sentences to address the problem of single-document summarization.

Automatic Text Summarization methodologies for single document (Fig 2) are listed below:

Naive-Bayes Methods [8]: The classification function categorizes each sentence as worthy of extraction or not, using a naive-Bayes classifier.

Rich Features and Decision Trees [9]: The importance of a single feature, sentence position claimed in []. Just weighing a sentence by its position in text, which termed as the "position method", arises from the idea that texts generally follow a predictable discourse structure, and that the sentences of greater topic centrality tend to occur in certain specifiable locations.

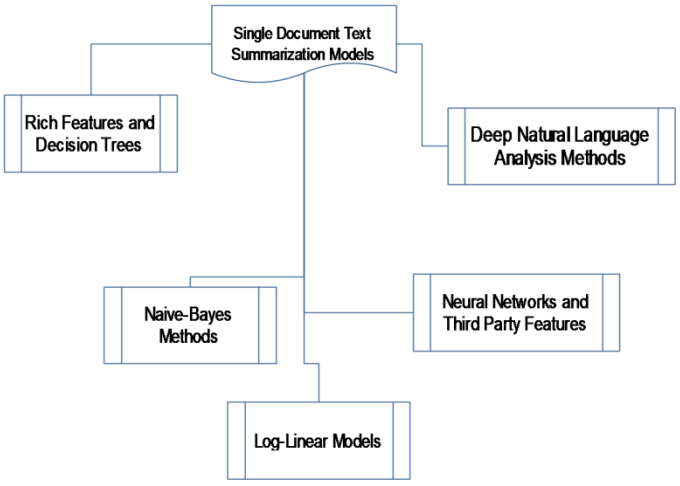

Fig 2: Single Document Text Summarization methodologies

Hidden Markov Models [10]: In contrast with previous approaches, that were mostly feature-based and non sequential, Conroy et al [] modeled the problem of extracting a sentence from a document using a hidden Markov model (HMM).

Log Linear Models [11]: Osborne [11] claims that existing approaches to summarization have always assumed feature independence. The author used log-linear models to obviate this assumption and showed empirically that the system produced better extracts than a naive-Bayes model, with a prior appended to both models.

Neural Networks and Third Party Features [12]: Neural nets and the use of third party datasets to tackle the problem of extractive summarization, outperforming the baseline with statistical significance.

Deep Natural Language Analysis Methods [13]: The main approach is using a set of heuristics to create document extracts. Most of these techniques try to model the text's discourse structure.

Multi-Document Text Summarization Models

Extraction of a single summary from multiple documents has gained interest since mid 1990s, most applications being in the domain of news articles. Several Web based news clustering systems were inspired by research on multi-document summarization, for example Google News[14], Columbia NewsBlaster [14] or News In Essence[16]. This departs from single-document summarization since the problem involves multiple sources of information that overlap and supplement each other, being contradictory at occasions. So the key tasks are not only identifying and coping with redundancy across documents, but also recognizing novelty and ensuring that the final summary is both coherent and complete.

Automatic Text Summarization methodologies (Fig 3) for Multi documents are listed below:

Abstraction and Information Fusion [17, 18]: A full multi-document summarizer is built by concatenating the two systems, first processing full text as input and filling template slots, and then synthesizing a summary from the extracted information.

Topic-driven Summarization and MMR [19]: The idea is to combine query relevance with information novelty; it may be applicable in several tasks ranging from text retrieval to topic-driven summarization. The maximal marginal relevance (MMR) is simultaneously rewards relevant sentences and penalizes redundant ones by considering a linear combination of two similarity measures.

Graph Spreading Activation [20]: The topic is represented through a set of entry nodes in the graph. A document is represented as a graph as each node represents the occurrence of a single word (i.e., one word together with its position in the text). Each node can have several kinds of links that are adjacency links (ADJ) to adjacent words in the text, Same links to other occurrences of the same word, and Alpha links encoding semantic relationships captured through Wordnet and NetOwl. Besides these, Phrase links tie together sequences of adjacent nodes which belong to the same

phrase, and Name and Coref links stand for co-referential name occurrences.

Centroid-based Summarization [21]: The first stage consists of topic detection, whose goal is to group together news articles that describe the same event. To accomplish this task, an agglomerative clustering algorithm is used that operates over the TF-IDF vector representations of the documents, successively adding documents to clusters and recomputing the centroids. Centroids can thus be regarded as pseudo-documents that include those words whose TFIDF scores are above a threshold in the documents that constitute the cluster. Each event cluster is a collection of (typically 2 to 10) news articles from multiple sources, chronologically ordered, describing an event as it develops over time. The second stage uses the centroids to identify sentences in each cluster that are central to the topic of the entire cluster.

Multilingual Multi-document Summarization [22]: Multilingual summarization is still at an early stage, but this framework looks quite useful for newswire applications that need to combine information from foreign news agencies. This framework consideres a preferred language in which the summary is to be written, and multiple documents in the preferred and in foreign languages are available. A translation system is first applied to translate the documents in foreign language to prefered. Then a search is made, for each translated text unit, to see whether there is a similar sentence or not in the documents in prefered language. If so, and if the sentence is found relevant enough to be included in the summary, the similar prefered language sentence is included instead of the foriegn-to-prefered translation.

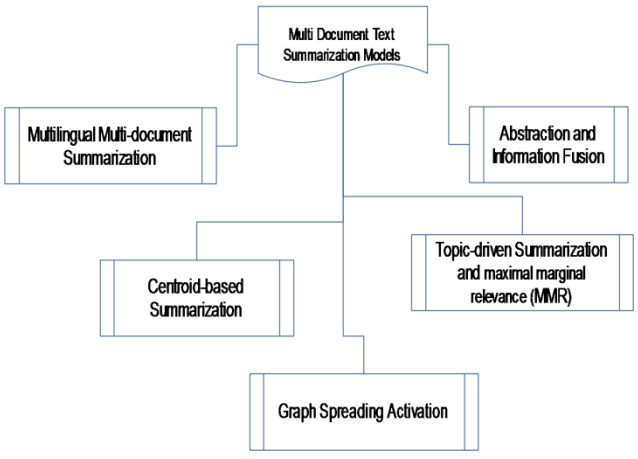

Fig 3: Multi Document Text Summarization methodologies

Current State of the Art

Algebraic Reduction in Automatic Text Summarization: Now-a-days data on whatever subject is widely available on various electronic media mainly internet. The count of different electronic media is increasing and the subject to browse is also varying. Due to this reason, the data provided by these media consists most of the unwanted matter regardless of the concerned subject. This problem of impeding unwanted data into the concerned data is referred to as "Information Overload". To solve this problem, Automatic text summarization methodology is introduced. This methodology concerns with the process of creating a short-hand edition of the concerned subject from the provided data.

In this context, many algebraic reduction techniques are considered to recognize required data and to gain only semantically needed data automatically from the provided text. This paper concerns with the ground work behind the authoritative techniques for automatic text summarization already proposed. Mainly the work on Non-negative Matrix Factorization (NMF) and Singular Value Decomposition (SVD) are considered. Moreover the real time applicability of these techniques is clearly explained in this paper. Apart from these, the merits and demerits of these techniques are also focused in this context.

The process of data reduction including identifying the concerned data and extracting it is primarily used when it is to be added to a summary. Summary is an extract from the original document to symbolize only the significant data within the document. As the summary needs to be short, simple and efficient covering all the main topics of the subject, it should be created effectively. But the data now available on electronic media are digital data, due to which they increase the quantity of overhead over the provided data. This overhead creates degrade in the quality of the data as it creates deprivation of the capability to take correct decisions, important deeds and data about the method.

Automatic Text Summarization concerns with this overhead and extracts significant data from the mixed data by straining out overhead. This process will creates an abbreviate edition of the original data. In this context large dimensional data is converted to small dimensional data while at the same time maintaining the significant data within the summary. There are many techniques that can be used to summarize the text one of which is algebraic technique. It is preferred much due to its capability to figure out the required data and its efficiency to calculate.

Nowshath Kadhar Batcha et al[23] presented a paper on "Algebraic Reduction in Automatic Text Summarization – The State of the Art", which mainly concerns with the process of deducting a whole set of wanted and unwanted data into required data with the use of some algebraic methods.

There are many considerable factors based on which summarization of the text can be performed. Some of them are: requirements of the real time applications, type of the text,

format followed in creating text, the size needed etc… Apart from these there are certain parameters based on which the summarization is done. Those are: semantic data and synchronization. Based on the range to scan the document to derive summary, automatic txt summarization is divided into 2 types. They are: shallow method and deeper method.

Shallow method is the technique which concerns only with the upper level of the document considering the key terms and their usage ratios. It is again divided into linguistic relied method and statistic relied method. Linguistic method takes into account language parameters such as words, phrases and lexical chains. Lexical chain is constructing a chain considering semantically connected words or phrases. The choice of the content to be added to the lexical chain plays a vital role here. In this paper the work over the algorithms to construct lexical chains and their utilization are mentioned. Statistical method considers the "term frequency" technique. Term frequency refers to the count of a term within the document. This count decides the eligibility to be within the summary or not. Deeper method deals with the process of creating summary using natural language. It is the normal process of creating the summary involving deeper considerations. In this method, the document is examined and the semantic connections between the terms are accounted.

Apart from authentic techniques, there are summarization techniques relied on algebraic deduction. They are: SVD relied method and NMF relied method. SVD is introduced to overcome the drawbacks of Vector Space Model (VSM) such as linguistic noise, inefficiency to identify semantic data. It is an average technique utilized by Latent Semantic Analysis to overcome those drawbacks. NMF relied methods are introduced to overcome the drawbacks of SVD such as the probability of negative values appearance.

Observation: The proposed subject mainly deals with the ground work done by the authentic algorithms. Also the current trends in automatic text reduction and how the authentic algorithms are modified to real time application algorithms are specified. Applications of SVD and NMF algorithms, also how the drawbacks of SVD is overcome by NMF are pointed in this paper. Most of all future work can be done by the hybrid model of NMF by including anaphora resolving along with the parameters like priority, initialization, looping and tolerance etc., will produce excellent outcomes making this technique much more efficient.

An Extractive Text Summarization Based On Multivariate Approach: Esther Hannah. M et al [24] addressed a method to automatically summarize a text with the help of multivariate statistical technique, where multivariate is a form of statistics encompassing the simultaneous observation and analysis of more than one statistical variable. The model they proposed a training methodology where the system trained by using manual summaries. The utilization of multivariate statistical technique for this task is justified by its ability to produce a model that resembles a relation. The model relied on primary subjective evaluation, in order to show that the approach is effective, efficient and promising

The paper has introduced the statistical approach to extractive text summarization where multivariate is used to generate the weight for every sentence. The texts are ranked to classify them as summarized or not. The steps followed by the authors in the extraction are as follows, they are:-

I. To bring out the early work done on the text summarization focusing on the contributions that laid the foundation for the research in this subfield of NLP.
II. To discuss the proposed work under the various subsections namely, pre-processing, feature extraction, comparison vector generation, weight generation and ranking.
III. Evaluation method has been used by the authors and the results are provided
IV. Conclusion of paper with providing scope for future work.

In the first step the authors discuss about some works that were previously in practise like MEAD (a state of sentence extractor) in DUC and some other computationally expensive extractions including NLP based methods where as the present system is much comfortable and cheap to get the result.

While coming to the next step, some probable subparts are introduced and the text is modeled by using two-phase classification "in" & "out" otherwise Boolean values '0' & '1' are assigned to the marked sentences respectively . The first subpart called Pre-processing is detailed by dividing it into four segments namely sentence segmentation, tokenization, stop word removal and word streaming and was explained by a system architecture figure.

Specifically the other subparts are implemented with the formulae to get the results. There are six formulae for subparts, each consisting one formula and thus are derived the six scores from these formulae depending on the keywords, number of articles, length, number of numerical data and the summation.

The feature subtraction part is derived from the sentence similarities, numerical values are derived from the numerical data, sentence relative strength from the number of articles and node similarities from the summation by using these formulae.

Then the author uses the compression vector generation to check whether the sentence matches the summary or not by using the in-out classification and selected sentences is weighted by using the weight generation technique through which the ranks are assigned in order to decide which sentence should be first and which one is last. Multiple linear

aggressions is a multivariate statistical technique, which examine the linear correlations between sentences & variables has been used in weight generation technique. The ranks are decided based on their weights and compression vector considering a formula.

This paper presents the work on evaluation in two methods namely intrinsic and extrinsic. Intrinsic mainly assess coherence and summaries while extrinsic assess impact of summarization.

The simulation result is obtained in final score which is derived by multiplying the each score with their respective weighted value obtained and then adding all the product values. To analyze the process the authors verified the system by considering 60 documents in which 30 are for training the system and 30 are for the testing the system. Precision, which is the average value of the documents (in percentage) is made and got the comparison with Microsoft word documents and presented in tabular forms. The comparison gives the assessment of the system by verifying how many documents are produced by either system.

Observation: Though the authors got the assumed results, At last the exact summarization is not given for those documents with low precision value, that is for less sized documents the summarization is big than needed and which could be verified by working on semantics way. This is the limitation of this proposal, which could be rectified further.

Document Relevance Identifying and Its Effect in Query-Focused Text Summarization: Tingting He et al[25] discussed about the impact of document relevance identification in query focused text summarization that focused on putting a spotlight on a crucial issue that specifies text summarization to be efficient enough to personify private matter and supplies analytic message to the end-users. User-feedback information is made use of to obtain appropriate records and also transductive inference SVM machine is also showcased. This above mentioned methodology can very well shun the prejudice of selecting the appropriate records aptly. Additionally, a sentence assortment policy through mining keywords is recommended. It can be computed through the usage of word co-occurrence window. Likelihood ratio is estimated via concerned characteristics and both the results are congregated to compute candidate sentences. The propositions cited are efficient to confine designs of a document set and comply with the query demands in a successful way. The new trend that is setting its sights on query-focused summarization is a unique scheme that is relative to current automatic summarization and information retrieval domain. It is also particularly proficient in accessing capable information from the unavailable data and is effective in sharing and analyzing data successfully. The methodology earlier specified first deals with the usage of a search engine that is relatively helpful in acquiring a prefaced yet ordered record list and involves selecting few related and unrelated records and labeling them. Then SVM ie. Support Vector Learning Method is applied for labeling the significance of the remaining records. Based on this, a sentence assortment policy is put to use with the of word co-occurrence scheme and likelihood ratio. The ultimate synopsis is then achieved by eliminating redundancy through Maximal Marginal Relevance (MMR). The record list that is chosen is segregated into 3 different types.

1. The labeled records that are related with the query,
2. Those labeled records that are unrelated with the query
3. The records that are not labeled at all.
4.

Identification of mysterious samples is attained by instituting set of methods and rules which can be accomplished through transductive inference training. There is also the incidence of a training program that houses 4 steps. Constraints are specified using the inductive method, an initial classification model is introduced, output of sign function of every mysterious sample is estimated, all samples are re-trained, mysterious sample pairs are explored which are labeled. Keywords are mined to relate itself to the query specified and the main gist of the record set is indicated. A matrix is assembled to witness the pertinent degree of the sample words with the expertise of word co-occurrence window. A signature term is derived by a likelihood ratio which is asymptotical in nature.

The query-related attribute and topic related attribute are calculated by combining together to estimate each word. If a sentence has a maximum number of keywords and if in case, they are vital, then they are required to obtain advanced scores and gets embraced in the final summary. After the crucial task of scoring sentences is done with, imperative sentences with premier score and least redundancy is mined and re-organized into the final summary. It is performed in 4 phases: 1. relevant documents acquiring; 2. keyword extracting; 3. sentences selection; 4. summarization generation. The summarization generation guarantees coherence of the summary. A record is then selected by the system that holds largest number of summary sentences as a frame of reference. Various experiments are performed with a data set that houses 10 query topics, a total of 400 texts which restored from the search engine. The estimation of the topic related attribute will be affected badly if it is not concerned with the query that is present in the relevant document set. As such, the degrees of few crucial words are decreased. Similarly, the result of employing an unrelated record deteriorates the set of records that have greater significance. Hence, it can be concluded that the author has recommended a sharp method of deriving keywords with the aid of two characteristic methodologies which are mainly user feedback information and transductive inference SVM machine learning technology where SVM stands for Support Vendor Machine.

The author has also offered another line for sentence selection strategy by means of extorting keywords. Subjectivity is eluded by appropriate records attainment and at

the same time, query can be insisted by gratifying and confining the main theme of the record by usage of sentence selection strategy. The experimental analysis proves that the primary labeled number is influenced by the exactness of related records. The exactness improves in reaction when the labeled number is augmented by steps.

Observation: Definite efforts need to be kept to analyze various algorithms to increase the most favorable value. Need to conduct experiments for mechanically settling on deciding the number of keywords required in the keywords extracting methodology. So that only the final summary can be analyzed, coherence and readability are not bothered with. The valuation techniques and analysis need to be concluded with of experiments on better text corpora.

Evaluation method of automatic summarization calculating the similarity of text based on HowNet: SUO Hong-guang et al[26] presented a better evaluation method for the automatic text summarization (ATS) which is based on identifying the equalities of the summary and the actual text document. Automatic text summarization refers to the process of reducing the quantity of the text preserving the actual content of the document. Once a summary is created, an evaluation technique is applied over the summary to validate it i.e. whether it is compatible to the actual text or not. There are many evaluation techniques available now-a-days. This paper proposes one of the evaluation techniques far better than the authentic and old fashioned evaluation techniques. This technique is relied over vector space model and it does the process of evaluating a summary relying on HowNet.

The proposed technique is introduced to provide a precise and effective alternative to the available automatic summarization algorithms. Here the techniques of various kinds of current automatic summarization algorithms along with their drawbacks are explained. Also, the benefits of the proposed technique over the available ones are clearly specified. This technique utilizes the HowNet in the vector space model to examine the actual content of the document. Also, this technique considers the parts of speech and further grammar which can be assumed to affect the meaning of the sentence. This analysis plays a key role in computing the priority of the terms to be included in the summary of the document. In this paper, this technique is proved to be better than PIR.

Now-a-days, there is a huge increase in the number of electronic media and the vast data provided in them. Due to this reason the data available on a particular subject may contain lot of overhead leading to us to move away from the actual content. Thus to solve this problem automatic text summarization technique is introduced. But, as this summary refers to the whole document, this summary must be a better reflective to the original document. Hence, to provide us with a better summary, evaluation method over the summarization technique is needed to perform.

The evaluation technique can be done in any of the two ways. Such as: exterior evaluation technique and interior evaluation technique. The exterior technique refers to evaluating the Automatic Text Summarization algorithm followed and how the summary, created, will act as in a document. The interior technique concerns only with the quality of the summary created. But both of these approaches have drawbacks. Exterior needs a lot of time and manpower where as interior faces the problem that ideal summary is impossible. Also another problem faced by interior technique is P/R defects i.e. the length of the line in the summary to the length of the line in the document ratio.

The proposed technique helps to overcome these drawbacks and provide a better evaluation technique. Many scientists use the interior technique in their summarization techniques i.e. to maintain the quality, efficiency, performance and consistency. Consistency refers to the extent to which the summary is fluent enough in the meaning and overall structure. The evaluation of the ATS includes four steps which are needed to be followed. They are: prior processing of summary, grasping conceptual characteristic term, computing the priority of characteristic term and comparing summary with original document.

The prior processing step includes eliminating spaces, missing words and stop words. Apart from these, the parts of speech of the words must be noted before proceeding to the summarization process. In the step of grasping conceptual characteristic term, the term is grasped using HowNet. This includes highest similarities with the preservation of the semantic data. In the step of computing priority of characteristic term, the priorities of the terms are calculated based on the number of times they appear in the document. For this purpose, it utilizes TF-IDF technique. In the step of comparing summary with the original document, Vector space model is utilized to find the similarities between the summary and the document.

Many experiments are conducted to prove that the proposed technique is better than the existing techniques. Two of them specified in this paper are: to compare 6 different types of word segmentation systems and evaluation outcomes of 3 various evaluation techniques. Thus, in this paper, the drawbacks of the existing techniques are specified along with their explanation. Also, it is shown how the proposed system overcomes their drawbacks and thus proved to be better. The proposed evaluation technique is based on Vector space model. It compares summary and document and to extract semantic data using the HowNet. This technique provided a better way to compute the priorities of terms used in the document and to decide which one to place in the summary.

Observation: However, this technique faces drawbacks such as: Difficulty to calculate the priority of the terms which is to be worked on in the future. But, the proposed system is proved to be simple, precise, efficient and better than the authentic techniques. Also, this technique gives better outcomes.

Evaluation of the Impact of User-Cognitive Styles on the Assessment of Text Summarization: Hien nguyen et al[27] presented a work which utilizes text-summarization

techniques in order to facilitate the users to find significant information quickly. Also since, the summary of a data in the user's preference can heavily influence the effectiveness and efficiency of a user's performance in an information finding task, the work of authors meets the important need to design a personalized text summarization system, which can take into consideration not only in what a user is urgently interested but also how a user perceives information. The later factor is referred as user's cognitive style in the paper. This proposed model specifically aims at studying the input of a user's cognitive styles when assessing multi content summaries. The authors have chosen two parameter's of a user's cognitive style as,

i. the analytic/wholist,
ii. Verbal/imagery dimension.

The study of the impacts has been done through the assessment of summary which was generated from a set of documents. In this paper type of document set refers to whether the set's content is loosely or closely related to the user's preference, the authors have utilized the document set type to explore the differences in the user's assessment of summaries that they generated from sets of different type. Moreover, the authors have explored the aspect of coherency ratings that were given to summaries from the two types of documents set which were considerably different from the analytic and wholist groups. The authors have further chosen two parameters which are directly related to the coherence of summary, the two parameters are: -

I. Graph Entropy
II. The percentage of standalone concepts.

Through this paper the authors have reported that these two factors and user's cognitive styles affect the user's rating on the coherency of a summary.

The authors have introduced text-summarization as an effective technique that is often used in combination with information retrieval and information-filtering application to help users save time finding critically relevant information and making timely decision. Text-summarization is being defined in the words of MANI and MAYBURY as the process of distilling the most important information from a source to produce an average version for a particular user and task.

For any text summarization system the two primary inputs as indentified by the authors are: -

I. Document: as the content of document are used to produce the key point of the summary.
II. Users; as users' interest and preference are used to determine which point should be included and how they should be provided.

Users-modeling technique is widely appreciated for their contribution in improving the user's efficiency and performance on information-seeking task. The user modeling research is based on the idea that each individual has different ideas and interest on what should be considered as key point.

Observation: The finding of this model has potential to improve summarization algorithms which can perform a better job of recognizing key information from collections. The authors have also reported that the wholist/analytic give significantly different coherency rating to summaries that where generated from the different types of document sets. Further the connectivity of document measured by graph entropy and the percentage of standalone concepts together with a user's cognitive style affected a user's coherency ratings. The result can be used to design a user centered text summarization system.

Post-Processing of Automatic Text Summarization for Domain-Specific Documents: Zengmin Geng et al [28] proposed a model called "Post-Processing of the Automatic Text Summarization for Domain-Specific Documents" that deals with the part of processing that is to be done once the automatic text summarization of a document is completed. Post processing of the automatic text summarization refers to the process done to reduce the ambiguity rising due to the inefficiency of the authentic text summarization methodologies to clearly refer to the actual meaning of the domain-specific files. Thus, post processing is done posterior to the text summarization to improve the quality of the summary produced.

The post processing of the automatic text summarization includes various methodologies followed to serve the purpose. They are: To remove repetitions, to avoid ambiguity, to align the whole file in groups of texts and formatting them and analyzing the lines written in the summary. Also it computes a knowledge base and uses it as the reference to substitute for the empty or left values with the domain-specific information. Thus, after examining and many trials over clothing field files, CTS is introduced. CTS is a software created to improve the process of Automatic Text Summarization. Moreover it is proved to be better than Microsoft Word 2003 in terms of performance, quality and preservation of the actual meaning. This software is compatible to be installed in mobiles, PCs, and Explorer web pages depending on the domain concerned.

Many algorithms to summarize text are already introduced into the market like DCM (Document Clustering Method), LSA (Latent Semantic Analysis), KFS (Key Frequency Statistic), MMR etc… All these techniques are based on the quantity basis and thus are termed as statistic techniques. These all techniques just count the frequency of the term appearing in the document and based on it, they give them priorities. The terms with higher priority are mentioned in the summary. Due to this process, the important context of the document may lose due to its lower priority.

Thus, to preserve the actual meaning of the document, a summary creation technique is developed called CTS (Clothing Text Summary). This algorithm deals with

examining the Automatic Text Summarization and develops summary relating to the clothing field. As mentioned earlier, this algorithm computes a knowledge base. The knowledge base is computed as a group of different data structures such as: Professional feature words database, Thesaurus database, Hierarchical class tree, domain expert and organization name database, special symbol database and relation database with other realms.

In professional feature words database, the words are counted based on their appearance and those are brought under the professional's vision to deal with them. Thesaurus database refer to reducing the number of different words and their count based on their meaning i.e. by substituting the words of the same meaning with a single word. In hierarchical class tree, a tree is constructed based on the generality of usage. It is the most significant data structure used in the construction of knowledge base. Domain expert and organization name database is concerned with the exclusive domains. Special symbol database refers to storing special symbols appearing in the document. Such special symbols are denoted in XML style. Relation database with other realms refers to storing the realms along with their connected realms. It is used to develop the quality of the ATS.

The summary is developed based on three major steps. Such as: prior processing, coarse summary and posterior processing. Many characteristics of a line are specified in this paper and with the process to compute their priority. Post processing includes removing the redundancy due to the repetition of the words. This is done by computing a threshold value and thus comparing lines in the document. Also text is formatted by grouping based on their priorities and an algorithm. Also line generalization is done which is a process of converting high standard data into an understandable version. Here, the process of substituting empty or left out fields is done by using some significant and basic words. Also the efficiency of the CTS is proved by considering 2000 documents and testing them under the supervision of 5 professionals.

Observation: Thus, this paper focuses on a posterior processing of the automatic text summarizing which apart from process of concise, also deals with improving the quality of the summary. It mainly deals with the algorithm called CTS and proving its efficiency over authentic algorithms. This algorithm follows many techniques for posterior processing. Also it is proven that CTS is far better than the Microsoft Word 2003. And due to its profitable characteristics, it is planned to implement this technique in the future projects and to extend CTS in various other fields.

The Design and Implementation of Domain-specific Text Summarization System based on Co-reference Resolution Algorithm: Shi Ziyan et al [29] proposed a text summarization model that concentrates on the concept of automatic abstracting which now-a-days is popular and well in demand. To gain advantage from the immense amount of information that is now available these days, automatic abstracting came into existence. Customary abstracting techniques have now been replaced by the statistic schemes and word-list in domain detailed field for advancement of domain detailed text summarization system. Referring that occurs during management of text, frequently leads to imprecise products. Co-reference resolution algorithm is brought into existence to resolve the issues that are caused by referring. By reanalyzing the imprecise products caused b referring in the actual text, term incidence and sentence magnitude are redesigned so as to obtain new abstracting conclusions. Summarization arrangement is classified into two classes mainly Linguistic approaches and Statistical approaches. The basic difference between the above mentioned two approaches is that, that the linguistic approach utilizes words, phrase and clause structural information making use of the linguistic reserves while the statistical approach makes use of title, frequency, location etc.

The drawbacks in the text summarization are feature sparseness and low performance. To decipher the disadvantages caused by text summarization, there is an implementation of co-reference resolution algorithm for extending domain detailed text summarization systems. Domain detailed text summarization systems with co-reference resolution algorithm has two prime elements, mainly original text summarization generating part, final result obtaining part. In the actual abstract receiving component, earlier undone work is accomplished, examined and the actual outcome is acknowledged. There are a total of 3 components that are available in basic text summarization system precisely, previous work, analyzing the article by statistical method and obtaining final result. For civilizing competence of a typical system, co-reference resolution is supplemented in a domain detailed text summarization system.

The first step involved is preprocessing, and then comes computation of term weight and sentence weight. The procedure that is to be followed while contemplating co-reference resolution is: Finding out the co-reference words, defining entity word set for every co-reference word that is present in the list, ruling out the entity words submitted by every co-reference words present in the entity words set of the respective co-reference word, replace the entity word for the co-reference word. The process of co-reference resolution algorithm is explained by Ziyan in a very innovative way. The co-referent in the route is outlined by a co-reference word list. There is a requirement of a term list in a domestic-concentrated region that can be made use of in the procedure of finding co-reference resolution. The entity words stated in the co-referent channel is established by assuming term list. A co-reference resolution algorithm is also prepared which enlist and encompasses all the issues mentioned of major concern in statistical abstraction of handling and contemplating text.

The assessment of an abstract system can be done with the help of 2 methods: experts evaluation method and automatic evaluation method is approved so as to authenticate probability and efficacy of the scheme that is presented in the paper. The expert evaluation method is helpful in analyzing

recall, precision and F-measure. The automatic evaluation method is responsible for computing resemblance between the abstracted text and the actual text given for the same. Both these method are effective yet efficient in their own ways. Recall component can be computed by the ratio of the number of sentences that are selected by both the system and the experts to the number of sentences that are selected by the experts. Similarly, precision can be analyzed by the ratio of the number of sentences that are selected by both the system and the experts to the number of sentences selected by the system.

Observation: There is an obligation for appraising the enhancing of our system which would require a typical abstract system. If a section houses two systems, then the first system is responsible for standard text summarization while the second system is responsible for computing Domain-specific text summarization System based on co-reference resolution algorithm. According to the outcomes obtained in the evaluation method, it is proved that the text summarization System based on co-reference resolution algorithm is not precise and does not meet the requirement. Hence, it is a demonstrated and verified concept that explains and projects the fact that time is a crucial factor in information recovery and extraction. The author made noteworthy efforts in clearing out the discrepancies but mentioned clearly that the future work includes addition of the disambiguation of the word senses for upgrading the occurrence of co-reference resolution algorithm.

A Statistical approach for Automatic Text Summarization by Extraction: Mahesh Chandra et al[30] focused on text summarization and proposed K-mixture semantic relationship significance (KSRS) technology to raise the eminence of the article synopsis results. The huge data availability in internet caused the necessity for search engines such as Google. The search on internet should be efficient and should be fast. For such searching we have a fast evolving branch of science. Research is also at its fast pace in implementing the better algorithms and techniques for the search.

In this background this paper is about Automatic Document Summarization relating both areas of physics and computers. The main theme of the paper is to develop a statistical approach for summarization of a document. The summarization process is to go through the document and minimize the size of the document covering all the aspects. This summarization process calculates the weights of the terms of a sentence. We do this using the algorithms KSRS (K-mixture Semantic Relationship significance) where weights are assigned by occurrence of nouns in that sentence. In general two types of techniques namely Statistical and Linguistic approaches are used in summary generation of a document.

Statistical approach is based on term frequency where rating is given on that basis. And linguistic approach is based on semantics like camas, semi-colons, quotes etc. The most basic methods used here are TD-IDF a traditional procedure to implement. In this process, the words in the capital letters headings are given utmost importance since their matching is comparatively easier while searching. The user should be able to search as fast as he could out of large pool of data available in internet. He also has to distinguish between the data he wants in the documents that result from search. Out of the two methods discussed, statistical method is efficient but not accurate where as linguistic method is accurate but not efficient due to time constraint.

The paper has the description of working which includes research issues, Preprocessing, Term weight determination, Tern relationship exploration and summary generation. In preprocessing, the abstraction of nouns and verbs which are important for summary generation is done. Next term weight generation is done on these abstracted objects by using formulae which are best described. For all the sentences, these values are calculated and then average is calculated to determine the final weight of the sentence in that document.

Two examples are described one to prove the working of the traditional TD-INS system and other to show the performance of KSRS. This is done by taking documents of average size of 800 words from different areas. Graphical representations are made comparing the evolution of summarization process with the original document.

Observation: So to conclude the proposed model met its aim of proposing the statistical process of summarization efficiently along with other processes like linguistic processes. It also showed the live result of the efficiency of both TD-IDS and KSRS by taking examples. The chance of taking the low summary proportions using the KSRS is also best explained.

Automatic Text Summarization Based On Rhetorical Structure Theory: Li Chengcheng [31] presented a new method called Rhetorical Structure Theory for effective automatic text summarization. This new method is based on natural language generation method for effective summarization of an article. The paper concentrates on text summarization using the rhetoric structure theory. This automatically shortens the document that a user is in need of and gives the summarized sentences. This theory extracts the rhetoric structure of the text and a compound that relates the sentences. All the process is best explained by author.

After this identification, the summarized text is converted to natural language which is user friendly. This type of summarization using the clauses and compounds of rhetoric structure is highly efficient. The main idea of this method is analyzing the candidate sentence identifying the rhetoric relations and forming the important part of sentence useful for final summarization.

Past systems based on the frequency of word generation i.e. the sentence is important because a key word is many times present in that sentence is inefficient. It lacks preciseness and recall. The RST system based on knowledge or script based analysis can efficiently rule out those backlogs.

The drive of the paper is explanation of rhetoric structure and summarization process basing on RST. In RST we have a sentence divided into nucleus and satellite. A nucleus is an important part of sentence and supplies a reader much information where as satellite independent of nucleus increases its understandability. Sometimes a satellite supplies more information than a nucleus.

Here, a RST tree is constructed by placing the nucleus as the root of the tree and satellites as leaf nodes. The summarization is done using the nodes i.e. nucleus. These nucleuses are also given weights based on script based analysis. Next follows the summarization method in which the tree starts its construction. For this the entire text is to be divided in to individual sentences which are meaningful. This can be done by dividing the sentences based on the camas, quotes and semicolons present in the sentences. Also the division is done by the presence of 'and' the punctuation marks present before and after and. This is then done into a graph, deletes the unimportant sentences and then summarizes the entire text.

RS tree construction is presents in the paper is well and after the construction, Nucleus filter statement is made i.e. unimportant statements are deleted and nuclei which best suits the document meaning is left out. At this stage, a sentence is logically, structurally understood well and the utility of sentence is known clearly. From this knowledge, the weights can be easily assigned to the sentences, lower weight sentences can be deleted. Now the system is all left with important information useful in the summarization. These sentences can be formed into complete, cohesive and readable summarization.

Observation: So as to conclude, the paper introduced the method of RST in summarizing the text of the document in such a large pool of data available in the web overcoming the drawbacks like recall and precision. But the paper should focus on the drawbacks like "it can not be applied on all documents like magazines. It's inefficiency of analyzing every sentence based on semantic evolution and the domain being limited."

Graph-Based Algorithms for Text Summarization: Khushboo S. Thakkar et al [32] proposed a graph based algorithm for text summarization. The process proposed initially constructs a graph from the given text. The graph is constructed by taking a sentence with more page value as a node and creating its child or links with connections between them in terms of either lexical symbols or words that are more similar. The connection between the sentences is established by relating words in the sentences. An edge that connects the sentences is identified and it is given the cost using the page ranking. The more similarities leads to less edge weight and the other in vice versa. The cost of each edge value is calculated by using the COS functions.

All over the graph, we have connections between them and a shortest path is calculated and is given as summarization. Starting from a node goes along the shortest path till the end node is reached. Likewise the search is made for N shortest paths. This is all entirely based on the path rank values of an edge. The results coming from the shortest path analysis are smooth and worthy.

But, the linking between the sentences is done only based on a word that matches. This is a common phenomenon for many sentences even though not that much related. In that case many unnecessary search results may occur. There the ranking algorithms are better replaced by shortest path algorithms.

Observation: To conclude the summarization process is very useful and efficient enough for coming through a large text. The page ranking values calculation, giving values to edges linking the sentences in constructing the graph and summarization based on the shortest path are best explained.

Conclusion

This literature review emphasizes extractive approaches to text summarization in recent literature. Also we attempted to derive the taxonomy of the automatic text summarization methodologies. This taxonomy is based on traditional models cited frequently in literature. Finally, recent trends in automatic evaluation of summarization systems have been briefed as current state of the art.

References:

[1] http://trec.nist.gov/.

[2] http://duc.nist.gov/.

[3] http://www.itl.nist.gov/.

[4] Radev, D. R., Hovy, E., and McKeown, K. (2002). Introduction to the special issue on summarization. Computational Linguistics., 28(4):399-408.

[5] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research Development, 2(2):159-165.

[6] Baxendale, P. (1958). Machine-made index for technical literature - an experiment.IBM Journal of Research Development, 2(4):354-361.

[7] Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM, 16(2):264-285.

[8] Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. In Proceedings SIGIR '95, pages 68{73, New York, NY, USA.

[9] Lin, C.-Y. and Hovy, E. (1997). Identifying topics by position. In Proceedings of the Fifth conference on Applied natural language processing, pages 283{290, San Francisco, CA, USA.

[10] Conroy, J. M. and O'leary, D. P. (2001). Text summarization via hidden markov models. In Proceedings of SIGIR '01, pages 406{407, New York, NY, USA.

[11] Osborne, M. (2002). Using maximum entropy for sentence extraction. In Proceedings of the ACL'02 Workshop on Automatic Summarization, pages 1{8, Morristown, NJ, USA.

[12] Nenkova, A. (2005). Automatic text summarization of newswire: Lessons learned from the document understanding conference. In Proceedings of AAAI 2005, Pittsburgh, USA.

[13] Barzilay, R. and Elhadad, M. (1997). Using lexical chains for text summarization. In Proceedings ISTS'97.

[14] http://news.google.com.

[15] http://newsblaster.cs.columbia.edu.

[16] http://NewsInEssence.com.

[17] McKeown, K. R. and Radev, D. R. (1995). Generating summaries of multiple news articles. In Proceedings of SIGIR '95, pages 74{82, Seattle, Washington.

[18] Radev, D. R. and McKeown, K. (1998). Generating natural language summaries from multiple on-line sources. Computational Linguistics, 24(3):469{500.

[19] Carbonell, J. and Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of SIGIR '98, pages 335{336, New York, NY, USA.

[20] Mani, I. and Bloedorn, E. (1997). Multi-document summarization by graph search and matching. In AAAI/IAAI, pages 622-628.

[21] Radev, D. R., Jing, H., Stys, M., and Tam, D. (2004). Centroid-based summarization of multiple documents. Information Processing and Management 40 (2004), 40:919-938.

[22] Evans, D. K. (2005). Similarity-based multilingual multi-document summarization. Technical Report CUCS-014-05, Columbia University.

[23] Nowshath Kadhar Batcha, Ahmed. M. Zaki, Algebraic Reduction in Automatic Text Summarization – The State of The Art, International Conference on Computer and Communication Engineering (ICCCE 2010), 11-13 May 2010, Kuala Lumpur, Malaysia

[24] M. Esther Hannah, Dr. Saswati Mukherjee, K. Ganesh Kumar, An Extractive Text Summarization Based On Multivariate Approach, 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)

[25] Tingting He, Fang Li, Liang Ma; Document Relevance Identifying and Its Effect in Query-Focused Text, 2010 IEEE International Conference on Granular Computing

[26] SUO Hong-guang, ZHANG Jing-jing; Evaluation method of automatic summarization calculating the similarity of text based on HowNet, 978-1-4244-6585-9/10, 2010, IEEE

[27] Hien Nguyen, Eugene Santos, Jr., Senior Member, IEEE, and Jacob Russell; Evaluation of the Impact of User-Cognitive Styles on the Assessment of Text Summarization, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, 10.1109/TSMCA.2011.2116001

[28] Zengmin Geng, Jujian Zhang, Xuefei Li, Jianxia Du, Zhengdong Liu; Post-Processing of Automatic Text Summarization for Domain-Specific Documents, 2010 International Conference on Communications and Mobile Computing

[29] Shi Ziyan, The Design and Implementation of Domain-specific Text Summarization System based on Co-reference Resolution Algorithm, 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010)

[30] Munesh Chandra, Vikrant Gupta, Santosh Kr. Paul; A Statistical approach for Automatic Text Summarization by Extraction, 2011 International Conference on Communication Systems and Network Technologies

[31] Li Chengcheng, Automatic Text Summarization Based On Rhetorical Structure Theory, 20IO International Conference on Computer Application and System Modeling (ICCASM 2010)

[32] Khushboo S. Thakkar, Dr. R. V. Dharaskar, M. B. Chandak; Graph-Based Algorithms for Text Summarization, Third International Conference on Emerging Trends in Engineering and Technology, 2010

Author's Biography

AUTHOR'S PROFILE

I, Mrs S. Suneetha received my MCA from Sri Padmavathi Mahila University, M.Tech from IETE, New Delhi. Presently working as HOD, Dept. of Computer Science, Hasvita Institute of Engineering and Technology. I am a member of IETE and is having 18 years of teaching experience at postgraduate level . My Research work is in the field of Text Mining and pursuing my Ph.D from JNTU, Hyderabad.