International Journal of Advanced Research and Development
ISSN: 2455-4030, Impact Factor: RJIF 5.24
www.advancedjournal.com
Volume 2; Issue 4; July 2017; Page No. 29-38

A survey on: Extractive text document summarization techniques

Chandra Shekhar Yadav, Rakesh Kumar, Prem Shankar Singh Aydav, Harendra Pratap Singh

School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India

Abstract
In modern word, popularity of the internet is growing day by day so, volume of data is growing exponentially on the web and a variety of information services increased therefore, to obtain any desired information became a challenging task. Text summarization may be a feasible and powerful to handle such type of problems. Even text summarization can help to the government for "Good Governance". In fact, e-governance is going to be a new approach of any government for Good governance. Summarized information also helps in many ways like in emergency decision support, policymaking, and government routine for government as well as for civil servants. For example Indian government recently launched a web portal www.mygov.in for all people, to give their contribution/views for policy making (like for Digital India, Green India, Skill development etc), to share views about different issues like the recent United Nations declared 21st June at 'International Day of Yoga'. Such type of initiative can help in better policymaking and this is possible only through Text Document Summarization.

Keywords: single document summarization, multi document summarization

1. Introduction
Nowadays, summarization of text document have very effective role in information retrieval (IR) because it presents a huge set of information in summarized form by considering the important and relevant sentences. In brief, according to M. Ramiz summarization in [1] is considered as three steps procedure (1) Text analysis (2) Summary representation known as Transformation, and (3) Synthesis- Generation of relevant summary. The first phase is the analysis phase, In this phase we have a task to analyze the given text document and to select some salient features. After analysis phase in transformation phase, transform the analyzed text into a summary representation based on selected features and the final phase which can be called synthesis (or sometime generation step) in which task to already represented summary are taken to produce more appropriate summary according to user need. Eduard Hovy and Chin-Yew Lin [2] introduced SUMMARIST system to create a racy automated Text Summarization system, three phase working of the system can be understood with equation as "Summarization = Topic Identification + Interpretation + Generation". The compression rate also plays an important role in summarization and effect of this (compression rate) can be seen in the quality of the summary. As user decreases compression rate, user finds more concise summary on the cost of information i.e. more information goes lost with decreased compression rate.

Definition1
Compression rate (C) can be defined, C = M/N: where M is the length of the summary (in words) and N is the length of given documents (also in words). Since here N is constant (for any set of document given), and M is the desired summary length which is vary as user needs so, we can say C $\propto$ M. According to [3, 4, 5] when C is between 5% to 30%, the quality of the summary is acceptable.

1.1 Summarization
Radev *et al*. in [6] has define a text summary as "a text that is produced from one or more texts, hat conveys important information in the original texts, and that is no longer than half of the original text and usually significant less than that".

1.2 Type of Summary
Basically type of summary can depends on (1) number of documents. (2) approach applied, and (3) user needs. According to number of document summarization may be single document or multi document, if we wants to classify according to approached then it may be Extractive, Abstractive (human like summary), if we wants to classify according to user need then it may be informative (query specific)or generic summarization.
Number of Documents: Single Document or Multi Document.
Technique: Extraction or Abstraction.
Detail: Indicative or Informative.
Content required by user: Generalized or Query-based.
Approach: Either Shallow or Deep, Domain specific, Template based, Statically or Soft computing.

1.3 Number of Documents: Single or Mult Document Summarization
Single document summarization: as clearly with name here summarizer system takes only one document for summarization and presents information in a shorter form. Multi-document summarization: It is just an extension of the Single Document summarization system. Multi document can be seen as a cluster of the same type documents. Multi document summarization can be done in two approaches (1) bag of words approach in which consider all document as one document [examples radev paper] and treat as a input, and, in second approach (2) we can generate summary for different clusters (or documents) and then integrate and transform in

desired summary. Since this (second approach) combines and integrates the information across all documents, according to [7] it needs to perform (1) Knowledge synthesis, (2) Knowledge discovery.

1.4 Technique: Extraction or Abstraction

Extractive: Extractive summarization selects (or extract) sentences (or part of a sentence) which are more informative in documents. Till now most of the work done is based on technique, that extract most informative textual elements such as keywords (significant terms which can be scored using TF × IDF score), or concepts (may be events or entities) with linguistic analysis, statistical analysis or hybrid approach. Abstraction: Abstractive methods require a deeper analysis (like linguistic, syntactic, semantic) of the sentences or text along with the ability to generate new sentences (which may be possible that not present in given document). Abstraction summarization provides an obvious advantage in (1) improves the focus of a summary (2) reducing its redundancy and (3) keeping a good compression rate.

1.5 Content required: Generalized or Query focused

Generic: This is also known as generic summarization in which summary, reflects the major contents of the document (s) given without providing any additional information or prior knowledge. Query focused summarization: Query focused summarization or query biased summarization is a special case of summarization, in which summary purposely demands, to be biased according to user needs that given in the form of query.

1.6 Detail: Indicative or Informative

Indicative summary provides ideas of what the text contains or what the text of, without conveying special information, on the other hand informative summary provides information about specific direction (content).

1.7 Approach used

Possible approaches are Statically, Linguistic, Machine learning or Soft Computing. In term of level in the linguistic space approach can be divided in two type Shallow and Deep approach. Shallow Approach: In this approach for most analysis done on the sentence level is syntactic, but important to note that, word level analysis may be a semantic level. Deeper Analysis: In deep analysis at least a sentential semantic level of representation is done. This produce abstract and the synthesis phase, so here generally involves Natural Language Generation from a semantic or discourse level representation. Nowadays Hybrid approaches also used.

2. Related work

In recent years, a number of text document summarization approaches have been proposed as well as evaluated. It is hard to properly categorize work done in past because a lot of area in overlapping in terms of approach. Model proposed, evaluation strategy used. Here we are classifying Automatic text document summarization according to Model proposed and technique used. The automation [8] of text summarization includes machine learning techniques, statics and phycology. So we can say that it is an interdisciplinary research area of computer science. There are two main categories of the text summarization techniques which is known as abstraction summarization and extractive summarization. The abstraction summarization is like actual summary of text document. The abstraction summarization requires the techniques known as fusion, compression and reformation of sentences. It may contain phrases, newly formed sentences and some words which are not present in the actual documents. In last few years a lot of researches have been done in in this area but we are still not near to abstractive summarization. The main challenges of the abstractive summarization are creation of different and new sentences, production various new phrases and the summary must contain same meaning as its root document. In extraction based summarization, the main purpose is to find out set informative sentences, subset of sentences or phrases. These informative sentences, part of sentences and phrases arte included in the summary of the document. The work can be divided in to two categories, the early work and the work in recent years. Following are the three early works which provides the main direction in area of text summarization.

FIRST: The early work in Document Summarization started on "Single Document Summarization", by H. P. Luhn at IBM in the 1950s [10]. Luhn proposed a frequency based model, frequency of word play crucial role, to decide importance of any sentence in the given document. At sentence level, he gave score (called significant factor) sentences are ranked according to significant factor, and top raking sentences added in summary.

SECOND: Another work of (P. Baxendale 1958) [11] at IBM, Introduced new statically based feature. He proposed a sentence position based measure. In his research, he found that, starting and ending sentences became salient sentences for summarization, but this is not better for every document like scientific research paper but good for newspapers summarization.

THIRD: (H. P. Edmondson 1969 [12]) also proposed an effective technique for document summarization. At first, Edmonson designed some rules for manual extraction, then rules were applied on 400 technical documents. Edmondson consider four features (1) sentences position, and (2) frequency of word, (3) presence of cue words and (4) the skeleton of document (to decide if the sentence is a title or a heading). The work was done almost manual. The author found that nearly 44% of the auto extracts match with the manual extracts. After early work lot of work done some are available here [6, 13], here in the next section we are presenting only work done in recent years.

J. Goldstein et.al. in [14] done some analysis and proposed some empirical property of summary, they found that summary length is independent of given document length and they also suggested that, constant summary length is better instead of C (compression rate- that deceases with document size increases specially with multi document) as given by [3, 4, 5]. The summary includes indefinite article (A. An) more frequently (62% times) compared to non summary sentences. Summary sentences which begin with an article (A / An / The) more frequently compare to non-summary sentences. Name of cities, state, countries and Days of week and Name entities (proper noun) also tends to more frequently in summary sentences. They given a long list of negative words (which also helps in finding of summary) that occur frequently in non-summary sentences such as according, adding, said, verb related to communication. Nowadays

number of researcher taking these words (positive words and negative words) as a parameter/ feature to find salient sentences for summarization.

2.1 Query focused summarization

Query focused summarization is a special case of document summarization, in which summary purposely demands, to be biased according to the user query. You Ouyang et.al 2011 [15] used SVR (Support Vector Regression) to calculate the importance of the sentences in a given document. For summarization of a given document a set of seven features is considered which widely divided in two categories (1) query dependent and (2) query independent. Three feature (1) word matching feature, (2) semantic matching feature, (3) Named entity matching feature are query depended and four features (1) TF-IDF feature, (2) Named entity feature, (3) Stop-word penalty feature, and (4) Sentence position feature, are query independent. The Nearly-true score has been assigned to every sentence. The training data is then used to learn mapping function from a set of seven defined features to ‘‘nearly-true” sentence and then learned function is used for prediction of the importance of the sentence, on the test data.

2.2 Graph based

In graph, there are two types of entities nodes and edges, same is also followed in the summarization process where nodes may be an entity such as a phrase, sentence, paragraph even a document and edges are relation among these nodes. Helen Balinsky 2011 [16] viewing automatic summarization in slightly different view, they thought summarization as a process of information compression (squeezing out unimportant information [6]). In all previous approaches (like Lex-Rank [17]), edges (or relation) between nodes (may be words, sentences, paragraphs) have been defined by some similarity threshold drive from standard Information Retrieval. The problem with these methods that, the range of ranking function (s) can be very limited in degree range, or possibility that only a small number of nodes can have small value of ranking function. To solve this difficulty they used different view and consider this graph as a "Kronecker Graphs" and an "Affiliation Networks". They used different graph based metrics for complex networks such as eigenvector centrality, barrenness centrality, Katz centrality, hubs and authorities, scale free network, power law etc, and find this helpful in document summarization. LexRank [17], TextRank [19], Document-sensitive graph model [18] are other graph based text summarization model.

2.3 Multi document summarization

In Multi Document summarization, a number of document for about specific event like a natural disaster, U.S. president Barak Obama visit in India are given for summarization task. Due to a number of documents, the main challenge of multi document summarization is a diversity of information, some information may be related or irrelevant to the central topic (theme) of given documents or in second words it may contain noise. The other challenge is coverage (how much information contains) and redundancy. Researcher done a lot of work in the multi document summarization field and many more like in [20] Considering this is as a global/multi optimization problem which requires simultaneous optimization of more than one objective function. Rasim M. Alguliev 2012 [20] has considered Multi Document Summarization as a "Quadratic Boolean Programming" (QBP) problem, in which the objective function is a weighted combination of (1) content coverage, and (2) for redundancy objectives. In another work [21] they proposed CDDS based summarization with two objectives diversity and coverage. Ordering of information (sentences) is also another challenge in multi document summarization. Due to improper ordering of information there is possibility that it can confuse the readers as well as can degenerate the readability of the summary. To maintain the association and order of sentences, Danushka Bollegala et.al [22] defined four criteria (1) chronology, (2) Topical-closeness (3) Precedence (4) Succession. These all four criteria are combined into a single criteria by using a supervised learning approach. They suggested that, combination of all these four parameter performing well for sentence ordering.

Another work is done by Ming Che Lee [23] for Text Document Summarization, as we know that similarity between sentences or paragraphs plays an effective role, and decides importance whether to include a sentence in the summary or discard. Similarity evaluation among of sentences is a laborious task because of (1) possibility of complex sentence present, and (2) in the sentences there any extra information is provided. They had been proposed "Transformed Vector Space" model. The "WordNet ontology" adopted for construction of the semantic vectors. The score of the vectors has been calculated using WordNet similarity measures instead of conventional methods as the frequency and probability of the words present in the sentence or document.

2.4 Redundancy

Redundancy can be defined as a multiplicity of sentences, sub sentences or information. Coverage and redundancy are reciprocal to each other. Summarization objective is maximum coverage and minimum redundancy, in other words for a constant length summary if we want to increase coverage we need to reduce redundancy. In [24,25,26] gave a simple approach to include a sentence in summary one by one based on modified cosine similarity threshold. To include sentences in the summary, first system (algorithm) selects the most top rank sentence, include it in the summary and then process repeated for remaining sentences. Next sentence is included in summary if similarity between sentences and summary is less than some defined threshold otherwise sentence not to include in the summary and algorithm stop when required summary length is reached. We using this model in our implementation to handle redundancy. Another model MMR popularly used (specially in with a given query) to reduce redundancy. The MMR (Maximal Marginal Relevance) criteria, "strives to reduce redundancy while maintaining query relevance in re-ranking retrieved documents and in selecting appropriate passages for text summarization". This technique gives better result for Multi Document Summarization. Rasim M. Alguliev et.al 2011[27] proposes an unsupervised text summarization model which can use for Single as well as Multi Document Summarization. In simple term, the problem is treated as a Multi objective problem, where the objective is to optimize three objectives (1) Relevance, (2) Redundancy and (3) Length. The summarization is consider as an "Integer Linear

Programming" (ILP) problem and to calculate the similarity among textual units (or sentences) two similarities first cosine based and second "Normalized Google Distance" (NGT) are used. According to them, the advantages of their model are that, it is able to directly discover important sentences, covers the important contents of the original documents with minimum redundancy.

2.5 Event based summarization

Event in defined in number of ways as [28] event is defined as an action and Name Entities, Vanderwade [29] defined event as a dependency triplets, according to Allen [30] event may be instance or durative, events are defined as event terms and associated event elements (location, origination, participants and time are event elements) and event term represents an action themselves, event element denotes action arguments. Mingli Wu 2002 [31] proposed independent event based approach, in that identified important content according to event frequency. Another work is done by Mingli Wu [32] by considering that, every event has its own internal structure, and meantime events generally relate to other events possibly through conditionally, temporally, semantically or spatially. N. Daniel *et al* 2006 [33] proposed a new direction based on events, for news document summarization. They detect whether sub-event can help in capturing essential information and they propose three algorithms. They findings about SAS (Sum of all scores) are important for future efforts to summarize by partitioning.

2.6 Abstractive summarization

Abstractive methods require a deeper analysis of the text and the ability to generate novel sentences, which provide an advantage in improves the focus of a summary, reducing its redundancy and keeps a good compression rate. Pierre-Etienne Genest, Guy Lapalme in [34] proposed a syntactic based approach for abstractive summarization, which relies on the concept of INIT (Information Items). INIT is the "smallest element of coherent information in a sentence". Identification of INIT is based on date, location, Subject–Verb–Object (SVO) triples. Selected INITs, are added in summary. This approach has some advantage of generating typical Short. Information focused sentences to produce a Coherent, Information rich and less redundant summary.

2.7 LSA based Summarization

LSA is a vector space approach (based on the singular value decomposition) that involves the projection of the given matrix $A_{M\times N}$ usually M >> N, to a reduced dimension representation A_r such that r < M. Input matrix A can be designed in any way, mostly A represents Words × Sentences matrix. Input matrix A is Term-Document matrix, which elements A_{ij} represent the weighted Term frequency of term i in document j. Using LSA, input matrix A is decomposed as follows Document are sentences present in text, represented in figure 1.

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

where U is a $M \times N$ column orthonormal matrix, which column are called left singular vector, $\sum_{N\times N} \approx S_{N\times N}$ is $N \times N$ diagonal matrix, whose diagonal elements are non-negative singular values, which are sorted in descending order as in equation (2), and V^T right singular matrix also an orthonormal matrix with size $N \times N$, which columns are called right singular vector. Let, Rank(A) = r, then S property can be express as equation (2), where S_i, 1<i<r and r ≤ N is representing an element in i[th] row or columns,

$$S_1 \geq S_2 \geq S_3 .... \geq S_r \text{ , and}$$
$$S_{r+1} = ...S_{n-1} = S_n = 0 \quad (2)$$

Some intial work for summarization task using LSA start with, Gong and Liu (2001) [35] uses V^T matrix for sentence selection, Murray *et al* (2005)[36] uses both V^T and S-matrix for sentences selection, Steinberger and Jezek (2004) [37] proposed another approach, they are also using both S and V matrix (transpose of V^T) for sentence selection. Cross method is proposed by Ozsoy *et al* (2011) [38], in which V^T matrix is preprocessed that represents only core sentences. In same paper Ozsoy *et al* (2011)[38] another Topic based approach proposed. In all these previos work summary is extracted based on sentence score. Detail of respective model is presented below.

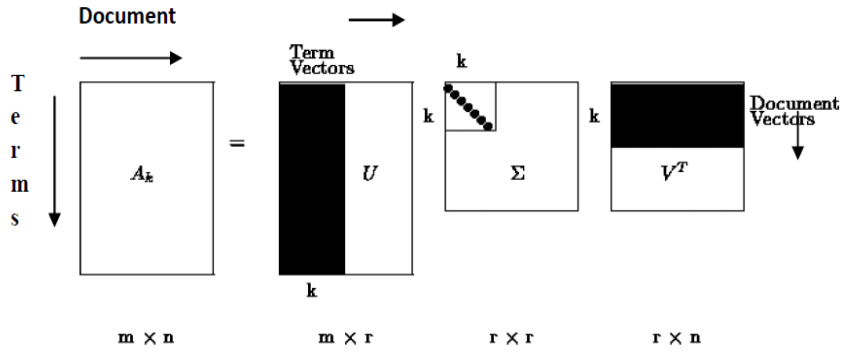

Fig 1: Showing Mechnism of LSA

2.7.1 Steps in Summarization using LSA

Here in figure 2 we are presenting a layout of LSA based summarization procedure, this is three step process, in first step we creat a Matrix A generally this is term docuemnt matrix, matrix is constructed based on different weighting creterias. in Step-2 LSA decomposition performed the output from this step is three decomposed Matrix, then in step-3 using these three matrix we analys our document based on this we extract some sentences, words and add them in summary until required length sumamry found.

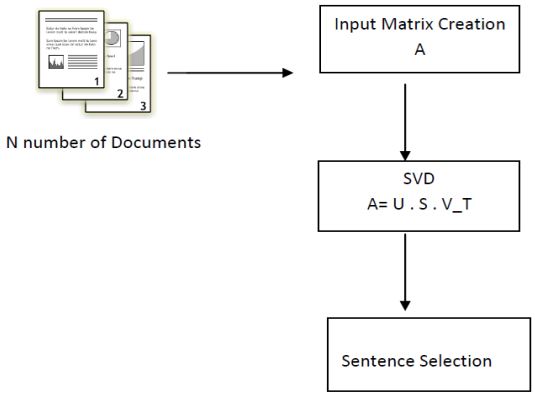

Fig 2: LSA based Summarization Procedure

2.8 Sentence selection Algorithms

Algorithm-1 Yihong Gong and Xin Liu 2001 Algorithm

Input: Document D i.e. content of words, sentences, paragraphs. K- Predefined length of summary sentences.

Output: Summary

Decompose the given document D into sentences {S_1, S_2,.. S_n}, and use these sentences to form the candidate sentence set S, and set K = 1.

1. Construct matrix a i.e. term by sentence matrix, for the given document D.
2. Perform the SVD on A to obtain different matrix, the singular value matrix Σ , and the right singular vector matrix V_T. In the singular vector space, each sentence i is represented by the column vector i = [vi1 vi2 vir] T of VT.
3. Select the k'th right singular vector from matrix VT.
4. Select the sentence S_i which has the largest index value corresponding to the k'th right singular vector, and include that in the summary.
5. If(K= = Predefined number), Terminate the operation else, increment K by one, and go to Step 4.

Salient about Yihong Gong and Xin Liu 2001[35]

1. New and unique approach.

Drawback

1. In this approach we summary is depends only on V_T, as per our convenient we can select number of rows i.e. reduction diminution of V_t. The maximum length of summary sentences will be the reduced dimension. If in case of automatic summarizer system user given a choice of summary-length > reduced dimension then, no way is proposed to select more sentences.
2. Claimed by Steinberger and jezek (2004), extraction of sentences which may belongs to less important concept.
3. Let some chosen a concept-k for sentence selection in this concept if two sentences have high values, Gong's approach choosing only one sentence. With one example If there is a case that, concept-k is related to 0.80 with sentence S_i and, 0.78 with sentence_S_i+1. As we know these are highly related then also this is not respected.
4. All chosen concepts considered at same level, but some may not so important in the V_T.
5. This is well know that concepts are independent to each other, so this is expected that sentences which are extracted also these concepts also independent, this is not possible in text data specially due to name entities, pronoun words and stop words. In simple words if we wants diversity in summary then our matrix A should be more clean i.e. noiseless data. So better output from this approach is depends only on Input matrix A, and we know pre-processing in text data challenging task.

Algorithm-2 Murray & Renals Approach 2005[36]:

Step 5: Instead of selecting only one sentence form V_T which depends on highest index value, select the number of sentences based on matrix S. The number of sentence selection from one concept is depends by getting the percentage of the related singular values over the sum of all singular values (or may reduced S).

Advantage

1) Overcome the problem of Yihong Gong and Xin Liu approach, which select only one sentence from each concept.

Algorithm-3 Steinberger & Jezek 2004: Algorithm

Input: Document D i.e. content of words, sentences, paragraphs. K- Predefined length of summary sentences.

1. Decompose the given document D into sentences {S_1, S_2,.. S_n}, and use these sentences to form the candidate sentence set S.
2. Construct matrix A i.e. term by sentence matrix, for the given document D.
3. Perform the SVD on A to obtain different matrix, the singular value matrix Σ , and the right singular vector matrix V_T. Find V, by taking transpose of V_T. In V, each sentence Si is represented by the row vector i = [vi1 vi2 vin]. (reduced space size is n)
4. For every sentence in reduced space n, by using V and S, compute sentence length. Sentence length S_k is computed as follow,
5. Length of Sentence i = $$S_i = \sqrt{\sum_{j=1}^{n} V_{ij} * \Sigma_{jj}}$$
6. Select highest length sentence S_i here 1 < i < N, and add to the summary.
7. If(K= = Predefined number), Terminate the operation else, increment K by one, and go to Step 5 to select next highest length sentence.

Advantage

1) Sentence selection is based on new reduced space, so consideration about only preferred concepts and highest length sentence.

Drawback

1. Long length sentences may dominate in calculating S_i.
2. Diversity in summary is not considered.
3. Even some sentences are somehow related to concepts, in preprocessing author put this to 0 is same as showing that respective sentence and concept are unrelated.

Algorithm-4 Cross Method

1. Decompose the given document D into sentences {S_1, S_2, S_n}, and use these sentences to form the candidate sentence set S.
2. Construct matrix A i.e. term by sentence matrix, for the given document D.
3. Perform the SVD on A to obtain different matrix, the singular value matrix Σ , and the right singular vector matrix V_T. In V_T, each sentence Si is represented by the coloumn vector i = [v_1i v_2i v_ni]. (reduced space size is n)
4. Preprocess V_t: select only core sentence find average sentence score for each concept, if for any sentence score in less than respective calculated average score, then set cell value 0 corresponding to that concept.
5. For every sentence in reduced space- n, by using preprocessed V_t and S, compute matrix multiplication W = V_t * S. Column of W is representing sentences and row representing concepts.
6. For all sentences find sentence length. Sentence length is given by,

$$S_i = \Sigma_{j=1}^{n} W_{ji}$$

7. Select largest length sentence in summary.
8. If(K= = Predefined number), Terminate the operation else, increment K by one, and go to Step 5 to select next highest length sentence.

Drawback

1) Diversity in summary is not considered.

Algorithm-5 TOPIC MODEL

1. Decompose the given document D into sentences {S_1, S_2,.. S_n}, and use these sentences to form the candidate sentence set S.
2. Construct matrix A i.e. term by sentence matrix, for the given document D.
3. Perform the SVD on A to obtain different matrix, the singular value matrix Σ , and the right singular vector matrix V_T. In V_T, each sentence Si is represented by the coloumn vector i = [v_1i v_2i v_ni]. (reduced space size is n)
4. Preprocess V_t: select only core sentence find average sentence score for each concept, if for any sentence score in less than respective calculated average score, then set cell value 0 corresponding to that concept.
5. As in matrix V_T rows are representing concepts/Topics. In this step this task is to find main topic.

a. X=Concepts * Concepts Matrix is created, | Concepts | = | reduced dimensioned space |

b. X[i][j]= Concepts[i] ∩ Concepts[j]

6. Now X is symmetric matrix, so we can find strength of a concept by adding either row or columns.
7. Select concept which has highest strength for sentence selection.
8. Corresponding to selected concept, select sentence from preprocessed V_T which has highest index value, and add this to summary.
9. If(K= = Predefined number), Terminate the operation else, Increment K by one, and go to Step 7 to select next highest strength concept and corresponding sentence from step 8.

Drawback

1. Still following Yihong Gong and Xin Liu approach from sentence selection i.e. select only one sentence from one concept.
2. Let some chosen a concept-k for sentence selection, in this concept if two sentences have high values, Gong's approach choosing only one sentence. e. With one example If there is a case that, concept-k is related to 0.80 with sentence S_i and, 0.78 with sentence_S_i+1. As we know these are highly related then also this is not respected.
3. Even some sentences are somehow related to concepts, in preprocessing author put this to 0 is same as showing that respective sentence and concept are unrelated.
4. This is said that Con-0 > Con-1 > Con-2.....>Con-n. (sign A >B showing A is preferred over B). property not considered.
5. Extracted Sentences are assumed to be implicitly diverse.

Advantage

1. Finding main topic nor not based on concept * Concept matrix is giving better results.
2. Higher concept score is showing that, a concept is much more related to other concepts.

2.9 Entropy based method

Yingjie and Jun (2013) [39] proposed a new way to create Input matrix A. weight of location A-IJ is given by weighted combination of local weight, global weight and neighboring term weight. LSA applied on input matrix A and from each topic two sentences are selected for the summary. Their approach combines term description with sentence description for each topic.

Ouyang and Qing-ping (2008) [40] proposed entropy based data summarization algorithm on real time data stream. The real-time data stream is a sequence of data items may be in form text, author considered that data stream arrives in some order i.e. repeated sequence. Now researchers are much interest in building of stream processing applications. In these applications, data are usually unbounded, continuous, huge volume, fast arriving, time various, or bursting. To process this input data stream in real time constraints, redundancy should be removed from the input stream. The key problem that how to reduce the redundancy in the data stream. Summarization algorithms can provide a way to this directions to drop overloaded input data.

Wenjuan *et al* (2010) [41] proposed summarization technique based on two features (1) Entropy, (2) Relevance. The entropy of a sentence is calculated according to Shannon information entropy and relevance between two sentences is calculated by overlapped words. Then they perform unsupervised summarization named Entropy and Relevance based summarization (ERBS) and supervised summarization utilizing Linear Regression and ELM regression.

Ravindra *et al* (2004) [42] also proposed summarization using entropy methods, which is applicable for both single and multi-document summarization. According to them entropy methods is not applicable to reduce redundancy from given text documents, so to solve this problem they proposed graph-based technique. In graph based redundancy removal, every sentence is represented as a node in a directed graph. A link is established from one node to another if at least three non-stop-words are common to them. They also discussed LSA for redundancy removal. To score a sentence, they used forward and backward entropy-based method. Top N sentences are selected for the summary.

Kennedy *et al* (2010) [43] also using entropy-based sentence selection, after sentence selection in summary they find entropy to measure uniqueness in the summary using Shannon entropy and based on that entropy they decide how much information is contained in that summary, and sentence add to the summary. Osborne (2002) [44] also using conditional maximum entropy for summarization.

2.10 Evaluation strategy

To find a good summary lot of work done, but to decide the quality of the summary still a challenging task. Research is done by Goldstein in [45] he conclusion that. "even human judgment of the quality of a summary varies from person to person". In their work, for a given document when a number of persons were called upon to select the most informative sentences, they found that there was only little overlap among the sentences picked by people. They also conclude that. "human judgment usually doesn't find concurrence on the quality of a given summary". Hence it is sometimes difficult to judge the quality of a summary.

For evaluation most researcher using the "Recall Oriented Understudy for Gisting Evaluation" (ROUGE) introduces by Lin [46] and this has been officially adopted by DUC for summarizer evaluation. ROUGE compares system generated summary with different model summaries (or can be called as a reference summary/standard summary). It has been considered that ROUGE is an effective approach to measure document summarizes so widely accept. ROUGE measures, overlap words between the system summary and standard summary (gold summary/human summary). Overlapping words are measured based on N-gram co-occurrence statistics, where N-gram can be defined as the continuous sequence of N words. Multiple ROUGE metrics has been defined for different value of N and different models (like LCS, weighted). Standard ROUGE-N is defined by:

$$ROUGE-N = \frac{\sum_{S\epsilon\{References\ Summaries\}} \sum_{gram_n \epsilon S} Count_{match}(N-gram_{\square})}{\sum_{S\epsilon\{References\ Summaries\}} \sum_{gram_n \epsilon S} Count(N-gram_{\square})} \quad (3)$$

Here N stands for the length of the N-gram, Count(N-gram) is the number of N-grams present in the reference summaries, and the maximum number of N-grams co-occurring in the system summary, the set of reference summaries is $Count_{match}$ (N-gram) ROUGE measures generally gives three basic score Precision, Recall, and F-Score. Suppose that, in some "text retrieval system", for given document "D", query "Q", we has been retrieved a number of documents for some Q. Now the question is, how we judge that, "how much accurate the system was" ?. If the set of documents relevant to the given query Q be denoted by {Relevant}, and the set of documents which are retrieved is given by {Retrieved}. The set of documents which are both Relevant and Retrieved can be denote with "{Relevant} n{Retrieved}". Now there are two basic measures Precision, Recall to decide the quality of text retrieval system.

Definition 2: Precision (P): "This is the percentage of retrieved documents that are in fact relevant to the query", (i.e., "correct" responses). It is given by

$$Precision(P) = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|} \quad (4)$$

Definition 3: Recall (R): "This is the percentage of documents that are relevant to the query and were, in fact, retrieved". It is given by

$$Recall(R) = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|} \quad (5)$$

Although theoretically, Precision and Recall are not related, in practice "high precision" is achieved generally at the cost of Recall and vice versa in mathematically formulation can given by this way, $P \propto 1/R$. It depends on nature of applications that which measure is useful for us. So we need a smart solution in this trade-off between Precision and Recall. We can define a single measure (F-score) to compare different systems. The F-score measure is generally used in an information retrieval system, it is given as harmonic mean of Precision and Recall.

Definition 4: F-Score: F-Score is generally given as

$$F - \text{Score} = \frac{(1+\beta^2) \times Recall \times Precision}{Recall + (\beta^2 \times Precision)} \quad (6)$$

where ß is constant given by $P_{lcs}(A,B)/R_{lcs}(A,B)$, A is reference summary and B is system summary. Other variation of ROUGE are; ROUGE-N, ROUGE-L, ROUGE-W, ROUGE S*, ROUGE SU*. In our evaluation we are using fourteen ROUGE measure (N= 1to 10, L, W, S*, and SU*) because only ROUGE-1 score is not sufficient indicator for summarizer performance.

If for us A and B are two given sequences, then "Longest common subsequence" (LCS) of A and B i.e. LCS(A,B) is a common sub-sequence with maximum possible length, and different ROUGE-L measures are given below where LCS (A,B) is the length of longest common sub-sequence below (this representation is for sentence level), and A is reference summary with length m, B system summary with length n.

$$R_{lcs} = \frac{LCS\ (A,B)}{m} \quad (7)$$

$$P_{lcs} = \frac{LCS\ (A,B)}{n} \quad and \quad (8)$$

$$F_{lcs} = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 .\ P_{lcs}} \quad (9)$$

Even LCS has nice property, but it can't differentiate their embedding sequence. Let X=[P Q R S T U V] is model summary (or reference summary) and Y1 = [P Q R S T A B] and Y2 = [P Q A R B S T] are two system generate summary, and we have to evaluate which summary either Y1 or Y2 is better. Y1 will be a better choice (because of the initial sentence sequence) but, if we are using ROUGE-L score then the score will be same. so we need to introduce some weight with sequence that version named ROUGE-W. In our experiment we considering as W=1.5.Major headings are to be column centered in a bold font without underline. They need be numbered. "2. Headings and Footnotes" at the top of this paragraph is a major heading.

3. Text Summarization Systems

Here we are presenting some tool available for summarization [47],

3.1 Mead

"This system was developed at the University of Michigan in 2001. It can produce both single and multi-document extractive summaries. The idea behind it is the use of the centroid-based feature. Moreover, two more features are used: position and overlap with the sentence. Then, the linear combination of the three determines what sentences are most salient to include in the summary. The system works as follows: MEAD uses the CIDR Topic Detection and Tracking system to identify all the articles related to an emerging event. CIDR produces a set of clusters. From each cluster a centroid is built. Then, for each sentence, three values are computed: the centroid score, which measures how close the sentence to the centroid is; the position score indicates how far is the sentence with respect to the beginning of a document; the overlap with the sentence or title of the document by calculating tf*idf between the given sentence. Then all these measures are normalized and sentences which are too similar to others are discarded (based on a cosine similarity measure). Any sentence that have not been discarded would be included in the summary".

3.2 Web in Essence

"This system was also developed at the University of Michigan in 2001. It is more than a summarization system. It is a search engine to summmarize clusters of related Web pages which provide more contextual and summary information to help users explore retrieval results more efficiently. A version of MEAD was used in the development of this Web-based summarizer, so that the features used to produce extracts are the same as the ones used in MEAD. The overall architecture of the system can be decomposed into two main stages: one behaves as a Web-spider that collects URLs from the Internet and then it groups the URLs into clusters. The second main stage is to create a multi-document summary from each cluster using the MEAD centroid-algorithm".

3.3 NeATS

"Was developed in 2001 by the University of Southern California's Information Sciences Institute. It is tailored to the genre of newspaper news. Its architecture consists of three main components: con-tent selection, content filtering and content presentation. The goal of content selection is to identify important concepts mentioned in a document collection. The techniques used at this stage are term frequency, topic signature or term clustering. For content filtering three different features are used: sentence position, stigma words and redundancy. To achieve the latter, NeATS uses a simplified version of MMR algorithm. To ensure coherence of the summary, NeATS outputs the sentences in their chronological order. From this system, iNeATS i.e., an interactive multi-document summarization system that provides a user control over the summarization process, was later developed".

3.4 GIS Texter

"This system was developed in 2002 and produces single and multi-document extracts and abstracts by template-driven IE. The system performs differently depending on working with single document or multi-document summarization. For single-documents, the most relevant sentences are extracted and compressed by rules learned from a corpus of human-written abstracts. In the stage, reduction is performed to trim the whole summary to the length of 100 words. When multi-document summarization has to be done, the system, based on Information Extraction (IE) techniques, uses IE-style templates, either from a prior set (if the topic is well-known) or by ad-hoc generation (if it is unknown). The templates generated by CICERO IE system are then mapped into text snippets from the texts, in which anaphoric expressions are resolved. These text snippets can be used to generate coherent, informative multi documents summaries".

3.5 NetSum

"Different from the other approaches previous shown, NetSum, developed in 2007 by Microsoft Research Department, bets on single document instead of multi-document summarization. The system produces fully automated single-document extracts of newswire articles based on neuronal nets. It uses machine learning techniques in this way: a train set is labeled so that the labels identify the best sentences. Then a set of features is extracted from each sentence in the train and test sets, and the train set is used to train the system. The system is then evaluated on the test set. The system learns from a train set the distribution of features for the best sentences and outputs a ranked list of sentences for each document. Sentences are ranked using RankNet algorithm".

4. References

1. Radev DR, Jing H, Stys M, Tam D. Centroid-based summarization of multiple documents. Information Processing & Management. 2004; 40(6):919-938.
2. Hovy E, Lin CY. Automated text summarization and the SUMMARIST system. In Proceedings of a workshop on held at Baltimore, Maryland. Association for Computational Linguistics. 1998, 197-214.
3. Hahn U, Mani I. The challenges of automatic summarization. Computer. 2000; 33(11):29-36.
4. Kupiec J, Pedersen J, Chen F. A trainable document summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. ACM. 1995, 68-73.

5. Mani I, Maybury MT. Advances in automatic text summarization. Cambridge: MIT press. 1998, l293.
6. Radev DR, Hovy E, McKeown K. Introduction to the special issue on summarization. Computational linguistics. 2002; 28(4):399-408.
7. Alguliev RM, Aliguliyev RM, Isazade NR. Multiple documents summarization based on evolutionary optimization algorithm. Expert Systems with Applications. 2013; 40(5):1675-1689.
8. Alguliev RM, Aliguliyev RM, Mehdiyev CA. Sentence selection for generic document summarization using an adaptive differential evolution algorithm. Swarm and evolutionary computation. 2011; 1(4):213-222.
9. Wan X. Using only cross-document relationships for both generic and topic-focused multi-document summarizations Information Retrieval. 2008; l(11):25-49.
10. Luhn HP. The automatic creation of literature abstracts. IBM Journal of research and development. 1958; 2(2): 159-165.
11. Baxendale P, Baxendale PB. Machine-made index for technical literature: an experiment. IBM Journal of Research and Development. 1958; 2(4):354-361.
12. Edmondson HP, Edmundson HP. New methods in automatic extracting. Journal of the ACM (JACM). 1969; 16(2):264-285.
13. Ganapathiraju MK. Relevance of cluster size in MMR based summarizer: a report, 2002.
14. Goldstein J, Kantrowitz M, Mittal V, Carbonell J. Summarizing text documents: sentence selection and evaluation metrics. In *proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1999, 121-128.
15. Ouyang Y, Li W, Li S, Lu Q. Applying regression models to query-focused multi-document summarization. Information Processing & Management. 2011; 47(2):227-237.
16. Balinsky H, Balinsky A, Simske SJ. Automatic text summarization and small-world networks. In Proceedings of the 11th ACM symposium on Document engineering. ACM. 2011, 175-184.
17. Erkan G, Radev DR. LexRank: Graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res.(JAIR). 2004; 22(1):457-479.
18. Wei F, Li W, Lu Q, He Y. A document-sensitive graph model for multi-document summarization. Knowledge and information systems. 2010; 22(2):245-259.
19. Mihalcea R, Tarau P. TextRank Bringing order into texts. Association for Computational Linguistics, 2004.
20. Alguliev RM, Aliguliyev RM, Hajirahimova MS. Gendocsum+ mclr: generic document summarization based on maximum coverage and less redundancy. Expert systems with applications. 2012; 39(16):12460-12473.
21. Aalguliev RM, Aliguliyev RM, Isazade NR. Cdds: constraint-driven document summarization models. Expert systems with applications. 2013; 40(2):458-465.
22. Bollegala D, Okazaki N, Ishizuka M. A bottom-up approach to sentence ordering for multi-document summarization. Information processing & management. 2010; 46(1):89-109.
23. Lee MC. A novel sentence similarity measure for semantic-based expert systems. Expert Systems with Applications. 2011; 38(5):6392-6399.
24. Yadav CS, Sharan A. Hybrid approach for single text document summarization using statistical and sentiment features. International Journal of Information Retrieval Research (IJIRR*)*. 2015; *5*(4):46-70.
25. Yadav CS, Sharan A, Kumar R, Biswas P. A New Approach for Single Text Document Summarization. In *Proceedings of the Second International Conference on Computer and Communication Technologies*. Springer India. 2016, 401-411.
26. Yadav CS, Sharan A, Joshi ML. Semantic graph based approach for text mining. In *Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference*. IEEE. 2014, 596-601.
27. Alguliev RM, Aliguliyev RM, Hajirahimova MS, Mehdiyev CA. Mcmr maximum coverage and minimum redundant text summarization model. Expert systems with applications. 2011; 38(12):14514-14522.
28. Filatova E, Hatzivassiloglou V. Event-based extractive summarization. proceedings of the 42th Annual Meeting of the Association for Computational Linguistics Workshop. 2004, 104-111.
29. Vanderwende L, Banko M, Menezes A. Event-centric summary generation. Working notes of DUC, 2004.
30. Allen JF. An Interval-Based Representation of Temporal Knowledge. In IJCAI. 1981; 81:221-226.
31. Li W, Wu M, Lu Q, Xu W, Yuan C. Extractive summarization using inter-and intra-event relevance. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics. 2006; 369-376.
32. Wu M. Investigations on event-based summarization. In Proceedings of the 21st International Conference on computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Association for Computational Linguistics. 2006, 37-42.
33. Daniel N, Radev D, Allison T. Sub-event based multi-document summarization. In Proceedings of the HLT-NAACL 03 on Text summarization workshop. Association for Computational Linguistics, 2003; l(5):9-16.
34. Genest PE, Lapalme G. Framework for abstractive summarization using text-to-text generation. In Proceedings of the Workshop on Monolingual Text-To-Text Generation. Association for Computational Linguistics. 2011, 64-73.
35. Gong Y, Liu X. Generic text summarization using relevance measure and latent semantic analysis. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR. ACM, 2001, 19-25.
36. Murray G, Renals S, Carletta J. Extractive summarisation of meeting recordings. In Proceedings of the ACL. 05, 2005.
37. Steinberger J, Jezek K. Text summarization and singular value decomposition. In: Proceedings of advances in information systems, ADVIS. Springer, 2004, 245-254.

38. Ozsoy MG, Alpaslan FN, Cicekli I. Text summarization using latent semantic analysis. Journal of Information Science. 2011; 37(4):405-417.
39. Yingjie W, Jun M. A comprehensive method for text summarization based on latent semantic analysis. In Natural Language Processing and Chinese Computing. Springer Berlin Heidelberg. 2013, 394-401.
40. Ouyang L, Qing-ping G. An Entropy-Based Data Summarization Algorithm in Data Stream System, 2008.
41. Wenjuan L, Fuzhen Z, Qing H, Zhongzhi S. Effectively Leveraging Entropy and Relevance for Summarization, 2010.
42. Ravindra G, Balakrishnan N, Ramakrishnan KR. Multi-document automatic text summarization using entropy estimates. In SOFSEM 2004: Theory and Practice of Computer Science. Springer Berlin Heidelberg. 2004, 289-300.
43. Kennedy A, Copeck T, Inkpen D, Szpakowicz S. Entropy-Based Sentence Selection with Roget's Thesaurus. In Third Text Analysis Conference, TAC, 2010.
44. Osborne M. Using maximum entropy for sentence extraction. In Proceedings of the ACL-02 Workshop on Automatic Summarization, Association for Computational Linguistics. 2002; l(4):1-8.
45. Goldstein J, Mittal V, Carbonell J, Callan J. Creating and evaluating multi-document sentence extract summaries. In Proceedings of the ninth international conference on Information and knowledge management, ACM. 2000, 165-172.
46. Lin CY. Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. 2004, 74-81.
47. Lloret Elena. "Text summarization: an overview." Paper supported by the Spanish Government under the project TEXT-MESS (TIN2006-15265-C06-01), 2008.