ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Extractive Techniques for Automatic Document Summarization: A Survey

Rajvardhan Oak

Second Year Student, Dept. of Computer Engineering, Pune Institute of Computer Technology, Pune, Maharashtra, India

ABSTRACT: The 21st century is witness to a major internet revolution. With this internet boom, a large number of documents are made available. The document size may range from a few pages to over a thousand pages. It is not practically possible to go through each and every document to extract the relevant information for a particular study or research. Document Summarization is a technique that generates a condensed version of a text document, preserving the key points, fundamental information and overall meaning. Presenting the reader with a summary of the document helps him in identification of the key ideas, and deciding if the document is relevant or not. The generation of a summary involves either extracting the key sentences of the document (extractive summarization), or retelling the document content in fewer words (abstractive summarization). In this paper, I present briefly the techniques used for extractive summarization.

KEYWORDS: Text Summarization, extractive summary, text processing, data mining, web mining

I. Introduction

With the advent of the internet, a large amount of information is made available in the form of documents across the web. Due to the sheer volume of the documents, and the amount of information contained in each document, manual summarization is not feasible. As a result, automatic summarization has become the need of the hour. Presenting the reader a summary of the document greatly facilitates the retrieval of essential and relevant information. Automatic document summarization is a field that is heavily researched today.

Formally, automatic text summarization may be defined as [6] the process of distilling the most important information from a source(s) and retaining only that information in order to produce an abridged version for a particular user(s) and task(s). It aims at extracting the gist of the document and present it to the reader in a condensed form, thus eliminating the need to read the entire document for the sake of a few points.

The most important advantage of using a summary is its reduced reading time. A good summary system should reflect the diverse topics of the document while keeping redundancy to a minimum. The document summary may be either generic (which summarizes the text as a whole), or query-centric (which analyses the text based on some search key and presents the relevant summary).

Text summarization may be classified into two types [5] [6]:

(1) *Extractive Summarization:* It involves selecting the powerful and meaningful sentences from the text and compiling a summary. The sentences are included in the summary as they are and without any change.

(2) *Abstractive Summarization:* It involves analyzing the document as a whole, interpreting the meaning of the key ideas and generating a conclusive summary. The sentences in this kind of summary may not be present in the source document.

A further classification [1] of automatic summarization may be seen in Fig.1

Generally, it has been observed that extractive summarization techniques are easier to implement, however they may not always produce an accurate summary. On the other hand, abstractive summarization techniques generate highly accurate summaries, but they require heuristic algorithms and are difficult to implement.

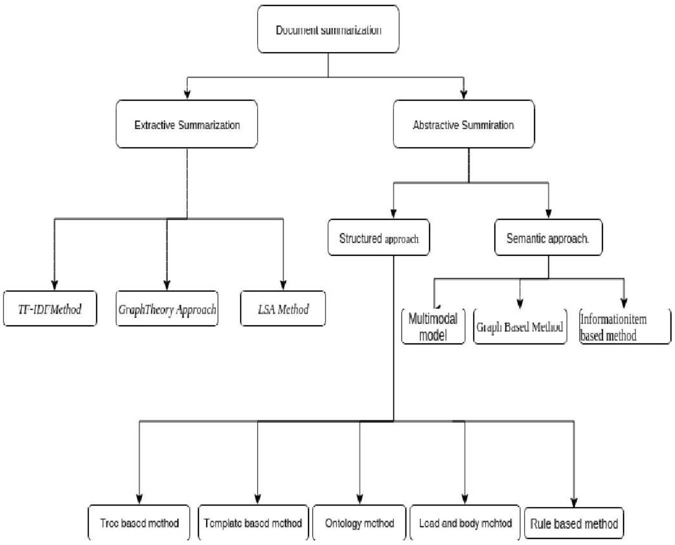

Fig 1

II. BACKGROUND

Extractive summaries are formulated by extracting key text segments (sentences or passages) from the text, based on statistical analysis of individual or mixed surface level features such as word/phrase frequency, location or cue words to locate the sentences to be extracted. The "most important" content is treated as the most relevant and is included in the summary. The sentences are copied as they are from the text. Search engines typically generate extractive summaries from web pages.

Extractive summarization involves a simple algorithmic approach. Each sentence in the document is analysed for the presence or absence of certain features, and decided whether it is important enough to be included as a part of the final summary [2]. Following are some of the features used as decision parameters for the selection of the sentences to be included in the summary:

(1) *Keyword Feature:* It involves determination of certain keywords by Morphological Analysis and NP-Clustering and Scoring. The sentences which contain the keywords are included in the summary. In another version of this feature, certain 'cue words' are defined, the presence of which determines the selection of the sentence.

(2) *Title Word Feature:* It is based on the principle that the words appearing in the title, headings and sub-headings of the document refer to important topics. Hence, the sentences containing these words are included in the summary.

(3) *Location Feature:* The sentences which are towards the beginning or the end of the text are considered important. The beginning will introduce the subject, and the sentences at the end will conclude an issue. Thus, these sentences are included in the summary.

(4) *Proper Noun Feature:* Proper nouns are names of people and places. This information is very important, and must be mentioned in the summary. Hence, the sentences which contain proper nouns are chosen for the summary.

(5) *Pronoun Feature:* The sentences which contain pronouns such as 'He', 'She', and 'It' are not included in the summary. Extracted out of context, they do not convey any meaning and may result in miscommunication of information.

The techniques of extraction usually employ one or more of the above features for deciding the validity of the sentence. Extractive summarization generally consists of two phases:

(1) *Pre-Processing Stage:* In this stage, the sentence structure is analysed and all the features are identified.

(2) *Processing Stage:* Based on the characteristic features observed in pre-processing stage, weights are calculated for each sentence. The weight is used as a decision parameter for the selection of a sentence in the summary.

III. TECHNIQUES OF EXTRACTIVE SUMMARIZATION

A. *TF-IDF method*

Term Frequency Inverse Document Frequency (TF-IDF) [2] [5] [11] method is a word distribution method used to determine what words in a corpus of documents might be more favourable to use in a summary. As the term implies, TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TF-IDF numbers imply a strong relationship with the document they appear in. In a single document summarization, the 'Bag of words' approach is used which models the document at the sentence level where sentence-frequency is the number of sentences in the document that contain that term. In this scheme, is composed of two components, namely, word/term frequency and inverse sentence frequency. Term frequency (T_f), indicate number of times a word appears in the text which measures salience of word within that document. Document frequency (D_f) indicate number of documents in which the word appears. The word/term frequency, TF_{ti}, of a word/term T_t is defined as the number of occurrences of the word/term T_t in sentence S_i. The inverse sentence frequency, ISF_t, of a word/term T_t is defined as:

$$ISF_t = \log(N/N_t) \qquad \ldots(1)$$

where N_t is the number of sentences in a document D in which the word/term T_t occurs. The weight WT_i is computed by:

$$WT_i = TF_{ti} \cdot ISF_t, \qquad \ldots(2)$$

t = 1,...,n, i = 1,...,N.

TF-IDF assigns to each term a weight in a document that is [4]:

i. Highest when t occurs many times within a small number of documents (thus lending high discriminating power to those documents).

ii. Lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal).

iii. Lowest when the term occurs in virtually all documents.

Once TF-IDF score has been computed for each word the next step is to calculate number of such thematic words per sentence. With this value sentences in the input text are ranked and highest scored sentences are picked to be part of summary.

B. *Graph Theory Approach*

A graph is a set of vertices connected by edges. It is mathematically defined as

$$G=(V,E) \mid E \subset V\times V \qquad \ldots(3)$$

In graph based approach [6][9][2], the document is modelled as a graph where the vertices denote the sentences. An edge exists between two vertices if there is a similarity between the corresponding sentences. The similarity measure is calculated on many parameters like content overlap and Term Frequency(TF). The vertices with a high rank have higher importance and should be included in the summary.

The affinity graph based summarization method consists of three steps[6]: (1) an affinity graph is built to reflect the semantic relationship between sentences in the document set; (2) information richness of each sentence is computed based on the affinity graph; (3) based on the affinity graph and the information richness scores, diversity penalty is imposed to sentences and the affinity rank score for each sentence is obtained to reflect both information richness and information novelty of the sentence. The sentences with high affinity rank scores[3] are chosen to produce the summary. The document graph is represented as a matrix M, where

$$M_{(m,n)}=F_{aff}(S_m,S_n) \qquad \ldots..(4)$$

where F_{aff} is a function that indicates the affinity of two sentences S_m and S_n in the graph. The graph is then diffused to get

$$M' = \sum_1^\infty \gamma^{t-1} M^t \ldots(5)$$

The information richness of sentences[3][1] is based on the principle that higher the rank of a sentence, greater is amount of relevant information in it. Also, sentences which are adjacent to rich sentences are considered as rich in information. We can define information richness function recursively as

$$F_{IR}(S_i)=\sum_{i\neq j} F_{IR}\left(S_j\right) \cdot M'_{j,i} + \frac{1-d}{n} \quad \ldots(6)$$

where ‘d’ is the damping factor generally set to 0.85.

As a result of this technique, subgraphs are identified, which indicate to the sentences related to a particular topic. This can generate query-specific summaries[1][2] as well as identify the sentences relating to a particular topic.

C. *Latent Semantic Analysis*

The LSA [2][1][7] method is perhaps the oldest method used for extractive summarization. It was first proposed by Deerwester et al as method for automatic indexing and retrieval in order to improve the detection of relevant documents. The basic intuition behind the use of LSA in text summarization is that words that usually occur in related contexts are also related in the same singular space. LSA transforms sentence vectors from a term-space of non-orthogonal features to a concept-space of lower dimensionality with an orthogonal basis. LSA method basically consists of three different steps:

(1) *Input Matrix Generation:* The sentence is represented as a matrix in which the row represents the words and the columns represent the sentences. The cell value represents the importance of the word in that particular sentence.

(2) *Singular Value Decomposition:* The input matrix isthen decomposed into three other matrices such that

$$A = U\,\Sigma V^T$$

where U and V are orthogonal matrices and Σ is a diagonal matrix whose diagonal elements represent the relative importance of each concept dimension in the basis of the concept space. It has been mathematically proven that any non-null matrix can be decomposed into three such matrices.

(3) *Sentence Selection:* From the decomposed matrices, the sentences are selected from the summary. The selection is done using several algorithms such as Gong and Liu Approach, Steinberger and Jezek approach, Ozsoy approach, etc.

The advantage of using LSA vectors is that conceptual (or semantic) relations as represented in the human brain are automatically captured in the LSA. It has the ability to collect all trends and patterns from each of the sentence. However, it is time consuming and may cause polysemy issues.

D. *Neural Networks Method*

A neural network is a system of programs and data structures that approximates the operation of the human brain. A neural network usually involves a large number of processors operating in parallel, each with its own small sphere of knowledge and access to data in its local memory. Extraction using neural networks [2] [10] involves three phases

(1) *Network Training:* The training phase involves training the neural networks to learn the types of sentences that should be included in the summary. This is accomplished by training the network with sentences in several test paragraphs where each sentence is identified as to whether it should be included in the summary or not. This is done by a human reader. The network thus ‘learns’ what sentences to include and exclude from the summary.

(2) *Feature Fusion:* Feature combining which is also called as feature fusion, applies to the neural network which give away the hidden layer unit activations into discrete values with frequencies. This phase finalises features that must be included in the summary sentences by combining the features and finding fashion in the summary sentences.

(3) *Sentence Selection:* It uses the modified neural network to generate the summary. In the Selection or pruning phase, the network can be used as a tool to filter sentences in any paragraph and determine whether each sentence should be included in the summary or not. This phase is accomplished by providing control parameters for the radius and frequency of hidden layer activation clusters to select highly ranked sentences.

Each document is broken down into a series of sentences. Each sentence is then represented as a vector [F_1 F_2 …..F_n] where F_1…..F_n represent different features of the document as given below [10].

F_1 : Paragraph follows title.

F_2 : Paragraph location in document.

F_3 : Sentence location in paragraph

F_4 : First sentence in paragraph

F_5 : Sentence length
F_6 : Number of thematic words
F_7 : Number of title words
F_8 : Numerical data

E. *Fuzzy Logic Approach*

Fuzzy Logic is the branch of mathematic in which each element in a set possesses a ‘degree’ of membership to the set. In other words, we can say that a particular element is more of a member than another element. Fuzzy logic system [1] [2] design usually implicates selecting fuzzy rules and membership function. This method considers each characteristic of a text such as sentence length, similarity to little, similarity to key word and etc. as the input of fuzzy system. Then, it enters all the rules needed for summarization, in the knowledge base of system. A value from zero to one is obtained for each sentence in the output based on sentence characteristics and the available rules in the knowledge base. The obtained value in the output determines the degree of the importance of the sentence in the final summary.

The fuzzy logic system consists of four components:

(1) *Fuzzifier:*In the fuzzifier, inputs are translated into linguistic values using a membership function to be used to the input linguistic variables
(2) *Inference Engine:* the inference engine refers to the rule base containing fuzzy IFTHEN rules to derive the linguistic values.
(3) *Knowledge Base:* It consists of the set of rules which are used for the decision making process.
(4) *Defuzzifier:* It performs action opposite to that of the fuzzifier. the output linguistic variables from the inference are converted to the final crisp values by the defuzzifier using membership function for representing the final sentence score.

In fuzzy logic method, each sentence of the document is represented by sentence score. Then all document sentences are ranked in a descending order according to their scores. A set of highest score sentences are extracted as document summary based on the compression rate. It has been proven that the extraction of 20% of sentences from the source document can be as informative as the full text of a document. Finally, the summary sentences are arranged in the original order.

F. *Machine Learning Approach*

Also known as the probabilityapproach [1] [3], this is based on the principle of Baye’s Theorem of inverse probability. In this process, a set of documents and their extractive summaries are given. Sentences are included or excluded in the summary by calculating the probabilities of their relevance, using Bayes Theorem as follows:

$P\ (s\in<S \mid F1, F2, ..., FN) = P\ (F1, F2, ..., FN \mid s\in S)\ \ *P\ (s\in S)\ /\ P\ (F1, F2,..., FN)$

where s is a sentence from the document collection, F1, F2…FN are features used in classification. S is the summary to be generated, and P (s∈< S | F1, F2, ..., FN) is the probability that sentence s will be chosen to form the summary given that it possesses features F1,F2…FN.

IV. AVAILABLE TOOLS

A. *MEAD*

MEAD [4] [11] is the most elaborate publicly available platform for multi-lingual summarization and evaluation. The platform implements multiple summarization algorithms (at arbitrary compression rates) such as position-based, centroid-based, largest common subsequence, and keywords. The methods for evaluating the quality of the summaries are both intrinsic (such as percent agreement, cosine similarity, and relative utility) and extrinsic (document rank for information retrieval). MEAD has been successfully used to evaluate an existing summarizer, test a summarization feature, test a new evaluation metric, test a short-query machine translation system. It has also been used in major evaluations such as DUC.

B. *SUMMARIST*

The goal of SUMMARIST [11] is to provide both extracts and abstracts for arbitrary English (and later, other-language) input text SUMMARIST combines symbolic world knowledge (embodied in WordNet, dictionaries, and similar resources) with robust NLP processing (using IR and statistical techniques). It works in three phases: topic identification, interpretation and generation.

V. COMPARATIVE STUDY

A comparison of the different techniques of extractive summarization may be seen in the following table:

Table 1. Comparison of Extractive Summarization Methods

Method	Advantage	Disadvantage
1. TF-IDF Approach	Good Heuristic for determining keywords	No semantic relation mapping
2. Graph Theoretic Approach	Can generate query specific summaries	Accuracy will depend upon selection of affinity function.
3. LSA Approach	Semantic relations are captured	Polysemy issues (Inability to capture multiple meanings of a word)
4. Neural Network Approach	High speed	Requires human involvement in the initial stages
5. Fuzzy Logic Approach	Compression ratio is as low as 20%	Overhead of designing membership function
6. Machine Learning Approach	Simple	Statistical data is required

VI. Conclusion and Future Work

Due to the Internet revolution, a plethora of information has been made available to us. It is not feasible to manually read each of the available documents thoroughly. Instead, a summary of the document will aid the reader in deciding the relevance of the text, or extract the collective gist of a number of documents easily. Extractive document summarization, although simple for implementation can cause ambiguity in the summary and may result in miscommunication. On the other hand, abstractive methods generate a highly accurate summary but require complex heuristic algorithms. As a result of the survey, we can conclude that the graph theory based approach is the best one, as it is simple to implement and also generates query-specific summaries which are highly important in search engine techniques. With the growing pool of information on the web, Automatic Document Summarization has become essential and is a branch of data mining that is the need of the hour has tremendous scope for research in the future.

VII. ACKNOWLEDGEMENTS

I express my gratitude and thanks to Prof S. S. Sonawane, Associate Professor, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India for her guidance and mentoring. She has been a great help throughout the entire process. This paper would not have emerged without her.

References

1. Rajvardhan Oak and Akshay Patel, “Automatic Document Summarization: A Survey of Techniques”, Proceedings of the National Conference on Recent Trends in Computer Engineering.

2. Vishal Gupta and Gurpreet Singh Lehal, “A Survey of Text Summarization Extractive Techniques”, Journal of Emerging Technologies in Web Intelligence, Vol. 2, No. 3, August 2010.
3. Kam-Fai Wong, Mingli Wu and Wenjie Li, “Extractive Summarization Using Supervised and Semi-supervised Learning”, Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)
4. Surajit Karmakar, Tanvi Lad, Hiten Chothani, “A Review Paper on Extractive Techniques of Text Summarization”, International Research Journal of Computer Science, Issue 1, Vol. 2
5. Archana AB, Sunitha. C, “An Overview on Document Summarization Techniques”, International Journal on Advanced Computer Theory and Engineering (IJACTE).
6. Mani I., Maybury M.T. “Advances in automated text summarization”, Cambridge: MIT Press, 1999.
7. Rasha Mohammed Badry, Ahmed Sharaf Eldin, Doaa Saad Elzanfally, “Text Summarization within the Latent Semantic Analysis Framework: Comparitive Study”, International Journal of Computer Applications, Vol. 81, No. 11
8. Mr. Sarda A.T., Mrs. Kulkarni A.R, ”Text Summarization using Neural Networks and Rhetorical Structure Theory”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6,
9. S S Sonawane, P A Kulkarni, “Graph based Representation and Analysis of Text Document: A Survey of Techniques”, International Journal of Computer Applications 96(19):1-8, June 2014.
10. Nikita Munot, Sharvari Govilkar, “Comparative Study of Text Summarization Methods”, International Journal of Computer Applications, Vol. 102, Issue 12
11. Rasim ALGULIEV, Ramiz ALIGULIYEV, “Evolutionary Algorithm for Extractive Text Summarization”, Intelligent Information Management, 2009, pp 128-138. doi:10.4236/iim.2009.12019

BIOGRAPHY

Rajvardhan Oak is a Second Year Engineering student at Pune Institute of Computer Technology, Pune. His interests lie in data mining, warehousing, text processing and cryptography. His other hobbies are reading and writing and learning new languages.