2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)
Dec. 22nd, 2017
(Iran University of Science and Technology) – Tehran, Iran

An Overview on Extractive Text Summarization

Shohreh Rad Rahimi
Department of Computer Engineering
Marlik Higher Education Institute
Nowshahre, Iran
Shohreh.radrahimi@gmail.com

Ali Toofanzadeh Mozhdehi
Faculty of Computer Engineering, Qazvin Branch,
Islamic Azad University
Qazvin, Iran
Mozhdehi@qiau.ac.ir

Mohamad Abdolahi
Iranian Academic Center for Education, Culture and Research (ACECR)
Mashhad- Iran
Mabdolahi512@yahoo.om

Abstract— **With the increasing of online information and recourse texts, text summarization has become an essential and more favorite domain to preserve and show the main purpose of textual information. It is very difficult for human beings to summarize manually large documents of text. Text summarization is the process of automatically creating and condensing form of a given document and preserving its information content source into a shorter version with overall meaning. Nowadays text summarization is one of the most favorite research areas in natural language processing and could attracted more attention of NLP researchers. There are also much more close relationships between text mining and text summarization. According to difference requirements summary with respect to input text, established summarization systems should be created and classified based on the type of input text. In this study, at first, the topic of text mining and its relationship with text summarization are considered. Then a review has been done on some of the summarization approaches and their important parameters for extracting predominant sentences, identified the main stages of the summarizing process, and the most significant extraction criteria are presented. Finally, the most fundamental proposed evaluation methods are considered.**

Keywords— Text summarization; Text mining; Fuzzy text summarization; Statistical text summarization; Text clustering.

I. INTRODUCTION

With an increasing amount of data available on the web, rising of news websites, publication of various electronic books, and a significant growth in the number of published articles in different fields of study, one of the main challenges for researchers of 21st century has been that of accessing accurate and reliable data. The widespread volume of information available on one hand and time limitation on the other has directed the researchers to the interesting area of summarizing texts and powerful system to summarize documents. Given that much researches paper has been done on the summarization subject and many articles have been published about it. But there is still a lot of weakness in summarization field and there is a large gap to achieve an efficient system that can acts like a human agent. Problems outlined in the Persian language are far more than other languages. The complexity of language and lack of precision tools are current problems facing the Persian language processing. Therefore, the review of operations and procedures performed on other languages, considering the semantic field and use the semantic relationship, using tools such as graph theory; statistical methods, fuzzy logic and data mining techniques can make a significant contribution to Persian language processing.

A text made up of components such as words, phrases and sentences connected completely and meaningfully together. One of the main areas of natural language processing is text mining, which means discover and extract new information from the documents. Text mining is analysis the documents to extract valuable hidden patterns from the text. It is also involves the detection of the connection between words and sentences, classify and summarize texts. The main propose in this research is an overview on text summarization.

More than half a century passed since the first research on the automatic text summarization (1). Studying on text summarization systems emerged in 1950ʺs which focused on the some basic features of the text such as the position of the sentences in the text due to lack of powerful computers and other problems in natural language processing. Since then, a lot of methods with powerful tools were presented to simulate text processing, such as human brain.

II. A REVIEW OF TEXT MINING AND ITS RELATIONSHIP WITH A TEXT SUMMARIZATION

A significant portion of available information is stored in text data. Text databases are growing rapidly due to increasing the amount of information in electronic form. Today, most of the data available in the industry, business and other organizations are stored in the form of text database (2).

978-1-5386-2640-5/17/$31.00 ©2017 IEEE

Text mining is a research field that has emerged from the combination of several other research areas such as data mining, natural language processing and information retrieval. But there are some differences between data mining and text mining.

Data mining often deals with structured data, but text mining involved with unstructured or semi-structured data. On the other hand, even if there is still some construction in the text, there is another reason that text mining is much more difficult than data mining. Unlike other data, such asnumerical data, there are some semantic concepts in text and it can be difficult to model them with traditional knowledge structures. For example, there are some difficulties such as words with same meaning and different dictation and different meaning with same dictation.

A. Text mining applications

There is a wide definition of text mining; As a result, there are different theories about its applications. Text mining is a new field in NLP but its software analyzes have been available since late 1990. Among the most common applications of text mining can be noted to the search engines. They can find the most relevant documents whether the user types a misspelled word or phrase. Some other applications of text mining that can be mentioned are:

Spam identification: Analyzing the title and content of email to determine whether it is spam.

Supervision: Monitoring the behavior of a person or group of people. There are some software that can detect and control people behavior from telephone, internet and other communication devices.

Aliases identification

Concepts relationship: Occurrence of some words dependent on some other words.

Search and Retrieval

Classification and clustering data

Text summarization

There are some big problems in text mining. A biggest problem is that any document is a very large set of words. If any word is assumed an element of a vector and considers different presence scenarios in the text, then we will find that we are dealing with a NP-Hard problem with high dimensional area. Word reduction is one of the preprocessing operations that lead to reduce the dimension of the problem. Preprocessing and word reduction should be so efficient and powerful, because eliminating words may be lead to some text noises such as grammatical error. Apart from reducing the size of the text, processing consists of some other steps as follows:

Case Folding: At this stage all words with big and small letters came in a uniform. Also, always try to converted uppercase letter to lowercase.

Stemming: This stage it root extraction of words without considering different scenarios such as singular or plural, the time and all prefix and suffix. This requires knowledge of a particular language and many algorithms have been proposed for each language. Examples of these types of words in the English language can be compressed and compression that convert to compress.

Stop Words: Some frequent words without concept. Examples of these types of words in the English language such as do, does, will, and so on.

N-grams: N-grams are a sub-set of N words come together. They should be protected, but it is better to come in the form of uniform.

Tokenization: Tokenization is a dividing the text to smaller units, which are often words. But in some cases it is difficult to define the word and its scope. A word as a unit of text has the following properties:

- Each set of continuous characters is always a word. These characters can even figure as well.
- Range of a word can be a whitespace character, symbols or end of the line.
- Sometimes whitespace characters and punctuation is not range word or token.

In languages such as English and Persian that words separated with whitespace, defining the boundary words is not always an easy task. For example acronyms that has point, words in hypertext links, and the words that are separated with symbolic characters and words with space character like New York (3).

One of the important functions of the text mining is text clustering. The main aim of text clustering is to place the similar sentences in same clusters. The number of clusters can be determined by the user or automatically by the program. The first step of text clustering is dividing the text to its component and separated the sentences.

B. Text mining and text summarization

It is obviously that it cannot be separated text mining and text summarization. Several studies have been conducted in the field of text mining and text summarization. All of them show that the text summarization is also kind of mining and exploring to find the important parts of a document. The main problems facing to text processing and summarizing are lack of qualified linguist expert work with programmers, grammar based of linguistic theories, and the lack of reasoning and thinking word processors machines.

The first problem can be overcome with employing linguistic expert. The second problem according to the semantics can be resolved by revising the language theories. But it is difficult to create machines with reasoning and thinking power.

On the other hand, dealing with high volume data on internet, especially when the goal understands the main concept of documents, using machines instead of humans can be easier, faster and more reasonable.

At the beginning consider the different views on the automatic text summarization and the summaries created by the man. There are three main views that are: no differences between human and machine summary, preference of human summary on machine summary and preference of machine summary on human summary.

The theory of the superiority of processing and summarizing man believes that human mind more powerful than machine processor. Decision-making power and choice of current machines are less than human mind. The theory of the superiority of machine processing and summarizing believes that in the near future computers can process the linguistic information with more speed and accuracy than human. The reason for this claim is the more quickly operation of math and machinery computing than a human.

There are three major advantages of automatic generation of summary by the machines. The advantages are: summary size is controllable, its content is predictable and it can be determined that any part of the summary related to which part of the original text.

III. Different criteria in the text summarization systems designing

There are various approaches to text summarization, some of which have been extant for about 50 years. Text summarization approaches can be categorized in different ways according to the various measures and features. For example, according to Hovy and Lin [4], the measures and features are related to input, output, purpose and result in different types of summary. In the following various methods have been considered in general classification (5).

A. Summarization based on output summary

Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences or paragraphs from the original text and gathering them into shorter text. The importance of sentences is decided based on some statistical and linguistic features of sentences, extract and placed in the output text. In this paper, the more emphasis is on extraction techniques.

An abstractive summarization attempts to extract the main concept of the text in clear natural language without necessity to use text phrases. Each abstractive summarization consists of comprehension part to interpret the text and find the new concepts and production part to generate new shorter text with most important information from the original document. In this method, sentences could be omitted or changed or even new sentences could be generated. It should be noted that this method is very complicated and even more complicated than machine translation.

B. Summarization based on details

Text Summarization methods also can be classified into another classes named Indicative and Informative summarization. In indicative summary the main idea of the text is presented and it is usually about 5 to 10 percent of the original text. This kind of abstraction is used to encourage the reader to read the original text. For example a brief summary of a movie or a story in the context of its advertising which only leads to further questions and encourage the readers to watch the film and read the story. Informative summary consist of the main abstraction and the important issues of the text. This kind of summary is between 20 and 30 percent of the original text and contains all the main points of the text.

C. Summarization based on contents

Another classification in text summarization is content base classes. This type of classification can be divided into two categories: Generic and Query-Based. Generic summarization is not depended on topic of text and it is assumed that the reader does not have any basic knowledge about the text. Generic summary contains all aspects and important issues of main text and readers are able to achieve a thorough understanding of the subject without prior knowledge of the text.

But in Query-Based summary, it is assumed that the reader has a general knowledge about the topic and just looking for specific information in the text. In this case based on the user's question, a related summary is created. Most of these summarization systems are extractive.

D. Summarization based on limitation

There is another category based on the limitations of input text. The mentioned group has three categories that which are: Independent, Domain dependent and Genre specific.

Independent summarization is something like generic summary, accepted every text of each field and generate a general summary regardless of the text scope or type.

Domain dependent summarization accepts texts with specific field of literature and type. There are many specific text patterns such as News, science text, fiction, sports. Generated summary of domain dependent systems are according to the input text type. Genre specific summarization is trying to summary much more specialized field. This group can be abstract specific literature such as sports news, texts related to the field of geography, political News and etc.

E. Summarization based on the number of input texts

Automatic summarization systems can be divided other two main categories based on their input text. The mentioned categories are Single document and Multi document summarization. The input of single-document summarization systems is only one text but multi-document summarization, which is very popular these days, is the improvement of single document summarization to collections of related documents. The main propose of multi-document summarization is summarizing texts and removing redundancy and considering the similarities and differences in the information content of different documents. Multi document text summarization accepts multiple documents with common scope in a different perspective and closely tied to answering systems and search-based summarization (6). There are two major approaches to summarize multiple documents. The first approach uses the usual methods of single document summarization and summarizes each document separately. Then combining all of the summaries together and tried to remove redundancy by overlap similar sentences to produce the final summary. There are some proposed methods too which behave created single summarize as inputs and then merge them to create main summary (7).

The second approach is specifically designed for multiple documents. In the mentioned approaches, all documents assumes as one document and all the important sentences are extracted from all of the documents together using methods like graphs or clustering. This approach is more challenging, intelligent and complicated. An example of the second approach is the SUMMONS system which extracts and combines information from multiple sources and passes them to a language generation component and produces the final summary. In general, the complexity of a single document model is far less than multi document models (8).

F. Summarization based on language acceptance

A text summarization system also can be divided to mono-language and multi-language. Mono language summarization only accepts documents in a specific language such as English, Persian. But a multi-language summarization adopted by different languages and can be able to summarize them.

IV. Similarity measures of extractive summarization

As mentioned earlier, extractive summary is using main text sentences to crate summary. This means that more important sentences are found and transferred to the summary. To determine the important sentences some parameters should be considered. After determining the parameters value, sum of them are intended as the final weight of each sentence. Finally sentences with the maximum weight based on compression rate are extracted and by ordering presence in main text transferred to the summary. Given that some mentioned parameters reduce the importance of a sentence their value is calculated as a negative (9). Some of The parameters and features are described below.

A. Content keyword feature

Keywords are often noun and determined by tf * idf criteria. Sentences containing keywords have more chances to exposure in summarized output. Other methods have been proposed for extracting keywords. Some of these methods are word analysis of morphological, statements extraction and scoring them, and noun phrases extraction, clustering and ranking them. Word morphological analysis plays an important role in natural language processing and helps to resolve the ambiguity in the words. The mentioned criteria study of root words, prefixes and suffixes attached to a word.

B. Similarity of sentence and title of document

Similarity of sentence and title of document is the number of common words between title and sentence. Sentences including title words have high importance and more chance to place in output summary. In most of the proposed methods the criteria is calculated as Eq. (1).

$$X_i = |Si \cap T| / \sqrt{|Si| * |T|} \qquad (1)$$

In Eq. (1), the numerator is the number of similar sentence and title words and the denominator is the square root of the product of the title length and sentence length.

C. Sentence location in the document

This measure is based on the assumption that sentences occurring in initial and end position of both text and individual paragraphs have a higher probability of being relevant to main topic of the document.

The experimental results showed that the best correlation between the automatic and human-made extracts was achieved using first and end sentences in summarized output. But the first sentence is more important than last one. By reading the first sentence of every paragraph can be understand its main topic. The Eq. (2) is used to calculate the importation rate of sentence location:

$$PSi = 1 + (0.5 * (PSi + C) * \log(0.5 * (PSi + C))) + ((1 - 0.5 * (PSi + C)) * \log(1 - 0.5 * (PSi + C)))$$
$$Xi = PSi / Maxj{=}1{:}n\ (PSi) \qquad (2)$$

PSi is effective value of the relative position of the ith sentence that calculated based on entropy. The C is a constant value between zero and one and Xi is a sentence location value.

D. Important words and phrases

There are some words and phrases that increase the sentence importance. Numbers (including date, time, price, percentage and weight) and proper names (including person, place, and time) have specific important information and determining their importance depends to texts type. The score rate of a sentence regard to the important words and phrases are obtained according to the Eq. (3) by dividing the number of these words on the total words in sentence.

$$Xi = |Simp| / |Si| \quad (3)$$

E. Capitalized words and acronyms

Capitalized words or acronyms are having special concepts. The words with capital letters are in the Latin alphabet languages and there are no such words in Persian. But there are acronyms in all languages and in Persian language can be used them to determine the significance of a sentence.

F. Cue-Phrase

Cue-Phrases are some phrases like “finally”, “as result”, “in this paper”. Statements containing these words certainly including important topic and must be preserved. There are some words that lead to increase the importance of a noun and its sentence. For example the words like Dr., Mr., and Miss.

There are some other words placed at the beginning of a sentence. These words lead to decrease the importance of the sentence. The mentioned sentences are come to complete the meaning of previous sentence and can be ignore it. Some example of them are “because”, “for example”, “therefore” and “thus”. The weight of statements containing these words in both positive and negative form calculated as Eq. (4).

$$Xi = (|Si| + |Spi| - |Sni|) / 2*|Si| \quad (4)$$

In the Eq. (4), Si is the number of words in sentences i, Spi is the number of positive Cue-Phrases and Sni is the number of negative Cue-Phrases.

G. Specific words

Specific words are certain words that have special meaning and may vary by keyword in the text. These words can be scientific issues such as "mathematics" psychology ". They increase the importance of sentence.

H. Sentences containing words with different fonts and font effects

Some of the words that are important or specific issues charged with capital letters started or bold, italic or underlined. They have a higher chance of exposure sentences in the output.

I. The continuity and similarity of a sentence with other sentences

For each sentence in the text resemblance with other sentences are considered and sentence with more resemblance has high score. The process is repeated for all sentences of text and sentences with higher scores are more likely to have a presence in summary. The similarity of two sentences is similarity between contained words, sentence length and other criteria mentioned before.

J. The similarity of a sentence with a paragraph topic sentence

In this section, the similarity of each sentence with paragraph topic sentence is considered. Usually the topic sentence in each paragraph is a first sentence in paragraph. Similarity measures in this section are the same criteria like in the previous sections.

K. Specific terms

Specific terms and phrases in text are marked by symptoms such as " ", << >>, etc. Review of the literature summarized by the human shows that the sentences containing these phrases are of more important. Sentences with specific terms scores as Eq. (5):

$$X_i = |S_{ip}| / |S_i| \quad (5)$$

S_{ip} is the number of specific terms in the sentence, S_i length of sentence and X_i scoring rate of the sentence based on mentioned criteria.

L. The length of sentences

Sentence length is an important criterion in text summarization. The best sentences for summary output are the ones with average length. Sentences containing less than a pre-defined number of words are not included in the abstract. But sometimes the effects of previous measures increases the short sentence importance and cannot be absolutely eliminated it. The Eq. (6) is determined and rated the effect of sentence length. The sentences out-of-threshold with negative rate is also considered.

$$Xi = - RSi * Log (RSi) – (1 – RSi) * Log (1 – RSi)$$
$$RSi = |Si| / Max\ j=1:n(|Si|) \quad (6)$$

In the Eq. (6) Si is sentence length and Max j=1:n(|Si|) is maximum sentence length in the document.

M. Pronouns

Sentences with pronouns have less importance than the one with nouns. Pronouns refer to nouns that described in the previous sentences. The position of the pronoun in the sentence is also important. For example, a sentence with pronoun at the beginning is less important sentence. But if the pronoun is placed at the middle and end of a sentence, it is a little more important sentence. The score of sentence including pronoun calculated as Eq. (7):

$$Xi = - (BPi + |SPi|) / (1 + |Si|) \quad (7)$$

In Eq. (7), the | Spi | is the number of pronouns in sentences, Si is the sentence length and BPi is a constant parameter. If the pronoun location is in the first three word in the sentence, BPi = 1, otherwise BPi = 0.

V. Different approaches to text summarization

Although summaries created by human are often abstractive, but the most successful machine summary are extractive. Because of the semantic ambiguity in natural language processing, a good extractive summary is much better result than abstractive summarization. In fact it is far away to create an ideal automatic abstractive summarization method. The first automatic text summarization and the most important new proposed methods are extractive systems that extract important sentences based on heuristic features such as sentence positions in the text, the frequency of words in it and some important keywords related to text (1) (11). To create a system with high-quality due to a different type of language and text input, various methods and algorithms in machine learning and natural language processing is proposed. Most of extractive summary methods are emphasis on key sentences. The differences between methods are the different algorithm used to detect, identify, extract and the way to put sentences in the summary. In the following the most important approaches will be discussed.

A. Statistical approaches

Using statistical methods to summarize the text is an effective approach that has been used in many articles. In statistical methods, the important sentences are selected based on word frequency, indicator phrases and other features regardless of the meaning of the words that mentioned in previous sections.

There are several methods to determine the key sentences such as, The Title Method (11), The Location Method (12), The Aggregation Similarity Method (13), The Frequency Method (14), TF- Based Query Method (15), and Latent Semantic Analysis (15). But the most common methods are Classifier Bayesian (Bayesian Classifier) and communication concept relations approaches. One of the most famous statistical methods to summarize the text is LSA method (16). LSA is an algebraic-statistical method that extracts hidden structures meaning of words and sentences (8). This approach is an unsupervised method that extracts text structures just by information of the words in sentence without the need of any other knowledge. The idea in LSV is words that they are shared between different sentences are the reason of meaning dependence. LSA also be able to show the meaning of words and vocabulary simultaneously. It uses algebraic singular value decomposition (SVD) method to determine the relationship between sentences and words. Addition to being able to model the relationship between words and sentence, SVD also can reduce the noises and lead to improve the accuracy.

Summarization algorithm based on LSA method consists of three steps: create the input matrix, applied the SVD method on the created matrix and sentence extraction.

LSA also has some limitations. The most important of them are as follows:

- The algorithm does not use the information about the words arrangement in sentences, grammar and morphology relationship. However, this information can be useful to better understand words and sentences.
- The algorithm does not use any word knowledge and word database.
- By increasing the number of different words and heterogeneous data, the performance of the algorithm greatly reduced. Performance reduction is due to the time and memory complexity of the SVD method.

B. Lexical Chain based approaches

Lexical chain produces a presentation of text contiguous structures. Basically, it uses the word net database to determine the connection between the terms and then creates a continuum between these terms. The first computational model of lexical chains was presented by Moris and Hirst in 1991 and the first use of lexical chains in text summarization was proposed by Barzilay and Elhadad (16) (17).

The scores that considered for terms are according to the type and number of the relationship of chains set. The final summary is included sentences with strongest chain connection. Lexical chains can be computed the semantically relation of words. It cans also identify the synonyms and hyponyms to place them in a group into the same lexical chain. Lexical chain also used for information retrieval and grammatical error corrections.

There are two drawbacks in lexical chain approaches. The first one is ambiguity in the word chain. If some of words have semantic ambiguity, the created chain also will have semantic ambiguity. The second drawback is the relation of created chain to main topic. All chains are not related to the main topic.

C. Graph based approach

Graph based approach provides text summarization methods using graph theories. After common preprocessing steps such as stemming and stop word removal, sentences in the documents are represented as nodes in a directed graph. Sentences are connected to each other by edges according to sentence. The basic idea of graph-based approaches is something like voting. As result when an edge connected a node to other node, it means that votes to it and whatever the input degree of a node is high, it has a higher priority. In the mentioned method there is also voting score. If a node has greater output degree, its importance increases. The importance degree of each node is calculated by S(Va) as Eq.(8) for all nodes in a graph recursively. The process continues until coverage and no change in S(Va).

$$S\,(V_a) = (1 - d) + d * \sum_{V_b \in In\,(V_a)} \frac{S\,(V_b)}{out\,(V_b)} \qquad (8)$$

In the Eq. (8), V is nodes, E is edges, in(Va) is the number of Va input edges, out(Va) is the number of Va output edges and d is an input parameter between zero and one. S(Va) is Va score and S(Vb) is Vb score. The algorithm can also be applied on an undirected graph, but the output summary is more different with complexity of time.

An effective graph based summary proposed algorithm is Text Rank algorithm (18). It uses an unsupervised method to extract key words and sentences with scoring to nodes based on previous mentioned similarity measures (16). The algorithm is started with an optional node values and recursively repeated until coverage to predefined threshold. The optional nodes value has no effect on the final scores.

D. Cluster based approaches

Some automatic summarization systems are used clusters to produce significant summaries. In this approach, different clustering algorithms are applied for dividing the text into tokens such as words, phrases, sentences and even paragraphs. Cluster-based approach is an extractive summarization and the most similar sentences are based on mentioned similarity measures are placed in the clusters. The biggest cluster is selected as main topic and its sentences are extracted and placed in summary output. Euclidean distance, Cartesian similarity, cosine and some other similarity measures are used for defining the similarity and dissimilarity of clusters (16).

A. Agrawal & U. Gupta, 2014 proposed K-means clustering algorithm to place important sentences within clusters and chose the biggest cluster as main topic (19). The documents are represented using term frequency-inverse document frequency (TF-IDF). In the context, term frequency (TF) is the average number of occurrence (by document) in the cluster. The topic is represented by words of which the value TF-IDF is higher in the cluster. The selection of the important sentences is based on the similarity measures of the sentences with the topic of cluster.

Two optimization ideas are proposed in their algorithm. The first one is a method to calculate sentence score and the second one is a way to get the optimal number of clusters. To calculating sentence score the algorithm uses TF-IDF and the sentence length (X).

$$\text{Score (x)} = \sum_t tf - idf_t \;/\, |x| \qquad (9)$$

One of the most important issues in k-means clustering is determining the optimal number of clusters. Given that in this method largest cluster is considered as main topic, input text size has a direct impact on determining the number of clusters. For example a big K lead to small clusters and as result small and disperse summary with very low correlation and small K lead to big and dense clusters and as result low compression text.

N.I.Meghana & M.S.Bewoor proposed another technique to create a query based summarization system using clustering methods (20). They used Expectation Maximization Clustering (EM) algorithm and implementation methodology is divided into two sections. The EM algorithm has been implemented in the first part and query-based summarization is done in second part. The proposed method has been used the Word net.

A. Fuzzy logic based approaches

Fuzzy logic based approaches consider each characteristic of a text as the input of fuzzy system. These methods only used fuzzy logic to detect and extract the important sentences. Fuzzy text summarization methods are differ based on differences in extracting features, fuzzy rules, linguistic variables, membership functions, fuzzification and defuzzification methods. Some changes in numerical values lead to produce better linguistic variables (21). F. Kyoomarsi, H. Khosravi proposed a method thatsummarizes the text in two stages (22). The first stage is pre-processing and the second stage is performing fuzzy analysis on text. The features that they extracted are: the number of common words including the sentence and title, the number of words in sentence, the similarity of a sentence with a paragraph topic sentence and similarity of sentence and first sentence in paragraph.

VI. INTEGRATION SIMILAR STATEMENTS IN EXTRACTIVE SUMMARIZATION

Redundancy in a series of sentences in texts, especially on the web, is one of the problems in natural language processing and on the other hand, creates a new research opportunities in text summarization area. There are many sentences with similar meaning and concept in every document. Because of determining the number of sentences in the summary by the user, statements with a similar concept are placed in summary. There are also some other important concepts in the main text and due to the limitation number of sentences in summary, they do not select and extract. So many sentences with one concept placed in summary and many other important parts are ignored. The problem is more common in multi document summarization (25). Similar concept sentences are also created the ambiguity in question answering systems.

There are some proposed methods to combine similar sentences to extract much more important sentences from the main text. One of the most important sentence fusions is dependency trees. Katja Filippova is proposed a directed acyclic graph (DAG) to sentence fusion (26). Clustering approaches are also can be used for combining similar sentences. In these approaches, small clusters are considered to place very similar sentences and then each cluster can be combined to one sentence.

VII. Assessment and Evaluation of Summarizing Methods

Due to extensive automatic summarization methods, assess the accuracy of created methods is important. It can be said that the process of evaluation is much more difficult than summarizing the text. In most cases there is not an ideal summarize of a document. The main problem in summarizing assessment is using extensive criteria and absence of a standard method for evaluation. Such problems are rare in other natural language processing cases.

Most of the developed methods are used two approaches for evaluation. The first approach is to judge by human and the second approach is compared to reference summary.

In the first approach summary generated by machine can be compare with the same summary generated by a man. The problem with this approach is the different personal tastes and opinions. To reduce the impact of the effects, automate summary compares with more than one summary generated by humans and it is time consuming.

In second approach, there is more than one original text with their summary in a dataset. The original document and its optimized summary is a good base for evaluate the proposed summarization algorithm.

One of the most famous evaluation methods is ROUGE-N. The mention evaluation method compares automate generated summary with summaries generated by the human (23) (24). ROUGE-N uses three evaluation criteria: Precision (p), Recall (R) and F1-measure and they are calculated as Eq. (10):

$$\begin{aligned} P &= (|sum\_ref \cap sum\_cand|) / |sum\_cand| \\ R &= (|sum\_ref \cap sum\_cand|) / |sum\_ref| \\ F1 &= 2PR / (P + R) \qquad (10) \end{aligned}$$

In the Eq. (10) sum_ref determines summary extracted by experts and sum_cand determines summary extracted by system.

Another evaluation criterion mentioned in ROUGE-N is calculated by the Eq.(11). The measure compares number of N-grams in machine summary and human summary.

$$\text{ROUGE-N} = (\sum S \in sum_{ref} * \sum N - gram \in s * \text{count\_match (N-gram)}) / (\sum S \in sum_{ref} * \sum N - gram \in s * \text{count (N-gram)}) \qquad (11)$$

In the Eq. (11) N is the number of N-grams, count_match (N-gram) is maximum number of N-grams which are in machine summary and human summary simultaneously. count(N-gram) is also the number of N-grams in human summary.

VIII. Conclusions

Text summarization is one of the most exciting research areas in natural language processing. It is an open research areas and a lot of researches are having been done about it. Text summarization can be classified into different groups and approaches and the most notably of them studied in this paper. In this study, at first, the topic of text mining and its relationship with text summarization are considered. Important designing criteria in text summarization systems are presented. Different approaches for summarization and important parameters needed to rate important sentences are introduced. Finally important evaluation approaches are introduced.

References

[1] H. Luhn, “The automatic creation of literature abstraction”, IBM Journal of research and development, Vol 2, pp.159-165, 1958.

[2] V. Gupta, G.S Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, Vol. 1, no. 1, 2009.

[3] G.O Makbule, I. Cicekli, F. Nur Alpaslan, “Text Summarization of Turkish Texts using Latent Semantic Analysis”, 23rd International Conference on Computational Linguistics, pp. 869–876, 2010.

[4] E. Hovy, C. Lin, ”Automated Text Summarization and the SUMMARIST System, In Advances in Automatic Text Summarization”, MIT Press, pp. 81-94, 1999.

[5] S. Gholamrezazadeh, M. Amini, B. Gholamzadeh, “A Comprehensive Survey on Text Summarization Systems”, IEEE, 2009.

[6] T. Hirao, Y. Sasaki, H. Isozaki, “An extrinsic evaluation for question-biased text summarization on qa tasks”. In Proceedings of NAACL workshop on Automatic summarization, 2001.

[7] J.V Goldstein, J. Mittal, J. Carbonell , M. Kantrowitzt, “Multi-document summarization by sentence extraction”. In Proc. of the ANLP/NAACL Workshop on Automatic Summarization, 2000.

[8] G.O. Makbule, “Text Summarization using Latent Semantic Analysis”, M.S thesis, Middle East Technical University, 2011.

[9] V. Gupta, G. Lehal, “A Survey of Text Summarization Extractive Techniques”, Journal of emerging technologies in web intelligence, VOL. 2, NO. 3, 2010.

[10] H. Edmundson, “New Methods in Automatic Extracting”, Journal of the Association for Computing Machinery, Vol16 (2), PP. 264-285, 1969.

[11] K. Youngkoong, S. Jungyun, “An Effective Sentence Extraction Technique Using Contextual Information and Statistical Approaches for Text Summarization”, Pattern Recognition Letters, 2008.

[12] M. Wasson, “Using leading text for news summaries: Evaluation results and implications for commercial summarization applications”, in Proc. 17th International Conference on Computational Linguistics and 36th Annual Meeting of the ACL, pp.1364-1368, 1998.

[13] W. al-sanie, “Towards an infrastructure for Arabic text summarization using rhetorical structure theory”, M.S Thesis”, Department of computer science. King Saud university, Riyadh, Kingdom of Saudi Arabia, 2005.

[14] G. Salton, “Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer”, Addison-Wesley Publishing Company, 1989.

[15] J. Bellegarda, “Exploiting latent semantic information in statistical language modeling,” in Proc. IEEE, Vol. 88, No. 8, pp. 1279-1296, 2000.

[16] N. Alami, M. Meknassi, N. Rais, “Automatic Texts Summarization: Current State of the Art”, Journal of Asian Scientific Research, Vol 5(1), pp 1-15, 2015.

[17] R. Barzilay, M. Elhadad, “Using Lexical Chains for Text Summarization”, the MIT Press, pp. 111-121, 1999.

[18] R. Mihalcea, “Graph-based ranking algorithms for sentence extraction, applied to text summarization”, in Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, 2004.

[19] A. Agrawal, U. Gupta, “Extraction based approach for text summarization using k-means clustering”, International Journal of Scientific and Research Publications, Vol 4, Issue 11, 2014.

[20] N.I. Meghana, M.S. Bewoor, M.S, S.H. Patil, “Text Summarization using Expectation Maximization Clustering Algorithm”, International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 4, pp.168-171, 2012.

[21] L. Suanmali, M. Salem, B. Salim, N. Salim, “Sentence Features Fusion for Text summarization using Fuzzy Logic, IEEE, pp.142-145, 2009.

[22] F. Kyoomarsi, H. khosravi, E. Eslami, M. Davoudi, “Extraction-based Text Summarization using Fuzzy Analysis”, Iranian Journal of Fuzzy Systems, Vol. 7, No. 3, pp. 15-32, 2010.

[23] M. Pourvali, A. Abadeh Mohammad, “Automated text summarization base on lexical chain and graph using of word net and Wikipedia knowledge base”, IJCSI International Journal of Computer Science Issues, No. 3, vol. 9, 2012.

[24] H. Shakeri, S. Gholamrezazadeh, M. Amini Salehi, F. Ghadamyari, “A New Graph-Based Algorithm for Persian Text Summarization”, Computer Science and Convergence, Lecture Notes in Electrical Engineering, 2012.

[25] R. Dragomir, J. Budzikowska, “Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies”, In Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization, pp 165–172, 2000.

[26] K. Filippova, M, Strube, ” Sentence Fusion via Dependency Graph Compression”, in Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp 177–185, 2008.