International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Research Paper of Extractive Text Summarization

[1]Arpit Jaiswal, [2]Prof. Shweta Nigam

[1]Student, [2]Faculty

[1,2]Department of Information Technology, Shatayu College of Professional Studies, Nagpur, India

ABSTRACT:

Text summarization may be a process of extracting or collecting important information from original text and presents that information within the sort of summary. Text summarization has become the need of the many applications for instance program, business analysis, market review. Summarization helps to realize required information in less time. This paper is an effort to summarize and present the view of text summarization from every aspect from its beginning till date. The 2 major approaches i.e., extractive and abstractive summarization is discussed intimately. The method deployed for summarization ranges from prearranged to linguistic. In Indian many languages also the work has being done, but presently they're in infancy state. This paper provides an abstract view of this scenario of research work for text summarization.

Introduction

Text summarization may be a process of extracting or collecting important information from original text and presents that information within the sort of summary. In recent years, need for summarization are often seen in various purpose and in many domain like news articles summary, email summary, short message of stories on mobile, and knowledge summary for businessman, officialdom , researchers online search through program to receive the summary of relevant pages found, medical field for tracking patient's medical record for further treatment. The automated summarization of text may be a well- known task within the field of tongue processing (NLP). Considerable achievements in text summarization are obtained using sentence extraction and statistical analysis. Text summarization approaches are often generally divided into two groups: extractive summarization and abstractive summarization. Extractive summarizations extract significant sentences or phrases from the first documents and group them to supply a summary without changing the first text. An extractive text summarization system is proposed supported POS tagging by considering Hidden Markov Model using corpus to extract important phrases to create as a summary. There are two different groups of text summarization: analytical and educational. Inductive summarization only represents the most idea of the text to the user. The standard length of this sort of summarization is 5 to 10 percent of the most text. On the conflicting hand, the informative summarization systems give brief information of the most text. The length of informative summary is 20 to 30 percent of the most text.

Problem Statement

1. Unstructured data is navigated to look and provides proper results.
2. We'd like to decrease the large files and focus key points to form summary.
3. Text summarization is required to travel through large data to offer the details as outcome.
4. This volume of text is priceless source of data and knowledge, which should be effectively summarized to be useful.

   During this problem, the most objective is to automatic text summarization for lighting more about processes.
5. Internet provides huge online information which is difficult to read and understand.

Extractive text summarization Techniques

1. Term Frequency-Inverse Document Frequency Method

Sentence frequency is defined because the number of sentences within the document that contain that term. Then this sentence vectors are scored by similarity to the query and therefore the highest scoring sentences are picked to be a part of the summary.

2. Cluster Based Method

Sentence selection is predicated on similarity of the sentences to the theme of the cluster (Ci). subsequent factor that's location of the sentence within the document (Li). The last factor is its similarity to the primary sentence within the document to which it belongs (Fi).

3. Text Theoretic Approach

Text theoretic representation of passages provides a way of identification of themes. After the common pre-processing steps, namely, stemming and stop word removal; sentences within the documents are represented as nodes in an undirected Text.

4. Machine Learning Approach

The summarization process is modeled as a classification problem: sentences are classified as summary sentences and non-summary sentences supported the features that they possess. -

The Classification probabilities are studied statistically using Naive Bayes Classifier rule.: P **(s є <S | F1, F2, ..., FN) = P (F1, F2, ..., FN | s є S) * P (s є S) / P (F1, F2,..., FN)**

4. LSA Method

This method involves training the neural networks to find out the kinds of sentences that ought to be included within the summary. -It uses three- layered Feed Forward neural network.

5. Automatic TS based on fuzzy logic

This method considers each characteristic of a text like similarity to title, sentence length and similarity to key word etc. as the input of the fuzzy system.

6. Query Based Extractive Text Summarization

In query based text summarization system, the sentences in a given document are scored based on the frequency counts of terms. -It uses Vector Space Model

Attributes of Extractive

Extractive text summarization features find appropriate sentences and select important sentences from the original document and include these sentences to the summary. Few features are described below that can be applied to select these important sentences:

A. Cue words features:

cue words or section are the group of words positioned around vital words like "summary", "reflects", "concludes", "purpose", "because" etc. That specifies the over-all content of the document and can be used as a indicator for the sentence to be included in the summary.

B. Keyword features:

Keywords have the key role in the criteria of selecting important sentences. Sentences that contain most of the keywords are considered to be included in the summary. The keywords can be the verbs, noun, adjectives and adverbs and are determined based on the TF-IDF method. These particular words can also be identified by their acronyms, capitalized or italicized property.

C. Title word feature:

Words contained in the title part are considered to be important words and sentences that contain these are necessary sentences and these sentences are included in the final summary. The sentences that are contained in the original document that has the stated words of the title of the document are contributed to the final summary as it can demonstrate the subject of the document.

D. Location or position word feature:

The location or position of sentences will determine whether the sentence is feasible to be included in the summary or not. Sentences that are placed in the beginning part of the document represents the introduction part of the original document and the sentences that are pointed towards the end part of the document can represent the conclusion part of the summary giving a proper meaning to the document.

E. Sentence position feature:

As most of the documents are hierarchically arranged containing sentences with more important content in the initial and conclusion part of the paraTexts in a document. Hence, sentences situated in the initial and edge parts are more likely to be contributed in the summary.

F. Proper noun feature:

Proper noun a name used for a particular person/individual which can be stated starting with Capital letter.

For example: Jane, Loan and Oxfam. Including the names of persons, places, concepts and organization in the final summary is very important to generate a shorten form of the document keeping the over-all meaning of the document constant.

G. Sentence length feature:

The length of the sentences is a essential feature in selecting sentences which one is to be included in the summary and which one is not to be. Shorter length text not having much words do not carry vital information so as the long length sentences consisting of many words but do not hold up any information are needed to be push aside.

H. Upper case word feature:

The words that contain only upper case or called abbreviation can be reflect on the summary and sentences holding these abbreviations can be included in the summary.

I. Similarity or cohesion feature:

Similarity feature is necessary to remove redundancies and reorder the segments to obtain a coherent summary. Similarity can be calculated among the sentences.

J. Term frequency:

Frequently occurring words rises the sentence score. TFIDF is used to calculate frequency of each word and the importance of the sentence increases for more number of times the word is visible in the sentence.

Related work

Scoring of phrases or sentences and obtaining summaries is that the most typical method utilized in automated extractive summarization. Sentence scoring is adopted within the majority of methods applied today. Scoring methods are classified as word scoring, sentence scoring, and Text scoring. In the word scoring methods, a scoring is made considering the importance of the sentences containing the frequency of a word in the text with words such as proper nouns, places and objects that are measured as a determinant being scored elevated.

In the text scoring methods, the formal properties of the words (emboldened, italicized, underlined) are taken under consideration. In addition, sentences starting with phrases such as temporarily, to conclude and As a consequence in the text are defined as sign phrases, and the sentences following these statements are noted as being significant sentences.

Similarly, evaluation is based on the title of the text to be summarized. Sentences that contain words found within the title are considered to be added to the summary, and their importance levels are increased accordingly. Sentence scoring methods also take under consideration the size of sentences, attaching greater importance to sentences of a bigger size. Points are assigned to sentences by determining the position of the sentence and whether or not it involves numerical values.

A multi-layered representation of documents, sentences and words was employed. The described documents with Texts in their study by utilizing link generation for automated document summarization. They defined the structure by revealing the text relationships within the documents, and evaluating the summaries by comparing them with those created by human hand.

A Text-based approach was presented in order to provide semantic continuity, with nodes corresponding to the terms of the documents and the edges reflecting semantic relationships between these nodes. Basically, a Text diameter calculation is performed for all the nodes in the Text, and the shortest and longest paths are described as the weakest and strongest bonds. While Text structures and documents were defined nodes and edges were created based on local similarities.

Random Walk has been used to obtain summaries of the main documents. A summarization system for the biomedical field was proposed. Using a system called Unified Medical Language, a Text was obtained based on concepts and relationships with a semi-dictionary-based application, and then the Page Rank algorithm was applied. INS proposed a new Text based on reinforced random walk.

Although most researchers have focused on extractive text summarization, some have worked effectively on abstractive summarizing. In the current revision, we introduce an innovative and tremendously simple text summarization system by taking Text-based automated document summarization studies one step further. The aforesaid examples from the literature are presented in Table 1 as a historical view of this research field.

Conclusion

Text summarization is growing as sub – branch of NLP as the demand for compressive, meaningful, abstract of topic due to large amount of information available on net. Precise information helps to look more effectively and efficiently. Thus text summarization is need and used by business analyst, marketing executive, development, researchers, government organizations, students and teachers also.

It is seen that executive requires summarization so that in a limited time required information can be processed. This paper takes into all about the small print of both the extractive and abstractive approaches alongside the techniques used, its performance achieved, alongside advantages and drawbacks of every approach. Text summarization has its importance in both commercial as well as research community.

As abstractive summarization requires more learning and reasoning, it is bit complex then extractive approach but, abstractive summarization provides more meaningful and appropriate summary compare to extractive. Through the study it is also observed that very less work is done using abstractive methods on Indian languages, there is a lot of scope for exploring such methods for more appropriate summarization.