Key-Phrase based Automatic Document Summarization

[1] Gosiya Kaleem.,[2]Md. Shahid, [3]Mukesh Rawat

[1,2] Department of CSE, Meerut Institute of Engineering and Technology, Meerut

Abstract : The information in today's world is big and rapidly growing at very high rate. It is becoming tough day by day to analyze this data and use this data correctly, which leads to data overload. Document Retrieval try to retrieve large number of Document that is beyond the reach of human analysis. Therefore, Document Retrieval is not perfect and required a something different such as abstract mechanism to reduce large data: the ability of summarize text. This research proposed automatic text summarization (ATS) . Summarization is a technique of shortening large textual information and showing concise and meaningful information.

A Text Summarization system need to analyze the context of the text and sentences, and separate the most essential text. The main purpose is to produce meaningful summaries of web documents, giving users a short, concise overview of the entire web document in a very short amount of time. This plays a very important role when a user has a large number of her web documents and cannot spend time reading them all.

Keywords – ATS, document retrieval, context

1. Introduction

The information in today's world is big and rapidly growing at very high rate. Especially if WWW provide the webpages or articles on newspaper, mail, database access. It is becoming tough day by day to analyze this data and use this data correctly, which leads to data overload. Document Retrieval from the past years reduces the work at various levels for the user. Document Retrieval is defined as, "collecting a group of documents from the search query and finding the sub-set of that data more relevant to search query and fetching the important and relevant documents". Many Search engines are built on this idea like; Google Search, Yahoo Search and AltaVista are few of them. The word information retrieval has gained widespread use, however the more appropriate term for this procedure is document fetching. The problem with the information overload is yet not resolved as the document contains important meaningful data. Document Retrieval try to retrieve large number of Document that is beyond the reach of human analysis, for e.g. if we write for the query "HiWE" in Google Search then it returns more than 29,100,000 results. Therefore Document Retrieval is not perfect and required a something different such as abstract mechanism to reduce large data: the ability of summarize text. This research proposed automatic text summarization (ATS) . Summarization is a technique of shortening large textual information and showing concise and meaningful information.

A Text Summarization system need to analyze the context of the text and sentences, and separate the most essential text. The main purpose is to produce meaningful summaries of web documents, giving users a short, concise overview of the entire web document in a very short amount of time. This plays a very important role when a user has a large number of her web documents and cannot spend time reading them all. The aim and objectives of the paper is mentioned below -

1. To Study various methods to reduce the size of the document for which a summary is going to be generated.

2. To study various sentence-scoring techniques which is used for finding important sentences in a document.

3. To find the process of scaling or normalizing the scores assigned to sentences to avoid giving importance to large sentences.

4. Reducing the summary size resulting in a more compact summary.

2. Literature Review

[1] Domain Specific Document Summarization by sentence extraction, It describing a method that involves sentence extraction and clustering for creating summaries from research papers. In this approach, sentences that are relevant to a specific category are extracted and grouped into clusters. From each cluster, a set of related sentences is selected to form a summary. The process involves two main steps: sentence extraction and clustering.

[2] Radev, Document clustering involves grouping similar documents together based on their content. In the context of this approach, documents are clustered into groups, and each cluster is represented by a centroid. A centroid is a representative document that summarizes the common themes and content of the documents within that cluster. Keep in mind that while this method can help in creating concise summaries, it might still face challenges such as ensuring that the selected sentences are coherent when placed together and handling situations where important information lies outside the cluster centroids.

[3] It describes a multi-document summarization approach using sentence extraction for user queries. This approach combines techniques such as passage segmentation, relevance identification, redundancy reduction, and summary cohesion to generate meaningful summaries from multiple documents in response to a user query. This approach generate summary related to the user query.

[4] Single document text summarization algorithm using semantic similarity

A high-level overview of a single document text summarization algorithm using semantic similarity. In this technique the semantic similarity between each sentence in the text using the vector representations. Cosine similarity is commonly used for this purpose. You can compare each sentence with every other sentence to create a similarity matrix. Calculate the importance score for each sentence based on its semantic similarity to other sentences.

[5] Classification of semantic similarity Edge Counting Methods Semantic similarity measures based on edge counting methods focus on quantifying the similarity between two texts by analyzing the structure of their semantic representation graphs. These methods often involve representing words

or concepts as nodes and their relationships as edges in a graph. The number of edges and their types can then be used to compute a similarity score. . However, the time complexity is very high to extract the summary.

3. Key points in sentence scoring

Sentence scoring

Sentence scoring refers to the process of assigning a numerical score or ranking to a sentence based on its relevance, quality, or other criteria. It is commonly used in text summarization.

The scoring process can vary depending on the specific task and the algorithms or models used. It may involve analyzing various linguistic features, such as word frequency, grammar, semantic meaning, or contextual information, to determine the score.

Sentence scoring can be helpful in a range of applications, including automated text summarization, chatbots, information retrieval systems, or content recommendation engines. By assigning scores to sentences, it becomes possible to identify and prioritize relevant or valuable information in a given text or document.

Using accurate sentence scoring technique

Selecting appropriate sentence scoring technique is a big task. Some of the sentence scoring techniques are mentioned below -

Frequency-based scoring: This approach assigns scores based on the frequency of words or phrases in a sentence. For example, sentences with rare or important words may receive higher scores.

TF-IDF (Term Frequency-Inverse Document Frequency) scoring: TF-IDF measures the importance of a word in a document relative to its occurrence in a corpus. Sentences containing words with high

TF-IDF scores are considered more relevant or informative.

Sentiment-based scoring: Sentiment analysis techniques can be used to assign scores to sentences based on the sentiment expressed within them. Positive or negative sentiment can contribute to a sentence's score.

PageRank-based scoring: Inspired by the Google PageRank algorithm, this approach considers the interconnectedness of sentences within a text. Sentences that are referenced or connected to by other important sentences may receive higher scores.

Normalized sentence scoring

The purpose of normalization is to ensure that the scores are comparable and consistent across different sentences or documents, even if they were scored using different criteria or algorithms. Normalization helps to standardize the scores and make them more interpretable or comparable. It can be done using various techniques, such as:

Min-max normalization: Using the method the scores is in the range, typically between 0 and 1. Calculated as below -

normalized_score = (sentence-score – sentence-min_score) / (sentences-max_score - min_score)

Z-score normalization: In Z-score normalization convert the scores to a mean of 0 and a standard deviation of 1. Calculated as mentioned below :

normalized_score = (sentences-score – sentences-mean) / sentences-standard_deviation

Decimal scaling: In this approach, the scores are divided by a suitable power of 10 to bring them within a specific range. For example, dividing all scores by 100 would bring them within the range of 0.00 to 1.00.

Selecting a best normalization technique is also important. So that all the sentences of varying lengths not have so much score differences.

The reduction of summary size

It refers to the process of condensing or shortening a text summary while still capturing the most important information. This is commonly done in text summarization tasks where the goal is to generate concise summaries that convey the key points of a longer document or piece of text.

Several techniques can be used to reduce the size of a summary:

Sentence fusion: Sentence fusion involves merging multiple sentences together to create more concise and coherent summaries. Redundancies or repetitive information across sentences are eliminated, leading to a reduction in summary size.

Sentence reordering: By rearranging the order of sentences or paragraphs, redundant or repetitive information can be eliminated, resulting in a more compact summary.

4. Research Methodology

a. HTML document converted into a plain text document using an HTML-to-text parser.

b. automate the process of extracting text content from a web page by providing a URL, parsing the HTML content, and then storing the extracted text in a document repository.

c. pre-processing the extracted text for further analysis or summarization. Cleaning and filtering out irrelevant words can help improve the quality of results.

d. Then process of extracting individual sentences and storing them with indices in a sentence repository.

e. Scoring words based on their frequency of occurrence in the document is a common approach to assess the importance of words within a given context. This can be used to identify the most significant words in the document.

Sum of word scores gives the sentence score.

f. Final score is calculated as below -

Calculate Sentence Scores: For each sentence, sum up the scores of its constituent words. This can be the term frequency score or any other scoring metric you have chosen.

Calculate Average Sentence Length: Calculate the average length of sentences in the document by dividing the total length of all sentences by the total number of sentences.

Evaluate Sentence Ratio: For each sentence, compute the ratio "average length / current length," where average length is the average sentence length calculated in the previous step, and current length is the length of the current sentence.

Final Sentence Score: Multiply the sentence's score (calculated in step 2) by the sentence ratio (calculated in step 4) to obtain the final score for each sentence.

g. Sentence fusion involves merging multiple sentences together to create more concise and coherent summaries. Each document summary can only contain four sentences. So if there are n documents, the cumulative summary will have n sentences. If the summary of (docs)1 have 4 or much lines , the reduction approach selects the highest 4 sentences with the highest scores.

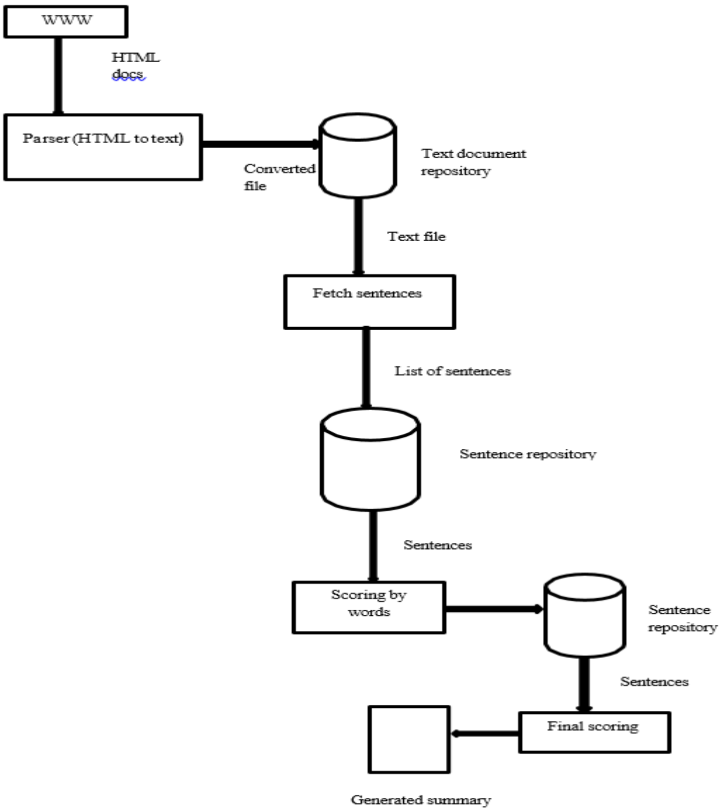

Figure 1: General Architecture for sentence scoring

References

[1] James Allan, Jaime Carbonell, George Doddington,“Domain Specific Document Summarization by sentence extraction”,Journal of King Saud University - Computer and Information Sciences, Volume 32, Issue 10, December 2020, Pages 1227-1228.

[2] Suad Alhojely, "A scalable summarization system using robust NLP",2020 International Conference on Computational Science and Computational Intelligence (CSCI),978-1-7281-7624-6/20/$31.00 ©2020 IEEE.

[3] Breck Baldwin and Thomas S. Morton, "Multi-Document Summarization using Sentence Extraction for user query", Appl. Sci. 2022, 12(9), 4479.

[4] Single Document Text Summarization Algorithm using semantic similarity, International Journal of Computer Applications (0975 – 8887)Volume 17– No.2, March 2011.

A. Siddharthan, A. Nenkova, and K. McKeown. Syntactic simplification for improving

[5] content selection in multi-document summarization. In Proc.of COLING, 2004.

[6] L. Vanderwende, H. Suzuki, and C. Brockett. Microsoft Research at DUC2006: Taskfocused

[7] summarization with sentence simplification and lexical expansion. In Proc. of DUC, 2006.

[8] Jaya Jayashree Jagadeesh, "Sentence Extraction Based Single Document Summarization", Article, January 2005, Research Gate.

[9] Logan Lebanoff† Kaiqiang Song† F, Scoring Sentence Singletons and Pairs for Abstractive Summarization, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2175–2189,Florence, Italy, July 28 - August 2, 2019.