Proceedings of the International Conference on Edge Computing and Applications (ICECAA 2022)
IEEE Xplore Part Number: CFP22BV8-ART; ISBN: 978-1-6654-8232-5

Information Extraction and Sentence Ordering in Multi-Document Summarization using Preference Learning

Anuj Kumar
Department of Computer Engineering & Applications
GLA University
Mathura Uttar Pradesh, India
anujkumar.gla@gla.ac.in

Atul Kumar Uttam
Department of Computer Engineering & Applications
GLA University
Mathura Uttar Pradesh, India
atul.uttam@gla.ac.in

2022 International Conference on Edge Computing and Applications (ICECAA) | 978-1-6654-8232-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECAA55415.2022.9936218

***Abstract*—**Multi-document summarizing is a process that automatically extracts information from many texts that are related to the same subject. For the purpose of information extraction, a technique that uses multi-document summarization which is based on phrase frequency is used. As a result of the phrases being picked from the documents depending on how important they are, the summary loses their coherence and the sequence in which the information is presented, which reduces the readability of the summary. A method of sentence ordering that is predicated on the chronological order of the phrases has been used in order to address this issue. According to the findings of this research, a multi-document summarizer that is based on a word frequency approach performs very well when it comes to the process of extracting relevant content units and increasing the readability of the summary via sentence sequencing.

Keywords— Information Extraction, Sentence sequencing**.**

I. INTRODUCTION

The world wide web now has an enormous quantity of online information, which is only expected to continue expanding. In spite of the fact that search engines were developed to process such a large quantity of data, they nonetheless make available a substantial number of outcomes in answer to the questions asked by users [10]. Under these conditions, the user is going to have a very tough time locating the document that he needs for his work. In addition, the majority of users are unwilling to go through the laborious process of reading through each of these papers. For this reason, systems that may automatically extract information are preferred. In this section, the reasons why automated summary is necessary, and the various methods of summarization have been discussed.

Automatic summarization is a process that involves extracting the most vital information from a source in order to create a shortened version that is suitable for a certain user or activity. This approach is referred to as "gathering the most significant information". The process of automatically summing text may be divided into two basic categories, called single document summarization and multi document summarization, depending on the number of source documents that are used.

Single Document Summarization [1] is much simpler, in comparison to the effort of summarising many documents. As opposed to summarising a single text, there is no need to worry about numerous languages, different input formats, different writing styles, or any other issues of that kind.

Multi-Document Summarization [1]: Two distinct dimensions may be used to define the multi-document summary: the dimension that is based on abstracts, and the one that is based on extracts. The obtained summary will only contain sentences that have been taken directly from the original source, but a conceptual description may include terms / expressions that are not included in the original text.

The goal of this paper is to find solutions for two issues that arise during the process of summarizing multiple documents into a single document: first, selecting the information which is most relevant to include in the summary; and second, arranging the information that has been selected during the first process in a way that improves the readability of the summary. The first part of the technique is referred to as the information extraction process, and the second part is referred to as the sentence ordering process. Together, these two parts make up the entire operation.

II. LITERATURE SURVEY

In the late 1950s, the first systems for text summarization were created [2], and they were distinguished by their focus on surface-level techniques. Later on, Edmundson presented the first entity-level technique that made use of syntactic analysis as well as the location feature. This approach focused on the entities themselves. In the 1970s, there was a resurgence of interest in the topic, which resulted in the development of expansions to the surface-level methodology, which included the use of cue phrases (bonus versus stigma items). At the end of the 1970s, the first discourse-based approaches on narrative grammars appeared, in addition to more comprehensive entity-level techniques. These developments [3] concurrent with one another.

In the 1980s, there was a proliferation of a wide range of different types of task, the most notable of which were entity-level techniques focused on Artificial Intelligence (AI). These methods involved the implementation of a variety of different sorts of work, such as scripts, logic and production rules, hybrid approaches, and semantic networks, amongst other types of labor. In addition, there was also a proliferation of a wide range of other types of work. A revival of the discipline may be said to have begun in the late 1990s. At this time, three different sorts of techniques were being investigated quite thoroughly [12], and the attention of both the government and the private sector was growing. The work that has been done during this time period has almost

solely concentrated on excerpts rather than abstracts, and there has also been a resurgence of interest in older studies.

The decade of the 2000s saw a change in the emphasis of research interests [5-8], with more attention being paid to multi-document summary. Recent years have seen the development of a wide array of approaches to the summary of several documents [15].

Every multi-document summarizing system has to make a decision in the order in which the output sentences should be presented to provide a summary. An uncontrolled predictive model for text structuring was proposed by Lapata [9]. This model learnt ordering restrictions from sentences that were characterized by a set of lexical and structural properties. This model was intended to learn how to structure text. Domain-specific content models have been presented by Barzilay and Lee [11] in order to capture topics and topic transitions for the purpose of sentence ordering. In order to successfully complete the evaluation job 5 that has been assigned to them, Barzilay and Lapata made use of the DUC 2003 multi-document summaries that were produced both by human writers and by machine summary systems.

III. PROPOSED METHOD

Figure 1 illustrates the basic structure of the Multi Document Summarization system, which is built on the principles of information extraction and sentence ordering. It is made up of two distinct procedures, namely the Information Extraction Module and the Sentence Ordering Module. The Information Extraction technique begins by accepting as input a collection of papers that have a common topic. It then picks the key phrases (or information) from those articles to be included in the summary. Finally, the method outputs the summary. A process known as "sentence ordering" is what gives the selected sentences the order that makes the most sense once the information extraction method has been applied to them.

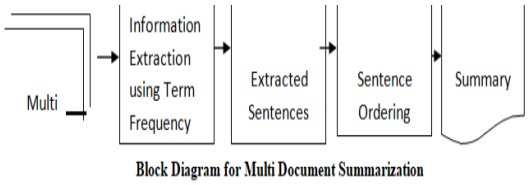

Fig.1: Multi Document Summarization

FREQUENCY-BASED SUMMARIZATION - In the first stage of the method for summarizing, a collection of source papers that are centered on the same subject, issue, or event are selected and given as input. The frequency of the word must be calculated in the following step. In the final stage of the procedure, each phrase is given a rating based on the terms it contains as well as the score that those terms are given, and the sentences that received the highest ratings are the ones that are selected to be included in the summary.

First, using the formula p(Wr) for each individual *i*, the likelihood allocation in excess of all the words that are included in the contribution must be determined. This calculation reveals that,

$$p(Wr) = nw/NW \quad (1)$$

where, nw is the total number of content word tokens in the input and NW is the overall number of occurrences of the word in the input. The chance that the given word will be found in the input is denoted by the symbol p(Wr).

Step 2: A weight is assigned to each sentence SW_j that is part of the input by multiplying the average probabilities of the terms in the sentence by that sentence's weight.

$$\text{Weight}(S) = \sum Wr_j * SW_j * p(Wr_j) \quad (2)$$

Step 3: The top-scoring statement which also includes the term with the greatest likelihood is selected.

Step 4: The probabilities of each word *wi* in the phrase that is selected in step 3, is revised using the formula,

$$pnew(Wr) = pold(Wr) \bullet pold(Wr) \quad (3)$$

Step 5. Move on to Step 2 if the length of the summary does not meet the expectations after the previous steps are finished.

The chronological ordering that is reflected by the chronology criteria [13], places stretch in the order that corresponds to the publication date in chronological order. Lin and Hovy's study was published in 2001, and McKeown and colleagues' study was published in 1999 [14]. In the following formula, the connection strength of arranging segments B after A, is defined by evaluating it based on a chronology criterion known as f_{chro} (A > B), which appears as follows:

Fig. 2: Connection strength based on chronology criterion

1. Timothy James McVeigh, 27, was formally charged with the explosion on Friday, according to a statement released by the Justice Department. The bombing took place at a federal facility in Oklahoma City and resulted in the deaths of at least 65 people.
2. A veteran of the Gulf War named Timothy McVeigh, who is 27 years old, was the first person to be charged with the bombing on Friday. McVeigh was caught for a driving violation immediately after the explosion on Wednesday.
3. Timothy James McViegh, 27, a suspect in the bombing that occurred in Oklahoma City, was taken into custody by federal officials. On Friday, official charges were brought against McViegh in connection with the incident.

Fig.3: Sample of Multi Document Summarization

The phrase resultant in the stretch is "Timothy James McViegh, 27, was formally charged on Friday with the bombing" which represents the frequent terms. The evaluation for the information extraction module is performed on the DUC-2002 records set using a phrase occurrence to summarize depending on, and the results are compared to those obtained by the MEAD summarizer which is a publically available toolkit. This data collection has a total of 59 clusters, each containing a unique combination of themes and documents. Documents pertaining to the same subject are included in each cluster. The clusters are organized into groups according to the total amount of documents that each cluster has.

The assessment findings for the stretch sequencing unit conducted out on the DUC-2002 records set for the word occurrence is to summarize depending on the MEAD summarizer. In order to attain this aim, metric Kendall's tau is tested both earlier than and subsequent to be valid sentence sequencing to the outline obtained by information mining.

Table 1. Assessment of Average Kendall's Tau (KT)

TauValues	Before	After
Term Frequency Based Summarizer	0.24	0.67
MEAD	0.13	0.63

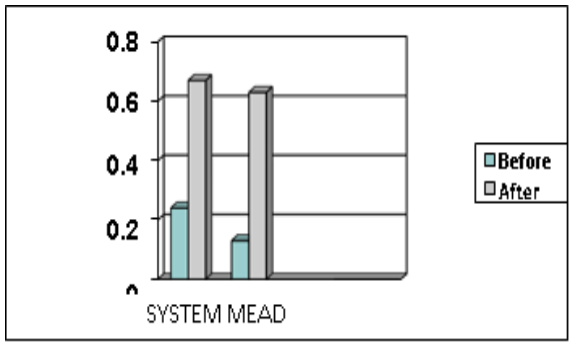

Fig. 4: Comparison of the average tau before and after applying Kendall's values

The figure presents a comparison of the average tau before and after applying Kendall's values to the summary sentences generated by the word frequency based summarizer and the MEAD summarizer. The summaries that are produced by the term frequency based summarizer have an improved average value of kendall's tau of up to 67 percent, whilst the summaries that are produced by the MEAD summarizer have an improved average value of kendall's tau of up to 63 percent.

IV. DISCUSSION

DUC-2002 document compilation, which is one of the document collections that the Document Understanding Conference (DUC) has made accessible has been used [14]. While the Defense Advanced Research Project Agency (DARPA), is in charge of founding and funding DUC, the National Institute of Standards and Technology (NIST) is in charge of maintaining it.

The DUC-2002 data are compiled from three hundred news stories covering a total of 30 distinct subjects. These pieces are sourced from the *The Associated Press, Financial Times, The Wall Street Journal*, and other comparable resources. This is a compilation of piece of writing from several newswire services and is made up of 59 different document clusters. A little less than half of these texts include sentences that were created by humans. Due to the fact that the primary emphasis of DUC is on the generation of abstracts, they do not make an appearance in the official DUC assessment. Nevertheless, the information is helpful for achieving the goals of this study. DUC 2002 data set is categorized in four categories.

1. Documents covering a single natural catastrophe incident and produced within a time range of no more than seven days maximum.

2. Papers relating to a single occurrence in any field, written during a span of time that is no longer than seven days.

3. Publications covering a same topic yet relating to a variety of different events (no limit on the time window).

4. Items that mostly provide information about one person and which give biographical information about that person.

The results on recall, precision, and f-measure are evaluated and compared to those obtained by the MEAD summarizer. It can be concluded that the word frequency based summarizing system has an accuracy that is roughly 14 percentage points greater than that of the MEAD summarizer. After applying the chronological sentence ordering method to both summarizers, it is discovered via experimentation that the ordering of both summarizers has been much enhanced as a result (67 percent). Because of this, extraction is much more difficult.

It is possible to arrange sentences according to the characteristics of their contents. The work might be expanded to include sorting the sentences according to the sorts of sentences they are. Because the sentences in the papers are selected for the summary based on how significant they are, the summary loses its coherence and the sequence in which the information is delivered, which makes the summary more difficult to understand. In order to find a solution to this problem, a way of organising sentences that is based on the chronological sequence in which the phrases occur has been applied.

V. CONCLUSION

According to the findings of this investigation, a multi-document summarizer that is based on an approach that considers word frequency performs very well when it comes to the process of extracting relevant content units and increasing the readability of the summary through the sequencing of sentences. This is determined by observing how well the summarizer performed in a series of tests.

References

1. Maybury, M.T. (2005). Karen Spärck Jones and Summarization. In: Tait, J.I. (eds) Charting a New Course: Natural Language Processing and Information Retrieval. The Kluwer International Series on Information Retrieval, vol 16. Springer, Dordrecht. https://doi.org/10.1007/1-4020-3467-9_7

2. Sarkar, Kamal. (2009). Improving Multi-document Text Summarization Performance using Local and Global Trimming. 272-282. 10.1007/978-81-8489-203-1_27.

3. H. P. Edmundson. 1969. New Methods in Automatic Extracting. J. ACM 16, 2 (April 1969), 264–285. https://doi.org/10.1145/321510.321519.

4. H. P. Luhn, "The Automatic Creation of Literature Abstracts," in IBM Journal of Research and Development, vol. 2, no. 2, pp. 159-165, Apr. 1958, doi: 10.1147/rd.22.0159.

5. Automatic Summarization [Amsterdam : John Benjamins Publishing Company, 2001.] - Permalink: http://digital.casalini.it/9789027299109.

6. Regina Barzilay and Lillian Lee. 2004. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, pages 113–120, Boston, Massachusetts, USA. Association for Computational Linguistics.

7. Gong, S., Qu, Y., & Tian, S. (2010). Summarization using Wikipedia. Theory and Applications of Categories.

8. Dragomir Radev, Jahna Otterbacher, Adam Winkel, and Sasha Blair-Goldensohn. 2005. News In Essence: summarizing online news topics. Commun. ACM 48, 10(October2005),95–98. https://doi.org/10.1145/1089107.1089111

9. Document Understanding Conference (DUC) PUBLICATION LIST, HTTP://WWW.NLPIR.NIST.GOV/PROJECTS/DUC/PUBS.HTML

10. U. Hahn and I. Mani, "The challenges of automatic summarization," in Computer, vol. 33, no. 11, pp. 29-36, Nov. 2000, doi: 10.1109/2.881692.

11. M. Zhong, J. Duan and J. Zou, "Indexing conceptual graph for abstracts of books," 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011, pp. 1816-1820, doi: 10.1109/FSKD.2011.6019789.

12. Mirella Lapata. 2003. Probabilistic text structuring: experiments with sentence ordering. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1 (ACL '03). Association for Computational Linguistics, USA, 545–552. https://doi.org/10.3115/1075096.1075165.

13. J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), 2019, pp. 1-3, doi: 10.1109/IconDSC.2019.8817040.

14. P. Raundale and H. Shekhar, "Analytical study of Text Summarization Techniques," 2021 Asian Conference on Innovation in Technology (ASIANCON), 2021, pp. 1-4, doi: 10.1109/ASIANCON51346.2021.9544804.

15. Jacob, I. Jeena. "Performance evaluation of caps-net based multitask learning architecture for text classification." Journal of Artificial Intelligence 2, no. 01 (2020): 1-10.