Key Phrase Extraction Based Multi-Document Summarization

Nidhi Chaudhary, Shalini Kapoor
Student
RGEC College
UPTU
Meerut, PA 250004

***ABSTRACT* - A summary text is a derivative of a source text condensed by selection and/or generalization on important content. The growth of the World Wide Web has spurred the need of an efficient Summarization tool. It is almost impossible to read whole of the document, it is very helpful if the summary of the document is available so, that the reader can notify whether the document is of his interest or not. Multi-document summarization is an increasingly important task: as document collections grow larger, there is a greater need to summarize these documents to help users quickly find either the most important information overall (generic summarization) or the most relevant information to the user (topic-focused summarization).**

I. INTRODUCTION

Automatic summarization is the creation of a shortened version of a text by a computer program. The product of this procedure still contains the most important points of the original text. The phenomenon of information overload has meant that access to coherent and correctly-developed summaries is vital. As access to data has increased so has interest in automatic summarization.

Multi-document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic. Resulting summary report allows individual users, so as professional information consumers, to quickly familiarize themselves with information contained in a large cluster of documents. In such a way, multi-document summarization systems are complementing the news aggregators performing the next step down the road of coping with information overload. The process of summarization is described below.

An example of the use of summarization technology is search engines such as Google. Technologies that can make a coherent summary, of any kind of text, need to take into account several variables such as length, writing-style and syntax to make a useful summary. Document summarization allows individual users, to quickly familiarize themselves with information contained in a large cluster of documents.

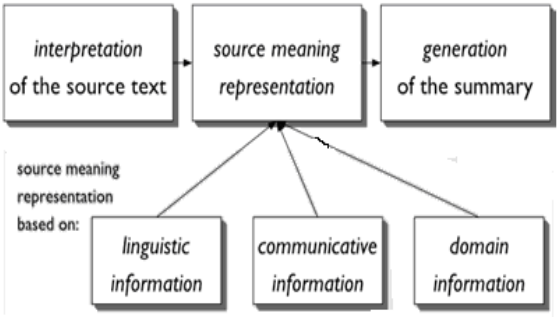

Fig 1 Summary Generation

Key benefits for text summarization are as follows:

- Multi-document summarization creates information reports that are both concise and comprehensive.
- Cut the time by pointing to the most relevant source documents.
- Limiting the need for accessing original files to cases when refinement is required.

- Automatic summaries present information extracted from multiple sources algorithmically.

There are different types of summaries depending what the summarization program focuses on to make the summary of the text, for example generic summaries or query relevant summaries (sometimes called query-biased summaries).

Summarization systems are able to create both query relevant text summaries and generic machine-generated summaries depending on what the user needs. Summarization of multimedia documents, e.g. pictures or movies are also possible.

Some systems will generate a summary based on a single source document, while others can use multiple source documents (for example, a cluster of news stories on the same topic). These systems are known as multi-document summarization systems.

II. RELATED WORK

A. Domain Specific e-Document Summarization Using Extractive Approach

we are using sentence extraction and clustering approach in our study. With sentence extraction approach sentences across all the research paper subtopics are clustered, following which, a small number of most related sentences are selected from each cluster of the particular category to form a summary. The sentence extraction strategy ranks and extracts representative sentences from multiple research papers. Radev described an extractive multi-document summarizer, which extracts a summary from multiple documents based on the document cluster centroids. Sentences extracted from the documents can describe part contents in a certain extent. Extraction-based summarization is a promising solution especially when the speed is concerned. In the sentence extraction strategy, clustering is frequently used to eliminate the redundant information resulted from the multiplicity of the original documents.

B. Multi-Document Summarization By Sentence Extraction

This paper discusses a text extraction approach to multi-document summarization, Multi-document summarization differs from single in that the issues of compression, speed, redundancy and passage selection. Conventional IR systems find and rank documents based on maximizing relevance to the user query some systems also include sub-document relevance assessments and convey this information to the user. Multi-document summarization capable of summarizing either complete documents sets, or single documents in the context of previously summarized ones.
Proposed multi-document summarizer works as follows:
1. Segment the documents into passages, and index them. Passages may be phrases, sentences, sentence chunks, or paragraphs.
2. Identify the passages relevant to the query using a threshold below which the passages are discarded.
3. Apply the MMR-MD metric. Depending on the desired length of the summary, select a
number of passages to compute passage *redundancy* using and use the passage similarity
scoring as a method of clustering passages.
4. Reassemble the selected passages into a summary document using one of the summary -cohesion criteria.

C. Multi-Document Update and Opinion Summarization

The area of Multi-document summarization can be subdivided into various domains like opinion summarization, update the summarization and query-based summarization etc. Various search engines like Yahoo, Google etc. provide a short snippet along with every search result for any query to given by the user. The automatic text summarization techniques are of mostly use in these real-world scenarios. Similarly, for many product there are numerous reviews available online and a summarized view of all those can be more informative to the user in more lesser time. Blogs news and product reviews are more important sources of opinions, in general. Because that the queries may or may not be posed beforehand, detecting opinions is somewhat similar to the task of topic detection for the sentence level. We look into automatic feature extraction mechanisms from product reviews. Further opinion summarization techniques which retrieves relevant information from the document set that are available, determines the polar orientation of each relevant sentence and last summarizes the positive-negative sentences accordingly.
This worked on two independent systems - one for query and update summarization and another is
opinion summarization.

A.) Update Summarization

MEAD toolkit provides the basic architecture above which different modules have been attached to different purposes. MEAD provides the simple interface and robust architecture where new modules can be added for rank and that choose sentences from all set of documents. Many modules of the whole summarization system and the flow are defined as:

- *Pretreating:* The documents that are in the form of xml format. The pretreating changes the format of all documents and then modifies the document a little part to removal of discrepancies.

- *Characteristic Scripts:* Characteristic Scripts are the modules that compute values of each various character of the set of sentences. That the system is summarization algorithms based on sentences these modules are essentially calculate values of various characteristic for every sentence available in the document pool. Some of the Characteristic used in our implementation of the summarization of the document–

  1.) Distance: Sentences having distance that are less than the specified threshold are imagined to be non-relevant for the summarization of the document.

  2.) Rank: This characteristic is relevant identifying important sentences are available in any document, these sentences that start of the paragraph are very important. Rank characteristic assigns to each sentence a value is calculated as,

  $$P(s) = 1/n$$

  where n is the number of the sentence available in the document.

  3.) Median point : A median point is a collection of words that are statistically important for a cluster of documents. As such, median point could be used for the classify relevant documents and also to identify salient sentences in a cluster.

Median point is a feature which is dependent on the words available in the sentence. The more important words it

contains, more preference it is in respect of the document cluster.

Then, for each word in this TF*IDF is calculated where IDF is defined as,

$$IDF(i) = log(P/ n_k)$$

Where, *P* is total number of documents and *nk* is the number of documents in which the word *k* is present. Now, for each sentence *Ck* the combined median point score is calculated as,

$$Ck = \Sigma\ Cw,k$$

Where, *Cw;k* is the TF*IDF score of the word *w* in the sentence k.

4.) Graph-based Lexical Centrality: It is completed by finding the most prestigious sentences.(Also, median point of a sentence is calculated in form of median point of words that it contains). This strategy is based on the basis of prestige in social networks, that all are inspired many ideas in computer networks and information retrieval. A cluster of documents can be viewed as a collection of sentences that are related to each other. Few sentences are more related to each other while few others may share only a little information. The sentences that are similar to other sentences in a cluster are more central related to the topic.

- *Default Classifiers:* This step is merging the different feature vectors they were already computed in the last step. In this, MEAD provides a default classifier which we are using this feature here. It is a user programmable in the sense that it allows to the user to assign different weights to different characteristic. In this we are using different combination of features to study the quality of summary generator.

Score(*Si*) = *feature*1 ¤ *weight*1 + *feature*2 ¤ *weight*2 + :::

- *Remove redundancy:* In this step we are removing redundancy from the extracting summary. In Multi-document summarization, the more documents may contain the same sentences which talking about more or less thing. We want to filter out all the sentences which are mostly similar to each other according to the documents. In this step, we are using the similarity feature to find out all the sentences that they sentences are considering for the summary. Each cross-sentence Informational Subsumption (CSIS) : - It refine that some sentences repeat the information available in other sentences and that may be omitted during summarization. If all the information content of each sentence is contained, then a becomes informationally redundant.
- *Processing after the summary extraction:* It includes several tasks like removing all unnecessary phrases and sentences from the summary because generally that the very important thing related with summaries is that clustering of information with as few of unnecessary parts as possible. So, we want to prune the sentences that are selected by the providing the priority to find out only the necessary parts of those in the summary. This step will allow us for including more sentences in the summary extraction to increase the information that are available.

B). Summarization by the help of survey

According to survey for summarization, summarizes survey of articles by telling the truth of sentiment polarities, related degree and correlated all events. We are defining decomposed the problem of survey of summarization into following steps:

Characteristic Extraction: we are identifying the character terms or phrases of the each document, we use this type of extraction information to identify the sentences that which contain important information about those characteristic.

Survey Identification: The aim of survey identification is to detect where in the documents survey are embedded. The survey sentence is the few part for complete semantic unit from which the survey can be extracted.

Opinion Classification: It is the work of finding positive and negative views, emotions and calculation. This phase uses a list of words that they define the semantic orientation. Most common polarity is assigned these words and this works as define the prior polarity for each word.

Summarization: This phase aims to generate a cross-document summary and at the last step we know the surveyed sentences and the specific characteristic they talk about, we can collect all the surveyed information from the corpus on a specific given heading.

4.) *Multi-Document Automatic Text Summarization Using Entropy Estimates*

Automatic extraction-based text summarization consists of a set of algorithms and mathematical operations performed on sentences/phrases in a document so that the most difficult task of identifying relevant sentences can be performed. The cognitive process that takes place when humans try to summarize information is not clearly understood, yet we try to model the process of summarization using tools like decision trees, graphs, word nets and clustering algorithms.

Many of the methods work well for specific sets of documents. Techniques that rely on features such as position, cue words, titles and headings perform poorly on unstructured text data. Methods that rely on word frequency counts obtained after analyzing the document collection to be summarized, perform well in cases such as multiple news reports on the same event, but fare badly when there is only one document to be summarized.

Sentence selection technique:

Any document set to be summarized is first classified as belonging to a particular domain. We use a database of documents clustered into various domains/ topics. Once the domain/topic has been identified, an entropy model for the various words and collocations in the identified domain is generated. Documents available in this identified domain constitute the training data set. The calculated entropy values are applied to each of the sentences in the document set to be summarized and a sentence ranking formula is computed.

III. PROPOSED METHODOLOGY

The methodology proposed takes input from a text file, and outputs the summary into a similar text file. The work of generating common summary of multiple similar documents is divided into three modules namely scoring of sentences in the documents, document summary generator and summary reduction. The general architecture of the proposed work is shown in the figure given below.

A. Description of modules Scoring

The most daunting task at hand was to generate an efficient scoring algorithm that would produce the best results for a wide range of text types. The scoring algorithm is divided into two parts: word scoring and sentence scoring. In word scoring algorithm words of the sentences of the text documents are scored using the following heuristics:

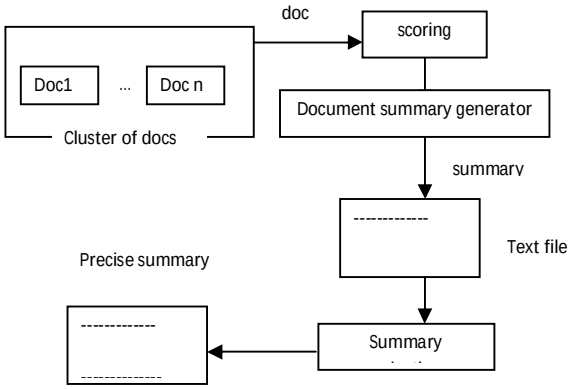

Fig 2 General architecture of multi-document summarization

1.) Remove Words: These are some insignificant words, that they are mostly used in English language, In this step we are removing the words Eg. I, a, an, of, am, the, et cetera.

2.) Catch Words: These are the words that they define the concluding sentences of a text. Eg. Summary, conclusion, thus, hence, etc.

3.) Common Words: more than 800 words of the English language have been defined as maximum frequently used words that they define meaning of a sentence.

4.) A Noun That Denotes A Particular Things; Usually Capitalized :A Nouns that denotes a particular things in most cases make the central theme of a used text. It provides semantics for summary, and given high importance while scoring sentences.

5.) Term words: The user has been given a chance to get a summary generated which contains a particular word that are mostly used, the term words.

6. Number Of Occurrence Of The Word: Basic scores have been allotted to each words, then their final score is calculated on the basis number of occurrence of the word in the document. Words in the document that are repeated more frequently than the others contain a more profound impression of the document, and that given a higher importance.

B. Sentence scoring

1. Initial Score: Using this above methods, a last word score is calculated, and then calculating the sum of word scores gives a sentence score. This gives extra long sentence a clear advantage that they break into smaller counterparts. it is not necessary defined of lesser importance.

2. Last Score: By multiplying the score that is obtained by calculating the ratio of "average length / current length" the above disadvantage can be nullified for a large extent, and a last sentence score is obtained.

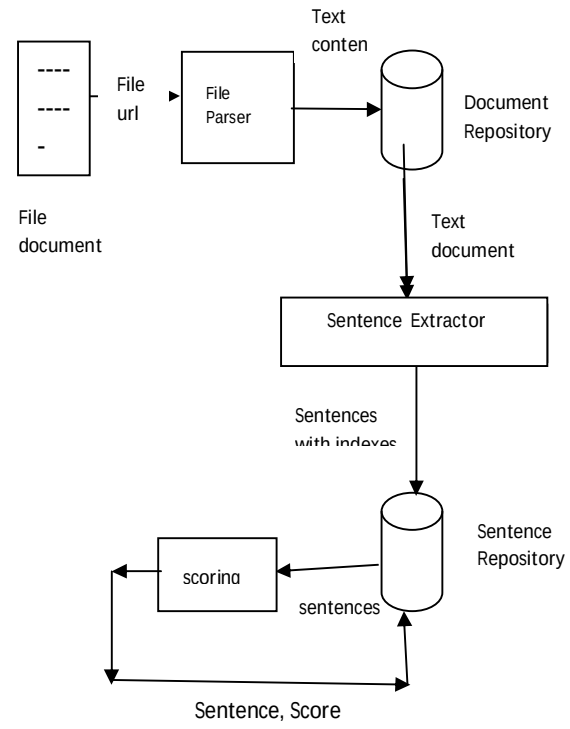

Fig 3 The proposed architecture for sentence scoring

The different modules of the sentence scoring are as follows:

- *File Parser:* This module extracts out the text from the document and stored them into text file in document repository.
- *Sentence Fetch:* This module extracts the sentences from the text file based on certain criteria such as the group of words ended with full stop(.), semi colon(;) or colon(:) considered as a single sentence and stored in a text file with indices in sentence repository.
- *Scoring:* This module is divided into two sub modules namely word scoring and sentence scoring. In word scoring the scores of the words are calculated based on certain criteria's such as

1. Stop Words: These are some insignificant words that are so commonly used in the English language, Eg. I, a, an, of, am, the, et cetera.

2. Catch Words: These are the words that they define the concluding sentences of a text. Eg. Summary, conclusion, thus, hence, etc.

3. A Noun That Denotes A Particular Things; Usually Capitalized: It provides semantics for summary, and given high importance while scoring sentences.

4. Term words: The user has been given a chance to get a summary generated which contains a particular word that are mostly used, the term words. the sum of word scores gives a sentence score. This gives extra long sentence a clear advantage that they break into smaller counterparts. it is not necessary defined of lesser importance. By multiplying the score that is obtained by calculating the ratio of "average length / current length"

V. CONCLUSION AND FUTURE SCOPE

The approach discuss above gives us a new idea for creating summaries from text of multiple similar documents by sentence scoring, there is also a new technique for reducing the summary size as the number of documents are increasing by selecting the sentences whose score are more as compared to the other sentences of the summary, but having certain limitations such as without the use of NLP, the generated summaries suffers from lack of cohesion and semantics, it is difficult to relate pronouns to their corresponding nouns in the summary. The possibilities are endless.

With Natural Language Processing:
a. Newspaper headlines can be generated.
b. Forms can be filled up.
c. Bio-data can be generated.

IV DOCUMENT SUMMARY GENERATOR

1.) Calculation of threshold value

Suppose, n(s) is the sum of the scores of the sentences of documents, then the threshold value is calculated as-

$$TH=n(s) / n$$

Where, n is the number of sentences of the document.

Now the sentences whose score are greater than or equal to the threshold value are considered as part of the summary. The sentences of the summary are arranged according to the decreasing order of their scores.

2.) Summary reduction

As the number of documents in the cluster is more so, the cumulative summary contains more number of sentences, summary reduction technique reduces the size of the cumulative summary.

Summary reduction is done by the following steps-

1. Suppose a sentence "x" having text "Ram is a good person" and a sentence "y" having text "Ram is a good person and going to Delhi", the sentence x is a subset of y, so, while generating a cumulative summary such sentences that are subset of other sentences are discarded.

2. The summary of each document must be consists of only four sentences, it means that if the number of documents=n, then the cumulative summary contains 4n sentences. If the summary of $(doc)_1$ contains more than 4 sentences, then the reduction technique selects the top highest scored 4 sentences from the summary of $(doc)_1$.

VI. REFERENCES

[1] *James Allan, Jaime Carbonell, George Doddington,Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study: Final report. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.*

[2]Chinatsu Aone, M. E. Okurowski, J. Gorlinsky, and B. Larsen. 1997. A scalable summarization system using robust NLP. In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pages 66-73, Madrid, Spain.

[3]Breck Baldwin and Thomas S. Morton. 1998. Dynamic coreference-based summarization. In Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP-3), Granada, Spain, June.

[4] Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In Proceedings of the CL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pages 10-17, Madrid, Spain.

[5] A. Siddharthan, A. Nenkova, and K. McKeown. *Syntactic simplification for improving content selection in multi-document summarization.* In Proc.of COLING, 2004.

[6] L. Vanderwende, H. Suzuki, and C. Brockett. *Microsoft Research at DUC2006: Taskfocused summarization with sentence simplification and lexical expansion*. In Proc. of DUC, 2006.

[7] X. Wan and J. Yang. *Improved affinity graph based multi-document summarization.* In Proceedings of HLT-NAACL, Companion Volume: Short Papers, pages 181–184, 2006.

[8] D.Zajic, B. Dorr, andR. Schwartz.*Automatic headline generation for newspaper stories.* In Proc. of DUC, 2002.

[9] D. Zajic, B. Dorr, J. Lin, C. Monz, and R. Schwartz. *A sentence-trimming approach to multidocument summarization. In Proc*. of DUC, 2005.