ISSN:0975-9646

A Survey of News Article Clustering and Summarization

Purabi Choudhury

Gauhati University Institute of Science and Technology
Jalukbari, Guwahati, Assam

***Abstract*— Reading newspaper is a very good habit. By reading a newspaper a reader may get various categories of information** for **e.g. business, sports, and entertainment and also be able to get some unknown information. In these days most of us prefer to read online newspaper. Because we can choose ourselves a news source from which we would like to read. But if a user wants to get information about particular news, he might have to face problem of finding the required information. The problem occurs due to the information overload i.e. high availability of online information. To get information, user has to provide the news query to the search engine. As a result, the search engine returns links to various news sources e.g. ndtv.com, times now, Hindustan times etc. A user exactly does not know in which link he might get their desired information. So he has to navigate them to every link until he is not satisfied. Searching information in these way is very time consuming and inefficient. To solve this problem we need a tool that can automatically retrieve all query related news articles from various news sources. From those selected news articles, the system will generate a coherent summary i.e. no redundant information is available. By reading it, a user will be able to get their required information as well as some other information that are related to the query.**

***Keywords*— Automatic Text Summarization, Single Document Summarizer, Multi-Document Summarizer, Extractive summarization and Abstractive Summarization techniques.**

I. Introduction

With the advancement of technologies in World Wide Web, huge amounts of rich and dynamic information's are available. With the help of a web search engines, a user can get their required information on that they are interested. Usually search engines returns many documents, a lot of which are relevant to the topic and some may contain irrelevant documents with poor quality. Due to high availability of online information, there is a problem of information overload. Information overload may cause difficulties for finding information on that user is interested. For example, if a user wants to know about particular news through online, user needs to provide a news query to the search engine. The search engine returns various links of news sources. By selecting one source it navigated to a page where various kinds of news article are present. For a user it is difficult to find only the query related news articles among all other news articles, because it consumes lots of extra time for reading. Besides that some sources include redundant information too. Automatic multi-document text summarization is one way to solve this type of problem by producing shorter presentation of original contents which covers non redundant and salient information that are extracted from a multiple news articles.

II. Background Survay

Automatic text summarization [1] is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. As the problem of information overload has grown, and as the quantity of data has increased, so has interest in automatic summarization. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. An example of the use of summarization technology is search engines such as Google.

A. Classification of text summarization

Depending on the number of documents a summarizer takes as input, there are two types of summarizer: Single Document Text Summarizer and Multi Document Text Summarizer

1) *Single document summarizer:* Summary is generated from a single document. Generated summary is coherent i.e. no redundant information is available.
   Text Teaser [2] is an application that extracts the most important sentences of an article. The purpose of the API is to provide a preview of what the article is all about. The API accepts the text of the article, and it will return the summary by extracting most important sentences in it.
2) *Multi Document Text Summarizer:* Summary is generated from multiple text documents. For example, given a collection of news articles, a multi document text summarizer is able to create a concise overview of the important events. One big issue of this type of summarizer is to reduce the redundant information. In the case of multi document summarization of articles about the same event, source articles can contain both repetitions and contradictions. Extracting all the similar sentences would produce a verbose and repetitive summary, while extracting only some of the similar sentences would produce a summary biased towards some sources. MultiGen[3] uses a comparison of extracted similar sentences to select the appropriate phrases to include in the summary and reformulates them as new text.

B. Techniques of Text Summarization

Text Summarization techniques can be broadly classified as Extractive and Abstractive

1) *Extractive Summarization:* This technique aims to pick out most relevant sentences from different documents

and maintain low redundancy of information in summary. In extractive-based summarization methods, picking out most important document regions like phrases, sentences, paragraphs etc. In this technique first, assigns a score to each sentence and then gives ranks to the sentences according to their scores. The sentence with highest score will get the top rank. The score for a sentence is calculated by using statistical features including sentence position, cue words, term frequency, document frequency, topic signature, etc. The highest ranked sentences from different documents can then be grouped into different cluster in terms of their similarity with other ranked sentences. Finally select representative sentences from each cluster to generate a simple and non redundant summary.

2) *Abstractive Summarization:* This method involves natural language understanding tool to generate summary from document(s). An abstract summary ([4], [5]) may contain words or phrases which mayn't be exist in the original document(s). The process of abstractive summarization is complex because it cannot be formulated mathematically or logically. The quality of abstractive summary depends on deep linguistic skills.

C. Previous researches on multi-document summarization:

The first stated of multi document by Radev and McKeown (1995) [6] developed SUMMONS to generate summaries of multiple documents on the same or related events, presenting similarities, differences, contradictions and generalizations among sources of information from realized as English sentences.

In 1998 [7], they improved their SUMMONS to combines it into a conceptual representation of the summary which selects information from underlying knowledge base. The structured conceptual representation of the summary, where information that appears in only one article is given a lower rating and information that is synthesized from multiple articles is rated more highly.

Kathleen R. McKeown. et al.(2001) [8] presented MultiGen and DEMS for Columbia multi-document summarization system built on the observation that depending on the intended

Purpose of the summary and on the types of document summarized. This technique focused on the summarization of sets of documents that all describe the same event or news. They used an enhanced version of MultiGen to summarize the document.

MINDS [9] integrates multi-lingual summarization and multi document summarization capabilities using a multiengine, core summarization system and provides fast, interactive document access through hypertext summaries. Core summarization problem of MINDS is taking a single text and producing a shorter text in the same language that contains all the main points in the input text. It is using a robust, graded approach for building the core engine by incorporating statistical, syntactic and documents structure analyses among other techniques. This approach is less expensive and more robust than a summarization technique based entirely on a single method. The core engine is being designed in such a way that as additional resources, such as lexical and other knowledge bases or text processing and MT engines, become available from other ongoing research efforts they can be incorporated into the overall multiengine MINDS system.

Generic relation extraction (GRE) in 2009 [10] is a novel multi document text summarization approach, which aims to build systems for relation identification and characterization that can be transferred across domains and tasks without modification of model parameters.

D. Different Types of Extractive Text Summarization:

There are two types of summarizer; query based and non query based.

1. *Query Based Summarizer*: A query based summarizer generate summary based on the query provided to it. This type of summarizer collects various text documents related to the query and then generate summary. It reduces the extra time of finding the desired information.

a) The working steps of a query based summarizer

i. A user search query is provided to the system

ii. A document selector selects documents those are related to the user search query.

iii. Sentence extractor is used to pick out the most important sentences. Importance of a sentence is measured on the basis of similarity of the sentence with the query. Both statistical and semantic similarity measuring approaches are used.

iv. Sentence scoring procedure, calculates the score value of the extracted sentences according to their feature profile like term feature, position feature, centrality feature, proper noun feature, numerical data feature. Sentence total score is calculated by adding all these feature score value. And then ranking the extracted sentences according to their associated score value in their descending order.

v. Finally top highly ranked sentences are included to generate the summary.

2. *Non query Based news Article Summarizer:* A query based summarizer generates a summary on the basis of the query provided to it. But if someone is interested to get an overview of various categories of news like sports, business etc. Then this of summarizer is very much helpful for them to get that kind information. A non query based summarizer can generate several summary documents. Each summary document produces summary from a specific categories like sports, entertainment, business of news articles.

a) The working steps of a non query based summarizer

i. A Document classifier divides collection of various documents into different groups. A group contains all the related documents.

ii. For each group of documents

- A sentence extractor is used that will extract the most important sentences from it.

Importance of a sentence is measured on the basis of similarity of the sentence with the main theme or the headlines.

- Extracted sentences are then ranked according to their associated score value.
- Top ranked sentences are selected to include in the summary

III. Conclusions

Most of the text summarizer generates summary using extractive based technique because generated summary is simple and easy to understand. But the summary is often tend to be longer than average due to inclusion of some unnecessary parts of segments. On the other hand, abstract summary is very much similar to human generated summary and are also presented in concise form. An abstractive type of summarizer involves natural language processing and uses newer concepts that best describe the contents of original documents in fewer words.

In our future work, we aim to develop a summarizer that can generate summary of multiple related news articles using a extractive technique. Important sentences are extracted based on their syntactic and semantic features. Extracted sentences are scored according to their feature profile like term feature, Position Feature, Sentence Centrality Feature, Sentence with Proper Noun Feature and Sentence with Numerical Data Feature. Sentences with highest score will get top rank. Ranked sentences belonging to different news article are grouped into cluster. Finally choose few sentences from each cluster and get included to the summary.

References

[1] Automatic summarization From Wikipedia, the free encyclopedia

[2] (https://www.mashape.com/mojojolo/textteaser#!documentation)

[3] Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Simone Teufel, M. Yen Kan, andBarry Schiffman. 2001. Columbia multi-document summarization: Approach and evaluation. In Donna Harman and Daniel Marcu, editors, Proceedings of the First Document Under-standing Conference (DUC ’01), NewOrleans, Louisiana, USA. 1-21.

[4] Inderjeet Mani, (1999) “Advances in Automated Text Summarization”, MIT Press.

[5] Sun J, Shen D, (2005) “Web-page Summarization using Clickthrough data”, In proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[6] Dragomir R. Radev and Kathleen R. McKeown. 1995. Generating summaries of multiple news articles. In proceeding of SIGIR’95, Seattle, Washington. 74–82.

[7] Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple online sources. Computational Linguis-tics, 4:469–500.

[8] Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou,Simone Teufel, M. Yen Kan, and Barry Schiffman. 2001. Columbia multi-document summarization: Approach and evaluation. In Donna Harman and Daniel Marcu, editors, Proceedings of the First Document Under-standing Conference (DUC ’01), New Orleans, Louisiana, USA. 1-21.132). Menlo Park, CA: AAAI, 1998.

[9] Cowie, J., Mahesh, K., Nirenburg, S., and Zajaz, R.,“MINDS-Multilingual INteractive document summarization”, In Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization

[10] Ben Hachey, “Multi-document summarization using generic relation extraction”, Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing: Volume 1, 420-429, 2009.