A Review Paper on Extractive Techniques of Text Summarization

Surajit Karmakar, Tanvi Lad, Hiten Chothani
Department of Computer Engineering, KJSIEIT
Ayurvihar Complex, Everard Nagar, Sion, Mumbai 400022
Maharashtra, India

Abstract*— Automatic summarization 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. There are two types of summarization: extractive and abstractive. Extractive summarization methods simplify the problem of summarization into the problem of selecting a representative subset of the sentences in the original documents. Abstractive summarization may compose novel sentences, unseen in the original sources. In our study we focus on sentence based extractive document summarization. The extractive summarization systems are typically based on techniques for sentence extraction and aim to cover the set of sentences that are most important for the overall understanding of a given document.

***Keywords*— Text summarization, extract, abstract, summary, linguistic**

I. Introduction

Text summarization has become an important and timely tool for assisting and interpreting text information in today's fast-growing information age. Enormous increasing and easy availability of information on the World Wide Web have recently resulted in brushing up the classical linguistics problem [3] – the condensation of information from text documents. This task is essentially a data reduction process. The goal of automatic text summarization is condensing the source text into a shorter version preserving its information content and overall meaning. Text summarization is the process of automatically creating a compressed version of a given document pre-serving its information content. Automatic document summarization is an important research area in natural language processing (NLP). The technology of automatic document summarization is developing [4] and may provide a solution to the information overload problem.

Summary may refer to abstract summary, abridgement or executive summary. An abstract is a brief summary of a research article, thesis, review, conference proceeding or any in-depth analysis of a particular subject or discipline, and is often used to help the reader quickly ascertain the paper's purpose. [5] When used, an abstract always appears at the beginning of a manuscript or typescript, acting as the point-of-entry for any given academic paper or patent application.

An abridgement (or abridgment) is a condensing or reduction of a book or other creative work into a shorter form while maintaining the unity of the source.[6] The abridgement can be true to the original work in terms of mood and tone, capturing the parts the abridging author perceives to be most important; it could be a complete parody of the original; or it could fall anywhere in-between, either generally capturing the tone and message of the original author but falling short in some manner, or subtly twisting his words and message to favour a different interpretation or agenda.

An executive summary, sometimes known as a management summary, is a short document or section of a document, produced for business purposes, that summarizes a longer report or proposal or a group of related reports in such a way that readers can rapidly become acquainted with a large body of material without having to read it all. It usually contains a brief statement of the problem or proposal covered in the major document(s), background information, concise analysis and main conclusions. It is intended as an aid to decision-making by managers [7] and has been described as possibly the most important part of a business plan.

II. Literature Survey

Automatic text summarization can be classified into two categories based on their approach: summarization based on abstraction and summarization based on extraction. Systems for extractive summarization are typically based on technique for sentence extraction, and attempt to identify the set of sentences that are most important for the overall understanding of a given document. Extractive summaries [8] are formulated by extracting key text segments (sentences or passages) from the text, based on statistical analysis of individual or mixed surface level features such as word/phrase frequency, location or cue words to locate the sentences to be extracted. The "most important" content is treated as the "most frequent" or the "most favourably positioned" content. Such an approach thus avoids any efforts on deep text understanding.

Abstraction method, on the other hand, heavily utilizes computation power for natural language processing (NLP) with the inclusion of grammars and lexicons for parsing and generation of summaries. Abstraction involves paraphrasing sections of the source document. Fully abstractive approach [9] with a separate process for the analysis of the text, the content selection, and the generation of the summary has the most potential for generating summaries at a level comparable to human.

III. Motivation

Text summarization (TS) is the process of identifying the most salient information in a document or set of related documents and conveying it in less space (typically by a factor of five to ten) than the original text. Some of the reasons for motivation of text summarization are as follows:

- To keep up with the world affairs by listening to news.
- People base investment decisions on stock market updates.
- People even go to movies largely on the basis of reviews they've seen.
- With summaries, People can make effective decisions in less time.
- The motivation here is to build such tool which is computationally efficient and creates summaries automatically.

IV. Problem Domain

Automatic Text summarization has become an integral part of daily life due to the availability of large volume of information, that need to be summarized for humans so that they can read important contents in short time. It has been said for decades (if not centuries) that more and more information is becoming available and that tools are needed to handle it. Only recently, however, does it seem that a sufficient quantity of this information is electronically available to produce a widespread need for automatic summarization. It is important to extract crucial information or compact the whole data in order to minimize amount of time invested to review this huge information. One of the ways to deal with this problem is Text Summarization.

V. Problem Definition

There are longer sentences in extract than average length. Therefore, sometimes even the part of sentences which is not important is also included, which results in space consumption. Important or relevant information is usually spread across sentences, and extractive summaries cannot capture this [1] (unless the summary is long enough to hold all those sentences).

Pure extraction often leads to problems in overall coherence of the summary—a frequent issue concerns "dangling" anaphora. Sentences often contain pronouns, which lose their referents when extracted out of context. Worse yet, stitching together decontextualized extracts may lead to a misleading interpretation of anaphors (resulting in an inaccurate representation of source information, i.e., low fidelity). Similar issues exist with temporal expressions. These problems become more severe in the multi-document case, since extracts are drawn from different sources. [2] A general approach to addressing these issues involves post-processing extracts, for example, replacing pronouns with their antecedents, replacing relative temporal expression with actual dates, etc.

VI. Solution methodologies

A. *Term frequency - Inverse document frequency(TF-IDF)*

We combine the definitions of term frequency and inverse document frequency, to produce a composite weight for each term in each document. The TF-IDF weighting scheme assigns to term t a weight in document given by

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t$$

In other words, TF-IDF assigns to term t a weight in document d that is

i. Highest when t occurs many times within a small number of documents (thus lending high discriminating power to those documents).

ii. Lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal).

iii. Lowest when the term occurs in virtually all documents.

B. *Cluster based method*

Clustering is a process of grouping set of objects in such a way that objects in same group are similar to each other. Organization of documents is done in such a way that they address different topics in some sequence. This is also applicable to summaries. Sentence selection is based on cluster Ci. Another factor for selection is location of sentence Li. The last factor that increases the score of a sentence is its similarity to the first sentence in the document to which it belongs (Fi). The overall score (Si) of a sentence i is a weighted sum of the above three factors: Si =W1 *Ci + W2 *Fi+ W3 *Li where Si is the score of sentence Ci, Fi and Li.

C. *Graph theoretic approach*

Graph theoretic representation of text provides a method to identify these themes. After the pre-processing steps, stop word removal and stemming, sentences in the documents are represented as nodes in an undirected graph. Two nodes are connected with edges; if they have common words here every node is sentence.

D. Machine Learning approach

Given a set of training document and their extractive summaries, the summarization process is modelled as a classification problem: sentences are classified as summary sentences and non-summary sentences based on the features that they possess. $P (s \in < S \mid F1, F2,.., FN) = P (F1, F2, ..., FN \mid s \in S) * P (s \in S) / P (F1, F2,..., FN)$ where s is a sentence from the document collection, F1, F2…FN are features used in classification. S is the summary to be generated, and $P (s \in < S \mid F1, F2,.., FN)$ is the probability that sentence s will be chosen to form the summary given that it possesses features F1,F2…FN.

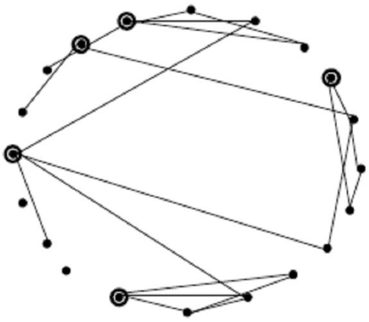

Fig. 1 Graph theoretic approach

E. Query based extractive text summarization

In query based text summarization [10], the scoring of the sentences of a given document is based on the frequency counts of words or phrases. Higher scores are given to the sentences containing the query phrases rather than the ones with single query words. The sentences with highest scores are then extracted for the output summary together with their structural context. Portions of text may be extracted from different sections or subsections. The resulting summary is the union of such extracts. In the sentence extraction algorithm, whenever a sentence is selected for the inclusion in the summary, some of the headings in that context are also selected.

The query based sentence extraction algorithm is as follows:

```
Algorithm:
1: Rank all the sentences according to their score.
2: Add the main title of the document to the summary.
3: Add the first level-1 heading to the summary.
4: While (summary size limit not exceeded)
5: Add the next highest scored sentence.
6: Add the structural context of the sentence: (if any and not already included in the summary)
7: Add the highest level heading above the extracted text (call this heading h).
8: Add the heading before h in the same level.
9: Add the heading after h in the same level.
10: Repeat steps 7, 8 and 9 for the next highest level headings.
11: End while
```

VII. CONCLUSION

This review paper discusses a few of the extractive methods of text summarization. An extractive summary is a selection of important sentences from the original text that briefly describes the original text.

Various methods of extractive approach have emerged in the past. But it is hard to say how much greater interpretive sophistication, at sentence or text level contributes to performance. Without the use of Natural Language Processing, the generated summaries may not be much accurate in terms of semantics. If the input documents cover multiple topics, it becomes difficult to generate a balanced summary. For this purpose, deciding proper weights of individual features is important as quality of final summary depends on it.

REFERENCES

[1] Jimmy Lin., “Summarization.”, Encyclopedia of Database Systems. Heidelberg, Germany: Springer-Verlag, 2009.

[2] Jackie CK Cheung, “Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection”, B. Sc. (Hons.) Thesis in the Department of Computer Science of the Faculty of Science, University of British Columbia, 2008.

[3] Jezek K., Steinberger J., Automatic Text Summarization, in Snasel, V. (ed.) Znalosti 2008, pp 1-12. FIIT STU Brarislava, UstavInformatiky a softveroveho inzinierstva (2008) ISBN 978-80-227-2827-0

[4] Rasim ALGULIEV, Ramiz ALIGULIYEV, Evolutionary Algorithm for Extractive Text Summarization, in Intelligent Information Management, 2009, pp 128-138. doi:10.4236/iim.2009.12019

[5] Gary Blake and Robert W. Bly, The Elements of Technical Writing, pg. 117. New York: Macmillan Publishers, 1993. ISBN 0020130856

[6] "Abridgment". m-w.com. Merriam-Webster.

[7] Definition of Executive Summary from Colorado State University

[8] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, “Optimizing Text Summarization Based on Fuzzy Logic”, in Proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008.

[9] Pierre-Etienne Genest, Guy Lapalme, Fully Abstractive Approach to Guided Summarization, in ACL '12 Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2, 2012, pp 354-358

[10] F. Canan Pembe and Tunga Güngör, “Automated Query-biased and Structure-preserving Text Summarization on Web Documents,” in Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, İstanbul, June 2007.