International Journal of Engineering & Technology, 7 (4.5) (2018) 728-731

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A brief review on text summarization methods

Rasmita Rautray [1] *, Lopamudra Swain [2], Rasmita Dash [3], Rajashree Dash [1]

[1] Department of CSE,Siksha 'O' Anusandhan, Deemed to be University,Odisha,India
[2] DDE student, Department of CSE, Siksha 'O'Anusandhan, Deemed to be University, Odisha, India
[3] Department of CS&IT, Siksha 'O' Anusandhan, Deemed to be University Odisha,India
**Corresponding author E-mail:*

Abstract

In present scenario, text summarization is a popular and active field of research in both the Information Retrieval (IR) and Natural Language Processing (NLP) communities. Summarization is important for IR since it is a means to identify useful information by condensing the document from large corpus of data in an efficient way. In this study, different aspects of text summarization methods with strength, limitation and gap within the methods are presented.

Keywords: *Summarization Steps; Methods; Summary.*

1. Introduction

With advances in information technology, people face the problem of dealing with tremendous amounts of information and need ways to save time and effort by summarizing the most important and relevant information. Thus, automatic text summarization has become necessary to reduce the information overload. Text summarization (TS) is the process of producing a condensed form of the original content or summary for human consumption. A summary to be generated from one or more texts conveys the most important and relevant information present in the original text(s). Which is significantly less than the original text(s) [1], [2]. The process of summary generation involves three stages such as: topic identification, interpretation or compaction and summary generation. Initially, to identify the main theme, text document is transformed into source representation by interpretation of the meaning; categories relevant and irrelevant information, and then reduce it into a shorter version in summary representation. At last, it merges the previously identified information to generate summary. The goal of TS problem can be achieved by selecting a subset of the most important sentences from the original text documents and summaries, by composing novel sentences and unseen sentences in the original sources, is also called extractive and abstractive summary respectively.

2. Summarization stages

Figure 1 illustrates the stages of summarization system such as topic identification, interpretation or compaction and summary generation. In topic identification, document(s) to be summarized are analyzed. In the next step, texts are processed based on types and methods appropriate for summary representation then sentences for summary is generated.

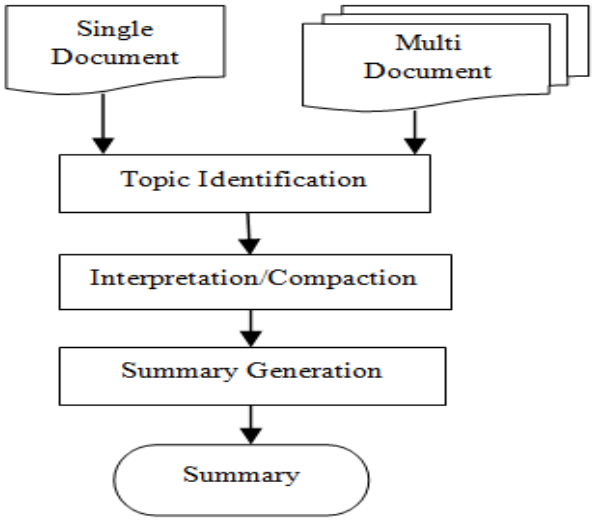

Fig. 1: Overview of Summarization System.

2.1. Topic identification

Most of summarization systems concentrate on this stage rather than other two stages. Here documents of different size to be summarized are analyzed for finding central topic from predefine schemas of what's important, is defined as top-down method. Alternatively, filtered out to retain only the M% (summary length with respect to the original text length) of most important by assigning score to each section of the document and rank them, is defined as bottom-up method [3]. The top-down method based system needs particular criteria of interest, which is used to focus certain type of information is called information retrieval (IR). IR is automatically retrieval of structure information from unstructured and/or semi-structured machine readable documents, which is relies on basic content management principle: "Content must be in context to have value". Contextualized content is ontologically categorized and semantically well-defined data. Alternatively, bottom-up method based system needs generic important metrics,

which is used to rate important contents, is referred as information retrieval [4]. It needs large number of text to be analyzed and operates at word level clues.

2.2. Interpretation/compaction

Topic interpretation or compaction is more complex task than the topic identification. The relevant part of the original document is converted into a shorter form, which can be easy or difficult. The easy method is compaction of document by presenting abbreviations in place of full forms, eliminating insignificant words etc. is called as extract summary. Whereas interpretation, compression, reformulation of document in a more general or new form by understanding the whole document is called as abstract summary, which is difficult [2].

2.3. Summary generation

Summary Generation is the final stage of summarization system, which addresses key sentences selection as summary sentence by merging and fusing information of the previous steps [1]. Thereafter it requires some additional operations to compress it further in order to increase the density of information. However, this is not so easy to handle, but most approaches only focus on first two stages [5].

3. Summarization methods

Text summarization is classified into two types namely Extract summaries and Abstract summaries, based on text analysis and summary generation, and the various methods for text summarization are either of above two. Extract summaries is formed from the input text(s) by reusing the important portion of the text and produce precisely as the summary. This summary consists of sentences those have already appeared in the text [5-7]. In contrast, Abstract summaries are created by regenerating the extracted content by analyzing the original text content and then present its comprehension in a human understandable form [1, 2]. To generate a high quality summary, various natural language processing techniques are used with the concepts such as language analysis, information inference etc. This section also discussed their strength, limitation and gaps between the methods in Table 1.

3.1. Statistical-based approach

The basic aim of statistical approach is to determine the most relevant sentences as summary sentences based on statistical features such as term frequency (tf), Term frequency-Inverse Document Frequency (tf-idf), word frequency, word weight etc. The features are used to assigning weights to the words present in the document and sentences in order to determine the relevant sentences as summary sentence [8].

3.2. Graph –based approach

The graph based approach has been pursued rigorously and proved to be better than other approaches. The basic concept of graph is to represent the connection or linking between the objects, where connections are based on their underlying relation. In text summarization, the graph represents the text structure along with sentence inter-relationship of the document. If inter sentence similarity is potentially higher than the threshold then a link is made between the sentences called as edges of the graph. Once graph is built, then central sentences are selected as summary sentences [9], [10].

3.3. Rhetorical-approach

In rhetorical approach based text summarization, creates a rhetorical relationships between different parts of the text and concept text are decomposed and extracted as summary is based on the rhetoric. The structure of approach presents connections and interpretations between different sections and subsections of the text logically and then, the summary is formed from the rhetorical connections between the identified text units [11], [12].

3.4. Linguistic-approach

This approach analyzes the words by considering the inter word connections and, original concept. The most common concepts used in linguistic approach are: Lexical chain and word net. Lexical chain presents text contiguous structure and some topics which are illustrated throughout the text by exploiting relations between words. For capturing the intuition that topics are expressed using not a single word but instead different related words. It is also used for information retrieval as well as grammatical error correction. But Word Net determines the use of relationship between words, where each word are sensed differently as well as word relationship are represented relationally by synonym, antonym etc. [1], [2].

3.5. Knowledge-based approach

Most of the document or article consists of contents related to a particular topic or event, since then researchers have put an effort in utilizing the background knowledge (i.e. ontology) for improving the summarization results .These topics or events generally belong to a particular domain containing own common knowledge structure. For improving the summarization result and improving the background knowledge a huge effort has been put by the researchers. For domain specific documents, ontology can be very useful. In ontology based text summarization, the key concepts pertaining to the documents is identified and sentences are selected based on predefined concepts of ontology [13], [14].

3.6. LSA-based approach

Latent semantic analysis (LSA) is most common application area of Singular Value Decomposition (SVD). SVD is a powerful mathematical tool, which can be used to find principal orthogonal dimensions of multidimensional data. In LSA based approach, SVD is applied on word matrices and groups of document that are semantically related to each other even if they do not share a common word. This method is applied for extracting the words and sentences from the topic and contents of the documents. The importance of LSA based approach for summarization is to capture the conceptual relations represented in human brain [15], [16].

3.7. Topic based approach

In topic based approach [13], the structure of the entire topic is characterized in terms of topic themes. Those are represented in terms of events that are reiterated throughout the document collection, and therefore represent repetitive information. Topics can be represented in five different ways such as topic signatures, enhanced topic signatures, thematic signatures, modeling the content structure of documents and template.

3.8. Concept obtained approach

The main idea of this approach is to obtain the concept of words based on Hownet [17], and use those concepts as feature, instead of word. To form a rough summary concept of vector space model is used, and then calculate degree of semantic similarity between the sentences for reducing its redundancy. This method goes through following stages:
Stage 1: To achieve text concept and establishing conceptual vector space model, use Hownet as a tool.
Stage 2: Evaluate concept importance using conceptual vector space model.

Stage 3: To select final summary, calculate importance of the sentence with reducing the redundancy.

3.9. Fuzzy logic based approach

In this approach, characteristics of each text such as length of each sentence, similarity too little, similarity to key word and etc are considered as the input of fuzzy system. Then, enter all the rules needed for summarization into knowledge base of the system. Afterwards, each sentence is assigned with a value from zero to one based on sentence characteristics and the available rules in the knowledge base. The obtained sentence value determines the degree of the importance of the sentence in the final summary. The important sentences are extracted using IF-THEN rules according to the feature criteria. [18].

Table 1: Strength, Limitation and Gaps of Different Methods

Method	Strength of method	Limitation of method	Gaps within the methods
Statistical method for summarization	Many possible compressions, ranked by sentence score can be produced.	Word order in the sentence cannot be changed, does not allow reorganization of the existing structure and Problem on handling coherence of text.	Implementation of unresolved anaphors, cataphores, synonyms and context dependent terms.
Graph based method for summarization	It allows flexibility in representation and lead to better overall results in content selection for multi-document summarization. The approach does not require language-specific linguistic processing.	Subtopic identification and information loss	Difficult to detect subtopic and information loss during summarization if the document has more than one idea
Rhetoric method for summarization	Construction of coherence structure based on rhetoric relation helps to find centrality of the textual units of the structure, will reflect their importance.	Tree structure representation requires to present rhetorical relations (structural features) between sentence segments of the document for non structural features.	Implementation of pattern matching algorithm for more rhetorical strategies like verbal irony & apophasis.
Linguistic method based summarization	It involves semantic and pragmatic analysis to find the main concepts of the documents.	Uses high quality linguistic analysis tools and linguistic resources. Requires much memory for saving the linguistic information. It is harder to implement	Implementation of centering theory with semantic transition pattern for multi document summarization
Knowledge based summarization	Summary result utilizes the background knowledge for relationship within the information of a domain.	It is harder to implement	Implementation on non-hierarchical ontology & anaphora resolution
LSA based summarization	Identify semantically important sentences.	It does not use word knowledge except input text. It does not uses information of word order, syntactic relations, or morphologies and the performance of the algorithm decreases with large and inhomogeneous data.	Searching and presenting web document summary in concise and coherent form
Topic based summarization	Sentences that appear in the topic are considered to be more important and are more likely to be included in the summary	Topic based approach cannot be effective if the document does not include any title information.	Obtaining more accurate prior information, for instance query expansion when the query is short
Concept based summarization	WordNet provides concept associated by each word in the text and frequency computation on concepts instead of a particular words.	Harder to implement	NA
Fuzzy logic based summarization	Identifying the sentence features by applying fuzzy rules.	Designing fuzzy rules, which have to cover all the relationships among the parameters of the document?	Implementation of fuzzy with different learning methods for text summarization

4. Conclusions

The research in text summarization has produced various kind of online as well as offline summarization systems which have been experimented in last few years. But, it still required a lot of improvement due to huge amount of data available in different forms and issues in evaluations of the generated summaries. To understand text summarization, we addressed a theoretical review of background information on text summarization system, especially focusing on various methods for summarization with respect to different aspects like strength, limitation and gaps within the methods.

References

[1] Gholamrezazadeh S, Salehi MA, Gholamzadeh B. A comprehensive survey on text summarization systems. *Computer Science and its Applications*. 2009 Dec, 10, pp. 1-6.

[2] Hovy E, Lin CY. Automated Text Summarization and the Summarist System, TIPSTER III Final Report (SUMMAC). 1998, pp. 197-214.

[3] Hovy E, Lin CY. Automated text summarization in SUMMARIST. MIT Press.1999, pp. 81–94.

[4] Hovy E, Lin C Y, Marcu D. Automated Text Summarization, SIGIR'99 Tutorial, Berkeley, CA, Tutorial: Automated. 1999.

[5] Lloret E, Palomar M. Text summarisation in progress: a literature review. 2012, 37, pp. 1-41.

[6] Partha L. Text summarization,June 13, 2002.

[7] Patil S R, Mahajan S R. Domain Specific e-Document Summarization Using Extractive Approach. *IJCA Proceedings on International Conference and workshop on Emerging Trends in Technology (ICWET)*. 2011, 11, pp. 36-41.

[8] Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*. 2(2), 159-65.

[9] Wan, X., Yang, J., & Xiao, J. (2007). Towards a unified approach based on affinity graph to various multi-document summarizations. *In: Proceedings of the 11th European conference*, 297–308.

[10] Fellbaum, C. (1998).WordNet: an electronical lexical database. *The MIT Press*, Cambridge.

[11] Kamyar, H., Kahani, M., Kamyar, M., & Poormasoomi, A. (2011). An Automatic Linguistics Approach for Persian Document Summa-

rization. *InAsian Language Processing (IALP). 2011 International IEEE Conference*. 141-44.

[12] Lin, C.Y., & Hovy, E. (2000). The Automated Acquisition of topic signatures for text summarization. *Proceeding COLING '00 Proceedings of the 18th conference on Computational linguistic*. 1, 495-501.

[13] Li, S., Ouyang, Y., Wang, W., & Sun, B.(2007). Multi-document summarization using support vector regression. *In: The document understanding workshop (presented at the HLT/NAACL)*. Rochester. New York USA.

[14] Priya, G. P, & Duraiswamy, K. (2012). An approach for concept-based automatic multi-document summarization using machine learning.International Journal of Applied Information Systems (IJAIS). 3(3), 49-53.

[15] Zamanifar, A., Minaei-Bidgoli, B., & Sharifi, M. (2008). A New Hybrid Farsi Text Summarization Technique Based on Term Co-Occurrence and Conceptual Property of Text. *In Proceedings of Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing. IEEE*. Washington, DC, USA, 635-639.

[16] Wang, M., Wang, X., & Xu, C. (2005). An Approach to Concept Oriented Text Summarization. *In Proceedings of ISCIT'05. IEEE international conference, China*. 1290-1293.

[17] Suanmali, L., Binwahlan, S. M., & Salim, N.(2009) Sentence features fusion for text summarization using fuzzy logic. *Hybrid Intelligent Systems, 2009. HIS'09. Ninth International Conference on*.IEEE, 1.

[18] Das, D., & Martins, A.F. (2007). A survey on automatic text summarization. Literature Survey for the Language and Statistics II course at CMU. 4, 192-5.