A comprehensive analysis on extractive automatic text summarization

Dr.V.R.Angela Deepa[1], Dr.P.Alagambigai[2]

[1]Assistant Professor,[2] Assistant Professor

Department of Information Technology, Women's Christian College,Chennai.

[1]angelrajan.research@gmail.com,[2]alagambigai@yahoo.co.in

Abstract

The modern technology demands the maintenance of the increasing data which are in structured and unstructured form. The text documents collected in the various platforms occupies a massive space in the architectural structure of the computer system both physically and virtually. Apparently the users demand the summarizing of the collected documents for easy access and usage. To enable this automatic text summarization came into phase. The automatic text summarization condenses the text documents into meaningful phrases and textual messages which helps the user to understand the conceptual ides behind each core values. The importance of automatic text summarization stands as a helping source in the growing data. This paper discusses the basic blocks of the automatic text summarization and its feature in identifying the intricate properties of the meaningful text through various approaches.

Keywords: Text summarizations, Natural language processing, Machine learning, Graph representations, Extractive, Abstractive

1.Introduction

The process of producing the meaningful text by condensing the source text for easily understanding of content with ease access is termed as automatic text summarization. This process grabbed the attention of the researchers ever since 1950 [1] but found its importance in the recent years due to the tremendous growing of data. Currently, to handle the massive data produced by the online sources like social media networks the text summarizations are grounded upon Natural language processing techniques [2] for better production of results.

Generally, the process of summarizations is categorized under three prime phases. [3]

- ✓ Interpretation:

This phase mainly focus on the input source texts are converted to appropriate semantic representations with the help of natural language understanding a part of natural language processing.

- ✓ Transformation:

This phase centers the output of the interpretation phase. Through the mapped text data, the elements of the important data are identified for better summary of the source text.

- ✓ Generation

This phase is the edited readable content of the source text. Natural language Generation (NLG) plays an important role in bringing out the summarized targeted text.

The primary aim of the automatic text summarizers is to excerpt the concise content of the given document with minimum repetition of the sentences and maximum presentation of the summarized data. Thus this paper gives an elaborate description on the various approaches that have been used in developing text summarizers. The paper is organized as recent works in the field of text summarizers followed by various approaches used in the field of text summarizers followed by conclusion.

2. The Dimensional Property of the Automatic Text Summarization

The following section gives an elaborate description on the various works carried out in the automatic text summarization process based on the dimensional features. The features are based on the nature of the document, requisite and the type of the expected output. [4][5].

2.1 Nature of the document

Millions of data are collected via the online sources like Facebook, twitters, blogs, webpages, e-mails, news blogs and stream of articles from various domains like finance, scientific findings, research data etc., These textual data can be either in a single type or multi-document. In single type document the final summary is based on the single document [6]. The summarizers play with the single document for the summarized data.

In multi- document the final summary obtained is based on the multi documents available with the same contextual meaning and values [7]. Therefore, the summarizers are grounded on the multi documents to pick the sentences for the final summary of the document. Repetition of sentences is the prime factor to be taken into consideration. Similarly, domain specific documents can be used under this category for easy summarizations of data.

2.2 ***Requisite***

The following are few mentionable summarizations techniques based on the radius of the user preference.

2.2.1 Generic summarizations

Generally, the process of summarization which satisfies a group of generic audience with less target on the preferential user is termed as Generic summarizations. Apparently these process are nonspecific that gives a surmise based output. [7]

2.2.2. Query based summarizations

These summarizations are based on the query given by the user. The human understandable query in a form of phrases or sentences are used to retrieve the summaries of the information. This plays a prime role in information retrieval systems and evidence based medicine [8].

2.2.3. Update summarizations

These summarizations give out the summary with an assumption that the user have already read the textual document. This process targets the multi-document type which brings out the summary of the diverse input document. [9]

2.3. ***Expected Output***

A vast amount of research has been carried out in the text summarizations based on two important models (i)Extractive (ii) Abstractive. As discussed earlier, the extractive models identify the sentences from the given document based on the higher frequency count to give out the final summary of the textual document. In contrary, the abstractive model compresses the textual content of the document in accordance to the user preferential quotient.

3.Features of Automatic Text Summarizations

The factor of identifying the sentences from the source input document to frame the composed summarized content tails an important part in the automatic text summarizations. The key features that support the extraction of sentences from the textual content is discussed below.

3.1 Probability of Word

The simplest way to count the occurrence of the words in the document. This measures the raw frequency of the given word with the times of occurrences. [10]

3.2 Term Frequency

Repetition of the words is the source of this features. The scores of each sentences are calculated based on the frequent occurrence of the words. [10]

3.3 Location

The point of indication of the start and the end of the sentences in a paragraph and the positional arrangement of few words are few intuitions that are taken into considerations when extracting sentences from the text documents [10]

3.4 Cue Method

Some textual content emphasis the importance of the sentences by referring the content with other sources, authors or words etc., in connection to their necessity. These are taken as parametric cues for the extraction of the sentences.

3.5 Length of the sentence

The length of the sentence determines the importance of the sentence. Very short sentences can be meaningless and very long sentences can also be irrelevant to the content. [11]

3.6 Title/Headline word

The header part of the document which consists of the topic heading can definitely reflect a positive vibe on the selection of the sentences as it denotes the core of the document. [10]

3.7 Similarity

The language understanding part with linguistic background is needed to find out the similarity between the title and the remaining content of the document. [11]

3.8 Grammatical notions like Proper nouns

This grammatical notion which refers the name of the person, place or organizations are essential in document summarizations. [11]

3.9 Vicinity

The distance between two word entities defines the selection of the appropriate sentences for the extraction of the summarized output. [11]

4.**Approaches Used in Extractive automatic Text Summarization**

This session brings out the various approaches used in the field of the automatic text summarizations. For past few decades the role of text summarizers forecast a prominent phase in the field of NLP. Though there are many models that performs the task of summarizations the extractive summarizations are comparatively used in high ratio. The general model of extractive summarizers is based on three important tasks

- Step 1: Construction of the intermediate representations
- Step 2: Score the sentences based on the intermediate representations
- Step 3: Selection of summary based on the top important sentences.

- Step 1: Intermediate representations:

This step involves the understanding of the content. It consists of two major categories

i. Topic representations
ii. Indicator representations

- *Topic representations*

It targets on the topic for the better understanding of the concept. The following are the few approaches used in the topic representations

- Topic Words:

This approach targets the topic words for the identification of the words from the input documents. Initially frequency threshold was used to represent the topic words. The advance level methods used log-likelihood ratio test, in which the explanatory sentences are trapped under the topic signature. The two ways of finding the results using log-likelihood ratio test are function of the number of topic signature which targets the longer sentences or the proportion of the topic signature that measures the density of the words. [12]

- Frequency driven approaches:

The two main methods used in this approach are probability of words and term frequency–inverse document frequency. The frequency of the word occurrences is the prime role in the identification of the word probability followed by the selection of the best sentences that contains the higher frequency of the words. Though this method was very helpful the term frequency–inverse document frequency method helped in the decisions of the wordlist. A weighing factor is used on the words to assess the importance of the words. Thus based on the conceptual content of the words and the low weighing factor the uncommon words are shortlisted and excluded [13].

- Latent semantic analysis:

An unsupervised method that tailors with the observed words for the summarizations of text. It constructs the sentence matrix based on the weightage of the sentences. It handles both single and multi-document type textual files. Thus the summaries are created based on the semantic representations of the words across the combinational documents. [14] [15] [16] [17] [18]

- Bayesian topic Model:

It is a probabilistic model that gives a detailed description on the document topic section. This model is powerful as it targets at higher proportion on the topic of the document. Summarizing the conceptual content of the multi-document which grounded on the description of the topic of the document produces quality results. [19].

- *Indicator representations*

Ranking the text sentences based on a set of features without relying on the topics of the input text. The following are the few approaches used in the indicator representations

- Graph representation

The similarity between the two sentences are identified using the formation of vertices and edges in the connected graph. The similarity level between the sentences are calculated using the cosine similarity and the TFIDF. Thus the no of edges and co-edges created in the document determines the selection of the sentences for the text summarization. The notable disadvantage of this approach is that it fails to understand the semantic and syntactic representations of the sentences. [20].

- Machine Learning representation

This approach handles the summarizations process as a classification problem. [21] A supervised approach demands for a trained labelled data which is not possible for all the input source text files. Therefore, the semi supervised approach that requires mere labelled data with undefined raw data worked better for text summarizations. The machine learning algorithms like Naive Bayes, decision trees, support vector machines, Hidden Markov models and Conditional Random fields are few mentionable algorithms are quite effective in achieving summarized texts.

Thus the represented step 2-sentence scoring and step 3- sentence selection in the extractive models are straightforward steps which doesn't demand for the elaborate description.

5. Conclusion

Automatic text summarizations are helpful in the generation of summaries that are benefits the human race to comprehend the content of the single and multi-documents. Though the created summaries may not reflect the full essence of the textual hues, it supports in the easy understandable of the contextual data which helps researchers and other authors for facts in a limited period of time.

References

[1] Hans Peter Luhn. 1958. The Automatic Creation of Literature Abstracts. IBM Journal of Research Development, 2:159–165.

[2] Ani Nenkova and Kathleen R. McKeown. 2012. A Survey of Text Summarization Techniques, pages 43–76. Springer US, Boston, MA.

[3] Karen Sparck Jones. 1999. Automatic Summarising: Factors and Directions. In Advances in Automatic Text Summarization, pages 1–12. MIT Press.

[4] Basiron, Halizah, et al. "A review on automatic text summarization approaches." Journal of Computer Science 12 (2016): 178-190

[5] Horacio Saggion and Thierry Poibeau. 2013. Automatic Text Summarization: Past, Present and Future, pages 3–21. Springer Berlin Heidelberg, Berlin, Heidelberg.

[6] Juan-Manuel Torres-Moreno. 2014b. Single-Document Summarization, chapter 3. Wiley-Blackwell

[7] Ani Nenkova and Kathleen R. McKeown. 2011. Automatic Summarization. Foundations and Trends in Information Retrieval, 5(2–3):103–233.

[8] Sarker, Abeed; Molla, Diego; Paris, Cecile (2013). An Approach for Query-focused Text Summarization for Evidence-based medicine. Lecture Notes in Computer Science. 7885. pp. 295–304.

[9] Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the TAC 2008 Update Summarization Task. In Proceedings of the First Text Analysis Conference (TAC 2008), pages 1–16.

[10] Khan Atif, Salim Naomie, “A review on abstractive summarization Methods”, Journal of Theoretical and Applied Information Technology, 2014, Vol. 59 No. 1.

[11] Reeve Lawrence H., Han Hyoil, Nagori Saya V., Yang Jonathan C., Schwimmer Tamara A., Brooks Ari D., “Concept Frequency Distribution in Biomedical Text Summarization”, ACM 15th Conference on Information and Knowledge Management (CIKM), Arlington, VA, USA,2006

[12] Text Summarization Techniques: A Brief Survey, Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut, 2017

[13] Fachrurrozi M., Yusliani Novi, and Yoanita Rizky Utami, "Frequent Term based Text Summarization for Bahasa Indonesia", International Conference on Innovations in Engineering and Technology Bangkok (Thailand), 2013.

[14] Froud Hanane, Lachkar Abdelmonaime and Ouatik Said Alaoui, "Arabic Text Summarization Based On Latent Semantic Analysis To Enhance Arabic Documents Clustering", International Journal of Data Mining & Knowledge Management Process (IJDKP), 2013, Vol.3, No.1.

[15] Ragunath R. And Sivaranjani N., "Ontology Based Text Document Summarization System Using Concept Terms", ARPN Journal Of Engineering And Applied Sciences, 2015, Vol. 10, No. 6.

[16] Babar S.A. and Thorat S.A., "Improving Text Summarization using Fuzzy Logic & Latent Semantic Analysis", International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2014, Vol. 1 Issue 4.

[17] Josef Steinberger, Karel Je˘zek, "Evaluation Measures For Text Summarization", Computing and Informatics, Vol. 28, pp 1001– 1026, 2009

[18] Ozsoy Makbule Gulcin, Cicekli Ilyas and Alpaslan Ferda Nur, "Text Summarization of Turkish Texts using Latent Semantic Analysis", 23rd International Conference on Computational Linguistics (Coling 2010), pages 869–876, Beijing, 2010

[19] Inderjeet Mani and Eric Bloedorn. 1999. Summarizing similarities and differences among related documents. Information Retrieval 1, 1-2 (1999), 35–67

[20] Kumar Niraj, Srinathan Kannan and Varma Vasudeva, "A Knowledge Induced Graph-Theoretical Model for Extract and Abstract Single Document Summarization", Computational Linguistics and Intelligent Text Processing - 14th International Conference, 2013

[21] Sarkar Kamal, Nasipuri Mita, Ghose Suranjan, "Using Machine Learning for Medical Document Summarization", International Journal of Database Theory and Application, 2011