www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 12 Dec 2015, Page No. 15172-15176

Review on Automated Text Summarizer using Top K-Rules

Ms.Priya J.Patel[1] Professor.Pravin G.Kulurkar[2]

[1] RTMNU University, VIT Nagpur, Maharashtra, India
priyapatel2405@gmail.com
[2] RTMNU University, Department of Computer Science and Engineering,
VIT Nagpur, Maharashtra, India
pravinkulurkar@gmail.com

Abstract: ***In this paper we address the automatic text summarization task. Text Summarization was showed to be an improvement over manually summarizing the large data. It summarizes the salient features from the text by preserving the content and serves the meaningful summary. To design an algorithm that can summarize a document by extracting key text and attempting to modify this extraction using a thesaurus and to reduce a given body of text to a fraction of its size, maintaining coherence and semantics. This summarization method can be done in natural language processing approach integrated with rule mining.***

Keywords: *Automatic Summarization, Extraction, Natural Language Processing, Top K-Rules.*

1. Introduction

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. It is very difficult for human beings to manually summarize large documents of text. Text Summarization methods can be classified into abstractive and extractive summarization.

Abstractive summarization aims at paraphrasing the source document, similar to manual summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into smaller form. The importance of sentences is decided based on statistical and linguistic features of sentences. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. 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.

The summarization has been studied by the Natural Language Processing community for nearly the last half period. The simple definition provides three important aspects that characterize research on automatic summarization:

- Summaries may be produced from a single document or multiple documents.
- Summaries should preserve important information.
- Summaries should be short.

Automatic text summarization is a useful tool when there is a lot of textual information to be analyzed manually. Automatic summarization is used to condense the large amounts of textual data. This achieves the following benefits:

- Firstly, several redundancies can be removed. The user does not excess time reading repetitive data.
- Secondly, summarization allows you to remove data that is not necessary to the understanding of the document.

There are many methods to proceed with automatic text summarization. In this model an extractive technique to obtain the summary from the given text. This summary is then improved further by replacing a few parts of it using an abstractive technique. The extraction of sentences from the document is done keeping consistency in mind and therefore the summary maintains the core of the original document. The sentences are then ranked using a text- ranking algorithm and the final cluster or summary is formed.

The important functions of the summarizer are:

- Reducing a single document to a user-defined fraction of its original size while maintaining coherence.
- Choosing the most relevant and important sentences from the text.
- Improving the length of the summary by using a thesaurus to replace semantically related units.

In effect, we aim to extractive summarize a single English document, not more than 300 sentences long, to a fraction of its original size, while maintaining cohesion, and then use a lexical database to abstract the generated summary.

2. Background

2.1 What is TEXT SUMMARIZATION?

A summary can be defined as a text that is produced from one or more texts, that contain a significant portion of the information in the original text, and that is no longer than half of the original text. Text summarization is the process of

distilling the most important information from a source to produce a concise version for a particular user and task.

When this is done by means of a computer, i.e. automatically, we call this Automatic Text Summarization. Despite the fact that text summarization has traditionally been focused on text input, the input to the summarization process can also be multimedia information, such as images, video or audio, as well as on-line information or hypertexts. Furthermore, we can talk about summarizing only one document or multiple ones. In that case, this process is known as Multi-document Summarization (MDS) and the source documents in this case can be in a single-language or in different languages.

The output of a summary system may be an extract (i.e. when a selection of "significant" sentences of a document is performed) or abstract, when the summary can serve as a substitute to the original document. We can also distinguish between generic summaries and user-focused summaries. The first type of summaries can serve as surrogate of the original text as they may try to represent all relevant features of a source text. They are text-driven and follow a bottom-up approach using IR techniques. The user-focused summaries rely on a specification of a user information need, such a topic or query. They follow a top-down approach using IE techniques.

2.2 Process of Automatic Text Summarization

Traditionally, summarization has been decomposed into three main stages which is:

- **Interpretation** of the source text to obtain a text representation,
- **Transformation** of the text representation into a summary representation, and,
- **Generation** of the summary text from the summary representation Effective summarizing requires an explicit and detailed analysis of context factors. Three classes of context factors: input, purpose and output factors.

In other words, first clean the text file by removing full stop, common words (conjunction, verb, adverb, preposition etc.). Then calculate the frequency of each word and select top words which have maximum frequency. This technique retrieves important sentence emphasize on high information richness in the sentence as well as high Information retrieval. These related maximum sentence generated scores are clustered to generate the summary of the document. Thus we use k-mean clustering to these maximum sentences of the document and find the relation to extract clusters with most relevant sets in the document, these helps to find the summary of the document.

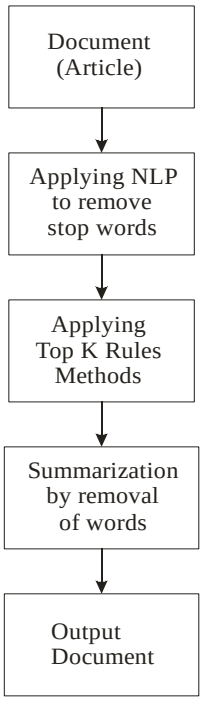

Fig: Design on automatic text summarization using top-k rules

3. Literature Review

3.1 Automatic Text Summarization Based on Rhetorical Structure Theory:

Li Chengcheng[16] presented a new method called Rhetorical Structure Theory for effective automatic text summarization. This new method is based on natural language generation method for effective summarization of an article. The paper focused on text summarization using the rhetoric structure theory. These automatically shorten the document that a user is in need of and gives the summarized sentences. This theory extracts the rhetoric structure of the text and a compound that relates the sentences. All the process is best explained by author.

After this identification, the summarized text is converted to natural language which is user friendly. This type of summarization using the clauses and compounds of rhetoric structure is highly capable. The main idea of this method is analyzing the candidate sentence identifying the rhetoric relations and forms the important part of sentence useful for final summarization.

Past systems based on the frequency of word generation i.e. the sentence is important because a key word is many times present in that sentence is inefficient and it lacks preciseness and recall. The RST system based on knowledge or script based analysis can efficiently rule out those backlogs.

The drive of the paper is explanation of rhetoric structure and summarization process basing on RST. A nucleus is an important part of sentence and supplies a reader much information whereas satellite independent of nucleus increases it's understand ability. Sometimes a satellite supplies more information than a nucleus.

Here, a RST tree is constructed by placing the nucleus as the root of the tree and satellites as leaf nodes. The summarization is done using the nodes i.e. nucleus. These nucleuses are also given weights based on script based analysis. Next follows the summarization method in which the tree starts its construction. For this the entire text is to be divided in to individual sentences which are significant. This can be done by dividing

the sentences based on the comas, quotes and semicolons present in the sentences. Also the division is done by the presence of 'and' the punctuation marks present before and after and. This is then done into a graph, deletes the unimportant sentences and then summarizes the entire text.
RS tree construction is presents in the paper is well and after the construction, Nucleus filter statement is made i.e. unimportant statements are deleted and nuclei which best suits the document meaning is left out. At this stage, a sentence is logically, structurally understood well and the value of sentence is known clearly. From this knowledge, the weights can be easily assigned to the sentences; lower weight sentences can be deleted. Now the system is all left with important information useful in the summarization. These sentences can be formed into complete, cohesive and readable summarization.

Observation: So as to conclude, the paper introduced the process of RST in summarizing the text of the document in such a large pool of data available in the web overcoming the drawbacks like recall and precision. But the paper should focus on the drawbacks like it cannot be applied on all documents like magazines. It's inefficiency of analyzing every sentence based on semantic progress and the domain being limited.

3.2 An Extractive Text Summarization Based on Multivariate Approach:

Esther Hannah[18] addressed a method to automatically summarize a text with the help of multivariate statistical technique, where multivariate is a form of statistics about the simultaneous observation and analysis of more than one statistical variable. The model they proposed a training methodology where the system trained by using manual summaries. The utilization of multivariate statistical technique for this task is acceptable by its ability to produce a model that resembles a relation. The model relied on primary subjective evaluation, in order to show that the approach is effective, efficient and promising.
The paper has introduced the statistical approach to extractive text summarization where multivariate is used to produce the weight for every sentence. The texts are ranked to classify them as summarized or not. The steps followed by the authors in the extraction are as follows:

- To bring out the early work done on the text summarization focusing on the contributions that laid the foundation for the research in this subfield of NLP.
- To discuss the proposed work under the various subsections namely, pre-processing, feature extraction, comparison vector generation, weight generation and ranking.
- Evaluation method has been used by the authors and the results are provided.
- Conclusion of paper with providing scope for future work.

In the first step the authors previously discuss about some works that were in practice like MEAD (a state of sentence extractor) in DUC and some other computationally expensive extractions including NLP based methods whereas the present system is much comfortable and cost-effective to get the result.
While coming to the next step, some probable subparts are introduced and the text is modelled by using two-phase classification 'in' & 'out' otherwise Boolean values '0' & '1' are assigned to the marked sentences respectively . The first subpart called Pre-processing is detailed by dividing it into four segments namely sentence segmentation, tokenization, stop word removal and word streaming and was explained by a system architecture figure.
Specifically the other subparts are implemented with the formulae to get the results. There are six formulae for subparts, each consisting one formula and thus are derived the six scores from these formulae depending on the keywords, number of articles, length, number of numerical data and the summation.
The feature subtraction part is derived from the sentence similarities, numerical values are derived from the numerical data, sentence comparative strength from the number of articles and node similarities from the summation by using these formulae.

Then the author uses the compression vector generation to check whether the sentence matches the summary or not by using the in-out classification and selected sentences is weighted by using the weight generation technique through which the ranks are assigned in order to decide which sentence should be first and which one is last. Multiple linear aggressions are a multivariate statistical technique, which study the linear correlations between sentences & a variable has been used in weight generation technique. The ranks are decided based on their weights and compression vector considering a formula.
This paper presents the work on evaluation in two methods namely intrinsic and extrinsic. Intrinsic mainly assess coherence and summaries while extrinsic assess impact of summarization.
The model result is obtained in final score which is derived by multiplying the each score with their respective weighted value obtained and then adding all the product values. To analyze the process the authors verified the system by considering 60 documents in which 30 are for training the system and 30 are for the testing the system. Precision, which is the average value of the documents (in percentage) is made and got the comparison with Microsoft word documents and presented in tabular forms. The comparison gives the evaluation of the system by verifying how many documents are produced by either system.

Observation: Though the authors got the assumed results, At last the exact summarization is not given for those documents with low precision value, that is for less sized documents the summarization is big than required and which could be verified by working on semantics way. This is the limitation of this proposal, which could be rectified further.

3.3 Evaluation method of automatic summarization calculating the similarity of text based on HowNet:

SUO Hong-guang[17] presented a better evaluation method for the automatic text summarization (ATS) which is based on identifying the equalities of the summary and the actual text document. Automatic text summarization refers to the process of minimize the quantity of the text preserving the actual content of the document. Once a summary is created, an evaluation technique is applied over the summary to validate it i.e. whether it is compatible to the actual text or not. This paper proposes one of the evaluation techniques far better than the authentic and old fashioned evaluation techniques. This technique is relied over vector space model and it does the process of evaluating a summary relying on HowNet.
The proposed technique is introduced to provide a precise and effective alternative to the available automatic summarization

algorithms. Here the techniques of various kinds of current automatic summarization algorithms along with their drawbacks are explained. Also, the benefits of the proposed technique over the available ones are clearly specified. This technique utilizes the HowNet in the vector space model to examine the actual content of the document. Also, this technique considers the parts of speech and further grammar which can be supposed to affect the meaning of the sentence. This analysis plays a key role in computing the priority of the terms to be included in the summary of the document.

Now-a-days, there is a huge enhance in the number of electronic media and the vast data provided in them. Due to this reason the data available on a particular subject may contain lot of overhead leading to us to move away from the actual content. Thus to solve this problem automatic text summarization technique is introduced. But, as this summary refers to the whole document, this summary must be a better contemplative to the original document. Hence, to provide us with a better summary, evaluation method over the summarization technique is needed to perform.

The evaluation technique can be done in any of the two ways. Such as: exterior evaluation technique and interior evaluation technique. The exterior technique refers to evaluating the Automatic Text Summarization algorithm followed and how the summary, formed, will act as in a document. The interior technique concerns only with the quality of the summary created. But both of these approaches have drawbacks. Exterior needs a lot of time and manpower whereas interior faces the problem that ideal summary is impossible. Also another problem faced by interior technique is P/R defects i.e. the length of the line in the summary to the length of the line in the document ratio.

The proposed technique helps to overcome these drawbacks and provide a better estimate technique. Many scientists use the interior technique in their summarization techniques i.e. to maintain the quality, efficiency, performance and consistency. Consistency refers to the extent to which the summary is flowing enough in the meaning and overall structure. The evaluation of the ATS includes four steps which are needed to be followed. They are: prior processing of summary, grasping conceptual characteristic term, computing the priority of characteristic term and comparing summary with original document.

The prior processing step includes eliminating spaces, missing words and stop words. Apart from these, the parts of speech of the words must be noted before proceeding to the summarization process. In the step of grasping conceptual characteristic term, the term is grasped using HowNet. This includes highest similarities with the preservation of the semantic data. In the step of computing priority of characteristic term, the priorities of the terms are calculated based on the number of times they appear in the document. For this purpose, it utilizes TF-IDF technique. In the step of comparing summary with the original document, Vector space model is utilized to find the similarities between the summary and the document.

Many experiments are conducted to establish that the proposed technique is better than the existing techniques. Two of them specified in this paper are: to compare 6 different types of word segmentation systems and evaluation outcomes of 3 various evaluation techniques. Thus, in this paper, the drawbacks of the existing techniques are specified along with their explanation. Also, it is shown how the proposed system overcomes their drawbacks and thus proved to be better. The proposed evaluation technique is based on Vector space model. It compares summary and document and to extract semantic data using the HowNet. This technique provided a enhanced way to compute the priorities of terms used in the document and to decide which one to place in the summary.

Observation: However, this technique faces the drawbacks as: Difficulty to calculate the priority of the terms which is to be worked on in the future. But, the proposed system is proved to be simple, precise, efficient and better than the authentic techniques. Also, this technique gives better outcomes.

3.4 Top K-Rules methods:

The current work on text summarization is limited to natural language processing based approaches. This approach is good for sentence level classification but might not be useful or might not give accurate results, when applied to an entire paragraph. Thus the efficiency of the existing system might be less as compare to the system which is based on top k-rules methods in rule mining.

In our proposed approach, we plan to integrate natural language processing with top k-rules methods so that, the advantages of both the techniques can be combined to create an automatic text summarizer. We will be using top k rules based approach to find out the support and confidence of text parts which appeared more frequently in the input dataset. This will allow us to find the best possible summary of documents which will be grammatically and content wise more accurate. Thus, the overall efficiency of the system will be increased.

4. Conclusion

In this paper, we have seen that Top k-rules are the best algorithm for mining, which shows that they are better than existing methods. Research on this field will continue due to the fact that text summarization task has not been finished yet and there is still much effort to do, to investigate and to improve. Definition, types, different approaches and evaluation methods have been exposed as well as summarization systems features and techniques already developed. Hence we propose the review on automated text summarization using top k-rules.

References

[1] Mihalcea, Rada, and Paul Tarau. "TextRank: Bringing order into texts." Proceedings of EMNLP.Vol. 4.No. 4. 2004.

[2] Ravi Som Sinha and Rada Flavia Mihalcea, "Using centrality algorithms on directed graphs for synonym expansion." FLAIRS Conference, AAAI Press, 2011.

[3] Blondel, Vincent D., and Pierre P. Senellart. "Automatic extraction of synonyms in a dictionary." vertex 1 (2011): x1.

[4] Sankar, K., and L. Sobha. "An approach to text summarization." Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies. Association for Computational Linguistics, 2009.

[5] George A. Miller (1995). "WordNet: A Lexical Database for English." Communications of the ACM Vol. 38, No. 11: 39-41. Christiane Fellbaum (1998, ed.) "WordNet: An Electronic Lexical Database." Cambridge, MA: MIT Press.

[6] Lin, C. Y. (2004, July). "Rouge: A package for automatic evaluation of summaries." In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop (pp. 74-81)

[7] Bird, Steven, Edward Loper and Ewan Klein (2009), "Natural Language Processing with Python." O'Reilly Media Inc.

[8]Sample Text Source: Grolier Electronic Publishing, Inc., 1995.

[9] H. Takamura and M. Okumura, "Text summarization model based on the budgeted median problem," in *Proc. 18th ACM Conf. Inf. Knowl.Manage.*, 2009, pp. 1589–1592, ACM.

[10] U. Hahn and U. Reimer, "Knowledge-based text summarization: Salience and generalization operators for knowledge base abstraction," *Adv. Automatic Text Summarization*, pp. 215–232, 1999.

[11] A. Molina, "A study on sentence compression for the automatic summarization," Ph.D. dissertation, Univ. d'Avignonet des Pays de Vaucluse (UAPV), Avignon, France, 2013.

[12] TAC, Tac 2011 guided summarization task guidelines. [Online].Available: http://www.nist.gov/tac/2011/Summarization/Guided-Summ. 2011.guidelines.html 2011

[13] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Celebi, S.Dimitrov, E. Drabek, A. Hakim, W. Lam, and D. Liu *et al.*, "Mead-a platform for multidocument multilingual text summarization," in *Proc.4th Int. Conf. Lang. Resources Eval.(LREC'04)*, 2004.

[14] D. Gillick, K. Riedhammer, B. Favre, and D. Hakkani-Tur, "A global optimization framework for meeting summarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'09)*, 2009, pp.4769–4772.

[15] D. Gillick and B. Favre, "A scalable global model for summarization," in*Proc. Workshop Integer Linear Program. Nat. Lang. Process.*, 2009, pp. 10–18, ACL.

[16] Li Chengcheng, Automatic Text Summarization Based On Rhetorical Structure Theory, 20IO International Conference on Computer Application and System Modeling (ICCASM 2010)

[17] SUO Hong-guang, ZHANG Jing-jing; Evaluation method of automatic summarization calculating the similarity of text based on HowNet, 978-1-4244-6585-9/10, 2010, IEEE

[18] M. Esther Hannah, Dr. Saswati Mukherjee, K. Ganesh Kumar, An Extractive Text Summarization Based On Multivariate Approach, 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)

[19] Osborne, M. (2002). Using maximum entropy for sentence extraction. In Proceedings of the ACL'02 Workshop on Automatic Summarization, pages 1{8, Morristown, NJ, USA.

[20] Nenkova, A. (2005). Automatic text summarization of newswire: Lessons learned from the document understanding conference. In Proceedings of AAAI 2005, Pittsburgh, USA

[21] Mani, I. and Bloedorn, E. (1997). Multi-document summarization by graph search and matching. In AAAI/IAAI, pages 622-628.

[22] Radev, D. R., Jing, H., Stys, M., and Tam, D. (2004). Centroid-based summarization of multiple documents. Information Processing and Management 40 (2004), 40:919-938.

[23] Evans, D. K. (2005). Similarity-based multilingual multi-document summarization. Technical Report CUCS-014-05, Columbia University.