Sentence based scoring method for extractive document based summarization review

Prashali Gupta
Department of Computer Science & Engineering
AIM & ACT, Banasthali University
Jaipur, Rajasthan
prashaligupta.15@gmail.com

Yogesh kumar Meena
Department of Computer Science & Engineering
Maliviya National Institute of Technology
Jaipur, Rajasthan
yogimnit@gmail.com

***Abstract*—Text Summarization is the process of reducing a text document in order to create a summary that retains the most important points of the original document. Text summarization can be single document text summarization and a multi document text summarization. This paper contains a large literature review in the research field of Text Summarization. In this paper we study about the single document text summarization. Here we study statistical or linguistic sentence scoring methods. Scoring sentence is extracted using different methodologies. Also study different type of summaries. Evaluate the summary using the ROUGE method. It is standard method for evaluating the system generated summary with the human generated summary.**

***Keywords*—Wordne; Lexical chain; Genetic; Fuzzy; ROUGE;**

I. INTRODUCTION

Text summarization is the process of automatically creating the shorter version of the text document or file. As we know the popularity of internet and a variety of information services is growing continue. Due to this reason obtaining the desired information in sort time period is becoming a serious problem.

Text summarization provides users with summaries of text document, allowing them to quickly understand the main idea of documents. Text summarization can be classified into two parts: extraction based and abstraction based text summarization.

In the extraction based text summarization, summary contain a set of most important sentences from a document, exactly as they appear. But in the abstraction based text summarization, summaries attempt to improve the coherence among sentences by removing duplicity. It also produces new sentences in the summary. Extractive summarization is much easier in comparison to abstractive summarization.

Generally extraction based text summarization method is performing in three steps:

- processing the original text data
- Scoring the sentence using different methods
- Select the high scoring sentences for summary generation threw different method

First step is the preprocessing step; in this step create the representation of the text. Normally, text divides into paragraphs, sentences, and tokens. In the second step scoring the sentences for finding the important sentences by which understanding the whole text. In the third step, the score provided by the previous step is used and generate the summary.

Traditional text summarizations have used three different techniques for extractive summarization. These techniques are: linguistic approaches, statistical approaches and hybrid approaches. All these approaches have some limitations. Linguistic approaches have some problem in using linguistic analysis tools and linguistic resources (WordNet, Lexical Chain, etc.); they require much memory and processor capacity because of additional linguistic knowledge and complex linguistic processing. This technique is very useful to understand the document for summary generation. On the other hand, statistical approaches is easy to implement and don't suffer from memory and processor capacity problem. These approaches can summarize texts using various statistical features (title, position, length etc.) But all the statistical features are not compatible for summary generation because some feature depend on particular format like if the document does not contain title then we can't calculate score based on the title. Last final approach is hybrid approach; it exploits best of both the previous method for meaningful and short summary.

Text summarization can also be classified into single document text summarization and multi document text summarization. A wide range of summary type depending on what is the requirement of the user, what type of input is given, etc. Different taxonomies is available, in which one of the most existing taxonomies [21], there is three part of context factors: input, purpose, output factors. Input factor is related to the source text, such as language, type of text (xml, database etc.), genre, and etc. Purpose factor is depending onto the user requirement and for what purpose it is using for example literary review or emergency alert. Finally the output factors, focus on the style and coverage, and normally derive from purpose factors. Another taxonomy purposed by [22] is also a similar taxonomy to the previous one. Here the types of summaries are classified in the same way as the input, output and purpose factors described in the previous taxonomy.

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors).*

It is also compulsory to understand the difference between generic and query-focused summaries which are commonly known as user focused or topic focused summaries. Generic type summaries are considered as a substitute for original text. Query focused summaries are based on needs and query of users. Board distinction is made between two types of summaries: "inductive" and "informative". Inductive indicates theme of content, as a result we get the brief idea of original text, where is informative summary elaborate the topic. Other summaries are also taken into account as "critical evaluative abstracts" which shows authors views about particular subject which includes review, opinions, feedback etc.

In the recent year new types of summaries also approved like textual, genres, sentimental-based summaries, update summaries etc. Language is very important in the summary generation and it can be divides into mono-lingual, multi-lingual, and cross-lingual summaries. It depends onto the number of languages used in the summary. When the language of original text and summary text is same, then this is mono-lingual summaries. And if different languages are involved, the summarization method is considered cross-lingual or multi-lingual. Explain with example, if it is able to deal with several languages, such as Spanish, English or German, and produces summaries in the same language as the input document was, we would have a multi-lingual summarization system. Beyond these approaches, if the summary is in Spanish, but the original documents are in English, the summarizer would deal with cross-linguality.

After generating the summary, compare the automatic generated summary with the human generated summary this process is called summary evaluation. Using summary evaluation we can find the correctness of the automatic generated summary. For evaluating the summary used the various methods, in these ROUGE method is the standard method. ROUGE stand for Recall-Oriented Understudy for Gisting Evaluation. It includes measures to automatically determine the quality of a summary by comparing it to summaries created by humans. The measures count the number of over- lapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans. ROUGE measures are: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S.

II. RELATED WORK

Review on automatic single document text summarization

For summarizing the document we need to extract the sentences. Sentences are extracting on the basis of their score which is provided by different methodology. In this paper, the scoring methods are grouped into different categories. Before reading the scoring method we need to process the data or text and convert into the desired form on which we can perform the scoring method. This step is called the preprocessing.

Preprocessing

Preprocessing is the first step of the summary generation. It divides into three parts. First is the segmentation, divide the text into paragraphs, sentences and words by using different delimiters like full stop (.). Second is removing the stop words, stop words are those which has no important meaning in the text. Stop words are predefined and are stored in an array or list and this is utilized for comparison with the words in the provided document. And the third and last step is word stemming. Stemming converts the word into its root form. It is generally removing the prefix or suffix of the words.

Scoring methods

Word scoring methods

Initially sentences are scored based on the word scoring. In this each word has a score and sentence score is the sum of all words appeared in that sentence.

For scoring the word different methodology is used like word frequency, tf/idf (term frequency/inverse document frequency). In the word frequency, as the name implies the more frequently a word occurs in the text, the higher its score. [1] 1st used term frequency counts to generate the summary of documents to find the relevant sentences. Text document also containing stop words, which not contain any semantic information such as "a", "an", "the", are not used for computing the term frequency. Many other techniques are also based on term frequency counts have been used in text summarization. Several statistical approaches, such as TF/IDF [5] approach is based on the term frequency. The idea behind the TF/IDF is that frequent terms in a document are important only if they are not very frequent in the whole collection. Different methods are used for calculating the TF/IDF score for each word. Further this score is used to calculate the score of sentence [27]. Some other method is also use for finding the score of word: Upper case feature [4], Proper noun [2], and Word co-occurrence [3]. According to the upper case feature, the words which contain one or more upper case latter have to provide high score. Because if the words contain upper case letter than it may be an important word or a proper nouns. Same case is in the proper noun. It is the specialization of the upper case feature. Word co-occurrence is measure the chance of two terms form a text alongside each other in a certain order. To implement this method using the n-gram, this is the contiguous sequence of n-term in the text. It gives higher score to the sentence that co-occurrence word appears.

Lexical similarity is also a method of scoring the world and based onto the Linguistic approaches. It is based on the assumption that important sentences are identified by strong chain. Words can be related by some relationships (synonymy, hyponymy, metonymy relations). Lexical Chains are used, based on [26] for creating the strong chain of words. First, we select a set of candidate words, generally nouns. Then the list of chains is searched and if a word satisfies the relatedness criteria with a chain word then the word is added to the chain, otherwise a new chain is created. This chain is implies that the

word occurs in the chain are related to each other by some relation so the sentences in which these word is also related to each other. The assumption is that important sentences are those that are identified by strong chains.

Sentence scoring methods

In previous section we read various word scoring methods and on the basis of these score calculate the sentence score. Now we analyze the feature of sentence itself and it was 1st time used by [18] summaries, which is produce by cue words identification. In general, the sentences started by "in conclusion", "our investigation" or "the aim of this paper" may be good indicators of relevant information [1][3][4]. Sentences which contain these words are assigned to the higher score.

Sentences are score on the basis of different features like presence of numerical data [1] [2] [4] [11]. The numerical data may be the date of event, money, quantity, etc., so that this kind of sentences is consider more importance. Sentence length is also a feature for scoring the sentence [2]. Sentences which are too sort or too long are not considering as optimal selection in the summary [11]. Sentence length is the number of words in the sentence. On the basis of the sentence length is scoring the sentences. Sentence position is also considered as a feature for scoring the sentences. Generally the sentence which is occur in the starting or in the ending of the document is consider to be more important than the others. According to the reference [11], 1st sentence in the paragraph is consider into the summary; [2] [5] assign score 1 to N to the starting N sentences and 0 to others, where N is a threshold value. In other methods, sentences which are more similar to the title or the sentences which contain the words into the title are considered to be more important [1] [2] [5]. Sentence resemble to the title is also a feature based on the title [1] [2] [5] [11]. Here finding the similarity between the sentence and title. Sentence which are more similar to the title consider being impotent for the summary.

[8] Has proposed methods for summarizing Hindi text. According to this find the score of the sentences based Statistical and Linguistic method.

Thematic feature [1] is based on thematic words. Thematic words are the most frequently occurring words in the document. The top n frequent words are considered as thematic words. The score for this feature is calculated by the following formula:

$$Thematic\ Feature = \frac{number\ of\ thematic\ words\ in\ sentence}{maximum\ number\ of\ thematic\ words}$$

Format based score (FBS) [1] is also a method for scoring the sentences. The importance of the sentences or headings is indicated by expressing the text in different text format e.g., Italics, Bold, underlined, big font size and etc. Sentence scoring formula:

$$FBS = \frac{number\ of\ word\ in\ sentence\ with\ special\ format}{tatal\ number\ of\ words\ in\ sentence}$$

Graph scoring methods

In Graph scoring method a graph is generated by connecting the sentence using some relationship. When a sentence is connecting to other sentence by link, there is a weight associated with this link. Weight is associated to the link is assign using sentence similarity method [2]. According to this method find the similarity between the sentences using the vocabulary overlap between the sentences.

Formally, let G = (V, E) be a document graph with a set of vertices V and a set of edges (or links) E where V = {S1, S2... Sn}; Si is sentence i in the document; and E is edge between the sentence. A graph can be represented as: an undirected weighted graph; a directed weighted graph with orientation of edges set from a sentence to sentences that follow in the text (called forward direction); or a directed weighted graph with the orientation of edges set from a sentence to previous sentences in the text (called backward direction) [20]. Using this graph we can find the "bushy path of the node" and "aggregate similarity" score for each sentence [2]. In the bushy path of the node, for each node (sentence) in the graph find all the connected link to that node is called the bushy path of the node. In case of aggregate similarity method calculate the sum of all the weight connected to that node.

[12] Proposed various methods for finding the similarity between the sentences. First method is "word form similarity (Sim1)" is measured by the number of same words in two sentences. It should be necessary to remove the stop word.

$$Sim_1(s1,s2) = 2 * ((samewrod(s1,s2))/(Len(s1) + Len(s2)))$$

Second method is "word order similarity (Sim2)" is used to describe the word sequence similarity between two sentences. And third method is "word semantic similarity (Sim3)" is used to define the semantic similarity between two sentences.

So the Sentence similarity between the sentence s1 and s2 is:

$$Sim(s1,s2) = \lambda1 * Sim_1(s1,s2) + \lambda2 * Sim_2(s1,s2) + \lambda3 * Sim_3(s1,s2)$$

Here λ1, λ2, λ3 is the constant, and satisfied the equation: λ1+λ2 +λ3=1.

Selection methods (Summary Generation)

After scoring the sentences using different methodology now we discuss the method for selecting the sentences for summary generation. Number of sentence or summary length is depending on the compression rate. Previously we read different type of summary depending onto the user requirement. For generating the summary in the previous papers different methodology is use, which is Genetic algorithm [2] [9] [24] [25], Fussy logic method [25], General Statistical method [25], and etc. Basic idea of the General statistical method is to select the sentences based on the score. Top high scoring sentences is selected for the summary, depend onto the compression rate.

[16] Proposed a different way for selecting the sentences using contextual information and statistical approach for text summarization. In this method first combine the two consecutive sentences into a bi-gram pseudo sentence and then apply statistical method (different scoring methods) for extracting the bi-gram pseudo sentences and then divide the selected sentences into two single sentences. Now again sentence extraction task is performed onto the selected sentences and final text summary is generated.

III. COMPARISON AMONG THE TECHNIQUES

At a glance comparison among the techniques of multi document text summarization has been shown in table 1:

TABLE I. COMPARISON AMONG THE TECHNIQUES OF MULTIPLE DOCUMENTS TEXT SUMMARIZATION

#	Researcher(s), Year, Reference	Category	Basis of procedure
1	Luhn, 1958	Word Scoring method	(Luhn) 1st used the scoring based on word frequency.
2	Satoshi et al, 2001	Word Scoring method (tf/idf)	The idea behind the TF/IDF is that frequent terms in a document are important only if they are not very frequent in the whole collection.
3	Prasad et al., 2002	Word Scoring method (upper case feature)	Words which start with the capital letters may be the important word, proper noun, name of any place etc. So it is important in the summary.
4	Fattah & ren, 2009	Word Scoring method (proper noun)	Proper noun is the extended version of the upper case feature.
5	Morris and Hirst, 1991	Word Scoring method (lexical chain)	Lexical chain method is based on linguistic approach. Here find the relation or similarity between the words.
6	Fattah & Ren, 2009	Graph Scoring Method (sentence similarity)	Find the similarity between the sentences and create the graph. Using graph find the aggregate score and bushy path.
7	Edmunson, 1969	Sentence Scoring method	(Edmanson) 1st used the cue phrases (in conclusion, etc.) based scoring. Sentences which contain these words have more importance.
8	ZHANG Pei-ying et al, 2009	Graph Scoring Method	Similarity between the sentences is finding using different ways (word form, word order, word semantic similarity). And add all score for finding the final score.
9	Kulkarni & Prasad, 2010	Sentence Scoring method (thematic feature)	Select the top n thematic words which are most frequently occur and on the basis of the thematic word scoring sentence.
10	Kulkarni & Prasad, 2010	Sentence Scoring method (format based)	Word in special format (bold, italic, underline) is also having some importance.
11	Fattah & Ren, 2009; Kulkarni & Prasad, 2010; Prasad et al., 2012; abuobieda et al., 2012	Sentence Scoring method (presence of numerical data)	The numerical data in the document generally brings about some important stats of the core idea of the document. It may be date time of any event.
12	Fattah & Ren, 2009	Sentence Scoring method (sentence length)	This method penalize to those sentence which are too sort. For find the score find the average length and multiply with sentence length.
13	Barrera & Verma, 2012; Abuobieda et al., 2012; satoshi et al., 2001; Fattah & Ren 2009	Sentence Scoring method (sentence position)	Sentence which are in the starting or ending of the document contain the important matter.
14	satoshi et al., 2001	Sentence Scoring method (sentence resemble to title)	Sentence which is more similar to the title is contain important part of the document.
15	Gupta et al 2011; tonelli & pianta 2011	Sentence Scoring method (word co-occurrence)	Word co-occurrence measure the chance of two terms form a text alongside each other in a certain order. It gives higher score to the sentence that co-occurrence word appears.
16	abuobieda et al., 2012	Sentence Scoring method (sentence length)	Sentences which are too short or too long are not consider into the summary. For solving this divide the sentence length with the longest sentence and find the scor.
17	satoshi et al., 2001	Sentence Scoring method (sentence length)	Here sentences panelizing which is sorter then the certain length (CL). Sentence whose length is less then CL is panelizes.

CONCLUSION

In this paper, concepts of single documents text summarization are reviewed that categorize different approaches in this ground. Here we study different type of summary generation, which depends on user requirement. Sentence is scored based on feature of the sentence, so sentence scoring depends on the document structure. Scoring based on the title feature; the document must contain the title. Same thing is in case of numerical data. No single feature can produce good result for text summarization so we need to use some other selection method(s) (summary generation) witch use different scoring method in combination [16] and generate the summary.

ROUGE [6] is a standard method for evaluating the summary. For check the usefulness and the trustfulness of the summary, use the Precision, Recall and F-measure.

REFERENCES

[1] .Kulkarni, U. V., & Prasad, Rajesh S. (2010). Implementation and evaluation of evolutionary connectionist approaches to automated text summarization. In Journal of Computer Science (pp. 1366–1376). Science Publications.

[2] Fattah, Mohamed Abdel, & Ren, Fuji (2009). Ga, mr, ffnn, pnn and gmm based models for automatic text summarization. Computer Speech and Language, 23 (1), 126–144.

[3] Gupta, P., Pendluri, V. S., & Vats. I. (2011). Summarizing text by ranking text units according to shallow linguistic features. In 13th International conference on advanced communication technology (pp. 1620–1625)

[4] Prasad, Rajesh Shardanand, Uplavikar, Nitish Milind, Wakhare, Sanket Shantilalsa, Jain, Vishal, Yedke & Tejas Avinash (2012). Feature based text summarization.International Journal of Advances in Computing and Information Researches, 1.

[5] Satoshi, Chikashi Nobata., Satoshi, Sekine., Murata, Masaki., Uchimoto, Kiyotaka., Utiyama, Masao. & Isahara, Hitoshi. (2001). Keihanna human info-communication. Sentence extraction system assembling multiple evidence. In Proceedings 2nd NTCIR workshop (pp. 319–324)

[6] Lin, Chin-Yew (2004). Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens (Ed.), Text summarization branches out: Proceedings of the ACL-04 workshop (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics.

[7] Tonelli, Sara., & Pianta, Emanuele. (2011). Matching documents and summaries using key-concepts. In Proceedings of the french text mining evaluation workshop.

[8] Thaokar, C. & Malik, L. (2013). Test Model for Summarizing Hindi Text using Extraction Method. Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013)

[9] Chatterjee, N., Mittal, A., & Goyal, S. (2012). Single Document Extractive Text Summarization Using Genetic Algorithms. 2012 Third International Conference on Emerging Applications of Information Technology (EAIT)

[10] Ali, M., Ghosh, M. K., and Abdullah-Al-Mamun. Multi-document Text Summarization: SimWithFirst Based Features and Sentence Co-selection Based Evaluation. 2009 International Conference on Future Computer and Communication.

[11] Abuobieda, A., Salim, N., Albaham, A. T., Osman, A. H., & Kumar, Y. J. (2012). Text summarization features selection method using pseudo genetic-based model.In International conference on information retrieval knowledge management (pp.193–197).

[12] ZHANG Pei-ying, LI Cun-he (2009). Automatic text summarization based on sentences clustering and extraction. IEEE 2009

[13] Ying-Qiang Wu Gang-Zhou Li-Qing Qiu. An Extensive Empirical Study of automated evaluation of multi-document summarization. First International Conference on Intelligent Networks and Intelligent Systems

[14] Radev, D. R., Jing, H. & Budzikowska, M. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies.

[15] KALLIMANI, J. S., Srinivasa K G, Eswara REDDY B (2010). Information Retrieval by Text Summarization for an Indian Regional Language.

[16] Youngjoong Ko, Jungyun Seo (2008). An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. In Pattern Recognition Letters 29 (2008) 1366–137

[17] Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 2(2), 159-165

[18] Edmundson, H.P. (1969). New method in automatic extracting. Journal ACM, 16(2), 264-285.

[19] Llorat, E., Palomar, M. (2011). Text summarization in progress: a literature review. Artif Intell Rev (2012) 37: 1-41 (Springer).

[20] Sornil, O., Gree-ut, K. (2006). An Automatic Text Summarization Approach using Content-Based and Graph-Based Characteristics. CIS 2006 (IEEE).

[21] Spärck Jones K (1999). Automatic summarizing: factors and directions. In: Advances in automatic text summarization. pp 1–14.

[22] Hovy E, Lin CY (1999). Automated multilingual text summarization and its evaluation. Technical report Information Sciences Institute, University of Southern California.

[23] Mani I (2001) Automatic summarization. John Benjamins Publishing Co. Amsterdam, Philadelphia, USA

[24] V. Qazvinian, L. S. Hasaanabadi, and R. Halavati, "Summarizing text with a genetic algorithm-based sentence extraction," International Journal of Knowledge Management Studies, vol. 2, no. 4, pp. 426–444, 2008.

[25] Suanmali, L., Salim, N., and Binwahlan, M. S. (2011). Fuzzy Genetic Semantic Based Text Summarization. 2011 Ninth IEEE InternationalConference on Dependable, Autonomic and Secure Computing.

[26] Morris, J. and Graeme H. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 17(1):21-48.

[27] Ferreira, R., Luciano de Souza Cabral, Lins, R. D., Gabriel Pereira e Silva, Freitas, F., George D.C. Cavalcanti, Lima, R., **S**teven J. Simske, Favaro, L. (2013). Assessing sentence scoring techniques for extractive text summarization.