International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7 Issue-5, January 2019

Two-Level Text Summarization with Sentiment Analysis for Multi-Document Summarization

Lavanya K C, C. Sivamani, Linnet Tomy, Ann Rija Paul

Abstract*: *Text summarization is way of reducing the text content of a document without the losing any information. People are likely to look multiple documents on a single topic because a one document may not include all the major details. The abstract/summary of multiple documents connected to a text will conserve the effort and time. Automatic text summarization is one of the area of natural language processing. Sentiment analysis is a machine learning method in which machine study and inspect the sentiments, opinions, etc about reviews about movies or products. This is extremely hard summarize by human. effective data from the very large document. In this research, we propose a novel method for multiple document summarizations using extractive method of summarization and sentiment analysis from online sources. At first, various document's URLs are fetched as input relate to a text and generate individual summaries. The sentiment analysis is tried on these generated separate summaries. The sentiment analysis says that whether these input documents have any dissimilar opinion about the topic. Lastly, a unique summary is generated from all these first level summaries. The performance of our proposed method evaluated by ROUGE metric.

Index Terms: Text Summarization, Sentence Extraction, Sentiment Analysis, Natural Language Processing, ROUGE

I. INTRODUCTION

By the wide exploration of internet, a huge collection of documents are available in the internet. Whenever people searches for a topic they compelled to read all the documents related to their search. This takes large amount of time and it is very tough to summarize manually. There comes the value of automatic multi-document summarization. Text summarization is the process by which reduces the size of the document without loss of its meaning. Text summarization process done by a machine is entitled as automatic text summarization. A good summary can give a rapid concise of a big document. In single document summarization, it only generates condensation of a single document. As the enormous changes in research and with the vast collection of documents are available in the internet, multi document summarization emerged. In multi document summarization,

it generates summary of multiple documents on the same topic. Hence multi document summarization is very difficult than the single document summarization. Because, when summarize multiple documents they overlap the information contained in the documents. This may leads to sacking in the generated summaries. Also an extra work is required to gather and arrange the information from different documents to generate a reasonable summary. In the area of document summarization, remarkable acquirements have been procured applying sentence extraction and sentiment analysis. Text summarization methods are categorized into different kinds based on their technology. They are mentioned below.

A. Abstractive versus Extractive

Abstractive summarization analyses the document and generates a summary as human generated summary. It considers advanced natural language processing methods [2] to generate the summary. Extractive summarization extracts the main sentences from the document and combined them to generate the summary.

B. Mono-lingual versus Multi-lingual

Mono-lingual text summarizer performs on only one language, such as English. Whereas, multi-lingual text summarizer performs with multiple languages, such as English, Spanish, Japanese, and Hindi.

C. Multi-document versus Single-document

The single document summarizer works on one document only and generates its summary. Multi-document summarizer works with more than two documents of common topic and generates a summary of them.

D. Generic versus Query-based

If people may search for the entire information rather than specific information then such type of information belongs to generic based summary. If a person requires distinct information rather than specific information, then it is termed as a query-based summary.

E. Indicative versus Informative

Indicative summaries generates summaries in terms of only important information from the document. These summaries are encourage the people to read the complete document. Whereas, Informative summaries contains all the important information in the document in a coherent manner.

Revised Manuscript Received on 30 January 2019.
***** Correspondence Author
Lavanya K C*, PG Scholar, Sahrdaya College of Engineering and Technology, Thrissur, Kerala, India,
Dr.C.Sivamani, Assistant Professor in BMIE, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamilnadu, India
Ms.Linnet Tomy, Assistant Professor in CSE, Sahrdaya College of Engineering and Technology, Thrissur, Kerala., (e-mail:
Mrs.Ann Rija Paul Assistant Professor in CSE, Sahrdaya College of Engineering and Technology, Thrissur, Kerala.,

In this research, we propose a novel method of two level extractive document summarization with sentiment analysis from online sources is introduced. If a person looks for a text on the internet, it gives a few links for this text. It is laborious and requires more time to read all these documents.

Consequently, there comes the importance of automatic text summarization to summarize all of these documents. This will give a summary of multiple documents in fewer sentences without loss of its meaning.

Initially, the main sentences from various articles related to a topic are extracted and generate the first-level summary separately. Sentiment analysis is performed on these individual summaries to check whether there is any difference in reporting the same concept. It will also give assistance to know the summary generated is whether positive, negative or neutral [3]. At last, a single summary is generated among all these individual summaries. This will provide the most important sentences from all the documents. The performance of this system can be evaluated using ROUGE metric. The rest of our paper consists as follows: Section II provides the background. Section III explains the details of the proposed methodology. Section IV describes the methods used for two-level summary generation and sentiment analysis. Additionally, the simulation results are there with ROUGE metric evaluation. Section V explains the conclusion and future enhancements.

II. BACKGROUND

The extractive text summarization takes the important sentences from the original document. Substantial research work has been performed on extractive summarization. Krishnaprasad et. al. done a Malayalam text summarizer based on extractive method which is a single document summarizer [4]. They used to calculate the frequency of each word in the document. Based on these frequency provide top ranks to highest priority words and compiled the top ranked words to generate the summary. The standard evaluation metric ROUGE is used to implement the performance. Feng et. al. construct a extractive based single document text summarizer by considering corpuses of news stories [5]. They utilized keyword-based method where they looked using keywords to bring the associated news stories about a topic saved in their corpus. Evaluations are performed by using ROUGE metric. Krzysztof et al. presented a extractive sentence based summarization for polish language [6]. They utilized Term frequency – Inverse document frequency method and Polish news corpus for summary generation. The evaluation metric ROUGE is used to check the performance.

Agarwal et. al. implemented a novel method by examining sentiment of tweets applying polarity based method where tweets were categorized sentiments of positive, negative or neutral [7]. They utilized n-gram, feature and tree kernel-based approach for categorizing the tweets and attained 71.35% accuracy applying SVM. Mirani and Sasi recognized sentiments of tweets with their correct locations by applying approach based polarity [8]. They used Naïve Bayes, SVM, Random Forest, KNN, Maximum Entropy Algorithms and Decision Trees and attained more than 95% average precision. In this research, an individual summary is generated from various articles to attain better results. Sentiment analysis is performed on these individual summaries to check how genuine are these sources of articles. Finally, the main sentences are extracted from these individual summaries for second-level summary generation.

III. METHODOLOGY

Extraction based summarization is applied to extract the main sentences or phrases from the input articles. The topics of the articles includes movie review, sports, science, health and politics. We consider all these topics to examine the outcome of this research. Many documents are accessible for every concept in the internet. Upon these we use two or three documents to generate the summary. The overall architecture of our proposed method is explained in Fig. 1. Initially, the input articles are given by fetching URLs to generate the individual summaries. Two or three individual summaries are generated using extractive summarization method. The summarization based on extractive method gets the remarkable sentences from the documents and arrange them in a chronological order by using sentence ranking method. After generate the first level summaries, sentiment analysis applied on them to check whether there is any difference in the idea of the documents. If there is any difference in viewpoints then sentiment analysis shows negative. If all the documents convey same idea then sentiment analysis shows positive. If some of the document said it is good and some of them are said it is bad then sentiment analysis shows neutral. Finally, main sentences from the individual level summaries are extracted and compile for second-level summary generation.

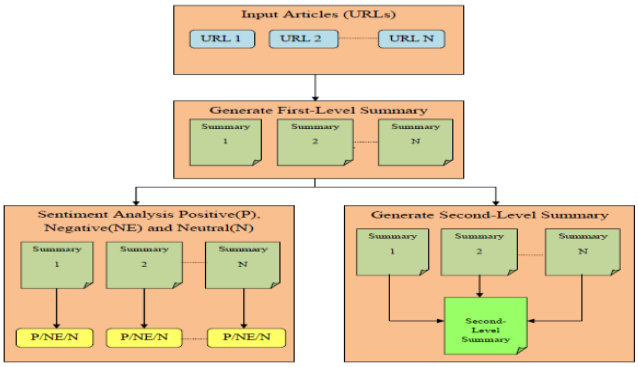

Fig 1. Overall architecture of two-level text summarization

A. Preliminary Processing

The pre-processing is the main step in NLP (Natural Language Processing) because it removes unnecessary contents from the data. It comprises of tokenization and regular expressions. The various steps involved in pre-processing are shown in Fig.2.

The user input the articles in the form of URLs and need to remove it for further processing. To remove stop words, HTML characters and punctuations are used. Then these articles are transformed into lower-case and the lower cased articles are changed into sentences and then sentences to words. The token is a sequence of words gather together in a text.

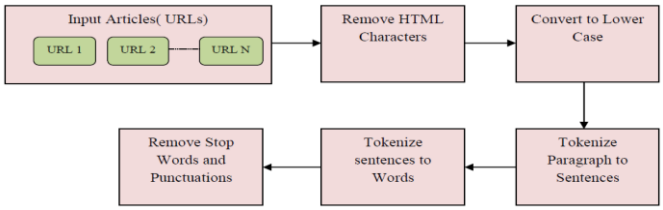

Fig 2. Pre-Processing

In another words, tokens are used in lexical analysis [9] as the words in the sentences. Tokenization is the process of replacing the data into words, phrases, symbols. Stop words are the commonly used words [10] and are ignored by contrasting the words with the stop words in dictionary. A sample text for pre processing is shown in Table 1. The table includes the elimination of HTML characters, stop-words is required. At last, the punctuations are removed.

Table 1. Steps for pre-processing a text sample

Pre-processing	Text Example: Title: Sentiment Analysis Sentiment Analysis is a machine learning approach!!! <a bref"http://www.sampletext.com"> here
Clean HTML characters	Title: Sentiment Analysis Sentiment Analysis is a machine learning approach!!! here
Convert to lower case	title: sentiment analysis sentiment analysis is a machine learning approach!!! here
Tokenize text into sentence	The sentences are tokenized from text
Tokenize sentences into words	Sentence: “sentiment analysis is a machine learning approach”!!! [token 1] “Sentiment” [token 2] “analysis” [token 3] “is” [token 4] “a” [token 5] “machine” [token 6] “learning” [token 7] “approach” [token 8]”!” [token 9] “!” [token 10] “!”
Remove stop words	Title: Sentiment Analysis Sentiment Analysis is a machine learning approach!!! here
Remove punctuation	Title Sentiment Analysis Sentiment Analysis is a machine learning approach

B. Individual Summary Generation

In our research, we used extractive summarization technique for generating the summary of online texts. The process of individual summary generation is explained in Fig. 3. A single document may include several subdivisions. These subdivisions are changed into a sequence of sentences and again converted these sentences into words.

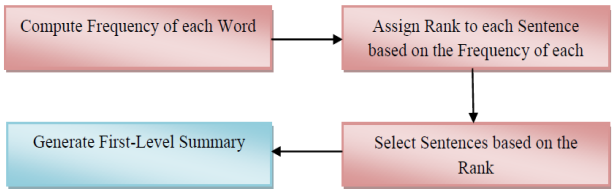

Fig 3. Steps for individual summary generation

In first step, each word’s frequency is calculated and put its threshold value. This will helps to know the occurrence of similar words in a document. The frequency of each word is calculated by using Term Frequency-Inverse Document Frequency (TF-IDF) method. Only extracts the words that have a frequency range in between the threshold value. The words with very low frequency as well as the words with very high frequency are eliminated. In the second step, assign rank for each sentence based on their word frequency. Repeat this process until all the sentences in the document get scored. The highest ranked sentences contain more meaningful information. For summary generation process, consider the first sentence in every paragraph is important and include them in the summary. Finally, arrange the top ranked sentences in a chronological order to generate individual summary.

C. Sentiment Analysis

In this research, Sentiment Analysis is used to know whether there is any difference in conveying the same information. This is performed on the first-level summaries and it has different opinions. They can be positive, negative or neutral sentiments. The execution of sentiment analysis is done by using the Text Blob library [11] that supplies a small text processing API by applying the NLP features such information extraction, part-of-speech, sentiment analysis, tagging, classification, etc. The feature of sentiment analysis feature gives polarity (i.e. Positive, Negative and Neutral) and subjectivity (i.e. Objective and Subjective). If the polarity closes to -1.0 then it is very negative and if polarity closes to 1.0 then it is very and positive. The score is within the range where 1.0 is very subjective and where 0.0 is very objective.

D. Second-Level Summary Generation

Users like to access multiple documents on a similar topic and each document may include minimum 30 sentences. This takes lots of time and energy to read multiple documents. If we get a summary of all these documents then it will save time and effort. The main purpose of second-level summary generation is to compile content supplied in various related documents from different online sources. In this research, the individual first-level summaries are extracted and combined them to generate the second-level summary using the same extractive method.

IV. SIMULATIONS AND RESULTS

In this research, the tool used is the open source Python language, since it is very simple to use and it gives a huge collections for effective visualization and statistical analysis [12]. The initial step of summary generation is get the URLs of online texts and it was achieved by applying “Beautiful Soup” and “requests” packages. Pre processing of these online articles is carried out by using The tokenization and Regular Expression. Regular Expression is performed using an inbuilt function in Python with Natural Language Toolkit (NLTK). Tokenization is performed using the nltk.tokenize package in python. The individual summary and the two level summary was generated by using this NLTK package.

The sentiment analysis is performed by using "Text Blob" package to check whether the articles have any different viewpoints. The performance of our proposed method can be evaluated using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [13].

A. Article Comparison

In our research, the documents are used in the area of Movie Reviews, Sports, Health, Science and Politics. Two or three related articles are taken for each topic.

The summary may be brief but it should contain the overall meaning of the article. In this research, the Less Sentences (less Sent.) and more sentences (more sent.) are the two sections of summaries used for the performance evaluation. We can set the length of summary to be needed by putting summary percentage calculation option. Less Sentences may have only 5 sentences and More Sentences may have 7 or 10 Sentences depends upon the length of the document. Table 2 shows the amount of sentences and words in the summary article (SA) and original article (OA). For the simulation of results, we use different sizes are used for topic simulation. i.e. Politics contain minimum of '8' sentences in the OA and Sports contain maximum of '90' sentences in the OA are applied.

Table 2. Original Article (OA) and Summary Article (SA) with Less Sentences and More Sentences Comparison

Topic	Words			Sentence		
	OA	SA (Less Sent	SA (More Sent.)	OA	SA (Less Sent	SA (more Sen
Politics						
Article 1	480	154	254	21	5	10
Article 2	500	131	27	34	5	10
Sports						
Article 1	242	123	146	13	5	7
Article 2	318	168	228	13	5	7
Article 3	192	130	138	8	5	
Health						
Article 1	339	154	203	17	5	7
Article 2	456	168	214	32	5	7
Science						
Article 1	721	128	246	33	5	10
Article 2	794	118	235	45	5	10
Article 3	457	125	234	19	5	10
Movie Review						
Article 1	1752	103	373	93	5	10
Article 2	1359	134	285	64	5	10
Article 3	1215	145	258	62	5	10

B. Sentiment Analysis

In our research, different online sources are applied to check whether they have different opinions about a single topic. The sentiment analysis gives the value in the scale of -1.0 is very negative and 1.0 is very positive. The average sentiment percentage of different article is shown in Table 3.

Table 3. Average Sentiments of summary of different articles

Topic	Article	Average Sentiment	
		Less Sent.	More Sent.
Politics	Article 1	-0.03	-0.06
	Article 2	-0.05	0.07
Sports	Article 1	0.16	0.18
	Article 2	0.05	-0.09
	Article 3	0.16	0.13
Health	Article 1	0.14	0.07
	Article 2	-0.02	-0.02
Science	Article 1	0.25	0.15
	Article 2	0.17	0.14
	Article 3	0.08	0.13
Movie Review	Article 1	0.27	0.18
	Article 2	0.12	0.92
	Article 3	0.24	0.25

C. ROUGE Metric Evaluation

ROUGE is the Recall-Oriented Understudy for Gisting Evaluation [13]. Normally, it is very tough to predict if the system generated summary is bad or good [4]. Two possible methods are there to evaluate the system-generated summaries. One is Human-based evaluation and the other is Machine-based evaluation. In human based evaluation, the judges examine the verbal skills and sentimental opinion. It is a complex task and takes huge amount of time. In addition, if there are more than two judges they might not ready to accept their determination. This judgment is carried out by matching the system generated summary with the human made summary. Machine-based evaluation does not require much more time and is a preferable process. Also, the Machine-based evaluation is not considering human partiality. It utilizes a similar technique to evaluate all the summaries. The ROUGE metrics automatically evaluate the caliber of the summary by comparing the system-generated summaries with human-generated summaries. The measures the n-gram overlapping between words and sequences. In this research, we use ROUGE 1 and ROUGE 2 techniques, to evaluate the n-gram overlapping between first-level summaries. The Precision, Recall and F-Measures are calculated by considering the more sentences (MS) and less sentences (LS).The ROUGE 1 result is the unigram(1-gram) showed excellent results than ROUGE 2 result that are bi-gram(2-gram). The better evaluation percentage shows by 'more' sentences than the 'less' sentences since 'less' sentences have only two valid sentences from the OA. It will not possible to implement all the valid content from the original document.

V. CONCLUSION AND FUTURE WORK

This paper introduces a novel approach Sentiment Analysis and Two-level Text Summarization for Multi-document summarization. The two-level summary gives meaningful information related to a similar topic from several online sources. The sentiment analysis gives the opinion of different websites. To achieve better results consider the original article that has more sentences. In future, we can use abstractive summarization method instead of extractive summarization method. Enhance this work applicable for mobile phone users for the people who utilize mobile phones to read the documents.

REFERENCES

1. P. Addala, "Text Summarization A Literature Survey". Available:https://www.scribd.com/document/235008952/Text-Summarization-Literature-Survey. [Accessed 09 April 2017].
2. Sharockman, "PunditFact checks in on the cable news channels", PolitiFact, 2015. Available: http://www.politifact.com/truth-o-meter/article/2015/jan/29/punditfact-checks-cable-news-channels/. [Accessed: 09- Apr- 2017].
3. Ansari, "Sentiment Polarity Classification Using Structural Features," IEEE International Conference on Data Mining Workshop 2015 (ICDMW), Atlantic City, NJ, 2015, pp. 1270-1273.
4. P. Krishnaprasad, A. Sooryanarayanan and A. Ramanujan, "Malayalam text summarization: An extractive approach," International Conference on Next Generation Intelligent Systems (ICNGIS) 2016, Kottayam, 2016, pp. 1-4.
5. Feng Li, Yan Chen and Zhoujun Li, "Learning from the past: Improving news summarization with past news articles," International Conference on Asian Language Processing (IALP), Suzhou, 2015, pp. 140-143.
6. K. Jassem and Ł. Pawluczuk, "Automatic summarization of Polish news articles by sentence selection," Federated Conference on Computer Science and Information Systems (FedCSIS), Lodz, 2015, pp. 337-341.
7. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R,
8. "Sentiment Analysis of Twitter Data," in Proc of ACL HLT Conf, 2011.
9. T. B. Mirani and S. Sasi, "Sentiment Analysis of ISIS Related Tweets Using Absolute Location," International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2016, pp. 1140-1145.
10. Jackson, P. and Moulinier, I. Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization. John Benjamins Publishing Co., 2002.
11. Manning, P. Raghavan and H. Schü tze, Introduction to information retrieval. New York: Cambridge University Press, 2008.
12. S. Loria, "TextBlob: Simplified Text Processing", Textblob.readthedocs.io, 2013. [Online].Available: https://textblob.readthedocs.io/en/dev/index.html. [Accessed: 09- Apr- 2017].
13. "Python", Python.org, 2001. Available:
14. https://www.python.org/about/. [Accessed: 09- Apr- 2017].
15. Chin Yew Lin, "ROUGE: A package for automatic evaluation of summaries," In Proceedings of Workshop on Text Summarization BranchesOut, Post-Conference Workshop of ACL, Barcelona, Spain, 2004.
16. R. Ebert, "Titanic Movie Review & Film Summary", Roger ebert, 1997. Available: http://www.rogerebert.com/reviews/titanic-1997. [Accessed: 21- Mar- 2017].
17. Tarun B Mirani, sreela Sasi, "Two-level Text Summarization from Online News Sources with Sentiment Analysis", International Conference on Networks & Advances in Computational Technologies (NetACT) |20-22 July 2017| Trivandrum