Two Level Text Summarization for Online News Source

Aneesha Joshi

Student, Department of Computer Applications, Christ Knowledge City College, Kerala, India

---***---

Abstract – *Text summarization is a process of reducing a set of data, and creating a subset of most relevant information within original content. The core benefit of the text summarization is that it minimizes the reading time and efforts. This paper presents a two-level text summarization method for online news sources; It is possible that user may miss most important information if the user read multiple articles from multiple news source. If a text summarizer is available then it will showcase most important information from multiple articles in one scan of all documents. Two level text summaries are generated using extractive-based approach. The first-level summary creates the summary of each article. The second level summary is used to create the summary of the combined first-level summaries of two/three related articles. Sentimental analysis is applied on the first level of summary to understand the variation from different news agencies. To evaluate the summarization ROUGE metrics is used.*

***Key Words*: Sentimental analysis, Natural Language processing, Text summarization, Extraction based summarization, ROUGE metric**

1.INTRODUCTION

Data growing faster day by day. Rapid increase in the published data or information, and the effect abundance is known as information explosion, as the amount of available data grows information handling become difficult which leads to information overload. To seek right information from the huge data thereby become difficult. For example: In context of news, to understand an occasion (an event/chain of events occurring at the definable time and place) people may consider different news articles from different news agencies. Relatively expansive number of newspapers are available thus reading all parts of an occasion turns into troublesome. Hence, we induce the text summarization method to gain important information from all documents by one scan.

Text summarization also known as a text reduction process. Automatic text summarization is a growing and interesting field in natural language processing. The intention is to create a well organized and easy summary having only main points outlined in document. The main benefits of the text summarization is that it minimizes the reading time and efforts. Text summarization methods are of different type. Text summarization are often categorized by the way it's done.

A. Extraction Vs Abstraction

Text summarization classified into two, extractive summarization, abstractive summarization. Extractive summarization is one which extracts all the important sentence and rank them. The highest ranked sentences are joined together to form a summary without changing the original content. Where Abstractive summarization generate summary that is similar to human generated summary by understanding the whole document. For this it uses advance natural language processing techniques.

B. Single document Vs Multiple document

Single document summary only takes one document as input and generate the summary. In multiple document summarization it input two or more document on similar topic then output the summary.

C. Generic vs Query based

Summary of entire information is known as generic summarization. Summary of specific information is known as query-based summarization.

D. Mono-lingual Vs Multi-lingual

Summary generation process applied on only one language us mono-lingual, Multilingual works with multiple languages.

E. Indicative Vs Informative

Indicative summaries indicate only important information. That is it highlight few important sentences in the document. Informative summarization replaces the original document *by providing concise information.*

In this paper an extractive based two level text summarizations from online news source with sentimental analysis is presented.

2. RELATED WORK

A. Extractive Summarization

The extractive summarization selects the most important sentence from the original document.

Extensive research has been done on extraction-based summarization. Krishna prasad et. al. implemented extractive text summarizer for Malayalam language [1]. It uses rank-based approach by providing a word scoring for each sentence supported their importance, and then selects top N ranked sentences, and generate a summary. The ROUGE metric is used to present the results. Feng et. al. built extractive text summarizion by using collection of stories . The research used keyword-based approach where they searched using keywords to extract the related news stories a few topic stored in their corpus. Evaluations are done u sing ROUGE sets. Krzysztof et. al. developed a sentence-based extractive summarization for polish language. They used TFIDF method and Polish news sources to get a summary.

B. Sentimental Analysis

Sentimental analysis is the way toward deciding if a piece of writing is certain, negative or unbiased. Sentimental analysis helps data specialists inside enormous undertaking check general assessment, direct nuanced measurable looking over, screen brand and thing reputation, and appreciate customer experiences. One challenge is to create technology to detect and summarize an overall sentiment. Twitter build models for classifying “tweets” into positive, negative and neutral sentiment. Agarwal et. al. presented a novel approach by analyzing sentiment of tweets using polarity-based approach where tweets were classifies d into positive, negative or neutral sentiment. They used to classify the tweets and achieved 71.35% accuracy using Support Vector Machine (SVM) by using unigram, feature and tree kernel-based approach. Mirani and Sasi uses polarity-based approach for identifying ISIS related tweets with their exact locations. They used SVM, Random Forest, Bagging, Decision Trees and Maximum Entropy Algorithms and achieved quite 90% average accuracy. In this research, a private summary is extracted from news articles from multiple sources to realize the simplest results.

3. METHODOLOGY

Extractive summarization is used to generate summaries by encountering the important sentence from the original document and by combining them. The news article topics may comprise politics, sports, science, health and entertainment news. The summarization system consists of 3 steps.

1. Pre-processing
2. Feature terms extraction.

3. Sentence ranking based on optimized weight.

The web crawling method is used to scan the URL that provided. After fetching the URL, it fed as the input of first level summary, where extraction method applied for ranking the important sentences. To find the variation among different news article sentimental analysis is applied in first-level summary, thereafter second-level summaries are generated by combining first-level summaries.

A. Preprocessing

It involves three parts, they are Sentence segmentation, Stop word removal, Tokenization.

Sentence segmentation is decomposing the documents into representing sentences along with their word count. In English language bound segments such as question marks(?), exclamatory mark(!), full stop(.) are used to segment the sentences.

Stop word removal is done due to their influences in summary generation, stop words are common words that convey less importance than the keywords.

Tokenization is splitting words by identifying special symbols, comma and space.

B. Feature extraction

The sentence are ranked after tokenization by taking two important features; frequency and sentence position value. Document is tokenized means it split into collections of sentences.

Frequency: Number of times a word occurred in the document is known as the frequency, frequency of the word has direct proportion to the significance. That is it has significant effect on the content of document.

Sentence position value: Importance of the sentence is decided by the position of sentence that is the first sentence defines the theme of the document whereas the last sentence conclude the document. Positional value is computed as the first sentence comprise the highest positional value where last sentence has lowest.

Sentence score: Sentences score is the linear combination of frequency, Sentence positional value, similarity with the title.

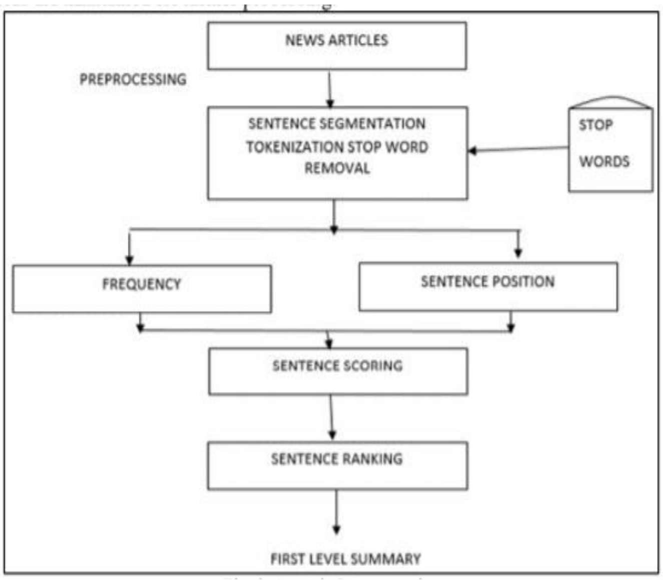

Fig: Preprocessing

C. Sentence Ranking

Sentences are arranged in descending order after scoring the sentences. Hence the highest ranked sentence will be in top position and lowest in bottom position.

3.1 First Level Summary

In this research, the extraction-based method is applied to get a summary of online news articles. Single news article may contain multiple paragraphs. These paragraphs are converted to sentences then from sentences towards.

Step 1: the frequency of every word is calculated and its threshold is about. In simple terms, what percentage times each word has appeared within the sentence is calculated. This word frequency is calculated by number of times the word appeared in a sentence. This is done because some words contain less value, and should not be removed during the stop words process.

Step 2: each sentence are going to be assigned a score supported the frequency of words. This process are going to be repeated until all the sentences of stories article are given a score. In simple words, the score may be a priority number which will be allocated to every sentence. Finally, the sentences with a top priority number are going to be selected and therefore the First-level Summary is generated.

3.2 Second Level Summary

People are likely to read quite one news story on an identical topic and every article may contain 30 – 40 sentences. This situation takes longer to read multiple news articles. In order to save lots of time and energy , a Second -level Summary will help the user to urge a far better idea of the content provided in various related news articles from different news channels. In this research, a Second-level Summary is generated from all the important sentences of the First-level summaries on a news topic using an equivalent extractive based approach. Two/ Three First level Summaries on a subject are wont to generate the Second -level Summary.

3.3 Sentimental Analysis

Many researches are available on sentiment analysis of social media such as twitter and so on; but very less research has been done summarizing news articles. The polarity of the sentence can be positive or negative or objective and subjective. [4]

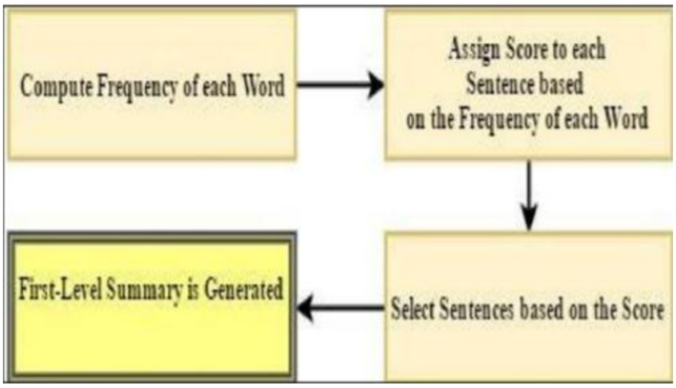
Compute Frequency of each Word
Assign Score to each Sentence based on the Frequency of each Word
Select Sentences based on the Score
First-Level Summary is Generated

Fig: Process for First-level Summary Generation

3.4 ROUGE Metrics

It is difficult to measure the quality of generated summary. There are two ways to examine the system generated, they are Human-based evaluation and Machine-based evaluation. In Human-based evaluation verbal skill and emotional viewpoints are considered. Whereas in Machine-based evaluation evaluates the summary using similar logics by using a program and provides the result. ROUGE which stands for Recall-Oriented Understudy for Gisting Evaluation [5]. The ROUGE metric provides the set of measures which will automatically evaluate the standard of the summary by comparing the summaries created by humans.

4. SIMULATION

In this research, Python language is used, because it is straightforward to use and provides a plethora of packages for better statistical analysis and visualization. Natural language tool kits are most popular tool available. The first step for generation of summary is to fetch the URLs of news articles and this was done by web crawler method. Search engines, use spidering for providing up-to-date data. Web crawlers won't create a replica of all the visited pages that may help fast search. Crawlers used to automate the maintenance tasks on an internet site, like checking links, validate HTML code. Also, crawlers can be used to gather query-based information from Web page. The Pre-processing of those news articles is completed by regular expression and Tokenization. Sentiment analysis can be performed by using NLTK library which is available in python. A sentiment analysis code gives the sentiment of a sentence supported a dictionary of words which tag as positive and negative and score them between -10 and 10. Using ROUGE metrics, summaries evaluated such that human made summaries compared with system generated First-level summaries. Thus summaries are evaluated

5. CONCLUSIONS

This paper presents a novel two-level Extraction based News Summarization. The two-level summary provides only the important content from different online news articles on a news topic in one place. Sentiment Analysis provides the view of varied news channels.

- To speed up the summary generation, a pre-processing approach is introduced, which retains the most important points of the original report.
- Extractive summarization is used for summarizing the document
- To check variations among different articles sentimental analysis is used.

ACKNOWLEDGEMENT

I would like to extend my heartfelt thanks to our, HoD, Ms. Geethu Krishna Kartha, Assistant Professor, Department of Dual Degree Master of Computer Applications for their valuable advices and guidance throughout my course of study. My most sincere thanks go to my mentor, Asst. Prof. Ms. Aimy Susan Thomas for her encouragement and support. I would like to show my sincere gratitude to Asst. Prof. Ms. Saranya Shaji for her guidance and idea, without her this paper would not be complete. She has given idea which helped me to accomplish writing this paper

REFERENCES

[1] Krishnaprasad, A. Sooryanarayanan and A. Ramanujan, "Malayalam text summarization: An extractive approach," 2016 International Conference on Next Generation Intelligent Systems (ICNGIS), Kottayam, 2016, pp. 1-4.

[2] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R, "Sentiment Analysis of Twitter Data," in Proc of ACL HLT Conf, 2011

[3] ATSSI: ABSTRACTIVE TEXT SUMMARIZATION USING SENTIMENT INFUSION Author Rupal Bhargava Yashvardhan Sharma Gargi Sharma

[4] Sentiment analysis algorithms and applications: A survey Walaa Medhat, Ahmed Hassa, Hoda Korashy.

[5] Chin Yew Lin, "ROUGE: A package for automatic evaluation of summaries," In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL, Barcelona, Spain, 2004.