Extractive Text Summarization Using Word Frequency Algorithm for English Text

AbinayaN[1], AnbukkarasiS[1] and VaradhaganapathyS[1]

[1]*Kongu Engineering College, Erode, Tamilnadu*

Abstract

Text summarization is the method of retaining the key information of article without losing its vital content. Summarization is essential in all the areas since a large volume of data is generated day by day. Because of the availability of huge data, it becomes difficult to extract the exact information. People lack patience to understand the content by reading the entire article. Summarization plays a major role in these times to provide available vital information fast and effectively. This can be done in two ways: Extractive Summary and Abstractive Summary. Extractive summary is simple compared to abstractive summary. While an abstractive summary creates new phrases, an extractive summary entails locating highly ranked sentences from the given text. Various techniques, including sentence ranking, Graph Based Modeling, RBF Models, and Sentence Similarity Measures, can be used for extractive summarization. This paper provides extractive text summarization for code mixed English text provided by ILSUM track of FIRE 2022. In this work, Word Frequency Algorithm is used for summarization and the ILSUM team measured the performance of the system by standard ROUGE metrics.

Keywords

Automatic Text Summarization (ATS), Natural Language Processing (NLP), WordFrequency Algorithm

1. Introduction

In this modern era, huge volume of text data is available in internet in the form of documents,e-books, news, movie reviews, articles etc. People find very difficult to obtain the significant information from the lengthy texts. We need a mechanism to identify the key information from the text, fast and effectively by reducing the time of reading. The fundamental problem in this digital world is how quickly the information can be compressed and located from the text. Automatic Text Summarization (ATS) helps to overcome this problem effectively [1]. Various approaches have been developed to generate two different summaries namely, extractive and abstractive. Former one is generated from the original text in the article whereas the later generate their own text which provides the information of original documents. Moreover, applications like search engines, news articles need summarizer as search engines tries to provide the snippet and news websites generate the headings based on the content [2]. Their application is also needed in many areas like library to summarize the content of magazine, e-books, journals etc.

Various machine learning algorithms under both supervised and unsupervised category are used for generating a good summary from a given text. The various issues that arise during summarizations are redundancy, ambiguity, key word identification, similarity etc. The approaches including word frequency, sentence scoring, sentence ranking are not much challenging for summarizer because of their statistical approach. The biggest challenge faced by summarizer is to identify the new features

Forum for Information Retrieval Evaluation, December 9-13, 2022, India
EMAIL: abi9106@gmail.com (A. 1); anbu.1318@gmail.com (A. 2); varadhaganapathy@gmail.com (A. 3)
ORCID: 0000-0002-8419-6201 (A. 1); 0000-0003-0226-8150 (A. 2)


that help to generate the summary and retain the semantics of the content [3]. The statistical approaches try to provide a good summary compared to other approaches implemented.

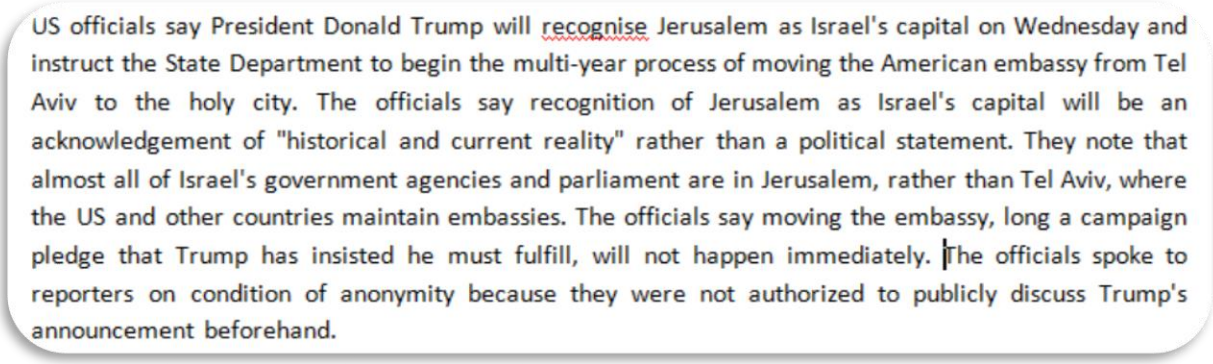
US officials say President Donald Trump will recognise Jerusalem as Israel's capital on Wednesday and instruct the State Department to begin the multi-year process of moving the American embassy from Tel Aviv to the holy city. The officials say recognition of Jerusalem as Israel's capital will be an acknowledgement of "historical and current reality" rather than a political statement. They note that almost all of Israel's government agencies and parliament are in Jerusalem, rather than Tel Aviv, where the US and other countries maintain embassies. The officials say moving the embassy, long a campaign pledge that Trump has insisted he must fulfill, will not happen immediately. The officials spoke to reporters on condition of anonymity because they were not authorized to publicly discuss Trump's announcement beforehand.

Figure 1: Sample Text from Dataset

The goal of this work is to summarize the English news article with normal statistical approach called Word Frequency algorithm and measure the performance of the system with ROUGE metrics. Figure 1 provides the sample text from the training dataset.

2. Related Works

Summarization based on the hypergraph transversal was done in [4]. The sentence of the corpus is considered as nodes and grouping the sentence having the same theme is mapped with hyperedges. This concentrates on achieving summary with minimal length and maximal content coverage without exceeding target length. This model outperforms other approaches by 6% of ROUGE-SU4 score. [2] searches for clustering of sentences based on semantic and lexical features. Doc2Vec and LDA are used for obtaining semantic features. This provides better performance on CNN/Daily Mail dataset with ROUGE-1 as 41.4.

The unsupervised approach has been completed combining clustering along with topic modeling. Topic modeling used Latent Dirichlet Allocation, while K-Medoids clustering used for summary generation. They evaluated their system on three different datasets DUC2002 Corpus, CNN/DailyMail and Wikihow [5].

[6] integrates word embeddings into deep neural network to enhance the quality of the summary being generated. They implemented ensemble techniques in three ways: BOW and Word2vec using majority voting, BOW combined with unsupervised neural networks and Word2vec combined with unsupervised neural networks. Summarization is also performed as binary optimization problem where quality of summary is based on sentence length, sentence position and relevance to the title. They use genetic operators and guided local search which improves the quality of the summary than other optimization techniques [7].

A model based on the rank fusion is implemented with four multidimensional sentences features like topic information, significant keywords, semantic content and position of the sentence [8]. This follows unsupervised model for generating scores and the weights are learned based on the labeled document. [9] proposed an idea on summarization based on combining fuzzy inference system, evolutionary and clustering algorithms. The summaries generated by this system are analyzed by the experts to know the performance.

Summarization is majorly done using sentence ranking. Each sentence used in the text is given with weights and are ranked depending upon weights. The sentences with highest rank are used in summary to accomplish good summary [10]. Similarly, summarization is performed on various

approaches considering features at word level and sentence level. The word feature includes content, cue phrase, case of the word, bias word and title of the word. The sentence level feature includes location, length, paragraph location and cohesion with other sentence [11].

3. Methodology

The methodology used in this work is Work Frequency algorithm which is implemented using Natural Language ToolKit (NLTK) library. Figure 2 shows the process of text summarization implemented in this paper. The steps involved in the proposed work are given below.

3.1. Preprocessing

The entire dataset provided by ILSUM track of FIRE 2022 have been imported into python dataframe. The text from the dataframe is processed for summarization. Text consists of various symbols and special characters are to be removed through preprocessing. This step also involves removing the stopwords from the given content. The list of words included in NLTK library is used to eliminate the stopwords from the text.

3.2. Sentence Score

The preprocessed sentences are tokenized to get the list of entire words used in the article. The weighted frequency for each word is calculated based on their occurrence. Equation (1) helps in calculating the weighted frequency for words that are tokenized.

$$\mathrm{WF} = \mathrm{Freq}_{\mathrm{word}} / \mathrm{Freq}_{\text{most occurred word}} \tag{1}$$

where

WF refers Weighted Frequency
$\mathrm{Freq}_{\mathrm{word}}$ refers the frequency of the current word for which WF is calculated
$\mathrm{Freq}_{\text{most occurred word}}$ refers the frequency of the word that is most occurred in the text

Each sentence score is calculated based on replacing the words with their weighted frequency and summing up all the WF for each sentence. Sentence Score for each sentence is calculated based on Equation (2).

$$\text{Sentence Score} = \sum_1^n WF \tag{2}$$

where

n refers the number of words in a sentence
WF refers Weighted Frequency

3.3. Generating the Summary

The average of all computed sentence scores is determined, and this average is used as a threshold value. Equation (3) gives the average of sentence scores. If the sentence score is more than the average score, it will be retained for the summary. This methodology is an extractive summarization technique which tries to retain the sentences of the text which has highest score and include the original sentences from the test into the summary. A threshold value can be modified to get different summaries. The sentences score that is above the threshold will be hold-on to generate summary.

$$Average = \frac{\sum Sentence\ Scores}{Total\ no\ of\ Sentences} \tag{3}$$

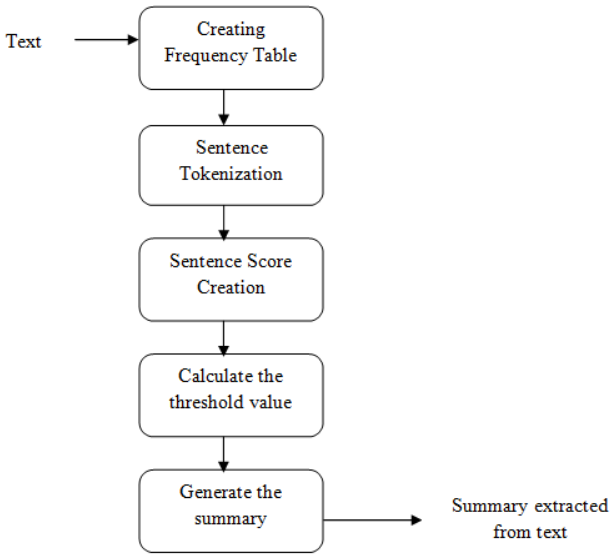

Figure 2: Proposed methodology of text summarization

4. Results

The size of dataset used in this work is showed in Table 1. The performance of the system was evaluated by organizing team using ROUGE metrics. The ROUGE-1, ROUGE-2 and ROUGE-4 are used for measuring the summary quality. Table 2 provides the measures of Precision, recall and F1-Score of our system. Figure 3 provides the graphical representation of the results achieved.

Table 1
Size of the Dataset

Dataset	No. of Articles
Train	12565
Validation	899
Test	4487

Table 2
Performance Measure of Test Data

ROUGE	F1- Score	Precision	Recall
ROUGE-1	0.34013	0.272376	0.5193
ROUGE-2	0.208011	0.164724	0.323263
ROUGE-4	0.170998	0.133656	0.272965

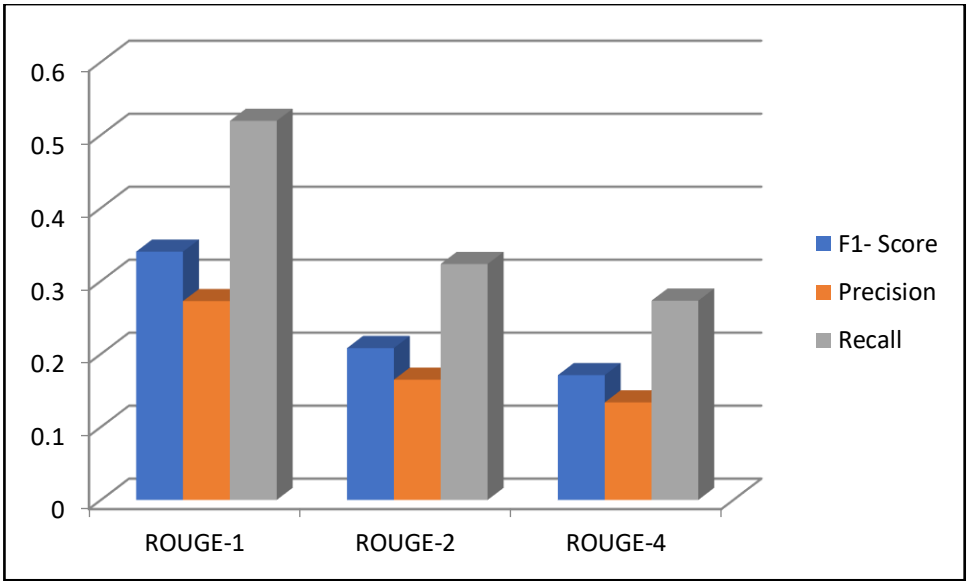

Figure 3: Performance measure for ROUGE metrics

5. Conclusion

The proposed system is used for summarizing the given information by retaining the vital information of the original text. In the proposed work, Word Frequency Algorithm is used to get the summary of the text by computing the weighted frequency for each words used in the content. With the help of weighted frequency, each sentence is assigned with a score and threshold value is computed. By changing the threshold value of the sentence score, different summary can be obtained. From the results, it is evident that the proposed methodology provides acceptable summary and it can be further improved by including lexicon information of the given text.

6. References

[1] Mengli Zhang, Gang Zhou, Wanting Yu, Ningbo Huang ,and Wenfen Liu, "A Comprehensive Survey of Abstractive Text Summarization Based on Deep Learning", Computational Intelligence and Neuroscience, (2022) doi:10.1155/2020/9365340.

[2] Ángel Hernández-Castañeda, René Arnulfo García-Hernández, YuliaLedeneva, Christian Eduardo Millán-Hernández, "Language-independent extractive automatic text summarization based on automatic keyword extraction", Computer Speech & Language, (2022) doi:10.1016/j.csl.2021.101267.

[3] AdhikaPramitaWidyassari, SupriadiRustad, GuruhFajarShidik, Edi Noersasongko, Abdul Syukur, AffandyAffandy, De Rosal Ignatius Moses Setiadi, "Review of automatic text summarization techniques & methods",Journal of King Saud University - Computer and Information Sciences, (2022), Volume 34, Issue 4, 1029-1046 doi: 10.1016/j.jksuci.2020.05.006.

[4] H. Van Lierde, Tommy W.S. Chow, "Query-oriented text summarization based on hypergraph transversals", Information Processing & Management, (2019), Volume 56, Issue 4, 1317-1338, doi: 10.1016/j.ipm.2019.03.003.

[5] Ridam Srivastava, Prabhav Singh, K.P.S. Rana, Vineet Kumar, "A topic modeled unsupervised approach to single document extractive text summarization", Knowledge-Based Systems,(2022), Volume 246, doi: 10.1016/j.knosys.2022.108636.

[6] Nabil Alami, Mohammed Meknassi, Noureddine En-nahnahi, “Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning”, Expert Systems with Applications, (2019), Volume 123, 195-211, doi: 10.1016/j.eswa.2019.01.037.
[7] Martha Mendoza, Susana Bonilla, Clara Noguera, Carlos Cobos, Elizabeth León, “Extractive single-document summarization based on genetic operators and guided local search”,Expert Systems with Applications, (2014), Volume 41, Issue 9, 4158-4169.
[8] Akanksha Joshi, Eduardo Fidalgo, Enrique Alegre, Rocio Alaiz-Rodriguez, ”RankSum—An unsupervised extractive text summarization based on rank fusion”, Expert Systems with Applications, (2022), doi: 10.1016/j.eswa.2022.116846.
[9] Pradeepika Verma, Anshul Verma, Sukomal Pal, ”An approach for extractive text summarization using fuzzy evolutionary and clustering algorithms”, Applied Soft Computing, (2022), doi: 10.1016/j.asoc.2022.108670.
[10] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," in: International Conference on Data Science and Communication, (2019), pp. 1-3, doi: 10.1109/IconDSC.2019.8817040.
[11] Abinaya N, Anand R and Arunkumar T, ”An Exhaustive Survey on Automatic Text Summarization Using Machine Learning Approches”, Webology, (2021), pp.1184-1190.
[12] Akash Panchal, url: https://github.com/akashp1712/summarize-webpage
[13] S. Satapara, B. Modha, S. Modha, P. Mehta, Findings of the First Shared Task on Indian Language Summarization (ILSUM): Approaches, Challenges and the Path Ahead. In Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022.
[14] S. Satapara, B. Modha, S. Modha, P. Mehta, FIRE 2022 ILSUM track: Indian Language Summarization. In Proc. of the 14th Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022.