Text Summarization Using Natural Language Processing

Islam MD Shariful
1720601
BCS,IIUM,Gombak,
Selangor, Malaysia
sinoyon780@gmail.com

Mohamed Moubarak
Mohamed Misbahou
Mkouboi 1820705,
BCS,IIUM, Gombak,
Selangor, Malaysia
mkouboimoubarak18@gmail.com

Abdella Mame Abdo
1714883
BCS,IIUM,Gombak,
Selangor, Malaysia
Abdellamame@gmail.com

ABSTRACT

Text summarization is a technique for synthesizing and condensing the form of a document on the fly. Manually summarizing important written materials is a difficult task for humans. The two most common ways for summarizing text are extractive and abstract. Though extractive summary is mainly concerned with what information should be included in the summary, the frequency with which words, phrases, and sentences from the source text should be utilized should also be considered. This research study presents a tf-idf model for summarizing big publications in order to provide the reader with a clear picture of the content.

KEYWORD: Text Summarization, Tf-Idf, Reuters, Sentence positions, F-1 Score, Rouge score, Blue score, Extractive.

Introduction

Text summarization is defined as the text that is summed up from one or more texts. Most of this new text contains a significant portion of information from the original text that we are summarizing, and it's not more than half of the original text. When we distill the most important information from a particular source, we are producing specific form text that can help us

understand our original text(s). Broadly speaking, Text summarization has two main groups when it comes to the output, which are Inductive and Informative summarization. Inductive summarization is mainly about the main point of the text, while informative summarization is more about giving specific information from the main text.

Furthermore, natural language processing continues to have a problem with summarizing large quantities of texts, until the introduction of Automatic Text Summarization which works on shortening large documents. These documents can come in the form of a Single Document where the text is small and can be handled by basic summarization models. The other type of input is Multi-Document, where the input is comparatively long and leads to increased complexity.

Problem statement

Automatic Text Summarization is a technique in which computer software shortens lengthy texts and provides summaries to convey the desired content. It is a prevalent challenge in machine

learning and natural language processing (NLP). Text summarizing is the process of constructing a concise, cohesive, and fluent summary of a lengthier text document, which includes highlighting the main ideas of the text.

Text identification, interpretation, and summary production, as well as analysis of the resulting summary, are some of the key issues encountered during the text summarizing process. Identifying important terms in the document and exploiting them to uncover relevant information to add to the summary are crucial jobs in extraction-based summarizing. There are two ways of text summarizing that are used:

- Extractive Summarization
- Abstractive Summarization

Project Objective

- To understand the concepts of natural language processing and create a tool for text summarization.
- To understand past works related to suspicious activities.
- To apply the Tf-Idf model to the problem of multi-document summarization.
- To understand, concentrate on creating a tool that automatically summarizes the document.

Motivation

Text summarizing is the process of constructing a concise, cohesive, and fluent summary of a lengthier text document, which includes highlighting the main ideas of the text. There are two ways of text summarizing that are used: Extractive Summarization is a term used to describe the process of extracting information from Summarization that is abstracted. Automatic text summarizing methods are desperately needed to deal with the ever-increasing volume of text data available online, both to help identify relevant information and to consume relevant information more quickly. Text summarization is a technique for condensing extensive passages of text. The goal is to develop a logical and fluent summary that only includes the document's major ideas. In machine learning and natural language processing, automatic text summarization is a prevalent challenge (NLP).

Related Work

The extractive-based summarization uses K-Means Clustering in conjunction with TF-IDF (Term Frequency-Inverse Document Frequency) for the summary. The study article also incorporates the concept of genuine K and splits the sentences of the input material using that value of K to produce the final summary. In the experiment, the K-means and TF-IDF methods were utilized for extracting text summarization with a specified K value, as these are two well-known approaches that are widely used for genuine K determination. It is obvious that the statistical measurements technique produces the greatest outcomes, and researchers employed two forms of comparative assessment to demonstrate this. One thing that became evident after all of these trials is that before summarizing the web crawled or corpus texts, the preprocessing stage should be thoroughly examined to ensure that any extraneous characters, keywords, tags, and punctuation are removed [1]. Gensim Word2Vec and the K-Means Clustering Algorithm are used to automatically summarize the text. For a single text, this study presents a sentence-based clustering approach (K-Means). This article

utilized Gensim word2vec for feature extraction, which is designed to automatically extract semantic themes from articles in the most efficient way possible. The K-Means clustering technique was used to cluster all of the texts in this study model. A phrase scoring system assigns a score to a sentence based on the presence of numerical values and nouns. These approaches were applied to databases of BBC news articles. The model performed better on business articles because they contain more numerical data, and the sentence score approach prioritizes numerical values [2]. The use of continuous vector representations for semantically aware sentence representations as a foundation for evaluating similarity evaluates alternative compositions for phrase representation on a standard dataset using the ROUGE evaluation measures. These research studies demonstrate that the assessed approaches increase the effectiveness of an innovative summarizing framework and strongly suggest the benefits of continuous word vector representations for automatic summarization. The findings of this article indicate that using word and phrase embeddings in summarization has a lot of promise [3].

Technical background

1. Position of sentence

One of the important aspects of texts is the position of every sentence in the text because this will impact the outcome of the summarization process. Based on this sentence position hypothesis, sentence position features are defined by the ordinal position of sentences. The effectiveness of these position features has been proven especially when it comes to generic summarization and query-focused tasks.

2. NLTK

NLTK Is a platform for building Python programs to work with human language data. NLTK refers to Natural Language Toolkit, which provides easy-to-use interfaces to over 50 corpora and lexical resources, such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum. In this paper, we use multiple packages from the NLTK corpus and routers.

3. Reuter Corpus

Corpus is one of the most important aspects of programming with NLTK, which is used in linguistics research or spoken material. Some Corpus work as a collection of words, like the British National Corpus (BNC), which contain 100 million word collection of samples of

written and spoken language from a wide range of sources.

4. Tf-Idf

Tf-Idf is an approach to measure single text terms. It refers to “term frequency-inverse document frequency”, which is used to score the importance of a word in a document based on how often it appears in that document and a given collection of documents. The formula for calculating tf and idf:

- TF(w) = (Number of times term w appears in a document) / (Total number of terms in the document)

- IDF(w) = log_e(Total number of documents / Number of documents with term w in it)

5. Evaluation

The application of Bilingual Evaluation Understudy (blue) is essential in our testing. As known, blue is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Its score is for comparing a candidate translation of text to one or more reference translations.

Methodology

In this project, we are trying to summarize the text by using real data. So that summarization can give more or less accuracy to analyze the real outcomes. We collect some articles and news from online sources.

Data collection: We collect from StraitsTimes and TheStar. The article is uploaded By Teoh Pei Ying and Nation on Wednesday, 26 Jan 20228:29 PM MYT

Tools: we will use google colab and Jupyter notebook.

Model: we are using some approaches such as the position of the sentence, top-down approaches, creative corpus, tf-idf method, etc, for text summarization. Using the tf-idf method we create a model summarization. We work retrieving articles or news and then we extract headlines and body. Then, we check the length of the sentences whether they are empty or not. If the article contains below than 5 sentences the model will print the existing article as it is. We evaluate our system using Rogue Score and Bleu Score.

Figure 1: Project WorkFlow

Results

Using Word Frequencies to calculate sentence score. We got the summarized sentence score of the given article as shown below.

'a 313 per…… .': 1.8461538461538463,
'During the.…. ': 1.7692307692307692,
'"the majority… .': 2.46153846153846,
'"the remaining …': 2.4615384615384,
'dr noor ……..': 3.8461538461538476,
'dr noor……..': 3.6153846153846168,
'he revealed……..': 3.3076923076923075,
'Health director……': 6.384615384615385,
'however, the…….': 1.5384615384615383,
'on the 4th……..': 3.230769230769231,
'other proposals ….': 1.923076923076923

After pushing our article1 into the Summarizer class to test using our tf-idf model, we got the summarized result of the given article as shown below.

```
# get top 5 sentences
import heapq
summary = heapq.nlargest(5, sentence_scores, key=sentence_scores.get)
summary=" ".join(summary)
print(summary)
```

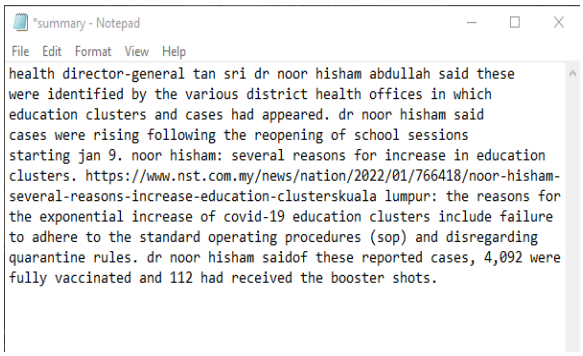

Figure 2: Summarization result.

Conclusion

Automatic Text Summarization is a prominent challenge in machine learning and natural language processing in which a computer program shortens lengthy texts and provides summaries to deliver the intended information. Summaries shorten reading time when studying materials, summaries make the selecting process easier. Indexing becomes more effective with automatic summarization. Automatic summarizing systems are less prejudiced than human summarizers. However, automated summarizing takes less time to complete than human summarizers.

REFERENCES

Khan, R., Qian, Y., & Naeem, S. (2019). Extractive-based Text Summarization Using K-Means and TF-IDF. *International Journal of Information Engineering & Electronic Business*, *11*(3).

Haider, M. M., Hossin, M. A., Mahi, H. R., & Arif, H. (2020). Automatic Text Summarization Using Gensim Word2Vec and K-Means Clustering Algorithm. 2020 IEEE Region 10 Symposium, TENSYMP 2020, June, 283–286. https://doi.org/10.1109/TENSYMP50017.2020.9230670

Kågebäck, M., Mogren, O., Tahmasebi, N., & Dubhashi, D. (2015). Extractive Summarization using Continuous Vector Space Models. 31–39. https://doi.org/10.3115/v1/w14-1504

Rahul, S. Adhikari, and Monika, "NLP based Machine Learning Approach for Text Summarization," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC).

I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand and P. K. Soni, "Natural Language Processing (NLP) based Text Summarization

A Survey," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021,

Samrat Babar, "Text Summarization: An Overview", Sanjeevan Engineering and Technology Institute Panhala, 2013, https://www.researchgate.net/publication/257947528_Text_SummarizationAn_Overview

Oguzhan Tas, Farzad Kiyani, "A SURVEY AUTOMATIC TEXT SUMMARIZATION", 2nd World Conference on Technology, Innovation and Entrepreneurship, 2017.