Nepali Text Summarization

Pitambar Mahato

Lumbini ICT Campus, Gaindakot

Abstract: The project presented here is the Nepali text summarization where I have used the extractive method of summarization technique. Now a day's text summarization plays the most important role to extract the hidden information from the large amount of articles/texts. The extractive method extracts the information from the document by calculating the importance of sentence/paragraph in the document and present the important information to the users. During this project, I have scraped the articles of the *setopati* to summarize them. The prototype of the summarizer system produces the good result for these articles.

Keywords: Extractive Summary, Abstractive Summary, Natural Language Processing, Artificial Intelligence

Abbreviations: AI → Artificial Intelligence, NLP → Natural Language Processing

TABLE OF CONTENTS

LIST OF FIGURES

1 INTRODUCTION

Text Summarization is the process of obtaining silent information from an authentic text document. In this technique, the extracted information is achieved as a summarized report and conferred as a concise summary to the user. It is very crucial for humans to understand and to describe the content of the text. Text summarization techniques are classified into abstractive and extractive summarization. The extractive summarization technique focuses on choosing how paragraphs and important sentences produces the original documents in precise form. The implication of sentences is determined based on linguistic and statistical features.

This project is also the Natural Language Processing related project that I have done during my college days. This project is able to summarize the news by calculating maximum frequency of the sentences. Data are collect by scraping setopati news website and performed text-rank algorithm to summarize the news.

Nepali news summarization is project that helps to summarize the news and provide the users to read the important news in easier and faster way. The Nepali text summarization provides Nepalese to read the contents of the web or any digital documents more smooth way by saving their time.

2 LITERATURE REVIEW

There are lots of researches on Automatic text summarization and various techniques are being developed. Various researchers have proposed new techniques using multiple methodologies for automatic text summarization.

A summary is a “text that is produced from one or more texts, that conveys important information in the original texts, and that is no longer than half of the original texts and usually significantly less than the original text. A summary is a decomposed and rebuilt version of its source only consisting the essence of the text. Because of the decomposing nature, summaries of low quality suffer from information loss [1].

Text summarization has its importance in both commercial as well as research community. As abstractive summarization requires more learning and reasoning, it is bit complex then extractive approach but, abstractive summarization provides more meaningful and appropriate summary compare to extractive [2]. In my project I have done it by using extractive method.

In 2003, Madhyastha, Harsha V., et. al [3] proposed a method which makes use of the syntactic structure assigned to the input text by the link parser and its work lies in the working of the rules for prediction of subject, object and their modifiers. In the subject prediction scheme, the linkage of each sentence is considered one by one. If the subject is in some other sentence, then it cannot be detected by this scheme.

In 2015, Luciano Cabral et al [4], proposed method for automatic summarization application which allows users to view summaries of news pages on Android-enabled mobile devices. The proposed method contains two approach first approach pre-processes web pages by reformatting or adapting them to a more appropriate way of viewing on small screens, without altering the original content Second approach selects the most salient and relevant content in a given page to the user, meeting their need for quickly grasping the fundamental information.

3 METHODOLOGY AND SYSTEM DESIGN

3.1 FLOW OF WORK

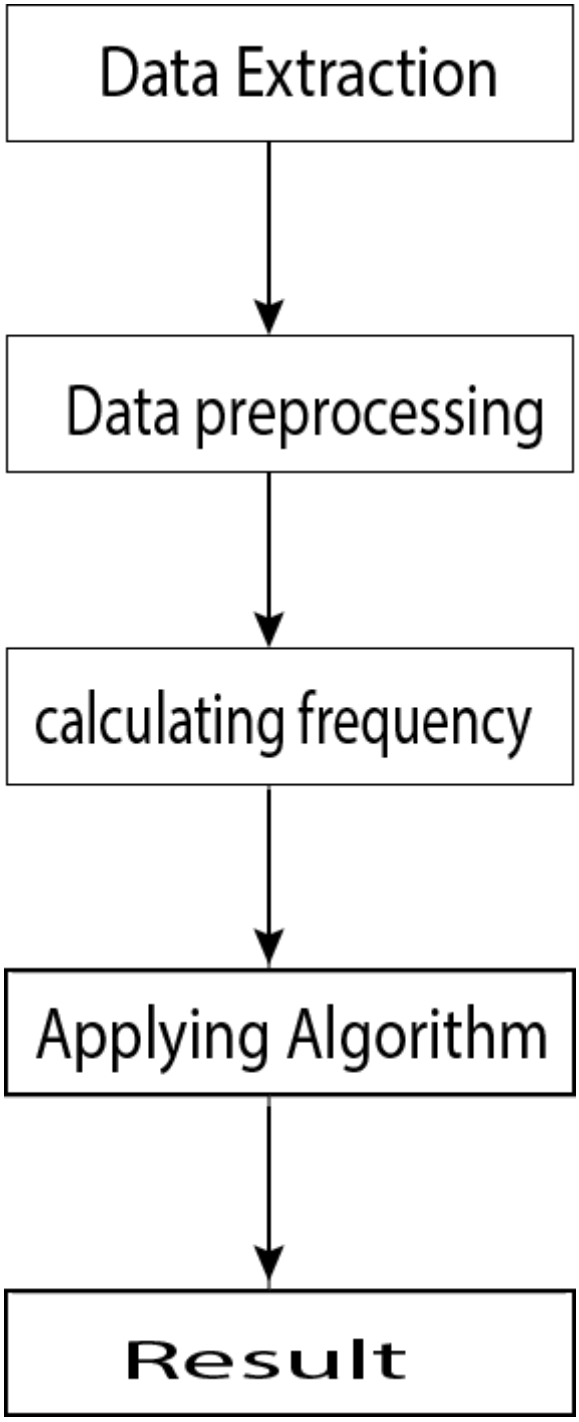

Figure 3.1: General Flow of Data

3.2 DATA COLLECTION

Data collection is an initial stage of this project our objective is summarized Nepali news so I collect Nepali news by scraping setopati. Web scraping method used for data collection. Web scraping is a term for various methods used to collect information from across the Internet. Generally, this is done with software that simulates human Web surfing to collect specified bits of information from different websites. Those who use web scraping programs may be looking to collect certain data to sell to other users, or to use for promotional purposes on a website. Web scraping is also called Web data extraction, screen scraping or Web harvesting. I use python programming language there is lots of package are available for scraping website I use Beautiful soup for scraping data from website.

3.3 PREPROCESS

This step cleans the document by removing HTML tags and noisy characters present in the document, by correcting spelling mistakes, grammar mistakes, and punctuation errors.

3.3.1 STOP WORDS

A stop word is a commonly used word (such as'छ', 'र','गरे को') that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. We would not want these words taking up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that we consider to be stop words. NLTK (Natural Language Toolkit) in python has a list of stop words stored in 16 different languages. We also used to stop words package for removing stop words from datasets.

3.3.2 TOKENIZATION

Tokenization describes splitting paragraphs into sentences, or sentences into individual words. Sentences can be split into individual words and punctuation through a similar process. In this project used splitting technique to tokenized data. For the tokenization process, I used the spacy library of the python.

3.3.3 REMOVE SPECIAL CHARACTERS

There are huge number of special character are present in our dataset. Special characters are those which may effect in analysis process or those which is not take in process. The dataset contain special character are number, '()', <>, '/', '-' etc. Special character was removed using regular expression.

3.4 FREQUENCY CALCULATION

Conventionally, histogram of words is the features for the NLP project problems. In general, we first build the vocabulary of the corpus and then we generate word count vector from each file, which is nothing, but frequency of words presents in the vocabulary.

3.4.1 WEIGHTED FREQUENCY CALCULATION OF WORDS

After calculating the frequencies of the words in the document, we have to calculate the weighted frequency of the words in the document through which we can calculate the frequency of each sentences and get the summary of the document.

3.4.2 FREQUENCY CALCULATION OF SENTENCES

After the calculation of the weighted frequency of the words in the document then the frequency of the sentences is calculated by adding the weighted frequencies of the words present in the sentences then the sentences having maximum frequency are selected for the summary of the document.

3.5 ALGORITHM

In this project, I use unsupervised machine learning algorithm. Unsupervised learning, in the context of artificial intelligence (AI) and machine learning, is a type of system in which only the inputs are provided. Input data are processed and then feed to the algorithm and then provide the result to the users. There are lots of advantages of unsupervised learning methods in the field of machine learning.

Algorithm:

Step 1: Input raw data

Step 2: Preprocess data

Step 3: Tokenize sentences

Step 4: Tokenize words

Step 5: Calculate the Weighted Frequency of each words in the document

Step 6: Calculate the Frequency of the tokenized sentences in the document.

Step 7: Find the most relevant sentences of the document.

Step 8: Get Summary of the document

4 RESULT ANALYSIS

4.1 RESULT

Result of the project is summarized news based on provided news in the system. For getting result number of pre-processing task is performing. After model built, the model was implemented in flask for showing how the algorithm works.

4.1.1 FINAL OUTPUT

After the completion of the data-preprocessing task the frequency of the words are calculated and then applied to the algorithm which provides us the summarized result of the news that was provided to the system.

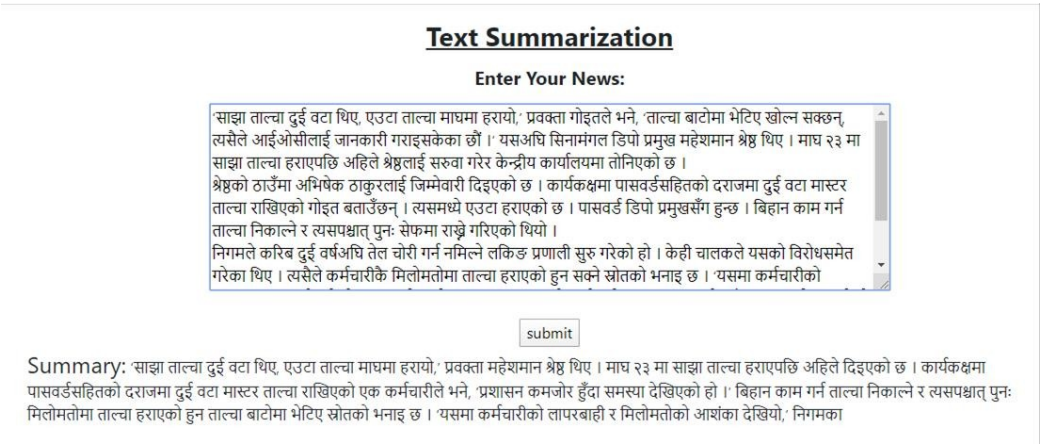

Figure 4.1: Output

4.2 CRITICAL ANALYSIS

Text Summarization is one of the most important process for the readers to read the documents in more reliable ways, which saves the time of the reader to read the documents. In the system, we need to give the input to the system a large document, which process the documents and then produces the summary result of the documents that we have provided.

The main objective of this project is to make the text summarization system for the Nepali text corpus. There have been many text summarization applications, which supports the summarization of the non-nepali text summarization, but for the summarization of the Nepali texts, the research process is still in process. There have been many different ways of the text summarization process but in this project, I used the extractive method for the summarization of the Nepali text corpus.

The extractive text summarization process works in such a way that the most important features of the texts are selected for the summarization of the whole documents. Those sentences/paragraph are selected whose frequency is greater than that of the other documents.

For the summarization process I built the web-based system by using the python programming language form the initial phase of the system. To make the system I used different libraries of the python programming language such as pandas, beautiful soup, nltk, spacy, and flask (for the web).

4.3 LIMITATIONS OF THE SYSTEM

There have been many advantages of the text summarization systems but sometimes due to presence of little amount of data the system will lose some semantic information during the summarization process and there will be the loss of important information. The system can summarize only one document at a time.

4.4 FUTURE WORK

This work can be further enhance to do some abstractive summarization techniques and also get some better results. The system can further extended to work for other languages.

To make system work more properly we can use the tf-idf scores instead of the frequency.

REFERENCES

[1] W. V. HOORN, "Automatic Text Summmarization as a Text Extraction Strategy for Effective Autoated Highlighting," 25 February 2018.

[2] D. K. Gaikwad and C. N. Mahender, "A Review Paper on Text Sumarization," 2016.

[3] J. T. and S. Thede, An Autoatic Text Summarizer, 2003.

[4] L. C. e. al, Automatic Summarization of News Article for Mobile Devices, 2015.