

Analysis, Design and Development of an Automatic Text Summarizer for Bangla Web Document

Mohammed Mahmudur Rahman

Dept. of Computer Science & Engineering, International Islamic University Chittagong, Bangladesh.

Corresponding Author: Mohammed Mahmudur Rahman

Abstract: Bengali is one of the ten most spoken languages in the world, with almost 200 million speakers. Growing online resources reveal a clear need for Bengali language applications, retrieval systems and automatic text summarization. Designing a system to produce human quality summaries is difficult and therefore many researchers have focused on sentence or paragraph extraction, which is a kind of summarization. In this research work an intelligent technique is introduced in order to summarize the Bangla texts with the help of machine supported technology. This system can be widely used in effective Bangla text summarization which helps to extract the focal or central parts of the documents at a quick glance. A good example of summarization system is the convention search engine like Google, to represent compressed description of the search results. Other examples are to summarize news to Bangla SMS or WAP, keyword directed news subscriptions of news etc. In our Bangla Summarization System, an algorithm is proposed to extract the Bangla summary which works on four passes. First two passes perform tokenization on the input Bangla sentences and word frequency calculation and next two passes perform sentence scoring and summary generation.

Keywords - expert system, extraction, morphology, natural language processing, text summarization

Date of Submission: 01-02-2019

Date of acceptance: 18-02-2019

I. Introduction

The growth of electronic texts is becoming increasingly common. Newspapers or magazines tend to be available on the World-Wide Web. Summarizing these texts can help users access to the information content more quickly. However, doing this task by humans is costly and time consuming. With the actual huge and continuously growing online text resources, it becomes necessary to help users get quick answers to their queries. Automatic text summarization is a solution for dealing with this problem. Automatic text summarization takes a text and produces a summary of the most important parts of the original text and this must be coherent with the original text. One approach to respond to the rapid growth of information is to use text summarization for faster and more efficient processing by human as well as computer agents. Summarization which is based on semantics would be a future success, but for now making summaries reduces to the task of Extraction.

II. Literature Review

Research work on automatic text summarization is going on. There has been substantial research work on automatic text summarization field on different languages since last two decade. Here is some presentation of the major ideas which are being worked on. Joel et al. present a document clustering and text summarization algorithm which was the development of the widely used TF-IDF model [1]. M. Sanderson proposed a query relevant summarizer that divides the document into equally sized overlapping passages and uses the INQUERY text search engine to obtain the passage that matches the user's query [2]. An algorithm is developed for automatic text summarization of technical documents by building and minimizing information network of the source document [3]. This text summarization algorithm creates text summaries by finding out the focal topic of the document and extracting the most the most relevant properties of that topic. The topic and its properties are represented as semantic networks and then converted to some sentences to create the summary. This summarization algorithm will extract the most important words from the document. It will find out a central word from the document, which will actually be the main theme. Sentences that refer the main topic will be treated as properties of the topic. There may be some other topics available in the document, but these are sub-topics of the main topic. Properties related to the sub-topics may also be found. The topics and the sub-topics along with their properties will be stored as the inheritable knowledge format. When the whole document has been processed and it has been completely represented in the inheritable knowledge graph format. This

approach is expected to be useful for summarizing technical documents. And it is expected to contain most important information with less redundancy.

In Bangla language this automatic text summarization field is in under development at present. Only a very few research works have been done in this field. A Corpus-Based Information Retrieval and Summarizer [4] is developed for Bengali Text. The goal of the summarizer is performed by document indexing and retrieves information based on key words using vector space retrieval method. It is implemented using Python Scripting Language that integrates powerful text processing modules like Regular Expression(RE), NLTK, String, database modules like MySQL, gadfly etc which are considered to be very effective and easy learning modules that gives opportunity to build Bhasa. Bhasa is capable of ranking the corpus files where query terms occur most frequently and summarizing the documents based on query terms by using vector space retrieval, term weighting. It uses a tokenizer to tokenize the documents and then performs document ranking and summarizations on the tokenized document. The tokenizer is capable of detecting different words, tags, abbreviations, and sentence boundaries and uses markups to represent words, sentences, heading and titles by syntactic and semantic analysis. In this case, automatically summarized text can sometimes result in dangling anaphors if the sentences are extracted by shallow linguistic analysis of the text. Text summarization system is developed in deferent language such as:

- A Text Summarizer for Swedish [5]
- A Persian text summarizer (FarsiSum) [6]

2.1 Text Summarizer for Swedish

The first text summarizer for Swedish-SweSum which is built on both statistical and linguistic methods as well as heuristic methods. The domain of SweSum is Swedish HTML tagged newspaper text. SweSum ignores HTML tags which control the format of the page but processes the HTML tags which control the format of text. The idea is that high scoring sentences in the original text are kept in the summary; the scores are calculated according to the criteria below:-

Position score: the sentences in the beginning of the text are given higher scores than the ones at the end. The formula is, $1/n$, where n is the line number, so called Baseline.

HTML tags which indicate sentences with bold text are given a higher score than the ones without bold text tagging, ditto title tagging. Bold text also indicates the beginning of a new paragraph in some of the Swedish news paper texts. Sentences containing numerical data are given a higher score than the ones without numerical values. Sentences which contain keywords are scored high so called Term frequency (tf).

2.2 FarsiSum

FarsiSum is an attempt to create an automatic text summarization system for Persian based on SweSum algorithm and technique which extract summary by using keyword extraction and sentence scoring. It is a web-based text summarizer that summarizes Persian newspaper text in HTML/text encoded in Unicode format. It uses same structure used by SweSum but it modifies some modules to be able to handle documents with Unicode content and UTF-8 encoding. The summarizer process starts when the user (client) clicks on a hyperlink (summarize) in the FarsiSum Web site:-

- The browser (Web client) sends a summarization request including a document to be summarized, through a HTTP server to the Web server where FarsiSum is located.
- The document is summarized in three phases using a Persian stop-list.
- The summary is returned back to the client through the HTTP. The browser then renders the summarized text to the screen.

This summarizer is developed in three passes. In first pass the tokenizer is modified in order to recognize Persian comma, semi colon and question mark. In second pass sentences ranking task is implemented. Here, there are no changes in the scoring algorithms used by SweSum. The sentences are scored and put into a ranking list. Each sentence is scored according to the position of the sentence and the scores of the words contained in it. The highest-ranking sentences are then identified and kept in the final summary.

2.3 Present State and Contribution

Last two decades has seen so many text summarization techniques but a very few of them was implemented. Most of them are extraction based text summarization algorithms. Some of these algorithms are used in some conventional search engines. A few text summarization techniques were implemented considering some specific languages. In Bangla language this automatic text summarization field is in under development at present. Some attempts were made to implement some desktop based Bangla text summarization system. As far we know there is not a single implementation of web based Bangla text summarization system. This is a very small step in text summarization area for Bangla language and we hope this proceed us to a greater and vast success one day. This achievement will lead us to a greater success in web search using Bangla words or corpus

if we couple this summarization technique with conventional search engine. We will be able to search Bangla documents using Bangla keywords.

III. Prospects

Our mother tongue Bangla is in the fourth position among the mostly spoken languages. Over 200 million of people all over the world speak this language. Bangla literature is highly rich with prose, poetry and proverb. A sentence or a set of sentences focus the excellent ideals, characteristics and moral displacement of human being or other important feature of the society. Bangla being highly morphological rich language needs to be treated differently than other language like English while designing a text summarizer for it. Growing online resources like online magazines, news papers, WEB sites reveal a clear need for Bangla information retrieval and text summarization system. We, the Bengali nation will be able to search Bangla web document using Bangla corpus like other languages.

One day such a text summarization system will be coupled with some conventional search engines. So the motivation for the research work is to analysis, development and design of an automatic text summarization system for Bangla WEB document. There are number of applications of this sophisticated text summarization system. Some are explaining below:-

- (a) As the amount of on-line information increase, more and more effort is dedicated to creating automatic summarization systems. So automatic Bangla text summarization system is used to extract the most important parts of original text at a quick glance.
- (b) Another application of 'Bangla Text Summarization System' is to summarize news to Bangla SMS or WAP (Wireless Application Protocol) –format for mobile Phone or Personal Digital Assistance.
- (c) To let a computer synthetically read the summarized text. Written text can be too long and tedious to listen to.
- (d) In Internet search engine like as Google, to present compressed descriptions of the search results.
- (e) In keyword directed news subscriptions of news which are summarized and sent to the user.

IV. Methodology

This system receives http request and contents travelling with the request and identifies the original text and takes them as input to another process. Basic idea is to generate sentence score considering some related criterion. Thus system identifies the most important sentences and generates summary taking obvious minimization into account. In this chapter the methodology of our proposed system is explained with corresponding procedure and flowchart.

4.1 Procedures of System

Our total proposed system consists of four passes. In text preparation pass tokenization is performed on the inputs of the hypertext in order to find out original text then necessary tokenization is done to find lines and each word separately for counting frequency of each word. In the second pass the frequency of word is measured by counting the number of that words are exists in the text. In this pass unnecessary words from key list are removed along with the pronouns. After the second pass the sentence score is obtained in the third pass. Sentence score of each sentence is measured by using the sentence score generation formula. Finally in the summary generation by extracting sentences pass, most important sentences are extracted on the basis of high rank. These passes are:-

Pass 1: Preparing text for processing

- a. Http contents are received using curl or received the bangla original text in text container.
- b. If the bangla original text is given then at first it converts into its Unicode.
- c. Perform the tokenization on this input to remove html tags and finding the original text.
- d. Perform the tokenization of this original text to divide the text into separate lines and find word boundaries.
- e. Another tokenization is done to find out bold or italic data.
- f. Replace the words with common synonyms.

Pass 2: Counting word frequency and finding the keyword

- a. Count the frequency of each word in the document.
- b. Sort them in descending order.
- c. Take only high frequency words and calculate the final word score (word score= (word frequency*keyword constant+ other criterion))
- d. List the key words and their corresponding scores.
- e. Take bold or italic words and keep them in list along with their default scores.
- f. Show the list of count

Pass 3: Generating sentence score

- See the position of the sentence and assign a default score according to its position.
- Take the original tokenized text and find words of each line to generate sentence score.
- Sentence score = $\sum \text{word score (for all keywords in the sentence) + other criterion.}$
Where, Word score = (word frequency) * (a keyword constant)
- Calculate average sentence length (ASL).
- Final sentence score = (sentence score * ASL) / no. of words in current line.
- Sort them in descending order.

Pass 4: Summary generation by extracting sentences

- Extract the sentences that have high scores.
- Eliminate those sentences with quotation marks, question marks and exclamatory marks.
- Consider the cutoff size and unit.
- Final result.

4.2 Flow Charts of the System

The proposed flowchart of the system is divided into two parts. The first flow chart fig. 1 shows first two passes that are text preparation & counting word frequency and finding keyword. The second flow chart fig. 2 shows another two passes that is sentence score generation and sentence extraction & summary generation.

4.2.1 Flow Chart for Preparing Text and Counting Word Frequency

At first http contents are copied using curl. Then tokenization is performed on this content to remove html tags to find actual text that is to be summarized. This text is taken as input to another function to be tokenized to find each separate line along with to find each word separately. Then words are replaced with common synonym. Frequency of each word is counted and sorts them in descending order. Here, high frequency words are considered as most important keyword that is used to produce sentence score for next phases.

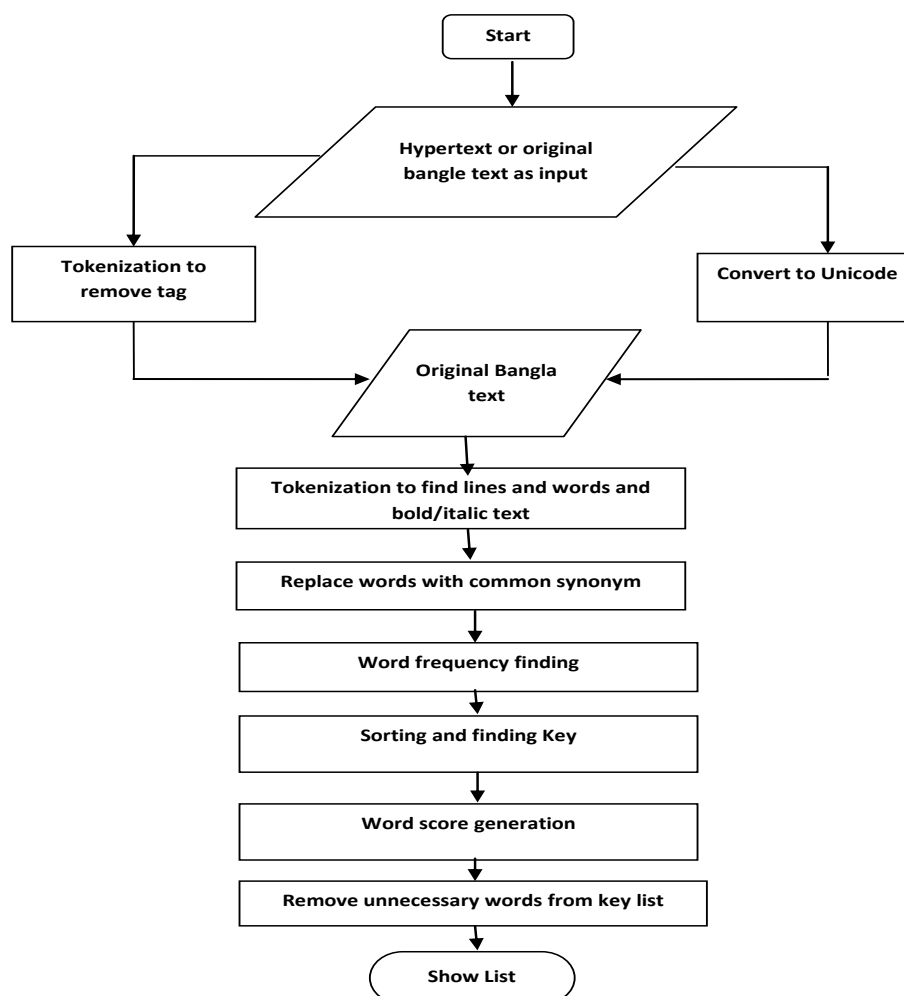


Fig. 1: Flow chart of preparing the text for processing and word score generation.

- Step 1: First take hypertexts located in the url as input using Unicode mapping.
 Step 2: Tokenization is performed in order to find out the original text removing html or other tags.
 Step 3: Another tokenization is done to find each line separately also to find separate words. In this step bold or italic texts are also identified.
 Step 4: Replace each word with its common synonym for counting keyword frequency.
 Step 5: In this step the frequency of word is measured from the text. This keyword would be different parts of speech such as nouns, adjectives and adverbs.
 Step 6: Sort the words in descending order according to their count and make a list of keywords.
 Step 7: Word scores are generated and stores to their corresponding keyword list.
 Step 8: If there any bold or italic words they are also kept in keyword list with their default values.
 Step 9: Here unnecessary stop words are removed from Keyword list if there any. This unnecessary word would be:-

Pronoun=["সে", "তর", "তাহার", "তাদের", "তাহাদের", "তিনি"]

Verb=["অনুমান করা", "প্রস্তাব করা"]

Article=["একটি", "টি", "টা"]

Preposition=["ভিতরে", "উপরে", "সাথে", "এর", "দিকে"]

Unimportant word=["যেমন", "উদাহরণ", "কথার কথা", "উপরন্তু", "অন্যদিকে", "ও"]

Useful wordlist=["এই বিষয়ে", "আলোচ্য অংশে", "পরিশেষে", "উপসংহারে", "প্রধান", "মূলত", "আসল", "উল্লেখ্য", "সংক্ষেপে"]

4.2.2 Flow chart for Sentence Score Generation & Summary Generation

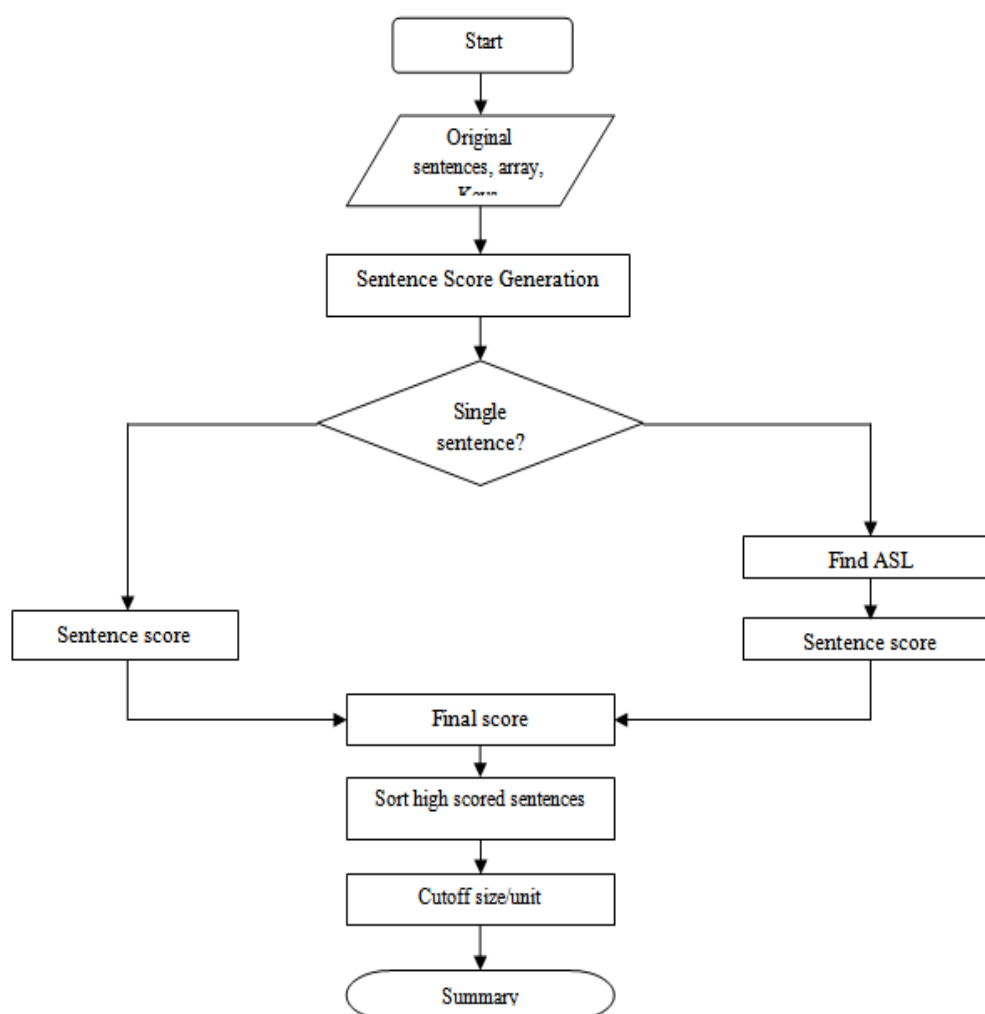


Fig. 2: Flow chart for sentence score generation, extraction & summary generation.

The original Bangla sentences and high frequency words from the previous phases are taken as input. Then each input sentences is tested to measure whether it is single line or multiple line sentences for scoring the sentences. For multiple line sentences, average sentence length is measured to generate the score. After sorting the each sentence score, the high score sentences are considered as most important for the input paragraph. Minimization is performed to get exact summary of the corresponding text to ensure high quality summarization which represent all key information through a minimum number of sentences. It finds out the central topic of the document and any important property that the topic bears.

Step 1: Take original Bangla sentences and array of original text where words are replaced with synonym and keyword from previous pass output as input in this step of the pass.

Step 2: Remove those lines from the original paragraph which does not contain any keyword. Actually this is an automatic process.

Step 3: Generate sentence score.

Step 4: In this step check each sentences whether it is single line or multiple line. This check is done comparing each sentence length with the average sentence length. If the sentence is multiple lines, find average sentence length (ASL) and then measure sentence score. If the sentence is single line it simply keeps the score of original one.

Step 5: Here all sentence score are sorted as descending order and take only high score sentences that represent the most important sentences in the given paragraph.

Step 6: Cut off size is considered and then the expected summary is produced as output.

V. Implementation

In order to develop it we have object oriented concept. Object oriented concept will assist us in layering the development process. According to our methodology this layering is extremely essential. This is implemented as an http client/server application. The summarization program is located on the server side and the client is a browser. The client interacts with the server using the HTTP protocol which helps in the accurate transfer of data from a browser and responses from the server. The architecture of the Summarization System is worked on four different passes using keyword, synonym handler, and stop word list. Here stop word list helps to eliminate unnecessary words from the keyword list. These stop words would be the most common verbs, pronouns, conjunctions, preposition and articles. The keyword of the keyword list helps to scoring the high rank sentences. Fig.3 shows the overall architecture of our Summarization System.

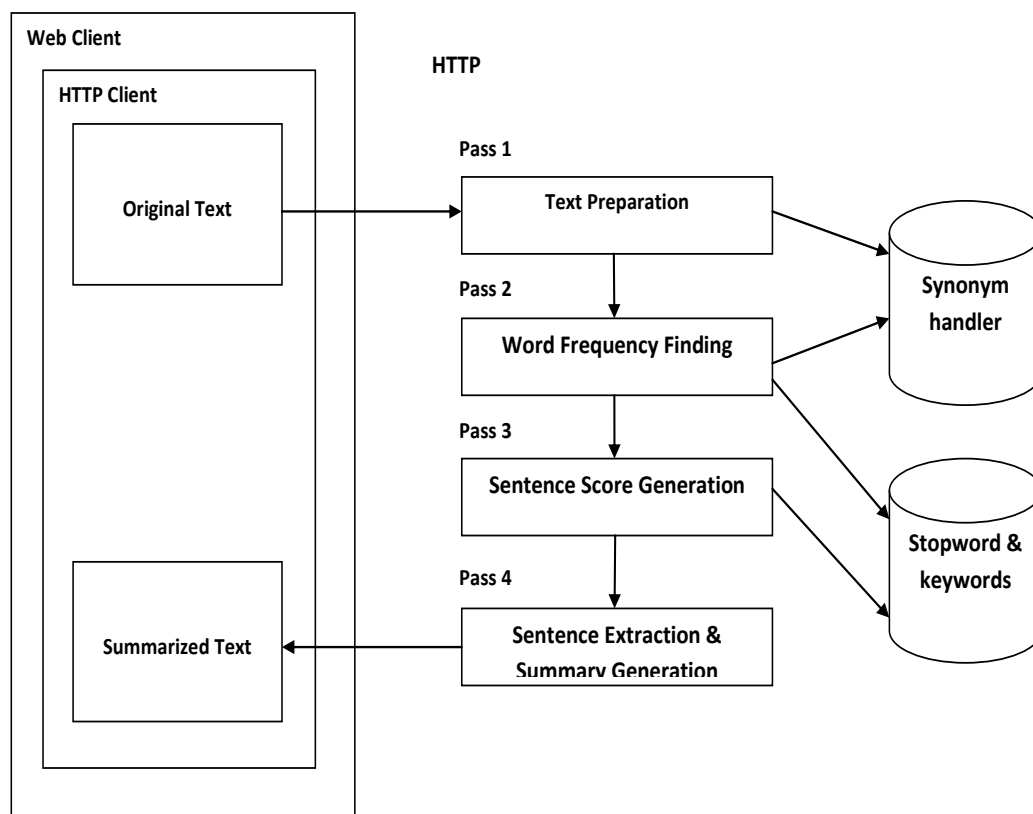


Fig. 3: Bangla Text Summarization System architecture.

5.1 Pass 1 & 2

Tokenize the hypertext and find out the original text. Then tokenize this original text to find individual lines and individual words. Using synonym handler words are replaced with their corresponding synonym. Then word frequency is found and high frequency words are assigned as keywords. Stop word list is used to eliminate unnecessary words from key word list.

5.2 Pass 3 & 4

Here each sentence is scored on the basis of how many keywords it contains and put into a ranking list. The highest-ranking sentences are then identified and kept in the final summary. A sample sentences-list produced in the third pass is shown below:-

Nr 1 Rank 12, পরিশ্রম মানুষের সৌভাগ্য ও উন্নতির নিদন।

Nr 2 Rank 7, মূলত অধ্যবসায় সৌভাগ্যের প্রসূতি।

5.3 Interface / Client Request Page

This interface enables the user to enter appropriate url of where the original Bangla text resides which is to be summarized. After entering the url appropriately in the text box client will fix percentage of how many sentences he or she wants to see in summary of the original text. There is a summarize button or hyperlink. If a client or user clicks on it after entering appropriate entry on text boxes he or she will see the output. Fig.4 shows the page where client will enter url and request the server where summarization algorithm resides through http to summarize the text.



Fig. 4: Client request page


5.4 Total Working Process

- At first the summarizer, residing at the server side will get the total hypertext or url contents and will tokenize this to remove all the html tags. This will be done to identify the original text. . Or the text container will get the bangle original text and convert it into Unicode, and then it will be done to identify the original text.
- Then this text will be taken as input to another tokenization function in order to find out the individual lines and individual word. For the input text “পরিশ্রম মানুষের সৌভাগ্য ও উন্নতির নিদন। সব কাজই পরিশ্রম সাপেক্ষ। মানসিক ও কায়িক পরিশ্রম ওতোপ্রোতভাবে জড়িত।”

Lines will be:

- পরিশ্রম মানুষের সৌভাগ্য ও উন্নতির নিদন
 - সব কাজই পরিশ্রম সাপেক্ষ
 - মানসিক ও কায়িক পরিশ্রম ওতোপ্রোতভাবে জড়িত
- Then it will be break into separate words like “পরিশ্রম”, “মানুষের সৌভাগ্য”, “ও”, “ উন্নতির”, “নিদন”
 - After finding words boundaries replacement with synonym will be done. For this connection with synonym handler is obvious.

- e. Then the frequency of keyword is counted from the input sentences on the basis of synonyms for scoring the each sentence.
- f. If there is any bold or italic text or word they will also be included in the key list along with their scores.
- g. Unnecessary words will be eliminated from key list.
- h. Sentence score will be generated on the basis of scores of the keywords that every sentence contains.
- i. High score sentences are considered as the most important sentences for the expected summary. So scoring each sentence the high sentences are taken and other low score sentences reject from the summary.
- j. Cutoff size will be considered to show the amount of summary.



An Extraction Based Automatic Text Summarizer For Bangla Web Document

Original Text in url	Sentence Scores & Rank
<p>পরিশ্রম মানুষের সৌভাগ্য ও উন্নতির নিদান। সব কাজই পরিশ্রম সাপেক্ষ। মানসিক ও কায়িক পরিশ্রম ওতোপ্রোতভাবে জড়িত। চেষ্টা ও চিন্তা, বুদ্ধি ও শক্তি এসবের সংযুক্তি ঘটিলে জগতের সব কাজই সিদ্ধ হইতে পারে। ব্যক্তিগত, সমষ্টিগত এমনকি জাতিগত সকল উন্নতির চাবিকাঠি পরিশ্রম। যে জাতি যত বেশি পরিশ্রম করতে পারে, সে জাতি ততো উন্নতি করতে পারে। পাশ্চাত্যের মানুষ যাবতীয় কাজকর্ম নিজেরাই করিয়া থাকেন। সেজন্য তাহারা জগতের বৃক্কে শীর্ষে আরোহণ করিয়াছেন। পৃথিবীতে যে সকল মহাপুরুষ সমাজ তথা দেশ ও জাতির মহাকল্যাণ সাধন করিয়াছেন তাহারা সকলেই অধিক অধ্যবসায় করতেন। মূলত অধ্যবসায় সৌভাগ্যের প্রসূতি।</p>	<p>Sentence 0 = 12 Sentence 1 = 7 Sentence 2 = 5 Sentence 3 = 5 Sentence 4 = 5 Sentence 5 = 4.30769230769 Sentence 6 = 2.35294117647</p>
<p>Summary পরিশ্রম মানুষের সৌভাগ্য ও উন্নতির নিদান। মূলত অধ্যবসায় সৌভাগ্যের প্রসূতি। সব কাজই পরিশ্রম সাপেক্ষ। মানসিক ও কায়িক পরিশ্রম ওতোপ্রোতভাবে জড়িত। ব্যক্তিগত, সমষ্টিগত এমনকি জাতিগত সকল উন্নতির চাবিকাঠি পরিশ্রম। যে জাতি যত বেশি পরিশ্রম করতে পারে, সে জাতি ততো উন্নতি করতে পারে। পৃথিবীতে যে সকল মহাপুরুষ সমাজ তথা দেশ ও জাতির মহাকল্যাণ সাধন করিয়াছেন তাহারা সকলেই অধিক অধ্যবসায় করতেন।</p>	

Fig.5: Showing original text from original text or url and resulted summary.

Fig5. Shows the output summary in the output division in output page for the corresponding input sentences in the input division. This output summary is automatically generated. Sentence rank and scores will be shown in Sentence Scores and Rank division.

5.5 Process of Converter

This converter works on to convert bangla text to Unicode. When here we write any text and submit then it will be converted into Unicode. This converted Unicode pass in the process where this text will be taken as input to another tokenization function in order to find out the individual lines and individual word. For the input text “পরিশ্রম মানুষের সৌভাগ্য ও উন্নতির নিদান। সব কাজই পরিশ্রম সাপেক্ষ। মানসিক ও কায়িক পরিশ্রম ওতোপ্রোতভাবে জড়িত।”

- Lines will be:
- i. পরিশ্রম মানুষের সৌভাগ্য ও উন্নতির নিদান
 - ii. সব কাজই পরিশ্রম সাপেক্ষ
 - iii. মানসিক ও কায়িক পরিশ্রম ওতোপ্রোতভাবে জড়িত

Then it will be break into separate words like “পরিশ্রম”, “মানুষের সৌভাগ্য”, “ও”, “উন্নতির”, “নিদান”,

After finding words boundaries replacement with synonym will be done. For this connection with synonym handler is obvious. Then the frequency of keyword is counted from the input sentences on the basis of synonyms for scoring the each sentence. If there is any bold or italic text or word they will also be included in the key list along with their scores. Unnecessary words will be eliminated from key list. Sentence score will be generated on the basis of scores of the keywords that every sentence contains. High score sentences are considered as the most important sentences for the expected summary. So scoring each sentence the high sentences are taken and other low score sentences reject from the summary. Cutoff size will be considered to show the amount of summary.

VI. Data Collection

Stop word list is a kind of word storage where unimportant words [4] that will not be included in the key words list though they occur frequently in text. These words are more frequent words [5] in bangla text. These words can't be treated as keywords.

Table 1: Stop word list in BanglaSum

Type	Word
Pronoun	সে, তার, তাহার, তাঁদের, তাহাদের, তিনি
Verb	অনুমান করা, প্রস্তাব করা
Article	একটি, টি, টা
Preposition	ভিতরে, উপরে, সাথে, এর, দিকে
Unimportant Words	যেমন, উদাহরণ, কথার কথা, উপরন্তু, অন্যদিকে, ও
Useful wordlist	এই বিষয়ে, আলোচ্য অংশে, পরিশেষে, উপসংহারে, প্রধান, মূলত, আসল, উল্লেখ্য, সংক্ষেপে
Important wordlist	উদ্দেশ্য, লক্ষ্য, বিষয় বস্তু, বস্তুত, বাস্তবিক, কারন, ফলাফল, আলোচনা, যৌক্তিক

6.1 Roles of Lexicon

A lexicon is a dictionary [3] of word where each word contains some syntactic, semantic and possibly some pragmatic information. This information in the lexicon is needed to help for determining the function and meaning of the word which is appended in the sentence. Each entry in a lexicon will contain a root form of the word. The inflected forms or derivation are obtained by the morphological analyzer.

Table 2: Typical Entries in Bangla Lexicon

Word	Type	Feature
আমি	Pronoun	1Pers, SG, Human, Personal
সে, তিনি	Pronoun	3Pers,SG,Human
মেয়ে, ছেলে	Noun	3Pers, SG, Human, Proper Noun
তুমি, তুই	Pronoun	2Pers, SG, Human
আপনি	Pronoun	2Pers,SG,Human
ভাল	Adjective	Qualitative, positive
মন্দ	Adjective	Qualitative, Negative
জাতি	Noun	3pers, Activity, collective
ধীরে	Adverb	Manner
বিদ্বান	Adjective	3pers,SG,human,personal

6.2 Synonym Handler

Synonym handler is kind of word storage database system. It can be treated as synonym dictionary. Words of the same meaning are reserved here for replacement. This contains two fields one called main word another is synonym. In synonym field words of same meaning of main words are organized separated by comma (,).

Table 3: Synonym Handler

Main word	synonym
জ্ঞানার্জন	জ্ঞানার্জন, জ্ঞান, শিক্ষা, বিদ্যার্জন
মানুষ	ব্যক্তি, মানব
পরিশ্রম	অধ্যবসায়
সৌভাগ্য	উন্নতি, শীর্ষে
মেরুদন্ড	অপরিহার্য অঙ্গ
সমষ্টিগত	জাতিগত
চরিত্র	চরিত্র, মদাচার
শ্রেষ্ঠত্ব	গৌরবের, ঐশ্বর্য
স্বাধীনতা	স্বাধীন, মুক্ত
অর্জন	লাভ, অর্জিত

VII. Experimental Results And Discussion

In order to prove that our project work meets the requirements of successful working, we collect some meaningful experimental data. These are some meaningful paragraph, from a very renowned Bangla grammar book [11]. From these meaningful paragraphs we decorate our synonym handler and find out the keywords. We have implemented our experiment data using php programming or script [12].

Experimental data are represented in this section from which experimental result has been generated. Every step of how data has been processed is shown here. We represented data or bangla characters using utf-8 character encoding. Here, is the page content where paragraph is embedded with html tags. Obviously, bangla paragraph are in Unicode format. The overall accuracy of the system is 68.29% and F1 score is 61.32%.

Corresponding content with html tag:

```
<html>
<head></head>
<body>
<p> <b>পরিশ্রম</b> মানুষের সৌভাগ্য ও উন্নতির নিদন। সব কাজই পরিশ্রম সাপেক্ষ। মানসিক ও কায়িক পরিশ্রম
ওতোপ্রোতভাবে জড়িত। চেষ্টা ও চিন্তা, বুদ্ধি ও শক্তি এসবের সংযুক্তি ঘটিলে জগতের সব কাজই সিদ্ধ হইতে পারে।
ব্যক্তিগত, সমষ্টিগত এমনকি জাতিগত সকল উন্নতির চাবিকাঠি পরিশ্রম। যে জাতি যত বেশি পরিশ্রম করতে পারে, সে
জাতি ততো উন্নতি করতে পারে। পাশ্চাত্যের মানুষ যাবতীয় কাজকর্ম নিজেরাই করিয়া থাকেন। সেজন্য তাহারা
জগতের বৃকে শীর্ষে আরোহন করিয়াছেন। পৃথিবীতে যে সকল মহাপুরুষ সমাজ তথা দেশ ও জাতির মহাকল্যান সাধন
করিয়াছেন তাহারা সকলেই অধিক অধ্যবসায় করতেন। <b>মূলত</b> অধ্যবসায় <b>সৌভাগ্যের</b> প্রসূতি।</p>
</body>
</html>
```

7.1 Pass 1 & Pass 2

7.1.1 Tokenization to Find Original Text

First a page content or hypertext is taken as input. Then we tokenize this to remove html and other tags and to find the original text where we will perform required processing.

Original text:

পরিশ্রম মানুষের সৌভাগ্য ও উন্নতির নিদন। সব কাজই পরিশ্রম সাপেক্ষ। মানসিক ও কায়িক পরিশ্রম ওতোপ্রোতভাবে জড়িত। চেষ্টা ও চিন্তা, বুদ্ধি ও শক্তি এসবের সংযুক্তি ঘটিলে জগতের সব কাজই সিদ্ধ হইতে পারে। ব্যক্তিগত, সমষ্টিগত এমনকি জাতিগত সকল উন্নতির চাবিকাঠি পরিশ্রম। যে জাতি যত বেশি পরিশ্রম করতে পারে, সে জাতি ততো উন্নতি করতে পারে। পাশ্চাত্যের মানুষ যাবতীয় কাজকর্ম নিজেরাই করিয়া থাকেন। সেজন্য তাহারা জগতের বৃকে শীর্ষে আরোহন করিয়াছেন। পৃথিবীতে যে সকল মহাপুরুষ সমাজ তথা দেশ ও জাতির মহাকল্যান সাধন করিয়াছেন তাহারা সকলেই অধিক অধ্যবসায় করতেন। মূলত অধ্যবসায় সৌভাগ্যের প্রসূতি।

7.1.2 Tokenization to Find Individual Line & Word

Example of Finding Lines:

“পরিশ্রম মানুষের সৌভাগ্য ও উন্নতির নিদন”, “সব কাজই পরিশ্রম সাপেক্ষ”, “মানসিক ও কায়িক পরিশ্রম ওতোপ্রোতভাবে জড়িতজড়িত”, “চেষ্টা ও চিন্তা, বুদ্ধি ও শক্তি এসবের সংযুক্তি ঘটিলে জগতের সব কাজই সিদ্ধ হইতে পারে”, “ব্যক্তিগত, সমষ্টিগত এমনকি জাতিগত সকল উন্নতির চাবিকাঠি পরিশ্রম”

Example of Finding Words:

“পরিশ্রম”, “মানুষের”, “সৌভাগ্য”, “ও”, “উন্নতির”, “নিদন”, “সব”, “কাজই”, “পরিশ্রম”, “সাপেক্ষ”, “মানসিক”, “ও”, “কায়িক”, “পরিশ্রম”, “ওতোপ্রোতভাবে”, “জড়িতজড়িত”, “চেষ্টা”, “ও”, “চিন্তা”, “বুদ্ধি”, “ও”, “শক্তি”, “এসবের”, “সংযুক্তি”, “ঘটিলে”, “জগতের”, “সব”, “কাজই”, “সিদ্ধ”, “হইতে”, “পারে”, “ব্যক্তিগত”, “সমষ্টিগত”, “এমনকি”, “জাতিগত”, “সকল”, “উন্নতির”, “চাবিকাঠি”, “পরিশ্রম”

7.1.3 Lexical knowledge for Synonym

For the “Summarization System”, knowledge about possible words and synonym in the input sentence is stored in the synonym handler. In the table 4 there is a simple synonym handler for “Summarization System”. Words found in the input sentences are stored in the ‘Main Word’ column and ‘Synonyms’ column store the synonyms of the main word.

Table 4: Words in the synonym handler

Main Word	Synonym
পরিশ্রম	অধ্যবসায়
সৌভাগ্য	উন্নতি, শীর্ষে
জ্ঞানার্জন	জ্ঞানার্জন, জ্ঞান, শিক্ষা, বিদ্যার্জন
মানুষ	ব্যক্তি, মানব

7.1.4 Frequency Count and Keyword Finding

Keywords are the most frequent words in the text. Sentences containing keywords scored higher than the ones with fewer keywords. Table 5 shows the frequency count of the keyword with synonyms for Bangla input sentences.

Table 5: Keyword Frequency with Synonym

Keyword	Frequency
পরিশ্রম, অধ্যবসায়	6
সৌভাগ্য, উন্নতি, শীর্ষে	4
সমষ্টিগত, জাতিগত	2

7.1.5 Assigning Scores to Keywords and Bold Words

In order to assign scores to words we test if frequency is greater than 2 times. Scores are multiplied with the frequency count of key words.

Table 6. Scores of Keywords

Keywords	Scores
পরিশ্রম	5
সৌভাগ্য	2
মূলত	2

7.2 Pass 3 and Pass 4

7.2.1 Sentence Scoring

Scoring of sentences is used to decide on the importance of each line in the document. Here, we set a line as single if its length is less or equal to average sentence length otherwise the sentence will be treated as multiple lines. The sentence score depends on the 'word score' which is the multiplication of 'word frequency' and a default first line constant value and final sentence score is the summation of 'word score'.

Word Score= (word frequency) *(a keyword constant) + other criterion

Sentence Score= \sum word score + other criterion

Sentence 1: পরিশ্রম মানুষের সৌভাগ্য ও উন্নতির নিদান।

As this is the first line default value 8 is set to it.

Sentence score 1: $8 + 5 + 2 + 2 = 17$

Sentence 2: সব কাজই পরিশ্রম সাপেক্ষ।

Sentence score 2: 5

Sentence 3: মানসিক ও কায়িক পরিশ্রম ওতোপ্রোতভাবে জড়িত।

Sentence Score 3: 5

Sentence 4: চেষ্টা ও চিন্তা, বুদ্ধি ও শক্তি এসবের সংযুক্তি ঘটিলে জগতের সব কাজই সিদ্ধ হইতে পারে।

Sentence Score 4: 0

Sentence 5: ব্যক্তিগত, সমষ্টিগত এমনকি জাতিগত সকল উন্নতির চাবিকাঠি পরিশ্রম

Sentence Score 5: 5

Sentence 6: যে জাতি যত বেশি পরিশ্রম করতে পারে, সে জাতি ততো উন্নতি করতে পারে।

Sentence Score 6: $(5+2)*8/13 = 4.30769230769$ (considering average sentence length)

Sentence 7: পাশ্চাত্যের মানুষ যাবতীয় কাজকর্ম নিজেরাই করিয়া থাকেন।

Sentence Score 7: 0

Sentence 8: সেজন্য তাহারা জগতের বুকে শীর্ষে আরোহন করিয়াছেন।

Sentence Score 8: 2

Sentence 9: পৃথিবীতে যে সকল মহাপুরুষ সমাজ তথা দেশ ও জাতির মহাকল্যান সাধন করিয়াছেন তাহারা সকলেই অধিক অধ্যবসায় করতেন।

Sentence Score 9: $5 \times 8 / 17 = 2.35294117647$ (considering average sentence length)

Sentence 10: মূলত অধ্যবসায় সৌভাগ্যের প্রসূতি।

Sentence Score 10: $2 + 5 = 7$

7.2.2 Average Sentence Length (ASL)

To avoid an inaccurately awarding long sentence with a high ranking, the sentence score is multiplied by the average sentence length (ASL) and divided by the number of words in the current sentence to normalize for length.

Word-count = number of words in the text.

Line-count = number of lines in the text.

Average sentence length (ASL) = Word-count / Line-count.

Sentence score = (ASL * Sentence score) / (number of words in the current sentence)

Word-count = 80.

Line-count = 10.

ASL = Word-count / Line count = $80 / 10 = 8$

Sentence 6: যে জাতি যত বেশি পরিশ্রম করতে পারে, সে জাতি ততো উন্নতি করতে পারে।

Number of words in the current sentence = 13

Sentence Score 6: $(5+2) \times 8 / 13 = 4.30769230769$

Sentence 9: পৃথিবীতে যে সকল মহাপুরুষ সমাজ তথা দেশ ও জাতির মহাকল্যান সাধন করিয়াছেন তাহারা সকলেই অধিক অধ্যবসায় করতেন।

Number of words in the current sentence = 17

Sentence Score 9: $5 \times 8 / 17 = 2.35294117647$

7.2.3 Most Important Sentences in the Text

After sorting the sentence score most important sentences are considered which score is high rank. Table 6 shows the most important sentences and its corresponding scores.

Table 7: Most important sentences in the paragraph

Sentences	Scores
পরিশ্রম মানুষের সৌভাগ্য ও উন্নতির নিদান।	12
মূলত অধ্যবসায় সৌভাগ্যের প্রসূতি।	7
সব কাজই পরিশ্রম সাপেক্ষ।	5
মানসিক ও কায়িক পরিশ্রম ওতোপ্রোতভাবে জড়িত।	5
ব্যক্তিগত, সমষ্টিগত এমনকি জাতিগত সকল উন্নতির চাবিকাঠি পরিশ্রম।	5
যে জাতি যত বেশি পরিশ্রম করতে পারে, সে জাতি ততো উন্নতি করতে পারে।	4.3076923

7.2.4 Summary Generation

After finding the most important sentences cutoff size is brought into account. Depending on this size numbers of lines are extracted to the final summary. So, the expected summary is:-

পরিশ্রম মানুষের সৌভাগ্য ও উন্নতির নিদান। মূলত অধ্যবসায় সৌভাগ্যের প্রসূতি। সব কাজই পরিশ্রম সাপেক্ষ। মানসিক ও কায়িক পরিশ্রম ওতোপ্রোতভাবে জড়িত। ব্যক্তিগত, সমষ্টিগত এমনকি জাতিগত সকল উন্নতির চাবিকাঠি পরিশ্রম। যে জাতি যত বেশি পরিশ্রম করতে পারে, সে জাতি ততো উন্নতি করতে পারে।

VIII. Conclusion

This work shows how processors of Bangla natural language apply several techniques to summarize Bangla sentences. It reads the input texts, processes the sentences, and generates an appropriate summary by ranking high score sentences that find out the focal topic of the document and any important property the topic bears. This summarization system is not comparable with human summarization. As this is an extraction based summarization system it doesn't generate summary the same way as humans do; it only generates summary that appear like human summary. As our summarization system is a web based application, we think it will play an important role for future development.

References

- [1] Joel Larocca Neto, Alexandre D. Santos, Celso A.A. Kaestner, Alex A. Freitas, "Document Clustering and Text Summarization"
- [2] M. Sanderson, "Accurate user directed summarization from existing tools", in Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM98), 1998.
- [3] Md Tawhidul Islam, Shaikh Mostafa Al Masum, "An algorithm for automatic text summarization of technical documents by building and minimizing information network of the source document" (ICCIT '05).
- [4] Md Tawhidul Islam, Shaikh Mostafa Al Masum, "Bhasa: A Corpus-Based Information Retrieval and Summariser for Bengali Text", International Conference on Computer and Information Technology" (ICCIT '05).
- [5] Hercules Dalanis, "SweSum - A Text Summarizer for Swedish" <http://www.nada.kth.se/nada/iplab/>
- [6] Martin Hassel, Nima Mazdak, "FarsiSum - A Persian text summarizer"
- [7] <http://www.nada.kth.se/iplab/hlt/farsisum/index-farsi.html>.
- [8] D. W. Patterson, "Introduction to Artificial Intelligence and Expert System", Printice Hall, India, 2002
- [9] M M Asaduzzaman & Muhammad Masroor Ali, "Transfer Machine Translation – An Experience with Bangla English Machine Translation System", ICCIT- 2003, pp (265-270).
- [10] Alferd V. Aho, R.Sethi and J D. Ullman, "Compilers Principles, Techniques and Tools", Second edition, Page 95-98, 171-174.
- [11] James Ignizio, "Introduction to Expert Systems" (1991), ISBN 0-07-909785-5.
- [12] ডঃ গোলাম সাকলায়েন, "একের ভিতরে পাঁচ" দ্বাবিংশ সংস্করণ।
- [13] Wrox - Beginning PHP5 Apache and MySQL Web Development
- [14] Massih R. Amini, Nicolas Usunier, and Patrick Gallinari, "Automatic Text Summarization Based on Word-Clusters and Ranking Algorithms"
- [15] Kaili Müürisep, Pilleriin Mutso "Estsum – Estinian Newspaper Text Summerizer"

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Mohammed Mahmudur Rahman. " Analysis, Design and Development of an Automatic Text Summarizer for Bangla Web Document" IOSR Journal of Computer Engineering (IOSR-JCE) 21.1 (2019): 13-25.