A Heuristic Approach of Text Summarization for Bengali Documentation

Conference Paper · July 2017

DOI: 10.1109/ICCCNT.2017.8204166

Sheikh Abujar
Daffodil International University
52 PUBLICATIONS **83** CITATIONS
SEE PROFILE

Mahmudul Hasan
Saitama University
33 PUBLICATIONS **71** CITATIONS
SEE PROFILE

M.s.I Shahin
Jahangirnagar University
4 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Syed Akhter Hossain
Daffodil International University
99 PUBLICATIONS **476** CITATIONS
SEE PROFILE

A Heuristic Approach of Text Summarization for Bengali Documentation

Sheikh Abujar
Dept. of CSE
Jahangirnagar University
Savar, Dhaka,Bangladesh
sheikhabujar@gmail.com

Mahmudul Hasan
Dept. of CSE
Comilla University
Comilla, Bangladesh
mhasanraju@gmail.com

M.S.I Shahin
Dept. of CSE
Jahangirnagar University
Savar, Dhaka, Bangladesh
msi.shahin71@gmail.com

Syed Akhter Hossain
Dept. of CSE
Daffodil International University
Dhanomondi, Dhaka, Bangladesh
aktarhosaain@daffodilvarsity.edu.bd

***Abstract*— Automated Text Summarization is a technique of summarizing any document or text automatically. Summarized text is the concise form of the given text. In Natural language processing many text summarization techniques are available for English language, but only a few for Bangla language. Bangla is one of the most taught and used language all over the world. Most of the text summarization techniques are implemented in two different ways, known as abstractive or extractive approach. This paper deal with the summarization of Bangla text based on extractive method. A new efficient extractive summarization method is proposed in this work. The other summarization tools developed for Bangla language seems not much appropriate from application point of view. The proposed analysis models are applicable for Bangla text summarization. In the proposed approach, basic extractive summarization is applied with new proposed model and a set of Bangla text analysis rules derived from the heuristics. Every Bangla sentences and words from original text is analyzed properly with Bangla sentence clustering method. This work proposed a new type of sentence scoring processes for Bangla text summarization. In the evaluation of this technique, the system reflects good accuracy of results, comparing to that of the human generated summarized result and other Bangla text summarization tools.**

Keywords—Bangla Text Summarization, sentence scoring, Sentence Analysis, Language Processing, NLP.

1. INTRODUCTION

A massive increase of information is a part of our life today. Information over Internet and offline are increasing so rapidly even more than the amount of printed data. Those electronic information like- news portal, blogs, e-books, etc., is very much difficult to summarize as the amount of information is large. In order to find out the essence of these information is very much pains-taking, nevertheless it is not feasible to sieve useful information from these large amount of data coming from any source of documents. The only way is to summarize these data is through automated text summarization process. This will classify all the data together and present the succinct information clearly. It will have to maintain a standard summarization procedure everywhere and more importantly it will help saving time and efforts both.

Text summarization is the process of automatically preparing a concise statement of any given text. English text summarization had a revolutionary research output and currently these developed summarizers' works accurately. In Bengali language, this research is not up to the mark and as a result there is no satisfactory summarizer at all which can be applied in Bangla text processing despite the content and users. A Bengali text summarizer has become now indispensable and very demanding. A large number of Internet, official or personal users will be benefited by using this Bangla text summarizer.

Summarization process could be done in two different ways: abstractive and extractive. The extractive approach find out the most used words and then score sentences from different perspective. In other words abstractive summarization clarify the contents and then improve the coherence among sentences by eliminating redundancies [1]. In extractive summary it will not add any other additional words or sentence into the summarized paragraph, however the abstractive summarize process may add new sentences into the summary. Therefore abstractive summarization is more difficult than extractive approach, however precise result will come after implementing abstractive method. Thus the use of extractive methods are much as it is easy to implement and except some cases, it works perfectly.

Extractive method follow few rules, initially it represents the whole document and separate into paragraphs, sentences and words respectively. Initially extractive method find out and removes the stop words. Now the document is ready to score sentences and then complete the summarization by selecting higher scored sentences [2]. In this research it has explained, how the Bengali sentences should be scored after removing stop words. Advanced sentence scoring methods are proposed here which will help to score sentences precisely.

In the text summarization a text document can be concise and the basic ideas of the topic also be realized whether it is relevant or not. Multiple news reports can be summarized and able to find out the relationships between those. Any trending topics can be show through graphical representation without analyzing all those information individually. Text summarization can be done for a single file or a set of files form different sources.

The rest of the paper is explained as follows: In section II, literature review is discussed, which contains the previous research summary of Bengali Text summarization done by other researchers. In section III, narrated proposed new extractive approach [14] [15] for Bangla text with quantitative assessments is discussed. Lastly, section IV and V details about the experimental results with discussion and conclusion respectively.

2. LITERATURE REVIEW

Extractive text summarization is basically formed in three different phases. Measuring the intermediate orientation of originl text is the exigent part of Text analysis. Scoring every sentences and finally consolidating those high scored sentences will produce a better extractive summary[1]. This section describes previous works, related to those phases and state of arts research.

Rafel, et al. explained all the basic requirements of extractive research approaches including those three features, are - Text analysis, sentence scoring and summarizing [1]. Here fifteen (15) different sentence scoring methods were explained. Those methods were assessed on news, blogs and different articles. Every word was scored in six different ways. In that case, word co-occurrences were analyzed using n-gram based process [8] and lexical similarities were identified. Sentence scoring methods rendered new features like, sentence centrality based on sentence similarity algorithm [9]. Aggregate similarities between sentences were used widely while evaluating performances. These were narrated in possible ways for improving sentence score results. In the work, possible reasons of occurring polysemy were mentioned.

Harsha, et al. commenced a hybrid summarizing technique for multi-text documents including all those basic requirements. The authors suggested for generating word graph and word net [4]. Through word graph, important nodes could be identified and for generating word graph, they have used heuristic rules. Word net is basically a lexical data dictionary, through which - words meaning and models could be provided. Different semantic meanings help to identify category of words. Vocabulary and set of synset will be provided through domain ontology [10].

Iftekharul et al. summarized Bangla text by sentence scoring and ranking [5]. Necessary text preprocessing measures for Bangla were discussed separately. A lightweight stemmer was introduced purpose of identifying canonical forms of words. Identifying various cue words from Bangla text and apprehend the skeleton of the document from title and header, improves the summary quality better perspective of Bangla text summarization. Here the performance of the summarization [11] is not up to the mark.

Bangla texts couldn't be analyzed like other language. Because grammatical rules and sentence patterns are very different here. The reason behind choosing a modified extractive approach for summarizing Bangla text, was that. Several approaches applied here, were already implemented in different extractive text summarization models. Though those are basic methods of extractive summarization. Such as sentence scoring, identifying word frequency, Sentence position, etc. Here, few new methods were introduced, such as repeated word distance, absolute deviation of sentences, frequent words percentile – which will be calculated by the rate of frequent words carried by each sentence as well as being used in other sentences, prime sentences identification – which will help to identify the most important sentences based on all above analysis. This was a hypothesis and found very good result after implementation. Other research were stated several methods for scoring sentences and words, though the individual sentences may have values based on the other related words being used in somewhere else on that document or being redirected to that sentence though some imitating words. Several times, same words were being used in different sentences in different forms. Either those sentences are internally linked or similar in semantic meanings. The relationship between those sentences were not identified for Bangla language yet. Similar sentences may not took position sequentially or in very near position. So standard deviation of sentences will help to identify the positional difference between sentences. From there, dependent and optimal sentences could be identified. Identifying Cue words or leading words, had already implemented in other languages, though it have several techniques. Here a method was proposed which found suitable for Bangla language. Several known topics were introduced here, in different way for Bangla language. Summarizer of English language or other language will not work for Bangla language, though the topics are very similar to cover. And few new topics were introduced as per required or necessity for the betterment of getting optimal solution. To overcome those limitations, a model is being proposed and discussed the output factors. And overall, this model provides optimal solutions for Bangla Documents.

3. PROPOSED METHOD

Bengali text summarization is an application of Natural language processing. This processing will be responsible for summarizing any Bengali text document and intelligently propose a summary. The most popular approaches for text summarization is Extractive summarization approach. Many researchers contributed in this area primarily for English language. The prime factor of this research area could be addressed by the following research question: How to identify those sentences, which contain the main gist of the given text? In general, there are three different approaches: (i) Word Scoring – evaluate every words to identify the most frequent and important words from the text; (ii) Sentence scoring – to identify relations among sentence and figure out the main leading sentences. In addition sentence position and redundancy detection will help to generate a precise summarization; and (iii) Graph scoring – analyze the relationship between words and sentences [1]. In addition distance between similar words and sentences will help to identify the uniform representation of sentences. Unnecessary sentences could be avoided through our proposed technique of words and sentences processing.

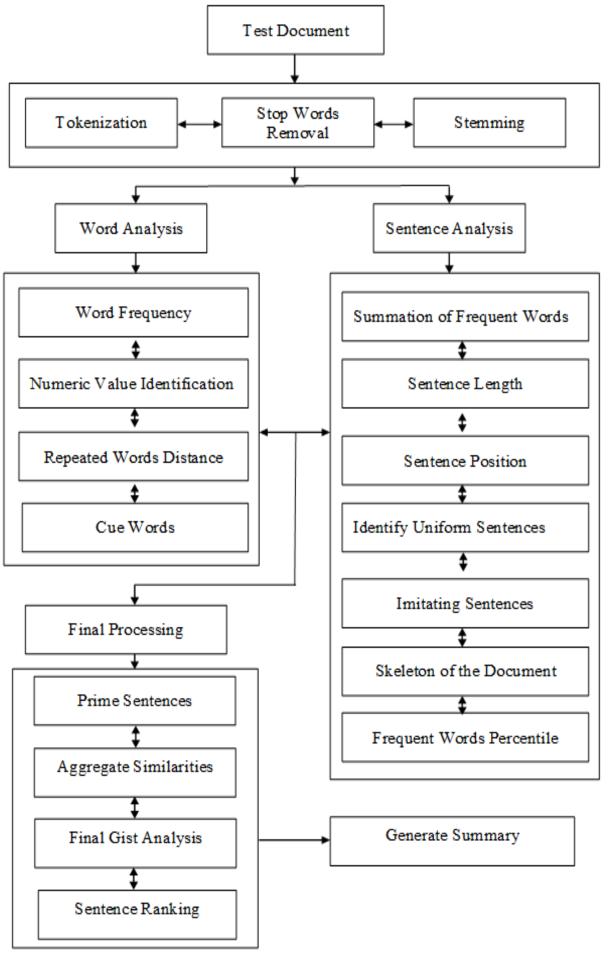


Fig. 1. Steps of Proposed Text summarization technique

The proposed Bangla text summarization technique will follow three different phases as shown in figure 1. The steps are: (i) preprocessing with Linguistic Analysis–Extract sentences from the text document and tokenize all those sentences with different segmentation process. Tokenization disjoins sentences into words, numbers, sentences and symbols, (ii) Prime sentences identification– the main leading sentences will be extracted from original set of documents through word analysis and sentence analysis process. Because words and sentence, both factors are equally important for preparing a qualitative summary and (iii) final processing – All those prime sentences will be evaluated again to increase the possibilities of representing accurate sentences as summarized output. Throughout our proposed rules and models, final processing features will generate a better quality summary from Bangla text. Detail of every part of our proposed model is explained in rest of the portion as follows. Entire Design of our proposed model is given in figure 1.

3.1 Preprocessing with Linguistic Analysis

Initially the document must be processed at least once before summarization. In Bengali language, sentences may not associated in any order. Before ranking those sentences or words individually, it is important to prepare those sentences to get better and accurate sentence score. Scoring methods was divided in two different segments - sentence scoring and word scoring. Both scores will help to identify most leading sentences as well most important keywords of the topic. It is being used to represent word graph.

Two different methods – Word Analysis and Sentence Analysis are used to score those sentences more accurately. Before that, the Linguistic Analysis part will be completed. Linguistic Analysis will be done though tokenizing every sentence, by removing stop words and finally stemming words.

3.1.1 Bangla Text Tokenization

Text summarization or document summarization is very similar, because ultimately it have to summarize many sentences, and sentences are combination of words. Sentences could be found in structured or non-structured form. For summarizing any text or document, it is important to identify the essence of those sentences. So, it is important to understand every words form every sentences. In that case, tokenization will separate those words from sentences and prepare a data set for further analysis. Analyzing every words individually will increase the possibilities to collect set of prime sentences for further Analysis. Tokenization will segment all those token words, for example – numbers, symbols, etc. A sentence is a set of tokens and important sentences always may not contain maximum high frequency words. Rather it may contain important information, such as – important dates, time, specific amount for any data, etc. For that findings, it is necessary to analyze those tokens individually and prepare a precise output set of prime sentences.

3.1.2 *Stop words removal*

Similar sentences could be joined though few conjunctive words. This kind of words will be eliminated in classification process. In Bengali language, words like এবং (And), কিন্তু (But), etc. are being used widely in sentences. For example: “বাংলাদেশ বনাম ভারতের ক্রিকেট ম্যাচ এর উদ্দেশ্যে বাংলাদেশ ক্রিকেট টিম আজ ভারত সফরে রওনা হয়েছে কিন্তু আগামীকাল ধর্মঘট এর কারনে খেলা স্থগিত করা হয়েছে”. Here two different sentences were introduced though the information category is same. The word –কিন্তু was used to add two different sentences together, only because of that information are interrelated. But the meaning of those sentences are not similar but dependent. Those two sentences are equally capable to become a prime sentence individually. So, here scoring process will start after eliminating stop words.

3.1.3 *Stemming*

Bengali is morphologically rich language [3]. Same words could be represented in different lexicon order. But the meaning of those words will not be changed a lot, because the root of those words are same. Through these root words, the relationship between sentences could assume easily. Word scoring is an essential part of extractive summarization approach, so if those words considered as different token then it would be difficult to found the similar words and set of similar tokens indeed. Similar tokens will help to find out the amount of detached words and provide a set of similar words. Through this, it is possible to trace the relationship among sentences and finally avoid sentence redundancy. Stemming will work like, consider there are few words as – খেলা, খেলি, খেলেছি, খেলব, খেলতে, খেলিয়াছি, খেলা ধুলা, etc. All these words are similar but just in different form of word – খেলা (Play). Stemming has three different phases like Root, Surface form and suffixes [3]. Table 1 represents the different segments of stemming analysis example.

Table1: Stemming segments

Root	Surface form	Suffixes
খেলা	খেলি, খেলো, খেলবো	◌ি , ◌ো, বো
বাংলাদেশ	বাংলাদেশের, বাংলাদেশকে	◌ের, কে
গাড়ি	গাড়িতে, গাড়ির, গাড়িটি	তে, র, টি

Suffixes are the additional parts of root word form. All these surface forms will be considered as a stemming cluster, which will be send for scoring word frequency. All those sentences containing same surface form may count as similar sentences and for final processing it will provide a set of similar sentence token. From where prime sentences will be picked for further analysis.

3.1.4 Word Analysis

Sentences consist of set of words. Words have the information of entire sentence. All words together represents an information. The analysis of words is more important to find out the words graph and words frequency and many other important parts, describe in rest of this section.

(a) Word Frequency – Words could be used redundantly. And the topic related words may use in maximum sentences. The number of every times a word is used in total text, is called word frequency. The stemming cluster tokens will be considered as a same rooted word and frequency for all those words will be added in the same root token. For example, Table 2 is representing a word frequency (after stemming) of a given text. Based on only this word frequency scores, is it possible to find out the topic, subject or context of the given text?

Table 2: Word frequency after stemming

Word	Frequency
ক্যামেরুন	4
ট্রেন	5
দেশ	3
নিহত	2
যাত্রী	3
৩০০	2
লাইনচ্যুত	2

The given word frequency scores are being generated after stemming, and the maximum frequency of words are listed here. The input text detail is given below in Table 3.

Table 3: Input text details

Source:	http://www.prothom-alo.com/international/article/1005109
Category of text:	General News report
Title of text:	ক্যামেরুনে ট্রেন লাইনচ্যুত হয়ে নিহত ৫৩
Total sentences :	8
Total words:	106

The given input text is a very small news article of 8 sentences, as well as 106 words. The maximum frequent words are: ক্যামেরুন, ট্রেন, যাত্রী, etc. All these are leading words of that given news text. Then the sentences those are containing these words are also leading sentences. And the Topic identification process for cross matching the leading sentences would be easier and accurate.

(b) Numeric value identification–Numeric values are always important. It holds significant information about the given text. Numeric values may address dates, years, any unit of amount, etc. In every text, numeric values could be found and those addressed precise information. The sentences containing numeric values will be considered as a high priority token for preliminary prime sentence clustering segment.

(c) Repeated Words distance – Similar words rarely used twice in a sentence. In a passage similar or related sentences could be found in contiguous tokens. From the sentence analogy matrix, simply it is possible to omit many redundant sentences. It will not be applicable if the distance of the words is very far. This calculation will be based on the number of total paragraph in the given text or the total length of entire text. Through repeated word distance analysis, some values will be collected for analyzing important words effect rate. The words effect rate is explained in Equation (1) as follows:

$$E_r = \left(\frac{\sum P_i}{\sum P_j} + \frac{\sum S_i}{\sum S_j}\right) + R_k \quad --- (1)\ Where, 1 < E_r < 5$$

The Words Effect rate will be calculated individually for every frequent word and will be recorded into word dataset. Example of effect rate sample is mentioned in Table 4. Here Repeating Nature is calculated as Low=1, Average=2 and High = 3.

Table 4: Effect rate of words considering repeated distance

Variable	Value
Word :	ট্রেন
Word frequency :	5
Number of total paragraph (P_j):	5
Word used in number of total paragraph (P_i):	4
Number of total sentences (S_j):	8
Number of total sentences used (S_i):	5
Repeating nature (R_k):	High
Maximum used in paragraph no:	2
Effect rate (E_r):	4.43

Here after calculation, The Effect rate of the word “ট্রেন” is 4.43.

(d) Cue Words –In Bengali language, information could be expressed by using more than one sentence. Semantic relation could be found in between of those linked sentences. Few semantic relation emphasize the summary or gist of those group of sentences. So these sentences could be accepted as a prime sentence and as representative of those groups of sentences. Words like – যেহেতু(Since), মোটকথা(In a Word),এছাড়াও(Also),জরুরী(Emergency),পরিশেষে(Afterward),অতঃপর (Hence) ,ইতিমধ্যে(Already)etc. are cue words in Bengali language. This words redirects towards the main leading sentence.

3.1.5 Sentence Analysis

Sentence Scoring is one of the best approaches to determining leading sentences and representative content. Sentences will be evaluated in many forms and the word analysis factors will help to identify the target sentences as well. Several steps of sentence analysis are discussed in rest of the part.

(a) Summation of frequent words – Set of frequent words in a sentence will be grouped together, and the total weighted values of those words will be calculated sentence by sentence. Maximum rate of weighted sentences will be considered as a representative sentence.

(b) Sentence Length – Amount of words in a sentence is considered as sentence length. Many small size sentences may achieve maximum weight in both word and sentence analysis. But the information in that sentence may not enough to represent the summation of a scenario. So, calculating mean value of those Sentence length will give a minimum acceptance rate for prime sentence, as well help to identify the Sentence position.

(c) Sentence Position– Position of sentences has very significant influence over the given content of documents. Different paragraph of input document may contain different types of information. First and last sentence of the document represents meaningful and significant information as well as for paragraphs. So, these two sentences will be considered as a prime sentence in extractive summarization approach, though the sentence frequency is low. Other sentences will be calculated through Gaussian distribution theorem. Sentence will be plotted in ascending order based on their Sentence length and as the absolute deviation factor, which will be calculated sequentially. It will help to omit less weighted sentences. The Absolute Deviation Factor (AD) is explained in Equation (2) as follows:

$$AD = \left|\left(\frac{|x-\mu|}{\mu}\right) - 1\right| - - - (2)$$

Where, x= is Sentence Length (which sentence to identified); μ= is the mean value of all Sentence Length.
AD= Absolute Deviation Factor

Table 5, contain population of Sentence Length and calculated values of Absolute Deviation Factor (AD) into equation 2. Given sentence Length information is taken from a dummy set of data.

Table 5: Absolute Deviation Factor (AD)

Sentence No	Sentence Length	Absolute Deviation Factor (AD)
1	5	0.39
2	8	0.62
3	9	0.7
4	11	0.85
5	13	1
6	16	0.77
7	18	0.62
8	18	0.62
9	19	0.54
10	27	0.077

The output data is developed based on Gaussian distribution theorem. The mean values will be in peak and rest of other values will be found as high as much near of the mean value. The figure 2 represents the plotting of sentences after calculating AD.

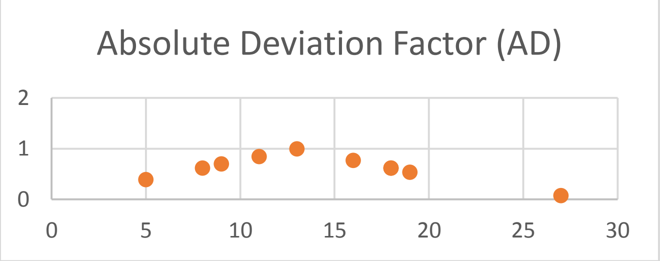


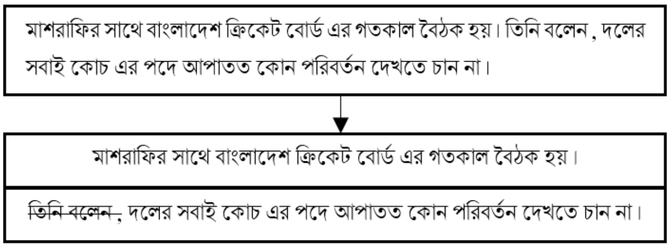
Fig. 2: Absolute Deviation Factor of Sentence

Mean of sentences position deviation could be calculated, it would be useful to identify average distance between sentences position. Based on that result a standard value of selecting minimum or maximum position could be defined.

(d) Uniform Sentences –Uniform sentences could be considered as identical sentences. Sentences that represent an information which already had discussed. Those linked to a chain set of information’s. In Bengali language those linked sentences could be identified, if few set of words matched. Words like –তিনিবলেন ,তারাবলেছেন ,এজন্যই ,সুতরাং ,এরফলে,cte . This set of information’s could be considered as a similar topic.

(e) Imitating Sentences – Sentences that reference in many other sentences is a set of Imitating Sentences. In a document, a topic is discussed in various passages in many forms. Those set of similar semantic data should be considered a group of similar data. And identifying those data will help to avoid data redundancy and omit the chance to represent similar data into summarize text. In extractive approach, one of the major problem is to identify the summarized sentences, either properly linked or not basis of morphological, linguistic and semantic rules. By identifying

those linking words, will help to extract related sentence, as well.



মাশরাফির সাথে বাংলাদেশ ক্রিকেট বোর্ড এর গতকাল বৈঠক হয়। তিনি বলেন , দলের সবাই কোচ এর পদে আপাতত কোন পরিবর্তন দেখতে চান না।

মাশরাফির সাথে বাংলাদেশ ক্রিকেট বোর্ড এর গতকাল বৈঠক হয়।

~~তিনি বলেন ,~~ দলের সবাই কোচ এর পদে আপাতত কোন পরিবর্তন দেখতে চান না।

Fig 3. Example of Imitating Sentences detection.

Basically here the Imitating Sentences detection was implemented to reconstruct the sentences if necessary. In figure 3, there are two sentences and both are linked together. If the second sentence is being selected as final summarization sentence with a high score? Then the problem is, second sentence extends some information of first sentence because of the word “তিনি বলেন” (He says). It directly indicates that subject is stated in previous sentence. In abstractive method summarization, the subject will be identified and placed instead the word “তিনি বলেন”, but in extractive approach it is difficult to identify the subject to what it states for. So here the uniform sentences will be prepared by eliminating the linking words, like –তিনি বলেন, সুতরাং, তাহলে, এজন্য, এই কারনে, etc.

(f) Skeleton of Document –Document title and headers contain the main idea of the given text. Those words could affect the weighted words. So extracting skeleton data of document and comparing those with existing weighted data is important. It sometimes help to identify the important sentences, which are really leading information’s.

(g) Frequent Word Percentile – Every sentence apprehends information, as well as treated like individually important factor. Frequent words associated in a sentence must affect the entire sentence. The more frequent words is found in a sentence, the more it increases the possibilities to get selected as prime sentence. But the only criteria is not carrying most frequent words, as the sentence length is a factor too. Sometime through this process, few small sentences get prioritized. Though, those are not appropriate. That’s why this module will be implemented on only those sentences, which full fill the minimum sentence length. A percentile value will be added to every sentence, as per combination of frequent words weight and sentence length. The frequent word percentile and effect rate is explained in Equation (3) as follows:

$$E_w = \frac{F_w + T_w}{100} - - - (3) \ \ Where,\ F_w = \frac{W_k}{W_n}$$

Here E_W= Effect Rate; T_W = Total weight of frequent words; F_W = Frequent words percentile; W_k = Total number of used word; W_N = Total number of words.

Table 6: Effect rate of frequent words

Sentence No.	W_N	W_k	T_W	E_W
1	18	10	25	0.25
2	26	4	10	0.10
3	17	1	3	0.03
4	9	5	12	0.12
5	16	5	16	0.16
6	6	2	4	0.04
7	5	3	7	0.07
8	6	2	4	0.04

In table 6, sentences with maximum weighted value and total number of words produced maximum effect rate. For example, sentence number 1, 4 and 5 have maximum effect rate. Sentence number 6 and 8 have similar W_N values and sentence 3 have less weighted value then other sentence. But the effect rate is high for sentence 1, because the ratio of W_k and $T_W$is also higher than other sentence.

3.1.6 Final Processing

In this proposed method, words and sentences were clustered in three different ways. Through the deep analysis of input document by Word and Sentence Analysis, prime sentences were categorized for producing quality summarized output. Prime sentences will be grouped by Word and sentence analysis. And then those set of grouped sentences will be send for final processing. Here, those prime sentences will be evaluated again.

(h) Prime Sentences –Prime sentences are the set of sentences which scored maximum, in sentence scoring method. The sentence scoring method is explained in Equation (4) as follows:

$$S_k = E_w + T_w + AD + S_u + S_i + D_k + W_n + E_r + W_c \\ - - - - - - - - - - - (4)$$

Here, S_k is the score of every individual sentence. E_W is the effect rate of frequent words in a sentence. T_W holds the sum of every weighted words used in that sentence. AD defines the rate of absolute deviation factor of that sentence from the mean sentence length. S_U, S_i and D_k represents the value of Number of Uniform sentences, imitating sentences and skeleton of document respectively. From analysis, the value of Numeric value identification, repeated words distance and Cue words are stated as W_N, E_r and Wc. After calculating every sentences, most scored words will be ready for final processing. All these sentence are stated as prime sentences.

(i) Aggregate Similarities–The final set of prime sentences may contain similar sentences. For a better summarization, similar sentences should be omit. Those similar sentences will be identified and set as group of similar sentence.

(j) Final Gist Analysis – Sentences form aggregate similarities, will be considered as similar token. The main leading sentence will be extracted from those groups and prepare a Gist of final information. Lastly, the Gist tokens will

be crosschecked with the topic words. Finally, the sentences will be released for publishing.

(j) Sentence Ranking – Sentence will be ranked based on the sentence position. As those sentences were found in an order in the given input document. Then, the sentences will be demonstrated as final summary.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The aim of text summarization is to generate extractive summary. Here different topic based Bangla text document were tested. For experimental purpose, 3 different Bangla Text were tested by this system. In addition, those documents were tested by different human users too. The experimental result were stated in figure 4.

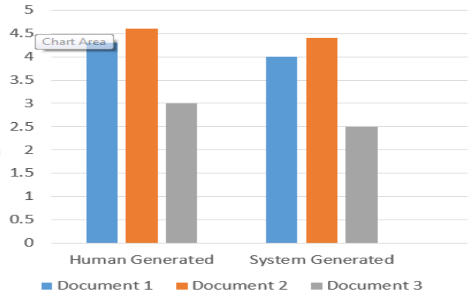


Fig. 4: Output statistics by system vs. human user

The experimental results were calculated out of 5. Different types of news and articles were tested through this system. Python and Natural Language Tool kit (NLTK version - 3) Library file was used to develop the system. Accuracy of output was evaluated by comparing the system generated summary with human generated summary. Example of a news article summarization is given below.

Original Text –

<u>ক্যামেরুনে ট্রেন লাইনচ্যুত হয়ে নিহত ৫৩</u>

পশ্চিম আফ্রিকার দেশ ক্যামেরুনে অতিরিক্ত যাত্রীবাহী একটি ট্রেন লাইনচ্যুত হয়ে কমপক্ষে ৫৩ জন নিহত ও ৩০০ জন আহত হয়েছে। গত কাল শুক্রবার দেশটির রাজধানী ইয়াউন্ডে ও অর্থনৈতিক কেন্দ্রস্থল দুয়োলা শহরের মধ্যে যাতায়াত করার সময় ছোট শহর ইসেকার কাছে লাইনচ্যুত হয়ে ট্রেনটির বগি গুলো উল্টে গেলে হতাহতের এ ঘটনা ঘটে। বিবিসির প্রতিবেদনে বলা হয়েছে, সম্প্রতি ভারী বৃষ্টিপাতের ফলে ভূমিধসের সৃষ্টি হওয়ায় দেশটি জুড়ে সড়ক যোগযোগ বিপর্যস্ত হয়ে পড়ে। এতে করে স্বাভাবিক অবস্থার চেয়ে ট্রেনে যাত্রীদের চাপ অনেক বেড়ে যায়। বার্তাসংস্থা এপি জানিয়েছে, স্বাভাবিক অবস্থায় গড়ে ৬০০ যাত্রী চলাচল করলেও ট্রেনটি ১৩০০ যাত্রী বহন করছিল। ক্যামেরুনের পরিবহনমন্ত্রী এডগার্ড অ্যালাইনমেবি জানিয়েছেন এ দুর্ঘটনায় ৩০০ জন আহত হয়েছে। নিহত ব্যক্তির সংখ্যা আরও বাড়তে পারে।

System Generated Output –

পশ্চিম আফ্রিকার দেশ ক্যামেরুনে অতিরিক্ত যাত্রীবাহী একটি ট্রেন লাইনচ্যুত হয়ে কমপক্ষে ৫৩ জন নিহত ও ৩০০ জন আহত হয়েছে। স্বাভাবিক অবস্থায় গড়ে ৬০০ যাত্রী চলাচল করলেও ট্রেনটি ১৩০০ যাত্রী বহন করছিল। নিহত ব্যক্তির সংখ্যা আরও বাড়তে পারে।

Human Generated Output –

ক্যামেরুনে অতিরিক্ত যাত্রীবাহী একটি ট্রেন লাইনচ্যুত হয়ে কমপক্ষে ৫৩ জন নিহত ও ৩০০ জন আহত হয়েছে। ভূমিধসের ফলে দেশটি জুড়ে সড়ক যোগযোগ বিপর্যস্ত হয়ে পড়ে। স্বাভাবিক অবস্থায় গড়ে ৬০০ যাত্রী চলাচল করলেও ট্রেনটি ১৩০০ যাত্রী বহন করছিল। নিহত ব্যক্তির সংখ্যা আরও বাড়তে পারে।

5. CONCLUSION AND FUTURE WORK

This work explains and introduced different approaches of extractive text summarization methods in relation to efficient Bangla text processing. The proposed technique is implemented for efficient Bengali extractive text summarization process. We developed a better summarizer for Bengali language. In extractive approaches, the scoring methods are mostly used. But here, scoring methods were implemented based on different association and linguistic rules. Relations between words and sentences were extracted. Here two steps summarization techniques were implemented. Selection of prime sentences is one of the most significant task. Topic of the document could be extracted from the word and sentence frequency.

Nonetheless there are lots of improvements required for an enhanced summarizer. Development so far, for English language is far better than other languages. Abstractive summarization, definitely provides better summary than the extractive one. But the problem is, implementing abstractive method requires many development phases. However, a hybrid approaches is the future of summarization in Bengali language. Furthermore, we wish for validate the proposed approach for producing new sentences taking into account the words with highest tradeoffs could be suitable for sophisticated summarization.

6. ACKNOWLEDGMENT

We would like to thanks, Department of Science and Engineering of four universities: Daffodil International University, Jahangirnagar University, Britannia University and Comilla University, Bangladesh for facilitating such joint research.

REFERENCES

[1] Rafael Ferreira et al. “Assessing Sentence Scoring Techniques for Extractive Text Summarization”, Elsevier Ltd., Expert Systems with Applications 40 (2013) 5755-5764.

[2] Ani Nenkova, Kathleen McKeown , “A survey of text summarization techniques”, Springer Science+Business Media, LLC 2012

[3] Amitava Das,Sivaji Bandyopadhyay, “Morphological Stemming Cluster Identification for Bangla”, Jadavpur University, Kolkata 700032, India, 2011.

[4] Harsha Dave and Shree Jaswal, “Multiple Text Document Summarization System using hybrid Summarization technique” NGCT-2015, Dehradun, India, 4-5 September 2015.

[5] Efat, Md Iftekharul Alam, Mohammad Ibrahim, and Humayun Kayesh. "Automated Bangla text summarization by sentence scoring and ranking." In International Conference on Informatics, Electronics & Vision (ICIEV), , pp. 1-5, IEEE, 2013.

[6] Anusha Bagalkotkar , Ashesh Kandelwal ,Shivam Pandey , S. Sowmya Kamath “A Novel Technique for Efficient Text Document Summarization as a Service” ICACC-2013, 29-31 August 2013, Kochi, Kerala, India.

[7] E Lloret, M Palomar, “Analyzing the use of word graphs for abstractive text summarization” IMMM 2011, October 23-29, 2011 - Barcelona, Spain

[8] Mariòo, José B., Banchs, Rafael E., Crego, Josep M., Gispert, Adrià, Lambert, Patrik, Fonollosa, José A. R., et al. (2006). N-gram-based machine translation. Computational Linguistics, 32(4), 527–549.

[9] Haque, Rejwanul, Naskar, Sudip Kumar, Way, Andy, Costa-jussa, Marta R., & Banchs, Rafael E. (2010). Sentence similarity-based source context modelling in pbsmt. In Proceedings of the 2010 international conference on asian language processing (pp. 257–260). IEEE Computer Society

[10] Lin, F. and Sandkuhl, K, A Survey of Exploiting WordNet in Ontology Matching, In IFIP International Federation for Information Processing, Volume 276; Artificial Intelligence and Practice II; Max Bramer; (Boston: Springer) 2008, pp. 341350

[11] K. Sarkar, “Bengali text summarization by sentence extraction,” In Proceedings of International Conference on Business and Information Management (ICBIM-2012), NIT Durgapur, pp. 233-245, 2012.

[12] El-Shishtawy, Tarek, and Fatma El-Ghannam, "Keyphrase based Arabic summarizer (KPAS)." In 8th International Conference on Informatics and Systems (INFOS), pp. NLP-7. IEEE, 2012.

[13] M. Kutlu, C. Cigir, and I. Cicekli, “Generic text summarization for Turkish.” The Computer Journal, vol.53, no.8, pp.1315-1323,2010.

[14] Atif Khan and Naomie Salim, “A Review on Abstractive Summarizati on Methods”, Journal of Theoretical and Applied Information Technology, Vol. 59, No. 1, January 2014.

[15] F. Liu, J. Flanigan, S. Thomson, N. Sadeh and N. A. Smith, “Toward Abstractive Summarization Using Semantic Representations” (2015).

[16] N. Kumar, K. Srinathan and V. Varma, “A Knowledge Induced Graph-Theoretical Model for Extract and Abstract Single Document Summarization”, In Computational Linguistics and Intelligent Text Processing , Springer Berlin Heidelberg, pp. 408–423, (2013).

[17] Abujar, Sheikh, and Mahmudul Hasan. "A comprehensive text analysis for Bengali TTS using unicode." *Informatics, Electronics and Vision (ICIEV), 2016 5th International Conference on*. IEEE, 2016.

[18] Mihalcea, Rada, Courtney Corley, and Carlo Strapparava. "Corpus-based and knowledge-based measures of text semantic similarity." *AAAI*. Vol. 6. 2006.

[19] Islam, Aminul, and Diana Inkpen. "Semantic text similarity using corpus-based word similarity and string similarity." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 2.2 (2008): 10.

[20] Mohler, Michael, and Rada Mihalcea. "Text-to-text semantic similarity for automatic short answer grading." *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009.

[21] Gomaa, Wael H., and Aly A. Fahmy. "A survey of text similarity approaches." *International Journal of Computer Applications* 68.13 (2013).

[22] Bär, Daniel, Torsten Zesch, and Iryna Gurevych. "DKPro Similarity: An Open Source Framework for Text Similarity." *ACL (Conference System Demonstrations)*. 2013.

[23] Huang, Anna. "Similarity measures for text document clustering." *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*. 2008.

[24] Bilenko, Mikhail, and Raymond J. Mooney. "Adaptive duplicate detection using learnable string similarity measures." *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003.

[25] Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi. "WordNet:: Similarity: measuring the relatedness of concepts." *Demonstration papers at HLT-NAACL 2004*. Association for Computational Linguistics, 2004.