A Rule Based Extractive Text Summarization Technique for Bangla News Documents

Article *in* International Journal of Modern Education and Computer Science · December 2018

DOI: 10.5815/ijmecs.2018.12.06

3 authors, including:

Rezvi Shahariar
University of Dhaka
5 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Muhammad Asif Khan
Federal Urdu University of Arts, Science and Technology
129 PUBLICATIONS **2,533** CITATIONS
SEE PROFILE

I.J. Modern Education and Computer Science, **2018, 12, 44-53**
Published Online December 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2018.12.06

A Rule Based Extractive Text Summarization Technique for Bangla News Documents

Partha Protim Ghosh
Institute of Information Technology, University of Dhaka
Email: bit0440@iit.du.ac.bd

Rezvi Shahariar
Institute of Information Technology, University of Dhaka
Email: rezvi@du.ac.bd

Muhammad Asif Hossain Khan
Department of Computer Science & Engineering, University of Dhaka
Email: asif@du.ac.bd

Received: 12 September 2018; Accepted: 17 October 2018; Published: 08 December 2018

***Abstract*—**News summarization is a process of distilling the most important information from a news document in a precise way. For the advancement of Internet nowadays almost all of the Bangla newspapers have their online versions, and people of this era like to read newspaper from website using Internet. But large amount of electronic news content is a burden for human to come out with valuable information. For mitigating this pain point, this paper proposes an automatic method to summarize Bangla news document. In this proposed approach, graph based sentence scoring feature is introduced for the first time for Bangla news document summarization. After analyzing vast amount of Bangla news document 12 sentence scoring features have been introduced for calculating score of a sentence. An improved summary generation method has also been proposed which remove the redundant information from summary. The result is evaluated using a standard summary evaluation tool called ROUGE, and found proposed method outperforms all existing methods used in Bangla news summarization.

Index Terms—Bangla news summarization, Extractive based approach, NLP, ROUGE, Sentence scoring features.

I. Introduction

Man is passionately curious to know the unknowns. They like to share their knowledge, current social, political incident to others for the development of society. They want to build a social bonding through some social issues and this bonding initiates by communicating with the mass people through some media. Newspaper is one of the most popular media to do this. Johann Carolus was the man who first published newspaper [1].

In the perspective of Bangladesh, the first newspaper of independent Bangladesh was “The Daily Azadi”. As time went on, the number of Bangla newspapers has also increased. Though Bangladesh is a small country but there are a great number of newspapers published each day. A radical change was happened to the information world after the invention and commercialization of Internet all over the world in the late 1990s. People started to publish news over Internet. Those newspapers which publish its news over Internet are called online newspaper. After the development of Internet in Bangladesh, online newspaper gain popularity day by day. The First Bangladeshi online newspaper is bdnews24.com started its journey back in 2006 [31]. As the time went on, its' number had been increased dramatically. Now, almost all of the Bangla newspapers publish their news in website. Day by day the size of electronic Bangla news data becomes huge. Thus, people who speak in Bangla, face with an overflow of Bangla news articles in daily life. With this vast electronic Bangla news data, people face some problems when someone has to find out relevant information within shortest time. The news which is published by news companies, all the contents of the news are not important but people have to read all the contents to come out with exact information from the various news. Researchers identified this problem few years back and devoted themselves on Bangla Natural Language Processing (BNLP).

Text summarization is an application area of Natural Language Processing (NLP). Text summarization is a process of finding out the most important information from the source document in a precise way. Actually, it represents the condensed information of a longer text. In the shorter text, all the important information of longer text should be present and it should not be more than half

of the original text. With the help of this text summarization process, it is possible to summarize a long news contents to a shorter version including all salient information. Basically, summarization process can be categorized into two types i) Extractive summarization ii) Abstractive summarization. Extractive summarization is a process where important sentences are selected from the original text. And these sentences will represent the whole document. On the other hand abstractive text summarization is a process where a semantic method is used to examine and interpret the original text and then a new brief text is presented on the basis of the information of original text.

News summarization system was proposed initially for English news content around five decades ago. After that, several researchers have enriched text summarization system. Bangla is 8th most spoken language in the world [7]. In Bangladesh we have overabundance amount of Bangla electronic news data. But this is a matter of great regret that there are only a few research works done on Bangla news summarization [10, 11, 12, 13, 20]. English news summarization system may not directly applicable for Bangla news content, because of different sentence structure; grammatical rules and so on. Research work for Bangla language is difficult because automatic tools are unavailable for Bangla language to facilitate research. Identifying subject and object is very difficult for Bangla sentence. Additionally, grammatical rule of this language is too much inconsistence.

In this challenging context, a new approach for Bangla news document summarization has been presented here using rule based approach with the following major contributions.

i. Introduced graph based sentence scoring features for Bangla news document, and use those features along with surface level and corpus level features.
ii. Most of the sentence scoring features have been introduced for the first time in Bangla news summarization which helped to generate summary more accurately such as aggregate similarity, bushy Path, keyword in sentence, presence of inverted comma, and special symbol
iii. Also introduce an improved summary generation procedure that helps to remove redundant information from the summary.

The rest of the paper is organized as follows: Section 2 describes literature review on Bangla news summarization. In section 3, proposed method is discussed in detail. Evaluation and discussion on results are illustrated in Section 4. Finally, the conclusion is drawn with future directions in Section 5.

II. Literature Review

News summarization was first introduced for English Language by H. P. Luhn [3] in 1958. Here, the significant factor of a sentence is derived from an analysis of its words. It was proposed that the frequency of word in a news article establishes a useful measurement of words' impact. The method of H. P. Luhn [3] was first extended by incorporating position of sentences and cue-phrases by P. B. Baxendale [28]. It is said that sentence can be important based on its position and containing certain cue-words (i.e., words like "important" or "relevant") or the words of heading. Today, various research works are available in the arena of English news summarization [18, 19]. The recent research works on English news summarization have also followed for modeling our methodology. This [19] is one of the recent work proposed by Hilario Oliveira, Rafael Ferreira in 2016. They have used several new sentence scoring features like: i) lexical similarity ii) sentence centrality iii) word co-occurrence iii) text rank. This paper also presents a comparative analysis among these features; those have been used to calculate sentence score. The aim of this paper is to investigate several shallow sentences scoring performance.

English news summarization procedure has reached at a mature stage. Except English language, news summarization of other languages, like Bangla and Hindi are not well defined. There are only few attempts conducted in the field of Bangla news summarization. In 2004, Islam and Masum [8] presented "Bhasa", a corpus oriented search engine and summarizer. It performed document indexing and information retrieval based on keywords using vector space retrieval model [21] for Unicode Bangla text. This was the first attempt on Bangla text summarization. A few years later, some techniques from the investigation of English news summarization systems were applied to summarize Bangla news by Nizam Uddin et al. in 2007 [9]. They proposed a technique by incorporating some existing methods of English news summarization as follows: (i) location method, (ii) cue method, (iii) title method, (iv) term frequency, and (v) numerical data. They have taken 40% higher ranked sentences from the input document as summary.

In 2012, Kamal Sarkar [10] proposed an easy-to-implement approach for Bangla news summarization like the method of Edmandson [25]. It has three major steps: (i) preprocessing, (ii) sentence ranking, and (iii) summary generation. In this method, thematic term has been utilized which is related to the main theme of a news document. The term which has TF-IDF (Term Frequency Inverse Document Frequency) values greater than a predefined threshold value is taken as thematic term. In 2013, Md. Iftekharul Alam Efat et al. [13] introduced a method for Bangla news summarization by sentence scoring and ranking. In 2015, Md. Majharul Haque and Zerina Begum proposed an automatic Bangla news document summarization method by introducing sentence frequency and clustering.

Another work on Bangla news summarization have proposed by Sumya Akter, Md. Palash Uddin and Shikhor Kumer Roy in 2017 [6]. This paper presented a method for summarization which extracts important sentences from a single or multiple Bangla news documents. They used sentence clustering approach to

generate summary from both single and multi-documents. We found a recent work on Bangla news summarization have done by Sheikh Abujar, Mahmudul Hasan and M.S.I Shahin in 2017 [17].

In 2015 Md. Majharul [32] proposes a Bangla news summarization technique using term frequency and sentence clustering. It also considers numerical figures. For summary generation, it divides the sentences into two clusters and takes half of the summary sentence from each cluster. But the problem with the work is that it evaluates the method only using 5 news articles. Again in 2017 Md Majharul [20] proposes Bangla news summarization by introducing sentence frequency and improve sentence ranking technique. The significant thing of this work is that, it gives the first sentence more priority if it contains any title word. On the other hand, it also consider numerical figure in words.

All the Bangla news summarization research works have used either surface level or corpus level sentence scoring features to generate summary only. This finding draws our attention to devise an approach based on both. Hence, the propose method uses graph based sentence scoring features for the first time in the history of Bangla news document summarization along with both surface and corpus level features. Moreover, we analysed our proposed method using ROUGE (Recall Oriented Understudy of Gisting Evaluation) which shows better result than the five latest existing methods found in the Bangla literature.

III. Text Summarization Technique for Bangla News DocumentS

In this section, first we would like to describe the proposed methodology that we used for Bangla news summarization. This process starts with taking an input of Bangla news document. The entire process of this proposed method has been divided into the following four sub processes which are document preprocessing, calculating sentence score using sentence scoring features, ranking the sentences, and selecting summary delineated below.

3.1 Document Preprocessing

Preprocessing is the first step of our method which is started from user input of a Bangla news document and goes through preprocessing of the document. Input news document is segmented to sentences based on the punctuation marks "।", "?", or, "!"" as the end point of a sentence. Then every sentence is tokenized into words. In this way, a word list is generated from an input news document. There are some words in Bangla language which are used to indicate the tense, adjective or for adapting grammatical structure. These words are called stop words. Stops words have less importance to represent a document. These words should be removed for further analysis. For identifying stop words, a list of stop words is kept in the system with which all the words of the input document are checked and removed which matched the stop words. The list of 398 stop words for Bangla language has been collected from [29]. In Bangla language, words are very much inflectional. So word stemming algorithm is applied to convert the words with different endings to a single word that is shown in Fig. 1.

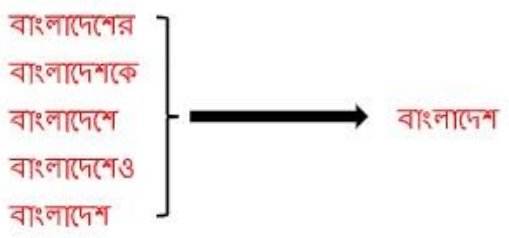

Fig.1. Word stemming in Bangla

3.2 Sentence Score Calculation

Sentence scoring features defines how important a sentence is, among all the sentences of a document. A sentence, which has higher score, is most important sentence in a document. Selecting features for a summarization method is most important part. In our work, more than 2000 Bangla news document have been analyzed to realize which features can represent a news document. After performing an analysis, 12 sentences scoring features have been selected to calculate sentence score. All these features can be categories into three types:

3.2.1 Graph Based Features

In this subsection, aggregate similarity and bushy path are discussed briefly.

3.2.1.1 Aggregate Similarity (F1)

Aggregate similarity is a graph based technique that shows how a sentence is relevant with other sentence of a document. It works based on centrality idea. Centrality idea identifies main discuss topic of a document by finding highest relevancy of a sentence among all the sentence of a news document. Relevant sentences have more information in common with other sentence of a news document. This technique computes the importance of a sentence (S_i) by calculating cosine similarity with all other sentences of that news document. Here, sentence (S_i) is considered as a vertex of a graph and computed cosine similarity with all other sentences. If cosine similarity between two sentences is greater than a threshold (0.16), an edge is created between them. The weight of that edge will be the similarity value. Total weight of a sentence will be the summation of all edge value, which is created. Thus highly connected vertices will represent central sentences that indicate the main discussion in a document. Aggregate Similarity is defined as:

$$\text{Aggregate Similarity of } s_i = \sum_{j=1, i\neq j}^{s} edge\_weight \quad (1)$$

Where S is the total number of sentence in a document and edge_weight is the similarity between the sentences s_i and s_j. If aggregate similarity score for a sentence is

greater than 1, then score should be normalized. Score can be normalized by the following equation:

$$\text{Normalized Score of } s_i = \frac{S1 - \text{Min}}{\text{Max} - \text{Min}} \quad (2)$$

Here, similarity score of current sentence is S_1, minimum similarity score is Min and maximum similarity score is Max within the document.

3.2.1.2 Bushy Path (F2)

Bushy path is another graph based method used to compute the salience of a sentence based on centrality idea. It is very similar to aggregate similarity method. It computes the importance of a sentence by calculating ratio of generating edge of a sentence in a news document. At time, to compute aggregate similarity edge was created between two sentences if similarity value is greater than the threshold. Here number of edges is counted for a sentence S_i in a document. This technique score is measured using the following equation:

$$\text{BushyPath}(s_i) = \frac{\text{NoE connected to Si}}{\text{MNoE connected to any sentence}} \quad (3)$$

Here, number of edges denoted as NoE and maximum number of edges is expressed by MNoE.

3.2.2 Corpus Based Features

In this subsection, only corpus based features are described. Term frequency inverse sentence frequency and keyword in sentences are mostly used corpus based sentence scoring features.

3.2.2.1 Term Frequency Inverse Sentence Frequency (F3)

The term frequency-inverse sentence frequency (TF-ISF) is used to measure the weight of terms as per their number of appearance in a document. Term Frequency (TF) measures how frequently a term exists in a document. Inverse sentence Frequency (ISF) measures how much descriptive a word is. This measure takes place by finding a word is common or rare across all sentences. If a word is appeared frequently in different sentences, that means it is important word for summary generation. TF-ISF score of a term is calculated by following equation:

$$\text{TF-ISF of term t} = \text{TF}(t_i) \times \log\left(\frac{S}{Sti}\right) \quad (4)$$

TF-ISF score of a sentence will be the summation of TF-ISF score of all term of a sentence which is determined by the following equation:

$$\text{TF-ISF of sentence } s_i = \sum_{tj \in T}^{s} \text{TF} - \text{ISF}(tj) \quad (5)$$

Here, TF returns the frequency of term t_i in the document, S is the total number of sentence in the document, T is the total number of term in a sentence s_i and S_{tj} is the total number sentences in which t_i occurs. In this case, also score of TF-ISF for sentence will be more than 1. Thus normalization procedure should be called for each score as it is used for aggregate similarity.

3.2.2.2 Keyword in Sentence (F4)

Keyword of a document is the highly frequent word. If a high frequent word is present in a sentence then this sentence will have the high probability to discuss main topic of news document. Here, top 10% high frequent word is selected as keyword.

$$\text{Keyword of sentence } s_i = \frac{\text{NKeys}}{\text{TNWowrds}} \quad (6)$$

Here, NKeys denotes the total number of keywords in a sentence and TNWords denotes the total number of words present in a sentence.

3.2.3 Surface level features

Different surface level sentence scoring features are used in the literature. Among them, sentence position, title word, cue word, numerical value, special symbol, presence of inverted comma, URL, and Email address are used extensively. Below these features are discussed shortly.

3.2.3.1 Sentence Position (F5)

The position of a sentence in a news document is one of the most effective features to select relevant sentences for summary. The sentence position feature is that the first sentence in a news document comprises the most relevant sentence of any document. And their importance decrease as the sentence goes further down in the document. Most of the time first sentence is the description of the title of a news document. So, the first sentence is the most important sentence for a news document. Proposed method also gives high importance to first sentence and importance goes down gradually. Sentence position score is measures by using the following equation:

$$\text{Sentence Position score of } s_i = 1 - \frac{i}{S} \quad (7)$$

Here, i is the i^{th} sentence in the document and i starting from zero and S is the total number of sentence in the news document

3.2.3.2 Title Word (F6)

Title word of a news document is most relevant word about document's discussing topic. It represents the theme a news document contains. In several existing methods [13, 16, 25], title words have been considered for sentence scoring. We have also observed from the analysis of 2000 Bangla news documents that title words convey the theme of the news document in the most cases. To compute the title words score of any sentence s_i, the following equation is used:

$$\text{Title Word Score of } s_i = \frac{\text{Wsi} \cap \text{Wt}}{|\text{Wt}|} \quad (8)$$

Here, Wsi denotes the set of words in the sentence and Wt is the set of words in title.

3.2.3.3 Cue Word (F7)

Cue phrase technique is one of the initial method uses for summary generation. In Bangla language, more than one sentence can be used for expressing information. There exist semantic relation between linked sentences. Cue phrase emphasize the gist of two sentences. The cue words can be as "মোটকথা" (in short), "অবশেষে" (at last), "ইতিমধ্যে" (already), "যেহেতু" (since), "পরিশেষে" (in summary) etc. Thus, sentences which contain any cue word has higher probability to select as summary sentence. Score is compute as:

$$\text{Cue word score of } s_i = \frac{\text{No.Cue words}}{\text{TNo. Cue Words}} \quad (9)$$

In this equation, No.Cue words denotes the number of Cue words in the sentence and TNo.Cue words denotes the total number of cue words in the document.

3.2.3.4 Numerical Value (F8)

Numerical figure is always important for representing significant information of a news document. Sentence containing numerical data are good candidate to be included in summary. In our proposed method, numerical figure identification pattern is used to identify numerical value. Numerical value is calculated by using the following equation:

$$\text{Numerical value of } s_i = \frac{\text{No.of Numerical Words}}{\text{Total Number of Words}} \quad (10)$$

In this equation, numerical value of S_i is obtained by dividing the number of numerical words by the total number of words in the sentence.

3.2.3.5 Presence of Inverted Comma (F9)

In Bangla, (“”, ‘’) quotation marks or inverted comma surrounding quotation, direct speech etc. contain important information. It is important especially for news document and articles where people give their speech. These speeches have a great chance to select in summary. Because people perception have to quote for represent news. Score of this technique is calculated as:

$$\text{Inverted comma score of } s_i = \frac{\text{No.Words in quotation Mark}}{\text{Total Number of Words}} \quad (11)$$

Here, Inverted comma score of a sentence is obtained by dividing the number of words in the quotation mark by the total number of words in the sentence.

3.2.3.6 Special Symbol (F10)

In this feature, different symbol like, %, different currency symbol are considered. Numerical value with currency has greater probability to select a sentence as a summary. Special symbol score is calculated as follows:

$$\text{Special Symbol Score of } s_i = \frac{\text{No.of Special Sym.}}{\text{Total Number of Words}} \quad (12)$$

Here, Special symbol score of S_i is obtained by dividing the number of special symbols in a sentence by the total number of words in the sentence.

3.2.3.7 Date Format (F11)

Dates are very important also for any news document. Presence of dates in the sentence increases the importance of the sentence, because date is more informative than any other words. Date format score is calculated by using following equation:

$$\text{Date Score of } s_i = \frac{\text{No.of date in sentence}}{\text{No.of dates in document}} \quad (13)$$

Here, Date score of S_i is obtained dividing the number of dates in a sentence by the total number of dates in the whole document.

3.2.3.8 Presence of URL/Email Address (F12)

Now-a-days use of Internet has widely spreaded. News document may have URL’s or Email address present in it, which provides more information about the document. So this valuable information should be present in summary. So sentences, which contain URL/Email, should give more priority while generating summary.

$$\text{URL/Email score of } s_i = \frac{\text{No.URLorEmail}}{\text{No.URLorEmailinDoc}} \quad (14)$$

Here, No.URLorEmail denotes the number of URL/Email in a sentence and No.URLor-EmailinDoc denotes the total number of URL/ Email in the whole document.

3.2.4 Sentence Total Score Calculation

For every sentence in the document, all the scoring features are applied. Total score of a sentence are the summation of all twelve features value (from F1 to F12) which is shown in the following equation:

$$\text{Sentence Total Score for } s_i = \sum_{k=1}^{12} F(k) \quad (15)$$

3.3 Sentence Ranking

After completion of total score calculation, every sentence of a news document will have a score. On the basis of assigned sentence score, sentence will be sorted in descending order. This sorted list is the rank list of sentences of that news document.

3.4 Summary Generation

This is final step of our proposed methodology. For summary generation, temporarily top 40% sentence will be extracted as summary from the rank list. This summary percentage is taken empirically. We have conducted an empirical testing over our methodology. We have extracted summary from the rank list as 30% to 45% and tested the result using ROUGE evaluation tool. We find out system gives better result when 40% sentences are extracted as summary. After selecting

sentences as temporary summary, cosine similarity is calculated among the selected sentence. If cosine similarity of any two sentences is greater than 0.6, smaller sentence is removed from the summary sentences and next top ranked sentence is selected as summary sentence. 0.6 similarity score means there are 60 percent similarity between two sentences. These similarities score have been used in recently publish journal paper for English news summarization [33]. The reason behind this action is to remove the almost similar sentence from summary that represent same information. On the other hand larger sentence represent all most all information of smaller sentence. That is why smaller sentence is removed. After performing this action, summary sentence will be arranged according to exact order of original document. These arranged sentences are treated as the summary of the document.

3.5 Pseudocode of Text Summarizing Technique on Bangla News Documents

Input:
Bangla news document
SW: List of stop words
CW: List of cue words

Output: SUMMARY: Summary Sentence

Begin:
Segmenting the news document into sentence according to punctuation mark: । , ?, !
Tokenize each sentence into word based on space after each word
Remove words from words list which is member of stop word (SW)
Convert all words into their base word with the help of stemming algorithm
TW← List of title words
KW← List of key words
N ← Number of sentences in the document

/*Variable Initialization*/
S_{AG_SM}← 0 // Aggregate Similarity Score
S_{BP}← 0 // Bushy path score
$S_{TF\text{-}ISF}$ ← 0 // term frequency and inverse sentence frequency score
S_N← 0 // Numerical figure score
S_{KW}← 0 // score of key word in sentence
S_P← 0 // sentence positional score
S_{Date}← 0 // score for presence of date in sentence
S_{CW} ← 0 // score for cue words in sentence
S_{IV_C}← 0 // score for presences of inverted comma in sentence
S_{TW} ← 0 // score for title words in sentence
S_{SS} ← 0 // score for presences of special symbol in sentence
S_{Email} ← 0 // score for presences of email in sentence
SCORE ←∅ // for containing all sentences score
SUMMARY ←∅ // Summary sentences of input document

For i ←1 to N do
S_{AG_SM}← Aggregate Similarity Score based on equation 1 and 2
S_{BP}← Bushy path score based on equation 3
$S_{TF\text{-}ISF}$ ← TF-ISF score based on equation 4 and 5
S_N← Numerical figure score based on equation 10
S_{KW}← key word Score based on equation 6
S_P← sentence positional score based on equation 7
S_{Date}← presence of date score based on equation 13
S_{CW} ← cue word score based on equation 9
S_{IV_C}← presence of inverted comma score based on equation 11
S_{TW} ← title word score based on equation 8
S_{SS} ← presence of special symbol score based on equation 12
S_{Email} ← presence of email address score based on equation 14
SCORE ← Total score of a sentence based on equation 15
End
Loop
Sort SCORES in descending order
SUMMARY ←Extract top 40% sentence as temporary summary from ordered list
n← number of sentences in temporary summary

For i← 1 to n do
Calculate similarity among sentences
If (similarity score>= 0.6 between any two sentences)
Remove small sentence between two sentences from SUMMARY
Add a top sentence to SUMMARY from remaining ordered list
End if
End
Loop
Sort SUMMARY in ascending order according to sentence position of input document
Return SUMMARY
End

IV. Evaluation and Result

Evaluating summarization method is a difficult task and the sophisticated way is yet to be achieved [3]. In this situation, several techniques have been applied to measure the quality of summary which is depended on the followings: a) importance of selected contents and b) presentation quality. Again, presentation quality can be assessed based on grammatical correctness and coherence. Considering all of these aspects, evaluation procedures are divided into two main categories as: a) intrinsic mode

and b) extrinsic mode [26]. Intrinsic evaluation of selected contents is usually done by comparing system generated summaries with model summaries written by human professionals. More specifically, evaluation is achieved by measuring the overlap between model summary and the automatically extracted summary as in ROUGE evaluation system [22]. In extrinsic evaluation method, the quality of summary is judged based on how it affects the completion of some other task. Proposed method is evaluated using intrinsic mode of summary evaluation.

4.1 Dataset

In the initial stage of Bangla news summarization there is no benchmark dataset for evaluating Bangla news summarization system. To mitigate these problem, some researchers created a standard dataset for Bangla news summarization evaluation by analyzing 3400 Bangla news documents. These documents had been collected from the most popular Bangladeshi newspaper the Daily Prothom-Alo. These news documents contain variety of news that cover a wide range of topics like political, sports, crime, economy, environment, etc. After analyzing these documents, 200 documents had been selected randomly. The model summary of these documents was generated by two groups of scholars of Bangla Language, each group has three members. So for each document 6 model summaries are generated and they selected 3 summaries randomly. These model summaries was compared with the system generated summary. These dataset has been used by several Bangla Natural Language Processing researchers in recent years. Point to be noted that several research works on Bangla news Summarization have been published based on these datasets [10, 11, 12, 20]. We have used this dataset [24] for evaluating our proposed method. We divide the dataset into two groups each contains 100 document randomly. Each document contains 3 model summaries.

4.2 Evaluation

Evaluation of summary is not an easy task, because principally there is no ideal summary of a news document. For evaluation of a summary precision, recall, F-measure evaluation metrics are used. It is noticeable that these evaluation matrices have been considered in several news summarization systems for Bangla [10, 11, 12, 13, 20], and English [19, 26].

If 'A' indicates the number of sentences retrieved by summarizer and 'B' indicates the number of sentences that are relevant as compared to target set, Precision, Recall and F-measure are computed based on the following equations:

$$\text{Precision (P)} = \frac{A \cap B}{A} \quad (16)$$

$$\text{Recall(R)} = \frac{A \cap B}{B} \quad (17)$$

$$\text{F} - \text{measure} = \frac{2 \times P \times R}{P + R} \quad (18)$$

4.3 Experiment and Results

To judge the efficiency of the proposed method, experiments have been conducted on 200 news documents. In each time, the system generated summary is compared with three model summaries of each news document and computed the average value of Precision, Recall and F-measure with ROUGE automatic evaluation package [23] which is shown in Table 1.

Table 1. Average of ROUGE-1 scores of the proposed Method

Data Set	Different Measurement Value		
	Average Recall	Average Precision	Average F_Measure
Data Set1	0.6637533	0.603662	0.6244975
Data Set2	0.6904979	0.5911405	0.6306442
Combined dataset	0.677126	0.597401	0.627562

The proposed method is compared with five existing modern methods [10, 11, 12, 13, 20] found in the Bangla literature which have been published in recent years. The reason of selecting these methods is: all of the methods have been evaluated with same data set for which the results have been varied from the respective results claimed by the corresponding authors of the existing methods [20]. Comparison results based on ROUGE-1 of two dataset have been depicted in Fig. 2 and Fig. 3 respectively where method 1 is presented in [10], method 2 is in [11], method 3 is in [13], method 4 is in [12], and method 5 is in [20]. For mean comparisons of proposed method with five existing methods, T-test has been performed at 95% confidence interval. Precision, recall, and F-measure of five existing methods have tested for statistical satisfaction.

In null hypothesis of T-test, we assume proposed method mean is less than or equal to each method mean and alternative hypothesis is proposed method mean is greater than from each method mean. For every time, t (calculated value) is greater than T (tabulated value). That indicates the rejection of null hypothesis and acceptance of alternative hypothesis. Thus it is easy to claim that proposed method achieved significantly better result than all the existing methods compared. Here significant improvement of proposed method has been shown from all the latest methods of Bangla news summarization.

4.4 Discussion on Results

In this proposed method, some innovative features have been introduced for getting better performance. Features like aggregate similarity, bushy path, key word in sentence etc. are newly used features in the field of Bangla news summarization. For these reasons, proposed method performed better than the previous methods.

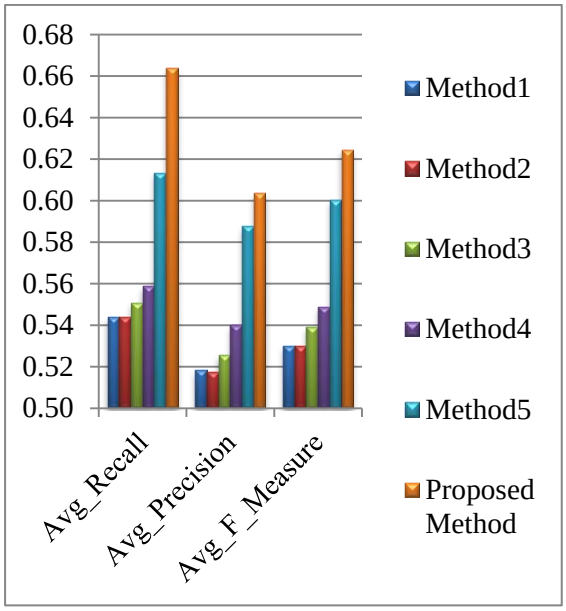

Fig.2. Comparison of proposed method with the five latest existing methods based on the average ROUGE-1 scores using dataset-1

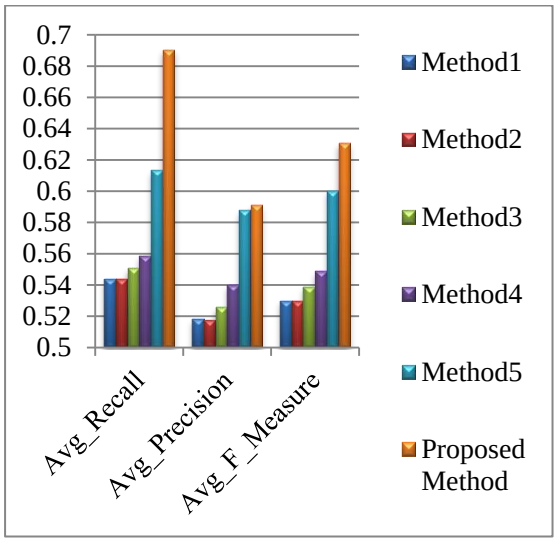

Fig.3. Comparison of proposed method with the five latest existing methods based on the average ROUGE-1 scores using dataset-2

In the previous subsection, the average of Recall, Precision and F-measure scores of ROUGE-1 have been shown for the proposed method. The comparison of the proposed method with the five latest Bangla news summarization methods has also been demonstrated for ROUGE-1 scores respectively. It has been found that the proposed method outperforms all of them. Now, the improvement of performance from the four latest existing methods is given in the following Table 2:

Table 2. Improvement of news summarization in the proposed method than the five latest existing methods

Method Name	**Improvement based on ROUGE-1 score**		
	Precision	Recall	F-Measure
Method1	15.27%	24.53%	18.41%
Method2	15.47%	24.48%	18.41%
Method3	13.65%	22.90%	16.44%
Method4	10.64%	21.20%	14.34%
Method5	1.66%	10.40%	4.54%

V. Conclusion and Future Work

A new approach has been illustrated here to summarize Bangla news document based on rule based approach. Though, there are many research works for English news summarization but these may not be directly applicable for Bangla because of the complexities of Bangla language in the structure of sentences, grammatical rules, inflection of words, and so on. Despite of these difficulties and challenges, in this work, an innovative method for summarizing Bangla news document has been introduced which produced extracted condensed news to the reader by using the tool produced using JAVA language. Thus, Bangla reader can save lot of time by using our approach to read only necessary news. In addition, graph based sentence scoring features are introduced for the first time for Bangla news summarization. On the other hand, corpus level and surface level sentence scoring features have also enhanced. Here a standard dataset is used for evaluation of proposed method. Proposed method shows significant improvement from all the latest Bangla news summarization methods. Evaluation has been done by measuring the similarity of system generated summaries with human professionals' summaries using ROUGE evaluation package. The average precision, recall and F-measure score for proposed method is 0.60, 0.68, and 0.63 respectively.

In our method we did not check any synonym. The words presented in different synonyms cannot be treated as same word because we do not have any tool for synonym identification. In future, we will try to address this issue. On the other hand, we will enhance sentence scoring features to make the system generated summary more closer to the human generated summary

ACKNOWLEDGMENT

We would like to give thanks to Information and Communication Technology (ICT) Division, Ministry of ICT, Government of the People's Republic of Bangladesh for supporting this research work.

REFERENCES

[1] First newspaper. retrieved from https://www.revolvy.com/page/Johann-Carolus [Online; accessed 5-may -2018].

[2] Ferreira, R., de Souza Cabral, L., Freitas, F., Lins, R. D., de Franҫa Silva, G., Simske, S. J., & Favaro, L. (2014). A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications*, *41*(13), 5780-5787.

[3] Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, *2*(2), 159-165.

[4] Haque, M. M., Pervin, S., & Begum, Z. (2013). Literature Review of Automatic Single Document Text Summarization Using NLP. *International Journal of Innovation and Applied Studies*, *3*(3), 857-865.

[5] Haque, M., Pervin, S., & Begum, Z. (2013). Literature review of automatic multiple documents text summarization. *International Journal of Innovation and Applied Studies*, *3*(1), 121-129.

[6] Akter, S., Asa, A. S., Uddin, M. P., Hossain, M. D., Roy, S. K., & Afjal, M. I. (2017, February). An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm. In *Imaging, Vision & Pattern Recognition (icIVPR), 2017 IEEE International Conference on*(pp. 1-6). IEEE.

[7] Chowdhury, M., Khalil, I., & Mofazzal, H. C. (2000). Bangla Vasar Byakaran. *Dhaka: Ideal publication*.

[8] Islam, M. T., & Al Masum, S. M. (2004, December). Bhasa: A corpus-based information retrieval and summariser for bengali text. In *Proceedings of the 7th International Conference on Computer and Information Technology*.

[9] Uddin, M. N., & Khan, S. A. (2007, December). A study on text summarization techniques and implement few of them for Bangla language. In *Computer and information technology, 2007. iccit 2007. 10th international conference on* (pp. 1-4). IEEE.

[10] Sarkar, K. (2012). Bengali text summarization by sentence extraction. *arXiv preprint arXiv:1201.2240*.

[11] Sarkar, K. (2012, August). An approach to summarizing Bengali news documents. In *proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 857-862). ACM.

[12] Sarkar, K. (2014). A keyphrase-based approach to text summarization for English and bengali documents. *International Journal of Technology Diffusion (IJTD)*, *5*(2), 28-38.

[13] Efat, M. I. A., Ibrahim, M., & Kayesh, H. (2013, May). Automated Bangla text summarization by sentence scoring and ranking. In *Informatics, Electronics & Vision (ICIEV), 2013 International Conference on* (pp. 1-5). IEEE.

[14] B. language. (2017) History of bengali language. retrieved from https://www.cs.mcgill.ca/rwest/link-suggestion/wpcd2008-09 augmented/wp/b/Bengalilanguage.html. [Online; accessed 05-May-2017].

[15] T. T. of Inida. (2017) Nearly 60% of indians speak a language other than hindi. retrieved from http://timesofindia.indiatimes.com/india/ Nearly-60-of-Indians-speak-a-language-other-than-Hindi/articleshow/ 36922157.cms. [Online; accessed 05-March-2018].

[16] Inshorts. (2017) Bengali is an official language in africa's sierra leone. retrieved from https://www.inshorts.com/news/bengali-is-an-official-language-in-africas-sierra-leone-1487699311123.[Online; accessed 06-February-2018]

[17] Abujar, S., Hasan, M., Shahin, M. S. I., & Hossain, S. A. (2017, July). A heuristic approach of text summarization for Bengali documentation. In *Computing, Communication and Networking Technologies (ICCCNT), 2017 8th International Conference on* (pp. 1-8). IEEE.

[18] R. B. System. (2017) Rule based system. Retrieved from http://www.j-paine.org/students/ lectures/lect3/node5.html. [Online; accessed 01-April-2017].

[19] Oliveira, H., Ferreira, R., Lima, R., Lins, R. D., Freitas, F., Riss, M., & Simske, S. J. (2016). Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. *Expert Systems with Applications*, *65*, 68-86.

[20] Haque, M., Pervin, S., & Begum, Z. (2017). An Innovative Approach of Bangla Text Summarization by Introducing Pronoun Replacement and Improved Sentence Ranking. *Journal of Information Processing Systems*, *13*(4).

[21] Wong, S. M., Ziarko, W., & Wong, P. C. (1985, June). Generalized vector spaces model in information retrieval. In *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 18-25). ACM.

[22] Lin, C. Y., & Hovy, E. (2003, May). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 71-78). Association for Computational Linguistics.

[23] R. 2.0. (2016) Java package for evaluation of summarization tasks with updated rouge measures. Retrieved from http://kavita-ganesan.com/content/rouge-2.0. [Online; accessed 25-May-2016].

[24] B. N. L. P. Community (2016) Dataset for evaluating Bangla text summarization system. Retrieved from http://bnlpc.org/research.php. [Online; accessed 8-August-2017].

[25] Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)*, *16*(2), 264-285.

[26] Hovy, E., & Lin, C. Y. (1999). Automated Text Summarization in SUMMARIST. Advances in Automatic Text Summarization, 81-94.

[27] Hariharan, S., Ramkumar, T., & Srinivasan, R. (2013). Enhanced graph based approach for multi document summarization. *Int. Arab J. Inf. Technol.*, *10*(4), 334-341.

[28] Baxendale, P. B. (1958). Machine-made index for technical literature—an experiment. *IBM Journal of Research and Development*, *2*(4), 354-361.

[29] Bangla Stop word list: Retrieved from https://github.com/stopwords-iso/stopwords-bn [Online; accessed 10-August-2017].

[30] Value Normalization: Retrieved from https://en.wikipedia.org/wiki/Normalization_(statistics) [Online; accessed 12-November -2017].

[31] Bangla News Paper list: Retrieved from http://www.24livenewspaper.com/bangla-newspaper[Online; accessed 5-March -2018].
[32] Haque, M. M., Pervin, S., & Begum, Z. (2015, December). Automatic Bengali news documents summarization by introducing sentence frequency and clustering. In *Computer and Information Technology (ICCIT), 2015 18th International Conference on* (pp. 156-160). IEEE.
[33] Oliveira, Hilário, et al. "Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization." *Expert Systems with Applications* 65 (2016): 68-86.

Authors' Profiles

Partha Protim Ghosh completed his BSSE and MSSE degree in Software Engineering from the Institute of Information Technology, University of Dhaka. His research interests include NLP, Text Summarization, and software engineering.

Rezvi Shahariar has completed both B.Sc. and M.Sc. degree in CSE from the University of Dhaka. He is now serving as Assistant Professor at the Institute of Information Technology, University of Dhaka. His research interests include Machine Learning, Data Science, NLP, Ad Hoc networking, and Security.

Muhammad Asif Hossain Khan has completed both B.Sc. and M.Sc. degree in CSE from the University of Dhaka. He also completed PhD from the University of Tokyo, Japan. Currently, he is working as an Associate Professor at the Department of Computer Science and Engineering, University of Dhaka. His research interests include NLP, Information Retrieval, Image Processing, and Machine Learning.

How to cite this paper: Partha Protim Ghosh, Rezvi Shahariar, Muhammad Asif Hossain Khan, "A Rule Based Extractive Text Summarization Technique for Bangla News Documents", International Journal of Modern Education and Computer Science(IJMECS), Vol.10, No.12, pp. 44-53, 2018.DOI: 10.5815/ijmecs.2018.12.06