Enhancement of Keyphrase-Based Approach of Automatic Bangla Text Summarization

Md. Majharul Haque, Suraiya Pervin
Department of Computer Science & Engineering
University of Dhaka
Dhaka-1000, Bangladesh
Email: mazharul_13@yahoo.com, suraiya@du.ac.bd

Zerina Begum
Institute of Information Technology
University of Dhaka
Dhaka-1000, Bangladesh
Email: zerin@iit.du.ac.bd

***Abstract*—An approach of automatic Bangla text summarization is presented here by enhancing an existing keyphrase-based method. The enhancement is accomplished with three steps as follows: (i) modifying the keyphrases selection process, (ii) including the first sentence in summary if it contains any title word and (iii) counting numerical figure which is presented in digits and words for sentence scoring. Step by step performance analysis of our proposed approach is portrayed for two datasets. Performance is measured with ROUGE (Recall Oriented Understudy for Gisting Evaluation) automatic evaluation package. The results, based on ROUGE-1 and ROUGE-2 scores, show that the proposed enhancement has significant influence for Bangla text summarization over existing keyphrase-based method.**

Keywords—Bangla text summarization; keyphrase; first sentence; numerical figure; ROUGE

I. INTRODUCTION

As long as Internet users are increasing, electronic contents (e-contents) are growing proportionally irrespective of language. The estimated size of the websites, which holds e-contents, was around 4.69 billion pages in 27 May, 2016 [1] and its size is increasing exponentially in each second. Users are encumbered with the huge volume of e-contents or texts, whereas they expect the concise information or knowledge within the shortest time. In such a situation, text summarization technique would be an indispensable solution because of generating a quick overview of an entire document within expected time of users [2]. The state-of-art-works in this field [2, 3, 4, 5, 6] have been focused on automatic text summarization in different languages starting with English. Automatic English text summarization technique was firstly proposed by Luhn [3] on the basis of term frequency, around five decades ago. With the increasing amount of texts, a notable development in English text summarization was proposed by Edmundson [4] by considering text title, cue-words and location of sentences. Still, the trend is being continued not only for English but also for Bangla text summarization [7, 8, 9]. As Bangla is the 7th most spoken language in the world [10], e-contents in Bangla are dramatically increasing throughout the cyber world. Therefore, an efficient Bangla text summarization technique is essential for researchers, international news agencies and individuals.

Unlike English which has seen a large number of systems developed to cater to it, other languages are less fortunate [11]. So far, few attempts have been made for Bangla text summarization [7, 9]. In 2004, Islam and Masum [12] proposed the first technique of Automatic Bangla text summarization, in which query terms were used for document indexing and information retrieval. After few years, some methods from the survey work regarding English text summarization systems were implemented to summarize Bangla text by Uddin and Khan [13]. They presented Bangla text summarization based on i) location method, ii) cue method, iii) text title, iv) term frequency and v) numerical data. Furthermore, other prominent researchers [7, 8] worked on some features (already implemented for English [4]) for Bangla text summarization and fine-tuned for better performance in [9].

Apart from the previous approaches, keyphrase based summarization method outperforms for both Bangla and English text, which was proposed by Sarkar in 2014 [2]. However, there are some limitations in sentence selection based on the frequency of keyphrases, term frequency and position of the sentences. Besides, this method [2] does not set the minimum length of keyphrases and that is why single word keyphrases may mislead for Bangla text summarization result as they always get higher rank in analysis. Additionally, general positional score of sentences can't differentiate the importance of the first sentence which can be very significant for Bangla news document. Moreover, some other proficient features can also be utilized in sentence ranking for better performance. Details of the keyphrase based method [2] and its limitations are discussed in section II.

In this paper, we have proposed some enhancements in the keyphrase based method [2] for Bangla text summarization and the results outperform over existing one. The enhancements include: i) setting the minimum length for keyphrases, ii) considering the first sentence specially and iii) counting numerical figure from words and digits for sentence scoring.

The rest of this paper is organized as follows: Section II describes keyphrase based method. Section III presents the proposed enhancements. Evaluation and results are depicted in section IV. Finally, this paper is concluded in section V with future works.

978-1-5090-2597-8/16/$31.00 ©2016 IEEE

II. KEYPHRASE BASED METHOD

Our enhancements rest on the foundation of keyphrase based text summarization method developed by Sarker in 2014 [2]. In this method [2], the keyphrases are extracted from any sentence as a sequence of words containing no punctuation mark and stop words. If any keyphrase consists of more than 5 words, it is not considered. Keyphrases with multiple words are segmented to single words, double words and so on to get more keyphrases. All the keyphrases are ranked using phrase frequency inverse document frequency (PF-IDF) and sentences are scored based on their position and term frequency. There are two phases for summary generation. In the first phase (phase-1), candidate summary sentences are selected which contain top ranked keyphrases. Phase-1 considers the sentences for selection which appear early (within position 5) in the document as per the author's experiment. From these candidate sentences, top scored sentences are selected as final summary sentences. If phase-1 fails to generate summary of user desired length, phase-2 is activated. In the second phase (phase-2), more summary sentences are selected based on the sentences' score from the rest of the sentences.

Point to be mentioned that the applicable fields of keyphrases are indexing [14], searching [15], summarizing [16, 17], etc. For the purpose of text summarization, mult-word keyphrases are used in [16] where redundancy is less emphasized [2]. Again, noun phrases are taken as keyphrases for Arabic text summarization in [17]. The keyphrase based method described in [2] is different from the others [16, 17] because sequence of words are extracted here as keyphrases and there is a specific way of redundancy elimination. In this method [2], when any sentence is selected for summary, selected sentence may contain one or more keyphrases. So, keyphrases which exist in the previously selected sentence(s), keep them apart so that they can't keep part in further sentences selection which results redundancy elimination.

It was stated that the keyphrase based methodology [2] outperforms LEAD baseline method (where first n words are selected as summary) and the methods in [7, 9]. The F-measure score was shown 0.4242 for Bangla text summarization [2].

However, there are some limitations in this keyphrase based method [2]. Here, keyphrases with multiple words are breakdown to single words, double words, etc. so that there will be more keyphrases. Again, keyphrases with single word (got from the breakdown of a keyphrase of multiple words) can have chance to appear again as single word and in the breakdown of other keyphrases. So, keyphrases with single word have more chance to be high frequent. As the keyphrases are ranked based on their frequency, keyphrases with length 1 (single word) are getting higher rank. Based on our investigation with 200 test documents, keyphrases with length 1 may not reflect any concept where reflecting concept was the principal reason of using keyphrases. In this regard, a minimum length of keyphrases should be proposed for better performance as enhancement in Bangla text summarization.

Again, sentences are scored based on their position and term-frequency only [2]. But, these features can't differentiate the first sentence of a document which can be very significant for Bangla news document (discussed in the next section). So, another enhancement has been proposed here to treat the first sentence specially.

III. PROPOSED ENHANCEMENTS

The keyphrase based method [2] has been enhanced in this paper and earned better performance as follows: (i) Modifying the keyphrases selection process, (ii) Considering the first sentence specially and (iii) Counting numerical figure presented in words and digits for sentence scoring. Details of the proposed enhancements are given below:

A. Modifying the Keyphrases Selection Process

In the existing method [2], keyphrases are selected as sequence of words from any sentence containing no punctuation mark and stop words where length of keyphrases can be from 1 to 5. Keyphrases are considered there by claiming that they contain the key concepts. But, it has been found in our observation of 200 test documents that single word keyphrases may not reflect any concept. We may think about the following sentence for example, "দ্বিতীয় বিশ্বযুদ্ধে অনেক মানুষ মারা গেছে" (ditiyo bishwjuddhe onek manush mara geche - Many people were died in the Second World War). Now, if a keyphrase is selected as "দ্বিতীয়" (ditiyo - Second), it is unable to reflect any meaningful thing. But if a keyphrase is selected as "দ্বিতীয় বিশ্বযুদ্ধে" (ditiyo bishwjuddhe - Second World War), it may contain a concept.

In this regard, the keyphrases with single word are ignored in the proposed enhancement. An experiment has been done with 200 news documents and 600 model summaries (three summaries are for each document) by removing keyphrases with length 1 (single word). In the experiment, the F-measure score is increased from 0.4513 (minimum length of keyphrases is 1) to 0.4625 (minimum length of keyphrases is 2). Here, F-measure has been calculated using (6) with our training dataset (discussed later in the section IV). Again, it has been found that performance is decreased if we set the minimum length of keyphrases as 3 or 4.

In the existing keyphrase based method [2], the score of a keyphrase is computed as the product of phrase frequency (PF) and inverse document frequency (IDF) when the length of keyphrase is 1. Otherwise, the score is computed as the product of phrase frequency (PF) and the logarithmic value of total number of documents in a given corpus.

But, after the proposed modification in keyphrases selection process, the score of each keyphrase is computed using (1):

$$SCORE_{pf*idf} = \begin{cases} 0, if\ plen = 1 \\ PF * log(N), if\ plen > 1 \end{cases} \quad (1)$$

where *Plen*: length of phrase in terms of words, *PF*: frequency of a phrase, *N*: the total number of documents in the corpus (a collection of documents in a domain under consideration). In (1), the score of keyphrase is set to 0 if the *Plen* is equal to 1 so that single word keyphrases will be ignored in the ranking.

In the keyphrase based method [2], sentences are primarily selected based on top ranked keyphrases (rank is calculated

using (1)) and sentences are finally selected based on sentence' score (calculated using term-frequency and position feature). But as per observation, the way of sentence selection and scoring can be optimized for the betterment of summary preparation which is discussed in the next two points *B* and *C*.

B. Considering the first sentence specially

In the existing keyphrase based method [2], the sentence score is depended on position and term-frequency. The positional score is the highest for the first sentence and the lowest for the last where the score is gradually decreasing from the first sentence. But in most of the time for news documents, the first sentence is much important than any other sentences as per our experiment which is explained in the lower part of this sub-section. So, general positional score (which is gradually decreasing) is not applicable for the first sentence of news documents. Again, some existing summarization methods emphasized on sentences those contain any title word [4, 8] and in news documents, the first sentence contains the full title often. So, an extra care is proposed here for the first sentence of the input document.

In the experiment with our training dataset (200 documents and 600 model summaries), it has been found that the first sentence is existed in the summary for 78% times. So, if the first sentence is always kept in summary, there will be wrong selection for 22% (100 - 78) times. But, after scrutinizing one step ahead, it has been found that if the first sentence contains any title word, it is existed in summary for 88% times where error rate is 12% (100-88). So, it is proposed here that the first sentence is selected in summary if it contains any title word.

C. Counting numerical figure presented in words and digits for sentence scoring

A new feature (counting numerical figure presented in words and digits) is recommended here for sentence scoring as an enhancement. In [13] numerical figure (in digits) was counted and shown that a sentence can be significant for containing numerical figure. But, the numerical figure can be presented in words which can't be identified easily like digits. We may consider the following two sentences for example, "করিমের জন্ম সাল ২০০৬। তাহার বয়স দশ বছর।" (korimer jonmo shal 2006। tahar boyosh dosh bochhor - Karim's birth year is 2006. He is ten years old.). Existing procedure [13] can find 1 numerical figure from the first sentence and unable to find any numerical figure from the second sentence as the numerical figure "দশ" (dosh - ten) is presented in words. So, a technique is introduced here to recognize numerical figure from both words and digits by checking the following conditions:

a) First part of the word is constituted with the following: ০(0), ১(1), ২(2), ৩(3), ৪(4), ৫(5), ৬(6), ৭(7), ৮(8), ৯(9) or "এক" (ek-one), "দুই" (dui - two), "তিন" (tin - three) … … … "আটানব্বই" (atanobboi – ninety eight), "নিরানব্বই" (niranobboi – ninety nine). While checking numerical figure from digits, decimal point (.) is also considered.

b) In the second part (if any), it contains "শত" (shoto - hundred), "হাজার" (hazar - thousand), "লক্ষ" (lokkho – million), etc.

c) The third part (if any) is suffix like "খানা" (khana-that), "খানি" (khani-that), "টি" (ti - this), "টা" (ta - this), etc.

If any word meets these three conditions, the word is tagged as numerical figure. We have experimented on 200 test documents for the proposed technique and found that numerical figure can be identified for 100% and 92% those are presented in digits and words respectively. The score for the existence of numerical figure is counted using the following equation:

$$S_{Nf} = \sum N_{digits} + \sum N_{words} \quad (2)$$

where S_{Nf} is the score of sentence for the existence of numerical figure in digits (N_{digits}) and words (N_{words}).

After considering all the proposed enhancements, the sentence score is computed finally as follows:

$$W_{final(i)} = W_i + S_{Nf(i)} \quad (3)$$

where $W_{final(i)}$ is the final score of the i-th sentence, W_i: the computed score of the i-th sentence according to the existing keyphrase based method using term-frequency and position feature [2], $S_{Nf(i)}$: the score for the existence of numerical figure based on (2). It is noticeable that incorporation of score for the numerical figure has upgraded the performance significantly (shown in section IV).

The remarkable point is that position of sentence and text title have been considered in several existing methods [4, 12]. And, numerical figure (in digits) has already been counted in [13]. But, to the best of our knowledge, (i) selecting the first sentence if it contains any title word and (ii) counting numerical figure from words for sentence scoring, have not been proposed in any existing method. So, it can be said that the proposed enhancements have brought something different.

IV. EVALUATION AND RESULTS

A. Dataset

From Bangla daily newspapers, 400 news documents (each document has 18 to 25 lines of Unicode text) have been collected as test corpus. These news documents contain variety of news that covers a wide range of topics like political, sports, crime, economy, environment, etc. Three human judges have generated summaries for each document. These human generated summaries are considered as reference/model summaries. These 400 documents-summaries are divided into two datasets as (i) randomly selected 200 documents with corresponding model summaries are taken as training set and (ii) other 200 documents with corresponding model summaries are treated as performance evaluation set. The evaluation set has been uploaded to internet so that other researchers may use this [18].

We have used human generated model summaries as there is no bench mark dataset for evaluating Bangla text summarization. Again, the dataset of 400 test documents is

around ten times larger than the evaluation dataset of some existing methods [2, 7, 9]. Moreover, some existing methods [2, 7, 9] were evaluated against one model summary only. But the proposed method is evaluated here with three model summaries of each test document. Here, someone may raise question for using human generated model summaries. But, the remarkable point is that human generated model summaries were also used for English text summarization methods despite the existence of bench mark dataset [19, 20] and for other languages where there was no bench mark dataset [7, 9, 11].

B. Evaluation

Evaluating the quality of a summary is a difficult problem, principally because there is no ideal summary [21]. For relatively straightforward news documents, human summarizers tend to agree only approximately 60% content overlapping [21]. In our proposed method, Precision, Recall and F-measure are brought into play as these have long application as important evaluation matrices in information retrieval field [22]. 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:

$$Precision\ (P) = \frac{A \cap B}{A} \tag{4}$$

$$Recall\ (R) = \frac{A \cap B}{B} \tag{5}$$

$$F\text{-}measure = \frac{2 \times P \times R}{P+R} \tag{6}$$

C. Experiments and results

We have developed the keyphrase based method [2] and incorporated the proposed enhancements with a server side scripting language named PHP (Hypertext Preprocessor). In the existing method [2], summary length is specified by user but in our implementation, one third sentences are selected as final summary. Performance has been measured after incorporating each proposed feature and clearly shown the step by step progress of performance in the fig. 1. Based on the result in fig. 1, it is apparent that every incorporated feature is important for better performance. Here, Precision, Recall and F-measure have been calculated using (4), (5), and (6) respectively upon training dataset (discussed in the beginning of this section).

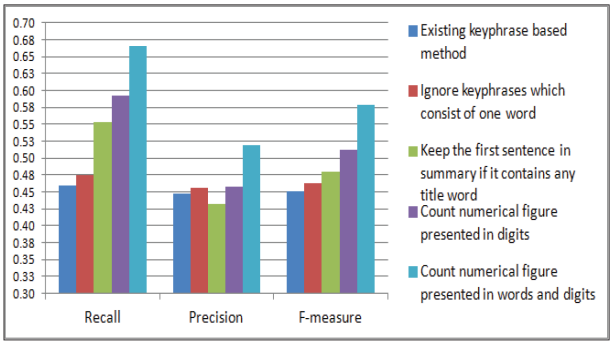

Fig. 1. Step by step improvement of performance for enhancements

TABLE I. COMPARISON ON THE BASIS OF ROUGE-1 SCORES FOR 200 DOCUMENTS WITH 95% CONFIDENCE INTERVAL

	Avg_Recall	Avg_Precision	Avg_F_measure
Proposed method	0.6819	0.5757	0.6166
Keyphrase based method [2]	0.5515	0.5603	0.5496

TABLE II. COMPARISON ON THE BASIS OF ROUGE-2 SCORES FOR 200 DOCUMENTS WITH 95% CONFIDENCE INTERVAL

	Avg_Recall	Avg_Precision	Avg_F_measure
Proposed method	0.6433	0.5459	0.5830
Keyphrase based method [2]	0.5075	0.5165	0.5060

In the fig. 1, the utilized features in each step of improvement include all the features of previous step(s) and better performance is obtained by combining all the proposed enhancements. Again, comparison has been turned between the existing method and its' enhanced version using the evaluation dataset of 200 documents of Bangla Unicode text [18]. ROUGE automatic evaluation package has been utilized here as it can be applied for Unicode text [23]. Average Recall, Precision and F-measure have been calculated based on ROUGE-1 and ROUGE-2 scores and displayed in table I and table II respectively. In the evaluation, it has been found that result of the keyphrase based method has been varied from the result claimed by the author in his paper [2] as it is evaluated with different set of data.

For the simulation, the proposed method and the existing method [2] have been implemented with a server side scripting language. Same list of stop words [24] have been used for implementing both the methods. Based on the evaluation results in table I and table II, it can be said that the method has shown better performance after incorporating the proposed enhancements (a. ignoring the single word keyphrases, b. considering the first sentence specially and c. counting numerical figure in words and digits for sentence scoring).

V. CONCLUSION AND FUTURE WORKS

A keyphrase based Bangla text summarization method has been investigated here in depth and proposed three enhancements. Explanation has been given for each proposed enhancements to clarify the importance. Step by step progress of performance has been demonstrated in the evaluation section. Moreover, an overview of several automatic Bangla text summarization methods has been given in the introduction part of this paper. In the overview, it has been indicated with reference that most of the incorporated features in various existing methods of Bangla text summarization were collected from the methods of English text. In this regard, the two introduced features for enhancements (i. considering the first sentence if it contains any title word and ii. counting numerical figure from words) have brought something different. Finally, it has been shown on the basis of ROUGE-1 and ROUGE-2 evaluation scores that the system after enhancements is performing better than the existing system.

Here, the enhancements have been proposed only for Bangla text summarization. In future, we hope to introduce more features for important sentence identification and adapt

the proposed enhancements for both Bangla and English text summarization.

Acknowledgments

This research work is funded by a Fellowship Scholarship from Information and Communication Technology Division, Government of the People's Republic of Bangladesh. There is also a valuable support from the Central Bank of Bangladesh.

References

[1] Kunder, M., "The size of the world wide web," online available at: www.worldwidewebsize.com/? (last accessed May-2016).

[2] Kamal Sarkar, "A Keyphrase-Based Approach to Text Summarization for English and Bengali Documents", International Journal of Technology Diffusion (IJTD), vol. 5, issue 2, pp. 28-38, April 2014.

[3] Hans P. Luhn, "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, vol. 2, no. 2, pp. 159-165, 1958.

[4] H. P. Edmundson, "New Methods in Automatic Extracting," Journal of the Association for Computing Machinery, vol. 16, no. 2, pp. 264-285, April 1969.

[5] Md. Majharul Haque, Suraiya Pervin, and Zerina Begum, "Literature Review of Automatic Multiple Documents Text Summarization," International Journal of Innovation and Applied Studies, vol. 3, no. 1, pp. 121-129, May 2013.

[6] Md. Majharul Haque, Suraiya Pervin, and Zerina Begum, "Literature Review of Automatic Single Document Text Summarization Using NLP," International Journal of Innovation and Applied Studies, vol. 3, no. 3, pp. 857-865, July 2013.

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

[8] Md. Iftekharul Alam Efat, Mohammad Ibrahim, and Humayun Kayesh, "Automated Bangla Text Summarization by Sentence Scoring and Ranking," International Conference on Informatics, Electronics & Vision (ICIEV), IEEE, pp. 1-5, 2013.

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

[10] Banglapedia, the national Encyclopedia of Bangladesh, Asiatic Society of Bangladesh, Dhaka, 2003.

[11] Aqil M. Azmia and Suha Al-Thanyyan, "A text summarizer for Arabic," Journal of Computer Speech & Language, Elsevier, vol. 26, issue 4, pp. 260–273, 2012.

[12] Md Tawhidul Islam and Shaikh Mostafa Al Masum, "Bhasa: A Corpus-Based Information Retrieval and Summariser for Bengali Text," In Proceedings of the 7th International Conference on Computer and Information Technology, 2004.

[13] Md. Nizam Uddin and Shakil Akter Khan, "A Study on Text Summarization Techniques and Implement Few of Them for Bangla Language," 10th International conference on Computer and Information technology, IEEE, pp. 1-4, 2007.

[14] Turney, P. D., "Learning algorithms for keyphrase extraction," Information Retrieval, vol. 2, no. 4, pp. 303–336, 2000.

[15] Wu, Y. F. B. and Li, Q., "Document keyphrases as subject metadata: Incorporating document key concepts in search results," Information Retrieval, vol. 11, no. 3, pp. 229–249, 2008.

[16] D'Avanzo, E. and Magnini, B., "A keyphrase-based approach to summarization: The LAKE system at DUC-2005," In Proceedings of DUC, 2005.

[17] Hamzah Noori Fejer and Nazlia Omar, "Automatic Arabic Text Summarization Using Clustering and Keyphrase Extraction," International Conference on Information Technology and Multimedia (ICIMU), Putrajaya, Malaysia, pp. 293-298, November 18 – 20, 2014.

[18] Bangla Natural Language Processing Community, "Dataset for Evluating Bangla Text Summarization System," online available at: http://bnlpc.org/research.php (last accessed September-2016).

[19] Rafael Ferreira, Frederico Freitas, Luciano de Souza Cabral, Rafael Dueire Lins, and Rinaldo Lima, "A Four Dimension Graph Model for Automatic Text Summarization," IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT), pp. 389-396, 2013.

[20] Jingqiang Chen and Hai Zhuge, "Summarization of scientific documents by detecting common facts in citations," Future Generation Computer Systems, Elsevier, vol. 32, pp. 246–252, 2014.

[21] Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown, "Introduction to the special issue on summarization," Journal of Computational Linguistics, MIT Press, vol. 28, no. 4, pp. 399-408, December 2002.

[22] Shanmugasundaram Hariharan, Thirunavukarasu Ramkumar, and Rengaramanujam Srinivasan, "Enhanced Graph Based Approach for Multi Document Summarization," The International Arab Journal of Information Technology, vol. 10, no. 4, July 2013.

[23] ROUGE 2.0 - Java Package for Evaluation of Summarization Tasks with Updated ROUGE Measures, online available at: http://kavita-ganesan.com/content/rouge-2.0 (last accessed May-2016).

[24] Indian Statistical Institute, "List of stop words for Bengali language," online available at: http://www.isical.ac.in/~fire/data/stopwords_list_ben.txt (last accessed May-2016).