Automatic Bengali News Documents Summarization by Introducing Sentence Frequency and Clustering

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

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

***Abstract*—A method has been proposed in this paper for Bengali news documents summarization which extracts significant sentences using the four major steps (a) preprocessing, (b) sentence ranking, (c) sentence clustering, and (d) summary generation. The noticeable feature of this method is the incorporation of the sentence frequency where redundancy elimination is a consequence. Another one remarkable aspect is sentence clustering on the basis of similarity ratio among sentences. The summary sentence selection is done from all the clusters so that there will be maximum coverage of information in summary even if information is found scattered in input document. Two sets of human generated summary have been utilized where one is to train the system and another is for performance evaluation. The proposed method has been found better while turning comparison with the latest state-of-the art method of Bengali news documents summarization. The results of performance evaluation show that the average Precision, Recall and F-measure values are 0.608, 0.664 and 0.632 respectively.**

Keywords—Documents summarization; sentence clustering; sentence frequency; redundancy elimination; similarity ratio

I. Introduction

The amount of available information increases rapidly with the development of information technology and wide use of Internet [1] for which a new era of information explosion is impending. The estimated size of the web in 2013 was around 3.82 billion pages [2] and this number is growing every day at a fast pace [3]. So the automatic summarization is needed to process the Internet data efficiently, scavenging useful information from it [3]. The goal of automatic text summarization is to condense the source text into a shorter version with preserving its information content and overall meaning [4, 5].

To this age, numerous types of research work have been accomplished by various researchers where we can be familiar with multiple ways of summary generation for English language [6]. But, a few works exist for Bengali language where most of them are somehow works on directly term-frequency and some other statistical measures [7].

However, a purely statistical method of producing extracts was suspected of being inadequate, and hence other methods were sought [8]. Again the term frequency based method will only select the sentences that contain frequent terms and the summary will contain similar sentences only. But, it may be happened that some significant sentences may exist in the document that will not contain frequent terms for which they will be discarded from selection.

In these regard, a method has been proposed here for Bengali news documents summarization with the following major contributions for which this method is different from others:

- Along with term frequency, this method introduces sentence frequency while sentence ranking. If one sentence contains 60% terms of any other, smaller sentence is removed and the frequency of larger sentence is increased between them.
- Sentences are clustered in different groups and selected from all clusters based on their volume. This feature will maximize the information coverage because of the participation of all the clusters in summary preparation.

The concept of clustering has already been incorporated in English text summarization [9] but in this proposed method the way of utilization with the similarity ratio adjustment is unique. Moreover, this method has applied sentence clustering for Bengali text summarization in the first time.

The rest of the paper is organized as follows: Section II describes related works about Bengali text summarization. Section III illustrates proposed methodology in details. Experiment and evaluation with the discussion on results are depicted in section IV. Finally, the conclusion is turned in section V with future works.

II. Related Work

The voice-over automatic i.e. computerized abstraction began around five decades ago by H. P. Luhn [10] in 1958 on the basis of term-frequency and it was first extended by P. B. Baxendale [11] by incorporating position of sentences and cue-phrases. Since then, the field of text summarization has witnessed continuous involvement of many researchers in the attempt to look for different strategies [5, 12]. H. P. Edmundson [13] in 1969 accomplished a notable progress by integrating title method, cue-phrase method and location method.

But, the noticeable point is that most of these research works have been conducted for English. Few attempts have been reported for Bengali language though it is the 7th international language in the world and mother language for Bangladesh [14]. It is also noticeable that online Bengali text is also increasing rapidly and there are a number of online newspapers such as “The Daily Prothom alo”, “Anandabazar Patrika”, “The Daily Jugantor”, etc. In this regard, the automatic Bengali news documents summarization is out of question.

In 2004, Islam and Masum [15] presented ‘Bhasa’, a corpus oriented search engine and summarizer. It performs document indexing and retrieves information based on key words using vector space retrieval method for Unicode based Bengali text. Corpus files can be ranked and summarized by this method as per the frequent appearance of query terms. A tokenizer has been used here which is able to determine different terms, abbreviations, tags and boundary of sentences, headings, titles and sentences using markups by semantic and syntactic analysis.

In 2007, Md. Nizam Uddin and Shakil Akter Khan [16] accomplished a survey on English text summarization system and implemented some existing features to summarize Bengali text. The features are 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. It has been found that 40% extract by this system has got the point 8.4 from human in the range of 0 to 10.

In 2010, Amitava Das and Sivaji Bandyopadhyay [17] offered a method for opinion summarization in Bengali. They have utilized subjectivity classifier [18] to determine subjective or factual sentences or documents for opinion mining. The system identifies the sentiment information in each document, aggregates them and represents the summary information in text. The aggregation is performed by using k-means approach and candidate summary sentences are selected by applying theme relational graph model. Standard page rank algorithm has been used here. In evaluation, the Precision, Recall and F-Score of this approach are calculated as 72.15%, 67.32% and 69.65% respectively. This system is mostly works on theme detection.

Somewhere, same procedure as like English text summarizer has been followed for Bengali language as proposed by Kamal Sarkar [7] in 2012. This is an easy-to-implement approach as like the method of Edmandson [13]. It has three major steps: (1) preprocessing, (2) sentence ranking, and (3) summary generation. This method is based on word-frequency, length of sentence and position of sentences for sentence scoring. The evaluation of this method has shown that average unigram based recall score is 0.4122.

In 2012, Kamal Sarkar [19] proposed another one method in the aim to provide an idea about the theme of a document without revealing the in-depth detail. This approach has four major steps (1) preprocessing, (2) extraction of candidate summary sentences, (3) ranking the candidate summary sentences, and (4) summary generation. This is also based on word-frequency, sentence position and sentence length that is similar to [7]. It was claimed that the features have been used here in more effective way for news documents summarization than [7]. Evaluation results showed that this system performs better than the lead baseline, baseline that uses term-frequency with position features and the method described in [7].

To the best of our knowledge, the method described in [19] is comparatively latest and better than any other state-of-the art methods for news documents summarization for Bengali language. In the evaluation section, the proposed method in this paper has been compared with the method described in [19].

III. Proposed Methodology

The proposed text summarization approach is described in the four steps as follows:

A. Preprocessing

At the preprocessing step, stop words such as “এ”, “এবং”, “আর”, etc. are removed as per the list of Bengali stop words [20]. Word stemming is applied to map the words with different endings to a single one such as "গ্রামের", "গ্রামে" will be "গ্রাম". For stemming procedure, lightweight stemmer for Bengali has been used that strips the suffixes using a predefined suffix list on a “longest match” basis [21]. For some words that cannot be stemmed with suffix stripping rules such as “কচলালেন”, look up table has been used as in [22] to get the root form. Bengali is very inflectional language for which stemming is required for calculating term frequency.

After stemming, term frequency for all terms is computed and the entire document is segmented to sentences. As per the analysis, it has been found that the sentences with length of less than or equal to 4 are very rare to be in summary, so they are deleted [19].

B. Sentence Ranking

For sentence scoring, values of some attributes are calculated for all the sentences at first and then sum-up all the attributes’ value to compute the score of each sentence. Three attributes are considered in this method as follows: 1) term frequency calculation for each sentence, 2) sentence frequency, and 3) existence of numerical data.

1) Term frequency calculation for each sentence (S_{TF}): It is well known that term frequency (TF) is the number of appearance of any term. It is estimated as follows:

$$\mathrm{TF(t)} = \frac{Number\ of\ times\ term\ t\ appears}{Total\ number\ of\ terms\ in\ document} \tag{1}$$

First attribute as term frequency for one sentence (S_{TF}) is calculated by summing up the TF of all the terms exist in the sentence.

2) Sentence frequency: Along with term frequency, this proposed methodology has introduced second attribute as sentence frequency (S_{SF}) which is based on the ratio of term overlapping. In this method, all the sentences are set to frequency 1 at first. If one sentence contains 60% or above terms of any other, the smaller sentence is removed and the frequency larger sentence will be computed as the summation of the frequency of both of the sentences. The containing ratio 60% is considered based on the threshold value of cosine similarity ratio [23].

3) Existence of numerical data: The third attribute is to count numerical data in each sentence (S_{Nc}). The value of S_{Nc} for each sentence is set to 0 (zero) at first and for the existance of each numerical data it will be incremented by 1.

After measuring all the attributes, the score of each sentence is computed as (2) where S_k is the score of K_{th} sentence:

$$S_k = S_{TF} + S_{SF} + S_{Nc} \qquad (2)$$

C. Sentence clustering

All the significant sentences may not contain frequent terms and not be similar to a central theme of the input document. So, coverage of information may be failed by applying the sentence selection process directly to the whole document. By considering this issue, it is proposed here that the sentences are clustered as per their cosine similarity ratio at first. If cosine similarity ratio among two or more sentences is equal to a minimum threshold point or above, they will be in a single cluster. This threshold point is adjusted on the training corpus for better result through experiments. The average F-measure for summarization performance is computed for the summaries of 15 test documents while clustering with different threshold points. The fig. 1 shows F-measure value by clustering from 0.00 to 0.59 similarity ratio (SR) and in each time the SR is incremented by 0.01. It is found that SR equal to 0.09 is the best minimum threshold point.

The range of SR is selected here from 0.00 to 0.59 for clustering. Because, 0.00 means no clustering and 0.59 is the highest limit of clustering as more than 0.59 SR is assumed as repetition while sentence frequency calculation. Actually in different threshold point, various numbers of clusters are generated for various documents for which the performance of summarization is varied. It has been found that the average value of F-measure is highest with the number of clusters constructed for SR equal to 0.09.

In this way all the sentences are clustered. But some sentences are there with no similarities with any other and keep these sentences to another one cluster. From the fig. 1, it has been found that the performance of clustering by the SR equal to 0.0 and from 0.50 to 0.59 are almost similar because most of the sentences are found with no clustering with SR from 0.50 to 0.59 and tends to be as like SR 0.00 or no clustering.

D. Summary generation

After clustering, sentences are selected from each cluster based on the volume of cluster. For example, if there are N sentences in the whole document and one cluster has C sentences and if summary will be of S sentences, the number of top scored sentences from each cluster (Ns) are selected as follows:

$$\mathrm{Ns} = \frac{\mathrm{C} \times \mathrm{S}}{\mathrm{N}} \qquad (3)$$

The number of summary sentences is kept as approximately one third of the total sentences according to the ratio of source document to summary mentioned in [24]. Now, all the selected sentences are ordered as per their order of appearance in the original document to display the final summary.

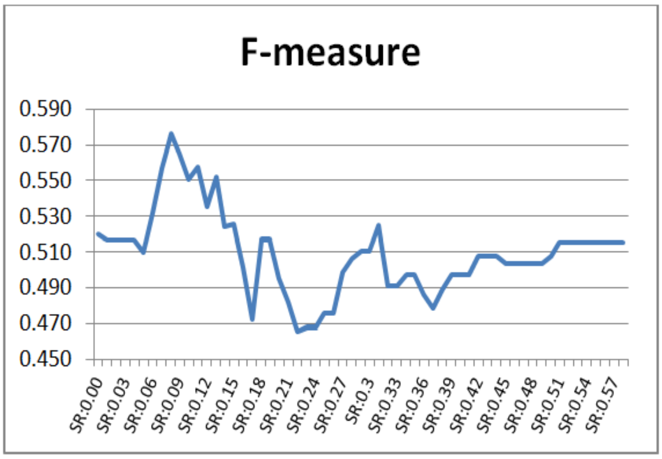


Fig. 1. Effect on summarization performance when the similarity ratio for clustering is varied.

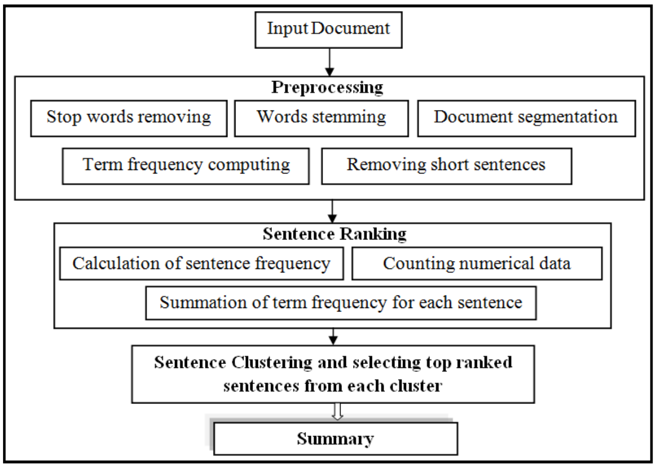


Fig. 2. Process flow of the proposed methodology.

At a glance process flow of the proposed method is given in the above fig. 2.

IV. Evaluation and Discussion on Results

For the evaluation of the proposed methodology, 40 news articles have been collected as test corpus from Bengali daily newspaper “The Daily Prothom alo” and “The Daily Jugantor”. From the test corpus, 20 news articles have been selected randomly. Three human (graduated in Computer Science and Engineering) generated summaries for each article and these summaries are considered as reference/model summaries. From these 20 document-summaries, randomly selected 15 document-summaries are considered as training set for adjusting the value of similarity ratio (SR) in the previous section. Other 5 document-summaries pairs are used for the evaluation of this proposed system.

A. Evaluation

Evaluating the quality of a summary is a difficult problem, principally because there is no obvious “ideal” summary and for relatively straightforward news articles, human summarizers tend to agree only approximately 60% content overlapping [24]. Even, summary generated by same person may be varied in different times for same article. In this regard, the summary of proposed system has been compared with three model summaries of 5 news articles and the results of evaluation is the average results of the comparisons. Precision, Recall and F-measure are brought into play here as these have long used as important evaluation matrices in information retrieval field [25]. 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:

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

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

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

B. Experiments and results

The proposed method has been implemented along with an existing Bengali text summarization method [19] with a server side scripting language named Hypertext Preprocessor (PHP). To judge the effectiveness of the proposed method, experiments have been conducted on several news articles. In each time, the system generated summary is compared with three model summaries of each article and compute the average value of Precision, Recall and F-measure through (4), (5) and (6) respectively. The proposed system has also been compared with an existing Bengali news documents summarization method [19] which was claimed to be better than two baseline systems and another one system illustrated in [7]. Point to be mentioned that same documents with corresponding model summaries have been utilized to calculate Precision, Recall and F-measure for the proposed method and the existing method of [19]. The results of evaluation and comparison have been depicted in the table I and II respectively.

As per the results in table I and II, the proposed method has shown promising outcome. It can be said that the proposed method has become more efficient for using some distinguished features such as sentence frequency, sentence clustering and considering the existence of numerical data in each sentence.

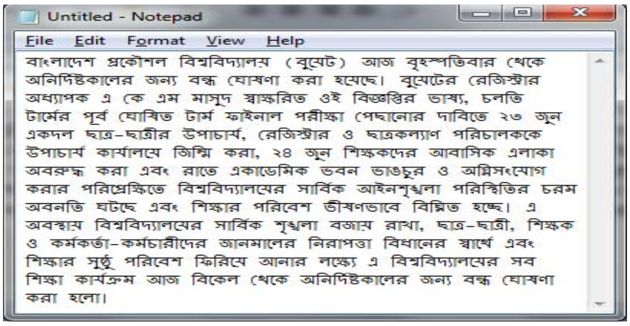
A model summary and a system generated summary have been given for example in the fig. 3 and fig. 4 respectively. In this regard, the input article is “বুয়েট বন্ধ, হল ত্যাগের নির্দেশ” has been taken from a Bengali newspaper “The Daily Prothom-alo”.

TABLE I. EVALUATION OF THE PROPOSED SYSTEM

Articles	P	R	F-measure
Article 1	0.570	0.600	0.590
Article 2	0.780	0.780	0.780
Article 3	0.500	0.670	0.570
Article 4	0.520	0.600	0.550
Article 5	0.670	0.670	0.670
Average	0.608	0.664	0.632

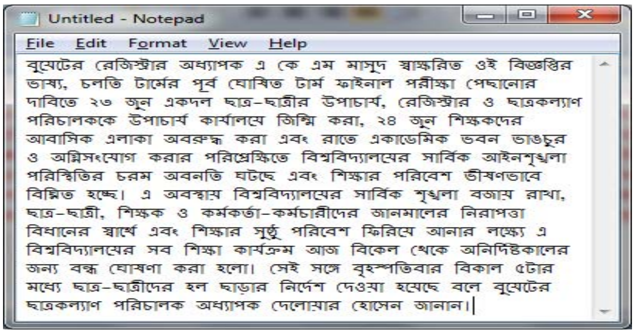
TABLE II. COMPARISON OF THE PROPOSED SYSTEM

Methods	P	R	F-measure
Proposed system	0.608	0.664	0.632
Existing system [19]	0.538	0.556	0.546



বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয় (বুয়েট) আজ বৃহস্পতিবার থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হয়েছে। বুয়েটের রেজিস্ট্রার অধ্যাপক এ কে এম মাসুদ স্বাক্ষরিত ওই বিজ্ঞপ্তির ভাষ্য, চলতি টার্মের পূর্ব ঘোষিত টার্ম ফাইনাল পরীক্ষা পেছানোর দাবিতে ২৩ জুন একদল ছাত্র-ছাত্রীর উপাচার্য, রেজিস্ট্রার ও ছাত্রকল্যাণ পরিচালককে উপাচার্য কার্যালয়ে জিম্মি করা, ২৪ জুন শিক্ষকদের আবাসিক এলাকা অবরুদ্ধ করা এবং রাতে একাডেমিক ভবন ভাঙচুর ও অগ্নিসংযোগ করার পরিপ্রেক্ষিতে বিশ্ববিদ্যালয়ের সার্বিক আইনশৃঙ্খলা পরিস্থিতির চরম অবনতি ঘটছে এবং শিক্ষার পরিবেশ ভীষণভাবে বিঘ্নিত হচ্ছে। এ অবস্থায় বিশ্ববিদ্যালয়ের সার্বিক শৃঙ্খলা বজায় রাখা, ছাত্র-ছাত্রী, শিক্ষক ও কর্মকর্তা-কর্মচারীদের জানমালের নিরাপত্তা বিধানের স্বার্থে এবং শিক্ষার সুষ্ঠু পরিবেশ ফিরিয়ে আনার লক্ষ্যে এ বিশ্ববিদ্যালয়ের সব শিক্ষা কার্যক্রম আজ বিকেল থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হলো।

Fig. 3. Model summary generated by human



বুয়েটের রেজিস্ট্রার অধ্যাপক এ কে এম মাসুদ স্বাক্ষরিত ওই বিজ্ঞপ্তির ভাষ্য, চলতি টার্মের পূর্ব ঘোষিত টার্ম ফাইনাল পরীক্ষা পেছানোর দাবিতে ২৩ জুন একদল ছাত্র-ছাত্রীর উপাচার্য, রেজিস্ট্রার ও ছাত্রকল্যাণ পরিচালককে উপাচার্য কার্যালয়ে জিম্মি করা, ২৪ জুন শিক্ষকদের আবাসিক এলাকা অবরুদ্ধ করা এবং রাতে একাডেমিক ভবন ভাঙচুর ও অগ্নিসংযোগ করার পরিপ্রেক্ষিতে বিশ্ববিদ্যালয়ের সার্বিক আইনশৃঙ্খলা পরিস্থিতির চরম অবনতি ঘটছে এবং শিক্ষার পরিবেশ ভীষণভাবে বিঘ্নিত হচ্ছে। এ অবস্থায় বিশ্ববিদ্যালয়ের সার্বিক শৃঙ্খলা বজায় রাখা, ছাত্র-ছাত্রী, শিক্ষক ও কর্মকর্তা-কর্মচারীদের জানমালের নিরাপত্তা বিধানের স্বার্থে এবং শিক্ষার সুষ্ঠু পরিবেশ ফিরিয়ে আনার লক্ষ্যে এ বিশ্ববিদ্যালয়ের সব শিক্ষা কার্যক্রম আজ বিকেল থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হলো। সেই সঙ্গে বৃহস্পতিবার বিকাল ৫টার মধ্যে ছাত্র-ছাত্রীদের হল ছাড়ার নির্দেশ দেওয়া হয়েছে বলে বুয়েটের ছাত্রকল্যাণ পরিচালক অধ্যাপক দেলোয়ার হোসেন জানান।

Fig. 4. System generated summary

V. CONCLUSION AND FUTURE WORKS

In this paper, a method for summarizing news documents for Bengali language has been proposed by introducing sentence frequency and clustering. A review study has also been portrayed to enumerate the basement of the exploration of automatic text abridgement for Bengali language. As per the results of evaluation, it can be said that the proposed system will help in getting precise information within a comparatively short time.

In future, we hope to introduce more features for sentence ranking. It is also expected to extend the proposed scheme to construct portable and language independent text summarization procedure.

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] Dongmei Ai, Yuchao Zheng, and Dezheng Zhang, “Automatic text summarization based on latent semantic indexing,” Journal of Artificial Life and Robotics, Springer, vol. 15, issue 1, pp 25-29, August 2010.

[2] Kunder, M., “The size of the world wide web,” online available at: www.worldwidewebsize.com/? (last accessed February-2014).

[3] Rafael Ferreira and Luciano de souza, “A multi-document summarization system based on statistics and linguistic treatment,” Journal of Expert Systems with Applications, Elsevier, vol. 41, issue 13, pp. 5780-5787, 1st October 2014.

[4] Yogan Jaya Kumar and Naomie Salim, “Automatic Multi Document Summarization Approaches,” Journal of Computer Science, vol. 8, issue. 1, pp. 133-140, 2012.

[5] V. Gupta and G. S. Lehal, “A Survey of Text Summarization Extractive Techniques,” Journal of Emerging Technologies in Web Intelligence, vol. 2, no. 3, pp. 258-268, August 2010.

[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, “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.

[8] G. J. Rath, A. Resnick and T. R. Savage, “Comparisons of four types of lexical indicators of content,” Journal of the American Society for Information Science and Technology, vol. 12, no. 2, pp. 126-130, April 1961.

[9] ZHANG Pei-ying and LI Cun-he, “Automatic text summarization based on sentences clustering and extraction,” 2nd IEEE International Conference on Computer Science and Information Technology, Beijing, pp. 167-170, August 2009.

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

[11] P. B. Baxendale, “Machine-made Index for Technical Literature -An Experiment,” IBM Journal of Research and Development, vol. 2, no. 4, pp. 354-361, October 1958.

[12] H. Saggion and T. Poibeau, “Automatic Text Summarization: Past, Present and Future,” Multi-source, Multilingual Information Extraction and Summarization, Springer-Verlag, Berlin, Heidelberg, pp. 3-21, 2013.

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

[14] “Banglapedia, the national Encyclopedia of Bangladesh”, Asiatic Society of Bangladesh, Dhaka, 2003.

[15] 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.

[16] Md. Nizam Uddin, 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.

[17] Amitava Das and Sivaji Bandyopadhyay, “Topic-Based Bengali Opinion Summarization”, International Conference COILING ’10, Beijing, pp. 232-240, 2010.

[18] Amitava Das and Sivaji Bandyopadhyay, “Subjectivity Detection in English and Bengali: A CRF-based Approach,” In Proceeding of ICON, Hyderabad, 14th-17th December 2009.

[19] 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.

[20] List of stop words for Bengali language. Online available at: http://www.isical.ac.in/~fire/stopwords_list_ben.txt (last accessed 12 July-2015).

[21] Md. Zahurul Islam, Md. Nizam Uddin, and Mumit Khan, “A light weight stemmer for Bengali and its Use in spelling Checker,” Center for research on Bangla language processing (CRBLP), 2007.

[22] Md. Redowan Mahmud, Mahbuba Afrin, Md. Abdur Razzaque , Ellis Miller, and Joel Iwashige, “A Rule Based Bengali Stemmer,” International Conference on Advances in Computing, Communications and Informatics (ICACCI, 2014.

[23] L. Gravano, P. G. Ipeirotis, H. V. Jagadish, N. Koudas, S. Muthukrishnan, L. Pietarinen, and D. Srivastava, “Using q-grams in a DBMS for approximate string processing,” IEEE Data Engineering Bulletin, vol. 24, issue 4, pp. 28-34, 2001.

[24] 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.

[25] 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.