Automated Bangla Text Summarization by Sentence Scoring and Ranking

Md. Iftekharul Alam Efat
Institute of Information Technology
University of Dhaka
Dhaka-1000, Bangladesh
Email: iftekhar.efat@gmail.com

Mohammad Ibrahim
Institute of Information Technology
University of Dhaka
Dhaka-1000, Bangladesh
Email: ibrahim_iit@yahoo.com

Humayun Kayesh
Institute of Information Technology
University of Dhaka
Dhaka-1000, Bangladesh
Email: hkayesh@gmail.com

***Abstract*—In Natural Language Processing (NLP) the document summarization is an area that is getting interest of modern researchers. Though there are many techniques that have been proposed for English language but a few notable works have been done for Bangla text summarization. This paper deals with the development of an extraction based summarization technique which works on Bangla text documents. The system summarizes a single document at a time. Before creating the summary of a document, it is pre-processed by tokenization, removal of stop words and stemming. In the document summarization process, the countable features like word frequency and sentence positional value are used to make the summary more precise and concrete. Attributes like cue words and skeleton of the document are included in the process, which help to make the summary more relevant to the content of the document. The proposed technique has been compared with summary of documents generated by human professionals. The evaluation shows that 83.57% of summary sentences selected by the system agreed with those made by human.**

I. Introduction

Bangla is one of the most spoken languages in the World and the national language in Bangladesh. Over time the number of Bangla documents is increasing in a large amount. Reviewing these documents and evaluating them from any specific perspective is a gigantic task for a reviewer. It would take a lot time and efforts. Therefore, a system that automates the manual and tiresome process of summarizing documents is necessary. It helps saving a lot of time by reviewing the documents.

English document summarization systems are already there and serving with satisfactory accuracy. But there is no complete system for Bangla document summarization. This can help someone evaluating a large amount of Bangla documents or writings and giving necessary information about the content of the documents. For example, a blogger posted interesting topic in a Bangla blog and got a large number of responses from the readers with thousands of comments. But he does not have enough time to review all those comments. An automated summarization system can help him in this case to get a digest of the responses from the readers. The work presented in this paper is intended to generate an extraction based summary from a Bangla document.

The rest of the paper is organized as follows: In section II, we discuss the previous research works in this area. Next in section III, we describe our proposed method for text summarization technique. Sentence scoring and summarization with pre-processing has been described in this section. Section IV illustrates the experimental results and discussion. Section V concludes the paper and provides direction for future work.

II. Related Work

The earliest work on single-document summarization proposed the frequency of a particular word in a document to be a useful measure of significance described by Luhn in [1]. Though Luhns methodology was a preliminary step towards the summarization, but many of his ideas are still found to be effective for text. In the first step, all the stop words were removed and rest of the words were stemmed to their root forms. A list of content words then compiled and sorted by decreasing frequency, the index providing an important measure of the word. From every sentence a significance factor was extracted that reflects the number of occurrences of significant words within a sentence, and correlation between them is measured due to the intervention of non-significant words. All the sentences are positioned in order to their significance factor, and the top positioned sentences are finally selected to form the automatically generated abstract.

Jing presented a sentence reduction system for removing irrelevant phrases like prepositional phrases, clauses, to-infinitives, or gerunds from sentences [2]. Their main contribution is determining less important phrases in a sentence using reduction decisions. The reduction decisions are based on syntactic knowledge, context, and probabilities computed from corpus analysis.

Edmundson et al. proposed a typical structure of text summarization methodology in [3]. They incorporated both word frequency and positional value ideas generated by two of his previous works. The first of two other features used was the presence of cue words (occurrence of words like significant, or hardly), and the second one was the skeleton of the document (the sentence is a title or heading). The sentences were scores based on these features to extract sentences for summarization.

All these research works are conducted for English, however the same procedure can be followed for other languages (e.g. Bangla) as proposed by Kamal Sarkar [4]. This approach used

978-1-4799-0400-6/13/$31.00 ©2013 IEEE

word frequency and sentence position in the document as significant features to rank the sentences in the document.

Another work on Bangla text summarization proposed by Amitava and Sivaji used features such as Part of Speech (POS), Title Words, First Paragraph Words, Words from Last Two Sentences, etc. [5]. They used theme clustering to create a reasonable set of clusters for a given set of documents and as a way of extracting sentences based on their importance which regulates the quality of the output summary.

Word stemming and its implementation are relatively easier in English text. In Bangla document, there are number of complex sentence available. The steaming of words from these complex sentences is too much difficult. Jubayer and Masud have proposed and implemented a method for building Bangla text corpus for Information Retrieval (IR) purposes [8]. They considered several criteria in their system, like priority based term frequency, random walk on graph algorithm, making metadata, etc to develop corpus.

Tawhidul and Mostafa proposed Bhasa, a corpus-based search engine and summarizer [9]. It uses vector space retrieval method on key words to perform document indexing and retrieving information. Bhasa prioritizes the corpus file based on terms frequency. The system used a tokenizer which is capable of detecting different words, tags, abbreviation, etc and then performed document ranking to summarize the tokenised document.

III. Proposed Method

The Bangla document summarizer is a Natural Language Processing (NLP) application which is proposed to extract the most important information of the document(s). In automatic summarization, there are two distinct techniques either text extraction or text abstraction. Extraction is a summary consisting of a number of sentences selected from the input document(s). An abstraction based summary is generated where some text units are not present into the input document(s). With extraction based summary technique, some more features are added based on Information Retrieval. However, the total system is alienated into three segments: pre-processing the test document, sentence scoring based on text extraction and summarization based on sentence ranking.

Input to a summarization process can be one or more text documents. When only one document is the input, it is called single document text summarization but in multi-document summarization the input is a group of related text documents. The text summarization can also be classified based on the types of users the summary is wished-for: User focused (query focused) summaries are adapted to the requirements of a particular user or group of users and generic summaries are aimed at a broad community of readers [6].

A. Pre-processing

In Bangla document summarization process, some pre-processing is needed before executing the sentence scoring algorithm. By the pre-processing, the documents are prepared for ranking and summary generation. The pre-processing done on the documents are as follows:

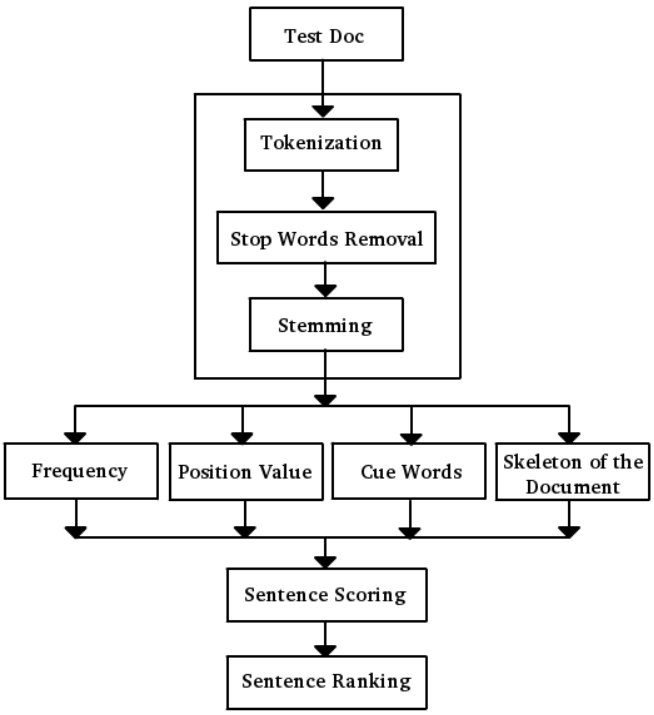


Fig. 1. Steps of the proposed text summarization technique

Tokenization A document is the combination of sentences and a sentence consists of some words. Here every word is considered as a token. A document is treated as a chain of tokens (marks).

Stop words removal In Bangla words like এবং(And), অথবা(Or), কিন্তু(But), etc. are used frequently in sentences which have little significance in the implication of a document. These words can simply be removed for classification process.

Stemming – A word can be found in different forms in the same document. These words have to be converted to their original form for simplicity. The stemming algorithm is used to transform words to their canonical forms, like বাংলাদেশে, বাংলাদেশের, বাংলাদেশকে, বাংলাদেশেও,etc. should be converted to their original form বাংলাদেশ. In this work, we use a lightweight stemmer that splits a word into its root form using a predefined suffix list [7].

B. Sentence Ranking and Summarization

After an input document is tokenized and stemmed, it is split into a collection of sentences. The sentences are ranked based on four important features: Frequency, Position value, Cue words and Skeleton of the document.

Frequency – Frequency is the number of times a word occurs in a document. If a word's frequency in a document is high, then it can be said that this word has a significant effect on the content of the document. The total frequency value of a sentence is calculated by sum up the frequency of every word in the document. The equation used to estimate the total frequency value of a sentence k is:

$$STF_k = \sum_{i=1}^{n} W_F \quad (1)$$

Where W_F (Word Frequency) is the total frequency of a word in the document, n is the number of words in a sentence and STF stands for Sentence Total Frequency.

Positional Value – The position of a sentence in a document has a considerable influence over the content of the document. The positional value of a sentence is computed by assigning the highest value to the first sentence and the lowest value to the last sentence of the document. The position value PV is calculated using the formula:

$$PV_k = \frac{1}{\sqrt{k}} \quad (2)$$

Where, k is the actual positional value of a sentence in the document.

Cue Words - Cue words are connective expressions (such as *therefore, hence, lastly, finally, meanwhile* or on the other hand) that links spans of communication and signals semantic relations in a text. This is one of the summarization strategies which involve the use of "Cue Words" to select important sentences. The examples of "Cue Words" in Bangla are মোটকথা, অবশেষে, ইতিমধ্যে, যেহেতু, etc.

Skeleton of the Document - The skeleton of the document consists of the words in titles and headers. These words are considered having some extra weights in sentence scoring for summarization.

Sentence Scoring - The final score is a Linear Combination of frequency, positional value, weights of Cue Words and Skeleton of the document. The formula used to produce the final score of a sentence k is as follows:

$$S_k = (\alpha \times STF_k) + (\beta \times PV_k) + \gamma + \lambda, 0 \leq \alpha, \beta, \gamma, \lambda \leq 1 \quad (3)$$

Where, α and β are two co-factors of Sentence Total Frequency and Positional Value respectively. On the other hand, γ and λ symbolizes the weights of Cue Words and Skeleton of the documents correspondingly. The values of α, β, γ and λ are 0 to 1.

Summary Making - After ranking the sentences based on their total score the summary is produced selecting X number of top ranked sentences where the value of X is provided by the user. For the readers' convenience, the selected sentences in the summary are reordered according to their original positions in the document.

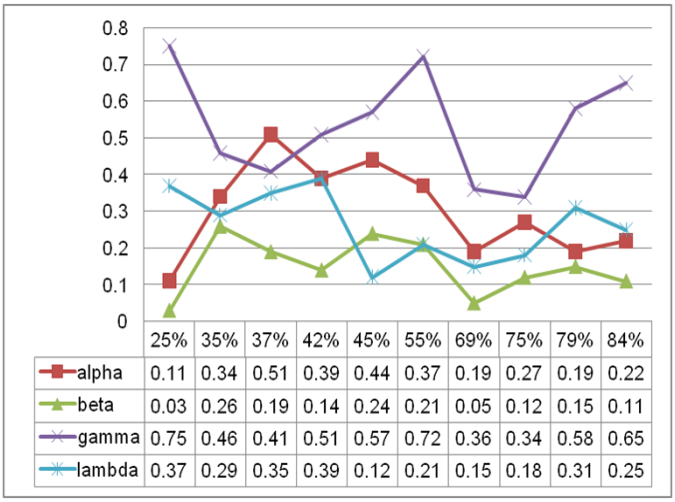
IV. Experimental Results and Discussion

As the text summarization in Bangla is a new field of research, there exists no standard dataset in this area. This is why, to test the accuracy of our summarization system, we collected 45 Bangla articles from different Bangla newspapers such as The Daily Prothom-alo, The Daily Ittefaq, The daily Jugantor, etc. The documents are typed and saved in the text files using UTF-8 format. For each document we consider only one reference summary generated by human professionals for evaluation. Evaluation of a system generated summary is done by comparing it to the reference summary.

Although it is difficult to summarize a document automatically according to human summarization technique, we identified and prioritized properties that are required to achieve an effective automated summarization technique. In our summarization technique, the total frequency of a sentence is more important than its positional value. If any cue word exists in the sentence then we need to consider it with high priority as a summary sentence.

For the most excellent results, it needs a fine tune of appropriate threshold value of the coefficient $(\alpha, \beta, \gamma, \lambda)$ factors. We have chosen 10 random documents from the 45 test documents those we had used to tune these parameters. Initially, we set the value of α to 0.1 as it only multiply with the total frequency of a sentence, so we give this value a little weight whereas the cue word feature γ is primarily given a weight of 0.7. Beside this, we also consider that positional co-factor β to 0.2 as the important sentences are naturally kept in the early portion of the document. To include the scoring more efficiently the last co-factor λ is weighted 0.4 as it compare to the documents with the headlines structure which generally moves to the most important key words of the document.

We implemented our summarization technique on the test data sets to tune the parameters by summarizing those documents and compared the accuracy with human summarization process. Figure 2 shows that the values of the co-factors are changed in the range of the primary assigned weight to find the accuracy. From the graph it is seen that the α curve moves like a sine curve where the amplitude is decreasing.



	25%	35%	37%	42%	45%	55%	69%	75%	79%	84%
alpha	0.11	0.34	0.51	0.39	0.44	0.37	0.19	0.27	0.19	0.22
beta	0.03	0.26	0.19	0.14	0.24	0.21	0.05	0.12	0.15	0.11
gamma	0.75	0.46	0.41	0.51	0.57	0.72	0.36	0.34	0.58	0.65
lambda	0.37	0.29	0.35	0.39	0.12	0.21	0.15	0.18	0.31	0.25

Fig. 2. Performance vs Co-factors graph

It means that the Sentence Total Frequency has only 22% impact of its total frequency in Bangla text summarization technique. In the other hand, β factor has not changed much from the range and almost saturated in the more accurate points. It has also 11% impact of its original positional value corresponding to this document. To contrast, the λ factor was initially changed

TABLE I
WEIGHT OF CO-FACTORS OF SENTENCE SCORING

Co-factor	Value
α (alpha)	0.22
β (beta)	0.11
γ (gamma)	0.65
λ (lambda)	0.25

TABLE II
ACCURACY MEASUREMENT WITH F_1 VALUE

Doc No.	N	k_h	k_m	r	$F_1 = 100\frac{2r}{k_h+k_m}$
1	172	31	35	25	75.76%
2	157	29	32	26	85.25%
3	166	35	34	30	93.75%
4	184	34	37	29	81.69%
5	145	32	29	23	75.41%
6	191	42	39	34	83.95%
7	178	32	36	31	91.18%
8	169	39	34	29	79.45%
9	188	35	38	33	90.42%
10	183	34	37	28	78.87%

due to varieties of documents types but finally in the end point it acts like an inundated point. The most frequently changed value of co-factor is γ because weighting the cue words in a small dictionary is harsh to determine. The value of γ is comparatively high (0.65) considering with other co-factors because line which contains titles and headers' words has more probability be selected as a summary sentence. Finally, we have set the weight of the co-factors are given in Table I.

We compared our algorithms' summaries with the human summaries, computing the following scores. For each document we let k_h be the length of the human summary, $k_m$the length of the machine generated summary and the r the number of sentences they share in common. We defined precision (P), recall (R) and F_1 as metrics to compare the two summaries by:

$$P = 100\frac{r}{k_h} \tag{4}$$

$$R = 100\frac{r}{k_m} \tag{5}$$

$$F_1 = 100\frac{2PR}{P+R} = 100\frac{2r}{k_h+k_m} \tag{6}$$

To measure the accuracy of our algorithm we tested on 10 documents with the above equation of F_1. For this we give the human summarize sentence line as input to compare with the sentences generated by the proposed system to find the F_1 value. Finally the average accuracy of Bangla text summarization is 83.57% corresponding with human generated summarization. The accuracy on basis of F_1 is given Table II where N is the number of lines in the document.

A. Example

The system generated and human professional generated summary of an example text taken from The Daily Prothom-alo is shown below:

1) Original Text: প্রায় চার ঘণ্টা বন্ধ থাকার পর আজ সোমবার দুপুর একটার দিকে ঢাকার সঙ্গে চট্টগ্রাম ও সিলেটের রেলযোগাযোগ স্বাভাবিক হয়েছে। হরতালের সমর্থনে হেফাজতে ইসলামের নেতা-কর্মীরা ব্রাহ্মণবাড়িয়া ও আখাউড়ায় রেলপথ অবরোধ করে চারটি ট্রেন আটকে রাখে। এতে সকাল নয়টা থেকে রেলযোগাযোগ বন্ধ হয়ে যায়। এ ব্যাপারে জেলার অতিরিক্ত পুলিশ সুপার জাহিদুল ইসলাম প্রথম আলো ডটকমকে জানান, ট্রেন আটকে রাখার বিষয়টি নিয়ে পুলিশ জেলা হেফাজতের আমির মওলানা মনিরুজ্জামান সিরাজীর সঙ্গে আলোচনা করে। পরে তিনি কর্মী-সমর্থকদের রেলস্টেশন অবরোধ কর্মসূচি থেকে সরে যাওয়ার নির্দেশ দেন। পুলিশ এ সময় রেললাইনের ওপর অবরোধকালে ফেলে রাখা জিনিসপত্র সরিয়ে নেয়। এতে দুপুর একটার দিকে রেলযোগাযোগ স্বাভাবিক হয়। রেলস্টেশন ও প্রত্যক্ষদর্শী সূত্র জানায়, সকাল নয়টার দিকে ঢাকা থেকে সিলেটগামী পারাবত এক্সপ্রেস ট্রেনটি ব্রাহ্মণবাড়িয়া স্টেশনের দিকে যাচ্ছিল। এর আগে সদর উপজেলার বড়হরণ রেলগেট এলাকায় ট্রেনে ইটপাটকেল ছোড়েন হেফাজতের কর্মীরা। এতে ট্রেনের অন্তত পাঁচজন যাত্রী আহত হন এবং ট্রেনের একাধিক জানালার কাচ ভেঙে যায়। সকাল সোয়া নয়টার দিকে হেফাজতের সহস্রাধিক কর্মী-সমর্থক লাঠিসোঁটাসহ ব্রাহ্মণবাড়িয়া শহরের রেলগেট এলাকায় অবস্থান নিয়ে পারাবত ট্রেনটি আটকে দেন। পরে তাঁরা রেললাইনের ওপর গাছের গুঁড়ি ও লাইনের স্লিপার ফেলে সেখানে বসে পড়েন। এতে ঢাকা-চট্টগ্রাম ও ঢাকা-সিলেট রেলপথে ট্রেন চলাচল বন্ধ হয়ে যায়। এ ছাড়া এই স্টেশনে তাঁরা চট্টগ্রাম থেকে ময়মনসিংহগামী নাছিরাবাদ ট্রেনটিও আটকে দেন। এর আগে আখাউড়া রেলস্টেশনে চট্টগ্রাম থেকে সিলেটগামী কুশিয়ারা এক্সপ্রেস এবং কসবায় নোয়াখালী থেকে ঢাকাগামী উপকূল এক্সপ্রেস ট্রেন থামিয়ে দেন হেফাজতের কর্মীরা। ব্রাহ্মণবাড়িয়া রেলস্টেশনের মাস্টার অমৃত লাল দেবনাথ জানান, হেফাজতের কর্মী-সমর্থকেরা অবরোধ তুলে নিলে চারটি ট্রেনই আবার যাত্রা করে। এতে ঢাকার সঙ্গে চট্টগ্রাম ও সিলেটের রেলযোগাযোগ স্বাভাবিক হয়।

2) Human Generated Summary in five sentences: প্রায় চার ঘণ্টা বন্ধ থাকার পর আজ সোমবার দুপুর একটার দিকে ঢাকার সঙ্গে চট্টগ্রাম ও সিলেটের রেলযোগাযোগ স্বাভাবিক হয়েছে। হরতালের সমর্থনে হেফাজতে ইসলামের নেতা-কর্মীরা ব্রাহ্মণবাড়িয়া ও আখাউড়ায় রেলপথ অবরোধ করে চারটি ট্রেন আটকে রাখে। রেলস্টেশন ও প্রত্যক্ষদর্শী সূত্র জানায়, সকাল নয়টার দিকে ঢাকা থেকে সিলেটগামী পারাবত এক্সপ্রেস ট্রেনটি ব্রাহ্মণবাড়িয়া স্টেশনের দিকে যাচ্ছিল। সকাল সোয়া নয়টার দিকে হেফাজতের সহস্রাধিক কর্মী-সমর্থক লাঠিসোঁটাসহ ব্রাহ্মণবাড়িয়া শহরের রেলগেট এলাকায় অবস্থান নিয়ে পারাবত ট্রেনটি আটকে দেন। এর আগে আখাউড়া রেলস্টেশনে চট্টগ্রাম থেকে সিলেটগামী কুশিয়ারা এক্সপ্রেস এবং কসবায় নোয়াখালী থেকে ঢাকাগামী উপকূল এক্সপ্রেস ট্রেন থামিয়ে দেন হেফাজতের কর্মীরা।

3) System Generated Summary in five sentences: প্রায় চার ঘণ্টা বন্ধ থাকার পর আজ সোমবার দুপুর একটার দিকে ঢাকার সঙ্গে চট্টগ্রাম ও সিলেটের রেলযোগাযোগ স্বাভাবিক হয়েছে। এ ব্যাপারে জেলার অতিরিক্ত পুলিশ সুপার জাহিদুল ইসলাম প্রথম আলো ডটকমকে জানান, ট্রেন আটকে রাখার বিষয়টি নিয়ে পুলিশ জেলা হেফাজতের আমির মওলানা মনিরুজ্জামান সিরাজীর সঙ্গে আলোচনা করে। রেলস্টেশন ও প্রত্যক্ষদর্শী সূত্র জানায়, সকাল নয়টার দিকে ঢাকা থেকে সিলেটগামী পারাবত এক্সপ্রেস ট্রেনটি ব্রাহ্মণবাড়িয়া স্টেশনের দিকে যাচ্ছিল। সকাল সোয়া নয়টার দিকে হেফাজতের সহস্রাধিক কর্মী-সমর্থক লাঠিসোঁটাসহ ব্রাহ্মণবাড়িয়া শহরের রেলগেট এলাকায় অবস্থান নিয়ে পারাবত ট্রেনটি আটকে দেন। এর আগে আখাউড়া রেলস্টেশনে চট্টগ্রাম থেকে সিলেটগামী কুশিয়ারা এক্সপ্রেস এবং কসবায় নোয়াখালী থেকে ঢাকাগামী উপকূল এক্সপ্রেস ট্রেন থামিয়ে দেন হেফাজতের কর্মীরা।

V. CONCLUSION AND FUTURE WORK

In this paper, we discussed extraction based Bangla text summarization method in a single document. The system selects those sentences which has most influence on the context of documents. The biasness of the system is reduced by a rigorous

pre-processing technique. In this pre-processing the total document is divided line by line and then with tokenization and word stemming the frequency count is measured. Finally, the sentence ranking method is used to measure the importance of the sentence comparing with the types of word used in this document. The performance of this proposed system is 83.57% in generating summaries that agree well with human generated summaries, despite using minimal natural language processing (NLP) information.

The proposed system performs well while the document is completely depends on a particular theme where the document is compact of using keyword frequently in the whole document. In the future, we would like to integrate multiple themes. Currently the values of four parameters or weights for frequency, position, cue words, and heading are determined by a trial and error process. This process can be automated using any learning model e.g. least square method. However, the overall performance of the system can be enhanced by adaptively adding more features in sentence scoring and ranking. In addition, there is a scope of using abstraction based summarization technique so that it will much relate to the theme of the document.

ACKNOWLEDGMENT

This research is completed with the support of the teachers of Institute of Information Technology, University of Dhaka. It would be a pleasure to thank *B. M. Mainul Hossain*, *Ahmedul Kabir* and *Shah Mostafa Khaled* for their valuable suggestion and direction contribution that helps us in many ways to complete this research.

REFERENCES

[1] H. P. Luhn, *The automatic creation of literature abstracts*, in IBM Journal of Research Development, volume 2, number 2, pages 159-165, 1958.

[2] Hongyan Jing, *Sentence Reduction for Automatic Text Summarization*, in Proceedings of the sixth conference on Applied natural language processing, pages 310-315, 2000.

[3] H. P. Edmundson, *New methods in automatic extracting*, in Journal of the ACM, 16(2):264, 1969

[4] Kamal Sarkar, *Bangla Text Summarization by Sentence Extraction*, in Proceedings of International Conference on Business and Information Management(ICBIM),NIT Durgapur, pages 233-245, 2012

[5] Amitava Das and Sivaji Bandyopadhyay, *[5] Topic-Based Bangla Opinion Summarization*, in Proceedings of the IEEE Second International Conference Social Computing (SocialCom), pages 675-682, 2010

[6] I. Mani, *Automatic summarization*, in Natural language processing, Amsterdam/Philadelphia, John Benjamins Publishing Company, 2001

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

[8] Jubayer Shamshed and S. M. Masud Karim *A Novel Bangla Text Corpus Building Method for Efficient Information Retrieval*, ISSN 2078-5828 (PRINT), ISSN 2218-5224 (ONLINE), VOLUME 01, ISSUE 01, 2010.

[9] 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