J Inf Process Syst, Vol.13, No.4, pp.752~777, August 2017
https://doi.org/10.3745/JIPS.04.0038

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

An Innovative Approach of Bangla Text Summarization by Introducing Pronoun Replacement and Improved Sentence Ranking

Md. Majharul Haque*, Suraiya Pervin*, and Zerina Begum**

Abstract

This paper proposes an automatic method to summarize Bangla news document. In the proposed approach, pronoun replacement is accomplished for the first time to minimize the dangling pronoun from summary. After replacing pronoun, sentences are ranked using term frequency, sentence frequency, numerical figures and title words. If two sentences have at least 60% cosine similarity, the frequency of the larger sentence is increased, and the smaller sentence is removed to eliminate redundancy. Moreover, the first sentence is included in summary always if it contains any title word. In Bangla text, numerical figures can be presented both in words and digits with a variety of forms. All these forms are identified to assess the importance of sentences. We have used the rule-based system in this approach with hidden Markov model and Markov chain model. To explore the rules, we have analyzed 3,000 Bangla news documents and studied some Bangla grammar books. A series of experiments are performed on 200 Bangla news documents and 600 summaries (3 summaries are for each document). The evaluation results demonstrate the effectiveness of the proposed technique over the four latest methods.

Keywords

Bangla News Document, Cosine Similarity, Dangling Pronoun, Pronoun Replacement, Sentence Frequency

1. Introduction

The amount of available information increases rapidly with the development of information technology and the use of the Internet [1]. The estimated size of the websites, which holds e-contents, was around 4.76 billion pages in November 19, 2016 [2] and it is increasing exponentially in every second [3]. Users are encumbered with the huge volume of electronic texts, whereas they expect the concise information within the shortest possible time. So, the automatic text summarization is needed to process the Internet data efficiently and scavenging useful information from it [3].

The state-of-the-art works in this field [4-7] have focused on automatic text summarization in different languages, started with English. Automatic English text summarization technique was first proposed by Luhn [8] using term frequency, around five decades ago. With the increase of online contents, Edmundson [9] offered a significant development in English text summarization by

Manuscript received December 22, 2016; first revision April 18, 2017; accepted May 30, 2017.
Corresponding Author: Md. Majharul Haque (mazharul_13@yahoo.com)
* Computer Science & Engineering, University of Dhaka, Dhaka, Bangladesh (mazharul_13@yahoo.com, suraiyacse@gmail.com, suraiya@du.ac.bd)
** Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh (zerin@iit.du.ac.bd)

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 [6,7,10]. As Bangla is the 7th most spoken language [11], 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.

To mitigate this burden of a large volume of text, very few research works have been accomplished for Bangla [6,7]. So, for the welfare of the large community of Bangla-speaking people, more research is necessary. But, research works for Bangla language is difficult for the followings:

- Based on our study, automatic methods are hardly available for Bangla language.
- The lexical database like WordNet in English [12] is not available for Bangla. Such a tool is going to be developed for Bangla with limited features [13].
- Subjects and objects of all the sentences are needed to be identified for proper recognition of structures of sentences which is also complex in Bangla than that of English. Because, the subjects in English sentences can be placed before the verb phrase, the auxiliary verb in the active voice and it can be appeared after the word 'by' in the passive voice. But, the subject may exist in the first, last or middle places of sentences (before or after the verb) for both active and passive voices in Bangla sentences.

Some significant problems have also been discussed in [14,15] about the research works on Bangla. Moreover, the scope of knowledge sharing is limited as there are a few researchers in the arena of Bangla language processing. Despite these difficulties, a procedure has been presented here for Bangla text summarization which also focuses on the problem of dangling pronoun in summary. In the output of text summarization, the existence of any dangling pronoun makes the information incoherent. So, the systems which have been developed for burden minimization from a large volume of text may deliver the wrong message if there is any dangling pronoun. Other than receiving a direction, the user will often be misguided with misinformation. In these circumstances, we have proposed a method with the following major contributions:

- Replacing pronouns by the corresponding nouns to minimize the number of dangling pronouns in summary.
- Introducing sentence frequency for sentence ranking and eliminating redundancy.
- Identifying numerical figures from the variety of forms (presented in words and digits) to assess the importance of sentences.

Point to be mentioned that part of this proposed method has been presented in [16].

To accomplish this research work, we have scrutinized 3000 Bangla news documents (news documents of around one month from the Daily Prothom-Alo which is the most popular newspaper of Bangladesh). Seven knowledgeable persons of Bangla language, who have completed four years graduation on Bangla language (their mother tongue is Bangla), helped us in this research work. After a detail discussion with those seven persons regarding the structures of sentences of Bangla language and analysis of news documents, we have used several rules in this proposed method.

The rest of the paper is organized as follows: Section 2 describes literature review on Bangla text summarization. Section 3 illustrates proposed method in details. Implementation model and algorithms are given in Section 4. Evaluation and discussion on results are depicted in Section 5. Finally, the conclusion is turned with future works in Section 6.

2. Literature Review

The automatic English text abstraction began around five decades ago by Luhn [8] in 1958. Since then, the field of text summarization has witnessed continuous involvement of many researchers in the attempt to look for different strategies [17-19]. Unlike English, which has seen a significant number of systems developed to cater to it, other languages are less fortunate [20]. So far, few attempts have been conducted for Bangla text summarization [6].

The first technique of automatic Bangla text summarization was proposed in 2004 by Islam and Masum [21] named "Bhasa". They claimed it [21] as a corpus-oriented search engine and summarizer. A few years later, some methods from the survey work regarding English text summarization systems were implemented to summarize Bangla text in [22] by utilizing (i) location method, (ii) cue method, (iii) text title, (iv) term frequency, and (v) numerical data.

In 2010, Das and Bandyopadhyay [23] offered a method for opinion summarization in Bangla other than generic text summarization [6,7]. The same procedure as like English text summarization system [9] was followed in the method of Bangla text summarization by Sarkar [6,7] in 2012. These methods [6,10] are based on word-frequency, length, and position of sentences. In 2012, Sarkar proposed another technique [7] by tuning the features of his previous method [6].

In 2014, Sarkar [24] presented keyphrases extraction based Bangla and English text summarization which is a variant of an existing method [25]. The algorithm for sentence selection and summary generation works in two phases. Phase-1 uses sentence position and document's keyphrases. If phase-1 fails to generate the summary of user desired length, phase-2 is activated and select more sentences.

To the best of our knowledge, to date, most of the incorporated features for Bangla text summarization systems have been taken from various existing systems of English text summarization. But, the proposed method in this paper has introduced some features (replacement of pronouns by the corresponding nouns, sentence frequency calculation, and numerical figures identification from words) different from both English and Bangla text summarization methods. Moreover, our approach has shown better performance from the four latest existing methods [6,7,10,24] based on ROUGE (Recall Oriented Understudy for Gisting Evaluation) [26] evaluation scores.

3. Proposed Method

The proposed Bangla text summarization approach is described in the following steps.

3.1 Preprocessing

Input document is segmented to the array of sentences S[S1, S2, S3, … , Sn] and all the sentences are then tokenized to words in this step. From the entire words, stop words are removed as per the list of 363 Bangla stop words [27]. The words in Bangla language are very much inflectional [7] for which word stemming [28] is applied to map the words with different endings to a single word.

Words in different forms	Words after stemming
"গ্রামের" (gramer), "গ্রামে" (grame)	"গ্রাম" (gram). In English: "village"
"হাঁটছেন" (hatchen), "হাঁটতেছেন" (hatitechen)	"হাঁটা" (hata). In English: "walk"

Fig. 1. Example of word stemming in Bangla.

3.2 Replacement of Pronoun by the Corresponding Noun

In our approach, eight forms of pronouns are considered for replacement: (i) তিনি (tini - he/she, (ii) তাকে (take - him/her), (iii) তাহাকে (tahake - him/her), (iv) সে (she - he/she), (v) ইনি (ini - he/she), (vi) উনি (uni - he/she), (vii) তার (tar - his/her), and (viii) তাহার (tahar - his/her). Only singular forms of pronouns which can be used for human are considered here. So, the corresponding noun should be an indicator of single human named entity.

We have observed from 3,000 Bangla news documents that the corresponding noun of any pronoun appears as subject or object of the two previous sentences for 88.63% times. For recognizing subject and object of any sentence, nature of every word is identified. The end to end process of the pronoun replacement is explained in the following.

3.2.1 General tagging

All the words are tried to tag as noun, pronoun, adjective, verb, etc. using lexicon database [29] and SentiWordNet [30]. As per our experiment with 200 test documents, 65.13% words can be tagged because lexicon database [29] and SentiWordNet [30] have limited number of predefined words.

Point to be mentioned that Bangla words (especially verbs) are very much inflectional. So, many verbs are left untagged as lexicon database and SentiWordNet have not covered the entire inflection. Though word stemming has been accomplished (in the previous step) to identify root forms of words, all the inflectional forms of verb could not be stemmed. In reality, the identification of verb is quite difficult because the verb can have a lot of suffixes in Bangla sentence. An English word "say" can be "saying", "said" and "says" for different tenses and persons. But in Bangla, this word has several forms for different tenses and persons. For example, the word "বল" (bol - say) can have three basic forms for the first, second and third person in the present continuous tense only. It can be "বলছি" (bolchhi - saying) for the first person, "বলছ" (bolchho - saying) for the second person and "বলছেন" (bolchhen - saying) for the third person. Moreover, there are three forms of the meaning of the word "you" in Bangla as "আপনি" (apni - you), "তুমি" (tumi- you) and "তুই" (tui - you) in significant, general and trivial forms respectively. For all of these meaning of "you" in Bangla, the forms of every verb are also different. Such as "আপনি বলছেন" (apni bolchhen – you are saying), "তুমি বলছ" (tumi bolchho – you are saying), "তুই বলছিস" (toi bolchhis – you are saying) where all the forms are given in present continuous tense and for the second person only. In this way, the word "বল" (bol - say) can have the following diversified forms: "বলে" (bole), "বলেন" (bolen), "বলিস"(bolish), "বলি" (boli), "বলছে" (bolchhe), "বলছেন" (bolchhen), "বলছ" (bolchho), "বলছিস" (bolchhis), "বলছি" (bolchhi), etc. [31,32]. So, the complexity of verb' recognition in Bangla can't be compared with English.

But, verb identification is essential for any language processing task as it is the chief word for any sentence [32]. So, if there is any word left untagged after tagging by using lexicon database [29] and SentiWordNet [30], we need to check the word if it is a verb. In this regard, a list of suffixes [31] as "ইতেছি" (itechhi), "তেছিলেন" (techhilen), "লেন" (len), "সেন" (sen), etc. is taken into account for ultimate checking. If the considered word has any of these suffixes [31], it is tagged as a verb.

The percentage of word tagging has been increased from 65.13% (the result of word tagging before considering the list of suffixes [31] for verb) to 66.73% after detecting verb using the list of suffixes. This is noticeable that the tagging in this step is a preliminary tagging and some tags may be updated in the next steps.

3.2.2 Special tagging

It is well known that word is the principal ingredient of language [32]. Without recognizing the nature of each word, it is hard to detect subjects and objects of sentences. A procedure is available for Bangla parts-of-speech tagging [29] but the procedure for identifying nature of words as acronym, initial, repeated words, numerical figure from digits and words, occupation, etc., does not exist. In this situation, nature of words has been identified as follows:

(1) *Checking for English acronym*: In Bangla news documents, there can have acronyms i.e. the words consist of some English letters that are written in Bangla. For example: "ইউএনডিপি" (UNDP), "আইএলও" (ILO), etc. For checking these words, all the English letters are written in Bangla such as "এ" (A), "বি" (B), "সি" (C) …. …. …. "ডব্লিউ" (W), "এক্স" (X), "ওয়াই" (Y), "জেড" (Z) and sorting them in descending order based on their string length. By this way, "ডব্লিউ" (W) is in the first place and "এ" (A) is in the last place and then match every letter of the words. Such as "ইউএনডিপি" (UNDP) is matched with "ইউ" (U), "এন" (N), "ডি" (D), and "পি" (P). The significant point is that sorting in descending order is done for ensuring the longest matching always. For instance, "এন" (N) is not matched with "এ" (A) at the first time rather it is fully matched with "এন" (N). In this technique, we got 100% success rate for detecting English acronyms.

(2) *Checking for Bangla* initial: As like English acronyms mentioned in the above point (1), there can be Bangla letters with spaces such as "আ স ম" (A S M). These letters are tagged as Bangla initial. The experiment shows that the correctness of finding initial is 100%.

(3) *Checking for repeated words*: In Bangla language, same words can be written for two times [31] as like "ঠান্ডা ঠান্ডা" (thanda thanda - cold cold). Some words are repeated partially such as "খাওয়া দাওয়া" (khawa dawa – eat drink). List of some other words has been collected from [32] where some irregular words are mentioned as repeated words as "দেনা পাওনা" (dena paona – payable receivable). If two consecutive words are matched with these listed words or repeated for two times (entirely or partially), they are tagged as repeated words. We have applied this procedure to 200 news documents and found 98% accuracy on the search for repeated words.

(4) *Checking for numerical figure*: Point to be mentioned that any numerical figure has only one written form in words and one written form in digits in English—e.g., '10' can be written as '10' (in digits) and 'ten' (in words) in English. But, it can have several forms as "১০" (10), "১০টি" (10 ti), "১০টা" (10 ta), "১০খানা" (10 khana), "১০খানি" (10 khani), "দশ" (dosh-ten), "দশটি" (doshti-ten), "দশটা" (doshta-ten), "দশখানা" (doshkhana-ten), and "দশখানি" (doshkhani-ten) in Bangla. In this way, every numerical figure in Bangla can be written in a variety of forms. So, it can be said that numerical figure identification from a text (presented in digits and words both) in Bangla is comparatively difficult and challenging than that of English. So, a technique is introduced here to recognize Bangla numerical figure by checking the following conditions:

   a) First part of the word is consisted with the following: ০(0), ১(1), … …, ৯(9) or "এক" (ek-one), "দুই" (dui - two), … … … "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), etc.

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

If any word meets these three conditions, the word is tagged as numerical figure. We have experimented on 200 test documents for this proposed technique and observed that numerical figure can be detected for 100% from both digits and words.

(5) *Checking for occupation*: There is a data file with 80 entries (collected from [33,34]) for the title of Bangladeshi occupation such as "মন্ত্রী" (montri - minister), "কৃষক" (krrishok - farmer), "ছাত্র" (chhatro - student), etc. Each word of the input document has been matched with these 80 entries and tagged as "occupation" if any match is found. Here, "মন্ত্রী" (montri - minister) will cover "খাদ্যমন্ত্রী" (khaddomontri – Food minister), "শিক্ষামন্ত্রী" (shikkhamontri – Education minister), and so on. In this way, if any word has suffix from the list of occupation or fully matched with the occupations' list, the word is tagged as occupation. It has been observed that the proposed system can identify occupation for 91% times.

(6) *Checking for the name of organization*: It has been detected from our analysis that name of an organization can be mentioned in Bangla news documents as:

a) The full name of organization is given which follows the acronym of the name enclosed in parentheses. Such as "দুর্নীতি দমন কমিশন (দুদক)" - "Durniti Domon Commission (DUDOK) – Anti Corruption Commission (ACC)".

b) The last part of the organization name have specific word such as "লিমিটেড" (limited - limited), "বিশ্ববিদ্যালয়" (bishawbiddaloy - university), "মন্ত্রণালয়" (montronaloy - ministry), "কোং" (kong - kong), etc. [35].

If there is any acronym according to the above point (a), enclosed with parentheses, count the number of letters in the acronym and then same number of words (immediately before the acronym) are tagged as a name of organization. Experiment with 650 acronyms (collected from [36,37]) shows that the accuracy of finding organizations' name using point (a) is 95.60%.

According to the point (b), if any of the words (mentioned in point (b)) is presented in the text, check three words immediately before the specific word. Three words are considered as it is observed in our analysis that organization' name is generally constituted with three words. If the organization' name is constituted with more than three words, selecting three words is considered enough to serve the purpose. If the three words are noun, named entity or any untagged words, consider them as the name of an organization. Name of organizations can be recognized for 87% based on the point (b).

(7) *Checking for the probable name of people and places*: A list of first name, middle name and last name with 7500 entries has been utilized in our method where most of them are collected from [38]. If any word is matched with these listed names, it is primarily tagged as a name of human. Part of name is identified at this point for more than 80% times which will be re-checked and full names will be detected from these parts of names in the sub-section 3.2.4.

Moreover, a table has been maintained with 700 entries for the list of division, district, upazila, and municipality as the name of places for Bangladesh [39] and 230 names of countries and their capital [37]. If any word is matched with these listed names of places, it is tagged as a place. In this way, around 82% names of places can be detected.

From the overall 31525 words (from 200 test documents), 5.80% untagged words are identified in special tagging which raise the words tagging from 66.73% (the result of general tagging in sub-step

3.2.1) to 72.53%. Some experimental results on the special tagging process are given in the Table 1.

The general and special tagging are static, but the words can be dynamic in nature depending on the surrounding words in sentences. In this regard, dependency parsing is introduced in the next sub-step.

Table 1. Experimental results of words identification from 200 test documents in special tagging process

#	Nature of words	Success rate (%)
1	English acronym	100
2	Bangla initial of name	100
3	Repeated words	98
4	Numerical figure from digits	100
5	Numerical figure from words	100
6	Occupation	91
7	Name of organization based on the number of letter in acronym which is enclosed in parentheses	95.60
8	Name of organization based on some specific last words	87
9	Name of human	80
10	Name of places	82

3.2.3 Dependency parsing of each word

The nature of words in sentences can be varied due to the effects of surrounding words. For this reason, dependency parsing has been incorporated so that any given tag can be updated and untagged words can be tagged with the help of previously tagged words as follows:

(1) List of adjectives has been collected from [31], and fully repeated words (mentioned in the special tagging process) have been treated as adjective [31]. The repeated words that are partially repeated are nouns. Adjectives are placed as neighboring words of nouns or verbs [31]. If any adjective (the adjective which is not tagged as repeated word in special tagging) has any suffix, it is treated as a noun. There can be consecutive adjectives where nouns or verbs are placed after these consecutive adjectives.

(2) Some words are placed before adjective, e.g., "অপেক্ষা" (opekkha - than), "চেয়ে" (cheye - than), etc. [33]. So, the word, which is placed after these words, is tagged as an adjective.

(3) The word, which is presented after adjective, is noun or verb. If the word after adjective has a suffix as "ইতেছি" (itechhi), "তেছিলেন" (techhilen), "লেন" (len), "সেন" (sen), etc. [31], it will be treated as verb otherwise noun [31,32].

(4) There is a list of articles as "টি" (ti - this), "টা" (ta - this), etc. which can be placed as a suffix with nouns or pronouns [32]. The list of pronouns is collected from [31,32]. So, if any word (except in the list of pronouns) has articles, it is considered as noun. There can be article along with number, occupation, organization as they are one kind of noun. So, a new tag is given for each of them as number with article, occupation with article, etc. if they contain any article.

(5) The word "গোটা" (gota) can be placed before numerical figure and "খানা, খানি" (khana-this, khani-this) can be placed after numerical figure [31].

(6) If there is numerical figure anywhere in the sentences, the next word of the numerical figure is a noun. If the numerical figure is the last word, the noun exists immediately before that.

(7) There may have comma separated words where the ending word may be separated by "ও" (o - and), "এবং" (ebong - and), "আর" (ar - also). In that case, all these words are same in nature.

(8) List of words is there as preposition/conjunction/interjection (অব্যয় - Obboy) in Bangla language. They are tagged as stop words. A list of 363 stop words has been collected [27] for Bangla language. These words can't have any dependency on surrounding words [31].

(9) The word which is placed after "দ্বারা" (dara - with), "দিয়া" (diya - with), etc., is verb [31].

(10) If there is any noun with suffix "র" (r), "এর" (er), there will be another noun after that. If the second one has the same suffix, this will follow another noun and so on. The final noun can be either subject or object of the considered sentence.

(11) The words as "ওহে" (ohe - hi) and "হে" (he - hi) follow the name of a person.

(12) The suffixes "কার" (kar) and "কের" (ker) are placed with the words which indicate time.

(13) If any of the words "যদি" (jodi - if), "যখন" (jokhon - when), "যার" (jar - whose), "যাকে" (jake - who), "যেখানে" (jekhane - where), "যেই" (jei - this), "যেইমাত্র" (jei-matro - when) exists in the initial position of any sentence, there will be two parts of the sentence. In that case, the former part is the secondary part and later part is the primary part which contains the principal subject and the principal verb.

(14) If any of the words "কখন" (kokhon - when), "কোথায়" (kothay - where), "কবে" (kobe - when), "কিভাবে" (kivabe - how) exists in the middle position of any sentence, there will be two parts of the sentence. In that case, the former part is the primary part, and the later part is the secondary part of the sentence where the primary part contains the main subject and the main verb.

(15) The word, which is placed immediately before "সমাহার" (somahar - combination), is noun where the previous word of the noun is a numerical figure.

(16) There are pairs of words as "যে-সে" (je-she -- who-he), "যা-তা" (ja-ta -- which-that), "যিনি-তিনি" (jini-tini -- who-he), "যাকে-তাকে" (jake-take -- whom-he), etc. From these pairs of words, if the first word exists, the second word is also existed [31].

(17) There can be sequence of words like "যে 'x' সে 'y'" (je 'x' she 'y' – who 'x' he 'y') or "যাকে 'x' তাকে 'y'" (zake 'x' take 'y' – whom 'x' he 'y') where 'x' and 'y' are two words of same nature. In these cases, 'x' and 'y' are any kind of designation or occupation. So, if we can identify the word 'x', we can also identify 'y' as same nature of the word and vice versa.

After dependency parsing, the tagging of words has been improved from 72.53% (the result of word tagging after special tagging in Section 3.2.2) to 79.50% in our experiment (Table 2).

Table 2. Experimental results of word tagging from 31,525 words of 200 documents

Word tagging in different phases	Number of tagged words	Tagging (%)
Tagging by list of words from [29,30]	20532	65.13
After considering suffixes for verb	21038	66.73
After special tagging	22865	72.53
After dependency parsing	25062	79.50

3.2.4 Finding the full names of human for the identification of subjects and objects

In the previous steps, all the processes of tagging are somehow depended on some lists of words. But, it is apparent that whatever the range of lists, there is a limitation. So, all the named entities can't be tagged based on the predefined lists, and some words can be wrongly tagged as named entity. Point to be mentioned that existing procedure of named entity recognition [35] has not been utilized in our proposed method. Because it is based on predefined lists of words only and the impact of surrounding words is ignored. Based on our observation, the full names of human is written for the first time for around 95% times, and then parts of the names may be used anywhere in the news documents. So, a mechanism, to recall the full name is necessary from the parts of all the names, which is another distinguishing feature of our approach.

By using the parts of a name, it is quite difficult to find out the full name. Because any single word may be utilized for multiple functions as like the word "সুরুজ" (Shuruz) may indicate for "sun" but "সুরুজ মিয়া" (Shuruz Miah) shows a name of a person as there is a recognizable last name "মিয়া" (Miah). So, multiple words are checked at a time in this step for getting all parts of all the names such as the first, last and middle name with or without any initial. In this regard, the following rules are brought into play based on our study of Bangladeshi news documents and Bangla grammar books [31,32] to identify named entities:

(1) Occupation exists before the name of a human in any text document. So, if any word has occupation tag without any article, check the immediate next four words. Four words are considered as there can be occupation and initial before the name, and full name has three parts usually (the first, last and middle name). From these four words, take the words as named entity that are tagged as the first name, middle name, last name, noun or any untagged word (at least one of the words should be tagged as part of name based on the step of special tagging).

(2) If there is any of the first, last or middle name available, there may have some other words to constitute the full name. So, if any word is found as the first, last or middle name, consider nearby two words also. Total three words are considered as there are three parts of a name (the first, middle and last parts of a name) [38]. From these three words, take the words as named entity those are tagged as name, noun, Bangla initial or any untagged words from special tagging. But, if no other words are existed with the considered word to form the full name, ignore the word.

(3) If there is a comma (punctuation mark) followed by a word with verb tag, there may have the subject before the verb. So, if any word is found as verb with an adjacent comma (punctuation mark) such as "বলেন," (bolen, - says,), "জানান," (janan, - inform,), "জানালেন," (janalen, - informed,), etc. or the word with verb tag is existed at the end of sentence, traverse from this word to the beginning of the sentence for collecting named entity as like the first two points of this sub-step.

(4) If there is any verb in text, go from this word to the beginning of sentence for collecting named entity as like the first two points of this sub-step.

(5) Some words are there as the verb like "কর" (do), "দেয়" (give), "খায়" (eat), "যায়" (go), etc. A name is presented before the verb where the name is treated as a subject. Provided that the word immediately before the verb is not an adjective.

(6) Based on our study, we have observed that some digits are enclosed in parentheses which indicate the age of a person immediately after the name. Such as: "আব্দুল বাতেন (২৪)" (Abdul Baten (24)). So, look for named entity immediately before such digits enclosed with parentheses as like the first two points of this step. In this regard, maximum three digits are taken into account as the indicator of age.

(7) The word that is presented immediately before the words "দ্বারা" (dara - with), "দিয়া" (diya - with), "দিয়ে" (diye - with), etc. is a noun which is an object [31]. If there are two words (both are nouns) before these listed words, the first one is an indirect object (personal) and the second one is a direct object (material).

(8) General structures of sentences can be as follows: (a) subject + object (personal) + object (material) + adjective of verb + verb or (b) subject + time related word + place related word + indirect object + direct object + adjective of verb + verb [31,32]. We may identify subjects and objects by following these structures.

(9) There are named entities immediately after the word "নাম" (nam - name) and immediately before the word "নামে" (name - named) or "নামের" (namer – name'). The experiment shows that this rule is correct for 98% scenarios.

(10) There may have wrongly selected human named entities in the previous points. For verifying these entities, the immediate previous word for each named entity is also considered. If the previous word is a number, word with an article, adjective or repeated words, the considered word is not taken as name of human. In this way, wrongly selected human named entities are removed from the list of collected names.

After finding all the named entities, a simple and well-organized mechanism has been incorporated here to keep them easily accessible. An associative array has been taken which means that the index of the array is not number but string. If a named entity is presented as "প্রধান শিক্ষক লতিফুর রহমান খান" (Head Master Lotifur Rahman Khan), it is placed in the array for five times based on the parts of the name as in Fig. 2.

Index		value
[প্রধান]	=>	প্রধান শিক্ষক আরিফুর রহমান খান
[Head]	=>	Head Teacher Arifur Rahman Khan
[শিক্ষক]	=>	প্রধান শিক্ষক আরিফুর রহমান খান
[Teacher]	=>	Head Teacher Arifur Rahman Khan
[আরিফুর]	=>	প্রধান শিক্ষক আরিফুর রহমান খান
[Arifur]	=>	Head Teacher Arifur Rahman Khan
[রহমান]	=>	প্রধান শিক্ষক আরিফুর রহমান খান
[Rahman]	=>	Head Teacher Arifur Rahman Khan
[খান]	=>	প্রধান শিক্ষক আরিফুর রহমান খান
[Khan]	=>	Head Teacher Arifur Rahman Khan

Fig. 2. Structure of associative array for keeping human named entities.

This mechanism of associative array has been utilized here so that full name can be recalled from part of name. That means, if the part of name is "খান" (khan) anywhere in the input document, the

associative array will be traversed for the index "খান" (khan) and get the full name "প্রধান শিক্ষক আরিফুর রহমান খান" (Head Teacher Arifur Rahman Khan).

3.2.5 Replacing pronoun

Though the recognizing name of human is an important task, some other works are also indispensable for replacing pronouns. Because from the identified named entities, subjects and objects are needed to be detected. Some rules are applied here for recognizing subjects and objects of all the sentences and determining the corresponding nouns of pronouns as follows:

(1) For the replacement, eight forms of singular pronouns are taken into account. Such as তিনি (tini – he/she), তাকে (take – him/her), তাহাকে (tahake – him/her), সে (she – he/she), ইনি (ini – he/she), উনি (uni – he/she), তার (tar – his/her) and তাহার (tahar – his/her). These eight forms of pronouns are considered in this method, and others are left as future work.

(2) Some cases are there in the sentence of Bangla language where a pronoun is available, but it is not a dangling pronoun. Such as "তাকে" (take – him/her) is followed by "যাকে" (zake - whom), "সে" (she – he/she) is followed by "যে" (ze - who), etc. In these circumstances, pronouns are not replaced.

(3) Get the named entities of the previous sentence as discussed in the Section 3.2.4. If no named entity is available in the previous sentence, look for the named entity in the second previous sentence. If only one named entity is found, replace the pronoun by this named entity. If there are more than two named entities in the previous sentence, keep the pronoun without replacing because this situation may make the subjects or objects plural which is not considered here. If there are exactly two named entities, it is needed to decide which one is a subject and which one is an object. Particular attention is given on the replaceable pronouns to determine the replacement by subjects or objects of the previous sentence. In this regard, following rules are applied:

   (i) It has been observed that named entities with the following suffixes are object: "কে" (ke), "রে" (re), "এর" (er), etc. [31].

   (ii) If there is a named entity after verb which is at the end of a sentence, this named entity is considered as subject and other (if any) is an object.

   (iii) If there is a named entity at the beginning of the sentence, it is considered as a subject. Provided that it has no similar criterion as like the above point (i).

   (iv) If a verb is presented with a comma, the named entity before the verb is considered as a subject. Provided that it has no similar criterion as like the above point (i).

   (v) If there are two named entities before the verb, generally the first one is a subject, and the other is an object. But, if the first named entity has suffix "কে" (ke), "রে" (re), "এর" (er) then the second one will be treated as a subject and the other is an object.

   (vi) A subject can be replaced by a subject and an object can be replaced by an object while pronoun replacement. The pronouns, such as, তাকে (take – him/her), তাহাকে (tahake – him/her), "তার" (tar – his/her) and "তাহার" (tahar – his/her) will be replaced by object of the previous sentence as these words are generally used as object of any sentence.

(vii) The pronouns, such as, তিনি (tini – he/she), সে (she – he/she), ইনি (ini – he/she), উনি (uni – he/she) will be replaced by subject of previous sentence as these words are used as subject of any sentence.

(viii) For ensuring replacement of pronouns in suitable format, the followings are carried out: a) if the pronoun is "তাকে" (take – him/her) or "তাহাকে" (tahake – him/her), a suffix "কে" (ke) should be added after the noun, b) if the pronoun is "তার" (tar – his/her) or "তাহার" (tahar – his/her), a suffix "এর" (er) should be added after the noun, c) if the pronoun is any of the following: তিনি (tini – he/she), সে (she – he/she), ইনি (ini – he/she) or উনি (uni – he/she), the noun should have no suffix.

In the proposed method, named entity may be existed as only one word where it is hard to determine that is a named entity or not. To overcome this difficulty, all the named entities of the input document, kept in an associative array in word by word (in Section 3.2.4). So, if the system needs to consider a single word is named entity or not, it will be searched in the associative array generated in Section 3.2.4. If the word is located as an index in the array, the value of the resultant index is used to replace the pronoun in Section 3.2.5 otherwise it is left without the replacement. Some analytical results (related to the replacement of pronouns by the corresponding nouns) are given in the following Table 3.

Table 3. Some analysis on 200 documents on identifying named entities, subjects and objects

Description of analysis	Experimental results (%)
Full names are given in news documents before writing only part of name	95
Parts of names identification	80
Full names identification	76.50
Recall full names from the parts of names	74.50
Categorize the named entities as subjects and objects	73
Corresponding nouns of a pronouns are existed within the 2 previous sentences	88.63

Table 4. Result on pronoun replacement for 200 Bangladeshi news documents

Total number of pronouns	Kept unchanged	Replaced correctly	Replaced incorrectly
255	60	183	12

According to the result in the Table 4, our procedure can replace 183 pronouns correctly from 255 pronouns. So, the accuracy of pronoun replacement is 71.80%. An example is given in figure below by using a part of document about replacing pronoun where bold lettered words are pronouns and the corresponding nouns.

A sample text has been given in Fig. 3 to illustrate the process of pronoun replacement where the sample event and names are imaginary. In the figure, the pronouns and nouns are presented with bold letter text. In the original message, there is one pronoun "তিনি" (tini - he) mentioned for three times. After applying our pronoun replacing technique, the pronoun "তিনি" (tini - he) has been replaced correctly for the first two times by the corresponding noun "আব্দুর রহিম সাহেব" (Mr. Abdur Rahim). For the third times, "তিনি" (tini - he) has not been replaced because the corresponding noun has not found within the two immediate previous sentences.

Sample text:
সংবাদ সংস্থার প্রধান **আব্দুর রহিম সাহেব** বলেন যে, সকল মানুষের প্রয়োজনীয় খবরের কাগজে ছাপানো উচিৎ। **তিনি** জানান, ভবিষ্যতে পত্রিকা হবে গণমানুষের জন্য। বর্তমান কর্মকান্ড নিয়ে **তিনি** বলেন, আমরা সবাই স্বচ্ছতা এবং জবাবদিহিতা নিশ্চিত করার জন্য কাজ করছি। বক্তব্যের মাঝে, সাবেক প্রধান আশুতোষ গুপ্তকে ধন্যবাদ জানান হয়। আশুতোষ গুপ্তের কাজের প্রশংসাও করা হয়। সবশেষে, **তিনি** সকলের শুভ কামনা করে বক্তব্য শেষ করেন।

Text after pronoun replacement:
সংবাদ সংস্থার প্রধান **আব্দুর রহিম সাহেব** বলেন যে, সকল মানুষের প্রয়োজনীয় খবরের কাগজে ছাপানো উচিৎ। **আব্দুর রহিম সাহেব** জানান, ভবিষ্যতে পত্রিকা হবে গণমানুষের জন্য। বর্তমান কর্মকান্ড নিয়ে **আব্দুর রহিম সাহেব** বলেন, আমরা সবাই স্বচ্ছতা এবং জবাবদিহিতা নিশ্চিত করার জন্য কাজ করছি। বক্তব্যের মাঝে, সাবেক প্রধান আশুতোষ গুপ্তকে ধন্যবাদ জানান হয়। আশুতোষ গুপ্তের কাজের প্রশংসাও করা হয়। সবশেষে, **তিনি** সকলের শুভ কামনা করে বক্তব্য শেষ করেন।

Sample text:
Mr. Abdur Rahim, head of the news agency, said that the required news of all people should be published in newspaper. **He** informed that in future newspaper will be for mass people. **He** said about present activities that we all are working to ensure transparency and accountability. Within the speech, ex. head Ashutosh Gupto was thanked. The work of Ashutosh Gupto was also praised. At last, **he** finished speech with seeking good for all.

Text after pronoun replacement:
Mr. Abdur Rahim, head of the news agency, said that the required news of all people should be published in newspaper. **Mr. Abdur Rahim** informed that in future newspaper will be for mass people. **Mr. Abdur Rahim** said about present activities that we all are working to ensure transparency and accountability. Within the speech, ex. head Ashutosh Gupto was thanked. The work of Ashutosh Gupto was also praised. At last, **he** finished speech with seeking good for all.

Fig. 3. Sample text for the example of pronouns' replacement.

3.3 Sentence Ranking

For sentence ranking, values of some attributes are calculated for all the sentences and then sum-up all the attributes' value to compute the final score of each sentence. Top scored sentences are assumed as top-ranked sentences and vice versa. The following attributes are considered in this method: (i) term frequency inverse document frequency, (ii) sentence frequency, (iii) numerical figure presented in both words and digits, (iv) title words, and (v) the first sentence.

3.3.1 Term frequency inverse document frequency score calculation ($S_{TF\text{-}IDF}$):

The TF-IDF score is calculated with the following equations:

$$TF - IDF_{(t)} = TF * Log(\frac{N}{DF}) \quad (1)$$

$$S_{TF-IDF(k)} = \sum_{t=1}^{T} TF - IDF_{(t)} \quad (2)$$

where, N is the number of documents in a corpus, DF indicates the number of documents in which the term t appears. $S_{TF\text{-}IDF(k)}$ means the TF-IDF score for k^{th} sentence which includes the summation of TF-IDF score of all the terms of sentence k.

3.3.2 Sentence frequency calculation (S_{SF}) and redundancy elimination:

This proposed procedure has introduced the second attribute as sentence frequency (S_{SF}) which is based on cosine similarity. Set of sentences S [s_1, s_2, s_3, … , s_n] has been found from preprocessing step. In this method, sentence frequency of each sentence is set as 1 (one) at first. If any sentence has cosine similarity 60% or more with any other, smaller sentence is removed and the frequency of larger

sentence is set as the summation of the frequency of both of the sentences. As there is a removal of sentence(s) on the basis of 60% or more similarity, this results redundancy elimination. The similarity ratio 60% is considered as per the threshold value of cosine similarity [40]. The sentence frequency (S_{SF}) is calculated using the Eq. (3):

$$\forall_{i\in\{1,\ldots,n\}} \begin{cases} \forall_{j\in\{1,\ldots,n\}} \\ \quad If\ (i=j)\ THEN \\ \qquad Continue \qquad // Same\ Sentence \\ \quad ELSEIF\ Sim(S_i,S_j) \geq 60\%\ AND\ len(S_i) > len\ (S_j)\ THEN \\ \qquad S_{SF(i)} = \ S_{SF(i)} + 1 \\ \qquad And\ Remove\ S_j\ from\ the\ Sentence\ Set\ S \\ \quad ELSEIF\ Sim(S_i,S_j) \geq 60\%\ AND\ len(S_j) > len\ (S_i)\ THEN \\ \qquad S_{SF(j)} = \ S_{SF(j)} + 1 \\ \qquad And\ Remove\ S_i\ from\ the\ Sentence\ Set\ S \end{cases} \tag{3}$$

where n is the number of sentences; $S_{SF(i)}$ is the sentence frequency of i^{th} sentence which is initially 1 for each sentence; $Sim(S_i, S_j)$ is the cosine similarity between sentences S_i and S_j; len(S) is the length of sentence. The cosine similarity between two sentences $S_i = [w_{i1}, w_{i2, \ldots,} w_{im}]$ and $S_j = [w_{j1}, w_{j2, \ldots,} w_{jm}]$ is measured as [41]:

$$Sim(S_i,S_j) = \frac{\sum_{k=1}^{m} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{m} w_{ik}^2 \cdot \sum_{k=1}^{m} w_{jk}^2}},\ i, j= 1, 2, \ldots, n \tag{4}$$

where w indicates the words in sentences and n is the total number of sentences.

It is noticeable that the scores of TF-IDF and SF have been boosted up for the replacement of pronoun. A noun can have better TF score, but if the pronoun form of the noun exists in any sentence, the sentence doesn't get any TF score for this pronoun. Moreover, pronoun and noun can't be matched, but after replacement of pronoun by the corresponding noun they can be recognized as same entities for sentence frequency calculation.

3.3.3 Counting the existence of numerical figure from digits and words (S_{Nc})

The third attribute is to count numerical figures for every sentence (S_{Nc}). The value of S_{Nc} for each sentence is set to 0 (zero) at first, and it is incremented by 1 for the existence of each numerical figure. In [22,42,43], 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, e.g., "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 procedures [22,42,43] can find one numerical figure from the first sentence, but unable to locate any numerical figure from the second sentence as the numerical figure "দশ" (dosh - ten) is written in words.

So, a technique is introduced to recognize Bangla numerical figure from both words and digits mentioned in the special tagging part in Section 3.2.2. All the sentences are segmented to words [w_1s_1, w_2s_1, … w_1s_2, w_2s_2, … w_ns_n] in the preprocessing step and count the numerical figure from digits and

words based on the following equations:

$$\forall_{i\in\{1,\ldots,n\}} N_{digits(i)} = Regexp(S_{(i)}, [0,1,2,3,4,5,6,7,8,9]) \quad (5)$$

$$\forall_{i\in\{1,\ldots,n\}} N_{words(i)} = Regexp(S_{(i)}, [FormatOfNumInWords] \quad (6)$$

$$\forall_{i\in\{1,\ldots,n\}} S_{Nc(i)} = N_{digits(i)} + N_{words(i)} \quad (7)$$

where n is the number of sentences; N_{digits} and N_{words} are the number of numerical figures in digits and words respectively; Regexp function returns the number of matches between the corresponding sentence and the given pattern in the second argument of this function. The pattern for matching digits is 0 to 9 and for words is FormatOfNumInWords (explained in the Section 3.2.2). Finally, both N_{digits} and N_{words} are summed up for all the sentences individually to get S_{Nc} which is the score of sentence for numerical figure.

3.3.4 Considering title words for sentence scoring

In existing methods [9,10], title words have been considered for sentence scoring. Because we have observed from the analysis of 3000 news documents that title words convey the theme of the news documents. The score of every sentence for containing title word is set to 0 (zero) at first and incremented by 1 for the existence of each title word.

For computing the title word score of any sentence (S_T), the title has been segmented to array of words TW[tw_1, tw_2, …, tw_n] and then proceed as Eq. (8):

$$\forall_{i\in\{1,\ldots,n\}} S_{T(i)} = match(S_{W(i)}, TW) \quad (8)$$

where n is the number of sentences in the input document, $S_{w(i)}$ is the array of words for i^{th} sentence, TW is the array of title words and *match* function returns the number of words matched with $S_{w(i)}$ and TW.

3.3.5 Treating the first sentence specially

In some existing methods [24,42,43], the sentence' score is depended on the position where the positional score is the highest for the first sentence and the lowest for the last. This score is gradually decreasing from the first sentence to the last. But, in most of the time especially for Bangla news documents, the first sentence is much important than any other sentences in our experiment which is explained in the lower part of this sub-section. So, traditional positional score (which is gradually decreasing as like [24,42,43]) is not applicable for the first sentence of news documents. Moreover, some existing summarization methods emphasized on sentences those contain any title word [9,10], and the first sentence contains the full title often in Bangla news documents. So, extra care is proposed 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 exists in 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 exists in summary for 88% times where the error rate is 12% (100 – 88). So, it is proposed here that the first sentence is selected in summary always if it contains any title word.

Point to be noted that this type of special care for the first sentence has been proposed for the Bangla news documents only and it may not be suitable for other types of text. After measuring all the attributes, the score of each sentence is computed using the following equation where S_k is the score of k^{th} sentence:

$$S_k = \begin{cases} w_1 \times S_{TF(k)} + w_2 \times S_{SF(k)} + w_3 \times S_{Nc(k)} + w_4 \times S_T, if \ k > 1 \\ \max(S_k) + 1, if \ k = 1 \ and \ S_{(1)} contains \ any \ title \ word \end{cases} \quad (9)$$

where 0 <= w_1, w_2, w_3, w_4 <=1; k = n, n-1, n-2, …., 1 and n is the number of sentences. The score of the first sentence will be set as the highest score + 1 if it contains any title word so that it will be selected always.

The values of coefficients w_1, w_2, w_3, and w_4 in the above equation are obtained by tuning them for the better results of summary generation. For selecting the optimal value of each coefficient, an experiment is done with a training data set of 200 documents and 600 model summaries (3 summaries for each document) generated by human. In the devised experiment, the summary is generated for each training document by setting the value of each coefficient from 0 to 1. Each time the experiment is run, the value is incremented by 0.01, and the system generated summaries are evaluated using the evaluation measure discussed later (in the evaluation and results section) in this paper. After generating all the summaries from the training documents, the average F-measure score for each value of the coefficient is calculated by comparing the system generated summaries and the corresponding model summaries. In this way, the optimal value is identified as 0.87, 0.09, 0.02, and 0.21 for w_1, w_2, w_3, and w_4 respectively. Figs. 4–7 are depicted to show the adjustment of w_1, w_2, w_3, and w_4, respectively.

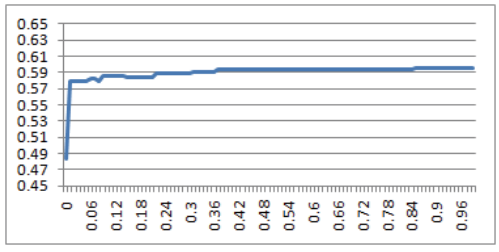

Fig. 4. F-measure for various values of w_1.

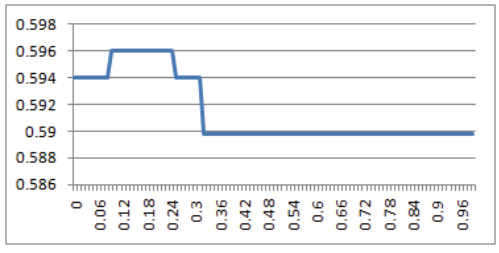

Fig. 5. F-measure for various values of w_2.

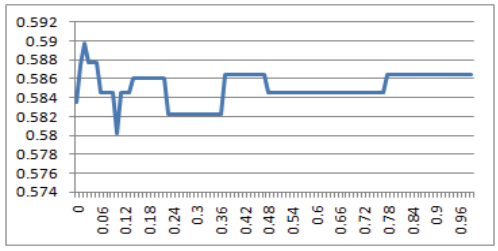

Fig. 6. F-measure for various values of w_3.

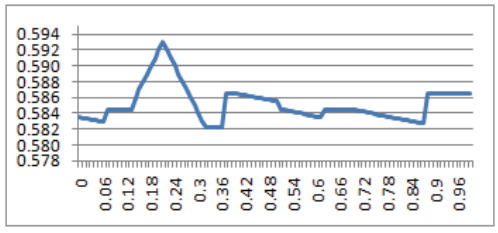

Fig. 7. F-measure for various values of w_4.

3.4 Summary Generation

After sentence ranking, one third top-ranked sentences are extracted as summary sentences as in the following equation:

$$\forall_{i\in\{1,\ldots,n/3\}}\ SumSen = SumSen \cup ExtTopScored(S) \tag{10}$$

where n is the number of sentences; ExtTopScored function extract top scored sentences from sentences' set S; SumSen is the set of summary sentences. The number of summary sentences is kept as approximately one third of the total sentences according to the ratio of source document to summary based on [44].

4. Implementation Model and Algorithm

4.1. Implementation Model

The proposed method has been implemented with the Rule-Based technique where we have utilized four components as like Rule-Based system [45] as follows:

- A rule base where in total 42 rules have been used for special tagging, dependency parsing, full named identifying and pronoun replacing.
- Inference engine to match rules and resolve the conflict based on the situation and priority of rules.
- Temporary working memory to hold the selected rules.
- Interfaces for taking input from one step and giving output to the other step.

For selecting the specified rule to apply, the current state has been observed for any word. Somewhere, we have used hidden Markov model [46], and somewhere, we have used Markov chain model [46]. Because, the states are known for some cases (Markov Chain Model), and in some cases, states are needed to be discovered (hidden Markov model).

4.2. Algorithms

Algorithm 1: Replacement of pronoun by the corresponding noun

Input:
S1: List of sentences from the segmentation of a Bangla news document
W: List of words with the corresponding tags after dependency parsing for all sentences
R: List of rules for replacing pronoun with the priority of rules
ListNE: List of full human named entities
NEA: An array for containing the parts of name as index and full name as value
PRO: List of pronouns for replacement
Output: S2: List of sentences after replacement of pronouns
Begin
TmpPRO = empty // Contains the list of pronouns of a sentence
NEs = empty // Contains the named entities of a sentence
TmpNEs = empty // Contains some named entities temporarily

For i = 2 to n // n is the total number of sentences in S1. Loop is started from second sentence
TmpPRO = Get the pronouns in the i^{th} sentence as per the list of pronouns in PRO
NEs = Get the named entities of the $(i-1)^{th}$ sentence
If TmpPRO = 'তিনি' OR 'ইনি' OR 'উনি' OR 'সে' Then // for subject
TmpNEs = Get the named entities from NEs without any suffix as subject
If Count(TmpNEs) == 1 // if the subject is singular named entity
Replace the TmpPRO by TmpNEs
End If
Else If TmpPRO = 'তার' OR 'তাকে' OR 'তাহাকে' OR 'তাহার' Then // for object
TmpNEs = Get the named entities from NEs with suffix as 'কে, রে, এর, রের, র'
If Count(TmpNEs) == 1 // if the object is singular named entity
Replace the TmpPRO by TmpNEs
End If
End If
Loop
S2 = S1
Return S2 // List of sentences after replacement of pronouns by the corresponding nouns
End

Algorithm 2: Sentence ranking and summary generation
Input:
S: List of sentences after replacement of pronouns by the corresponding nouns
WT: List of title words of the input document
Output: SUMMARY: Summary sentences
Begin
$S_{TF\text{-}IDF}$ = 0 // TF-IDF score
S_{SF} = 0 // Score of sentence frequency
S_{TITLE} = 0 // Score of title words
S_{Nc} = 0 // Score of numerical figure
SCORES = empty // contains the score of all the sentences individually
For i = 1 to n // n is the total number of sentences in S
$S_{TF\text{-}IDF}$ = TF-IDF score based on equation 2
S_{SF} = Sentence frequency score based on equation 3
S_{Nc} = Score for the existence of numerical figure from words and digits using equation 7
S_{TITLE} = Score for title words as per equation 8
SCORES[i] = Accumulated score of i^{th} sentence by equation 9
Loop
Sort SCORES in Descending order
For i = 1 to n // n is the total number of sentences in S
If SCORES[i] >= MAX 3 Scores from SCORES
SUMMARY = SUMMARY U S[i] // Keep the i^{th} sentence in summary
End if
Loop
Return SUMMARY
End

5. Evaluation and Results

5.1 Dataset

Since there is no benchmark dataset of Bangla news document to evaluate our proposed method, we have collected 3,400 Bangla news documents (each document has 18 to 25 lines of Unicode text) from the most popular Bangladeshi newspaper. These news documents contain a variety of news that covers a broad range of topics like politics, sports, crime, economy, environment, etc. We have analyzed 3,000 documents to understand the structures of sentences in news documents and identify the rules for replacing pronouns by the corresponding nouns. For other 400 news documents, three human judges have generated summaries for each document. These human-generated summaries are regarded 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 a training set for adjusting the value of w_1, w_2, w_3, and w_4 in the previous section and (ii) other 200 documents with corresponding model summaries are treated as a performance evaluation set. The performance evaluation set has been utilized for evaluating the proposed text summarization system as well as the efficiency measurement of the process of replacing pronoun. The evaluation set has been uploaded to the Internet so that other researchers may evaluate their systems with this [47].

Point to be mentioned that the dataset of 400 test documents is around ten times larger than the evaluation dataset of some existing methods [6,7,24]. Moreover, some existing methods [6,7,24] were evaluated against one model summary only. But, the proposed method in this paper has been evaluated with three model summaries of each test document. The remarkable point is that human generated model summaries were also used for English text summarization methods despite the existence of benchmark dataset [48,49] and for other languages where there was no benchmark dataset [6,7,20]. At last, ROUGE [50] has been applied, a widely used metric, to evaluate the automatically generated summaries of our proposed method. Updated ROUGE package has been utilized here as it can be applied to Unicode text [26].

5.2 Evaluation

Evaluating the quality of a summary is a difficult problem, principally because there is no "ideal" summary [44]. Even for relatively straightforward news articles, human summarizers tend to agree only approximately 60% content overlapping [44].

In this research, the summary of proposed system has been compared with three model summaries of 200 news documents each and the results of evaluation is the average results of the comparisons. The Precision, Recall, and F-measure are brought into play here as these have long used as important evaluation metrics in information retrieval field [51]. If 'A' indicates the number of sentences retrieved by summarizer and 'B' shows the number of sentences that are relevant as compared to target set, Precision, Recall and F-measure are computed as:

$$Precision\ (P) = \frac{A \cap B}{A} \tag{11}$$

$$Recall\ (R) = \frac{A \cap B}{B} \tag{12}$$

$$F\text{-}measure = \frac{2 \times P \times R}{P+R} \tag{13}$$

5.3 Experiments and Results

The proposed method has the following features:

(i) Pronoun replacing by the corresponding noun to minimize the number of dangling pronouns.

(ii) Sentence ranking by (a) term frequency inverse document frequency calculation, (b) sentence frequency measurement with redundancy elimination, (c) counting the existence of numerical data from digits and words, and (d) computing title word score.

(iii) Considering the first sentence especially if it contains any title word.

(iv) Tuning the coefficients of all the attributes listed in the above point (ii).

For justifying the significance of each feature of this proposed method, step by step progress of performance has been given in the Fig. 8. Precision, Recall, and F-measure have been calculated using (11), (12), and (13) respectively upon training dataset (discussed at the beginning of this section). In Fig. 8, the utilized features in each step include all the features of the previous step(s), and better performance is obtained by combining all the features.

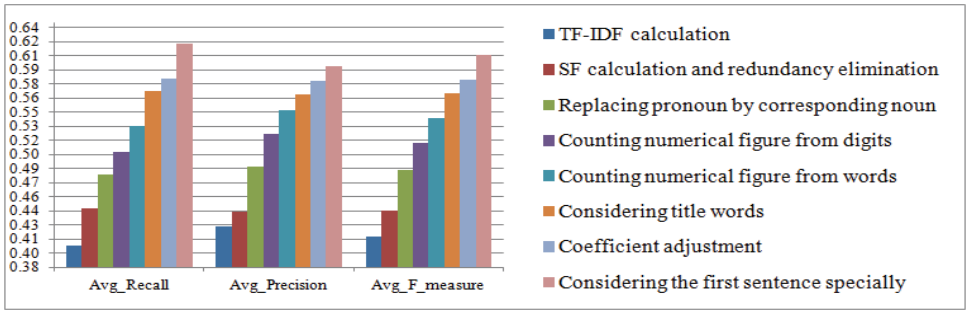

Fig. 8. Step by step improvement of performance for including each feature.

For the efficiency judgment 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 document, and compute the average value of Precision, Recall and F-measure with ROUGE automatic evaluation package (Table 5).

Table 5. Average of ROUGE-1 and ROUGE-2 scores of the proposed system for 200 documents with 95% confidence interval

	Avg_Recall	Avg_Precision	Avg_F_measure
Average of ROUGE-1 score	0.6134	0.5877	0.6003
Average of ROUGE-2 score	0.5924	0.5506	0.5708

The proposed procedure along with four existing methods [6,7,10,24] has been implemented with a server side scripting language named PHP (Hypertext Preprocessor). All the methods have been evaluated with same data set [47] for which the results have been varied from the results claimed by the

corresponding authors [6,7,10,24]. Comparison results, based on ROUGE-1 and ROUGE-2, have been depicted in Fig. 9 and Fig. 10, respectively where method 1 is presented in [10], method 2 is in [6], method 3 is in [7], and method 4 is in [24]. In our implementation for all the methods, one-third sentences are selected as a final summary, and the same list of stop words [2] is used.

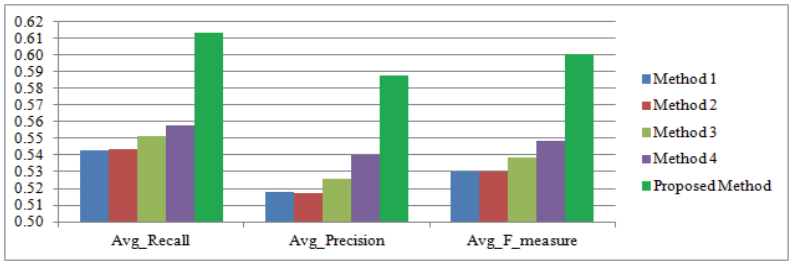

Fig. 9. Comparison based on ROUGE-1 scores of 200 documents with 95% confidence interval.

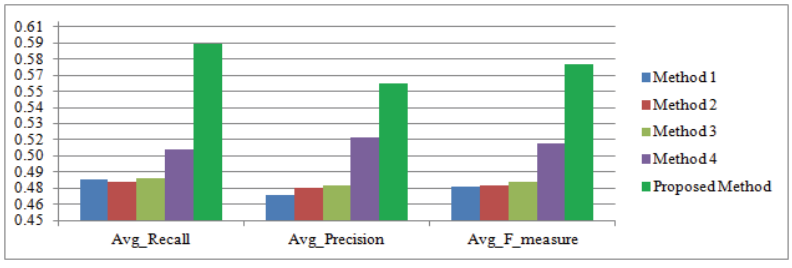

Fig. 10. Comparison based on ROUGE-2 scores of 200 documents with 95% confidence interval.

Another important aspect of the proposed method is to focus on the problem of dangling pronoun in summary for the first time. An analysis is given in the following Table 6 regarding the performance of dangling pronouns minimization as follows:

Table 6. Number of dangling pronouns in the output of Bangla text summarization systems for 200 news documents

#	Methods	Number of dangling pronoun
1	Method 1	80
2	Method 2	76
3	Method 3	81
4	Method 4	78
5	Proposed method	8

5.4 Discussion on Results

In our approach, some innovative features have been introduced for getting better performance. In Fig. 8, step by step improvement has been presented for incorporating each feature. After generating summary by using only term-frequency (the first step in Fig. 8), the F-measure score has been found as 0.4124 (using training dataset). Then, after incorporating each feature, the F-measure score has been raised as follows (Table 7).

Table 7. Percentage of performance improvement for including each feature

Features	Improvement (%)
Calculating sentence frequency	6.71
Replacing pronouns by the corresponding nouns	9.66
Counting numerical figures from digits	6.15
Counting numerical figures from words and digits	5.11
Considering title words	4.96
Tuning coefficients	2.47
Considering the first sentence specially	4.47

It has been found from the comparison results given in Figs. 9 and 10 that the proposed method outperforms the four latest existing methods. The improvement of performance from the four methods is given in Table 8.

Table 8. Improvement of performance in proposed method from the four existing methods

#	Methods	Improvement (%)	
		Based on ROUGE-1 score	Based on ROUGE-2 score
1	Method 1	13.28	20.85
2	Method 2	13.26	20.45
3	Method 3	11.48	19.83
4	Method 4	09.39	12.52

The proposed method has another remarkable feature to minimize the dangling pronoun in summary. Based on our analysis, only one dangling pronoun, in summary, is enough to deliver the wrong message. The evaluation result in the Table 6 shows that the proposed system has significantly minimized the number of dangling pronoun in the summary where the minimization rate is 90%, 89.50%, 90.12%, and 89.75% from the method1 [10], method2 [6], method3 [7], and method4 [24], respectively.

So, it can be said that the performance of the proposed system is better not only for higher ROUGE evaluation scores but also for minimizing dangling pronoun in summary to deliver an unambiguous message.

6. Conclusion and Future Works

There are a lot of research works for English text summarization, but these may not be directly applicable to Bangla text because of the complexities of Bangla language in the structure of sentences, grammatical rules, inflection of words, etc. The research for Bangla is also tough as there is hardly any automatic tool, no database for ontological knowledge of words and limited scope of knowledge sharing. Despite these difficulties, an approach of Bangla news document summarization has been proposed in this paper based on pronoun replacement and sentence ranking. The replacement of

pronoun by the corresponding noun has been accomplished with 71.80% accuracy. By this way, we have minimized the dangling pronoun in summary for 89.75% than the latest Bangla text summarization system. For replacing pronoun, nature of each word has been identified with general tagging, special tagging, and dependency parsing; subject and object of each sentence have also been recognized, and in total 79.50% words have been identified. It is expected that the process of replacement of pronoun will be helpful for any Bangla information retrieval procedure. For summary generation, sentences are ranked using term-frequency, sentence frequency, numerical figure, and title words. Moreover, numerical figures have been identified from words as well as digits for 100%, and the first sentence has been included in summary for containing any title word. All the features, utilized for sentence ranking, have been adjusted with tuning and step by step progress has been depicted for each. In the literature review section, it has been indicated with the references that most of the incorporated features in various existing methods have been taken from existing English text summarization systems. But, the proposed system has introduced some new features (replacement of pronouns by the corresponding nouns, sentence frequency calculation, and numerical figures identification from words). In the evaluation, the proposed system outperforms the four latest Bangla text summarization systems, and the performance has been increased to 9.39% (based on ROUGE-1 F-measure score) and 12.52% (based on ROUGE-2 F-measure score) than the latest existing method.

This research work has implications on Bangla language only. In future, we hope to make it language independent and introduce more features for sentence ranking to make the system generated summary close to the human generated summary.

Acknowledgement

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] D. Ai, Y. Zheng, and D. Zhang, "Automatic text summarization based on latent semantic indexing," *Journal of Artificial Life and Robotics*, Springer, vol. 15, no. 1, pp. 25-29, 2010.

[2] M. Kunder, "The size of the World Wide Web," 2016 [Online]. Available: www.worldwidewebsize.com.

[3] R. Ferreira and S. Luciano, "A multi-document summarization system based on statistics and linguistic treatment," *Journal of Expert Systems with Applications*, vol. 41, no. 13, pp. 5780-5787, 2014.

[4] M. M. Haque, S. Pervin, and Z. Begum, "Literature review of automatic multiple documents text summarization," *International Journal of IAS*, vol. 3, no. 1, pp. 121-129, 2013.

[5] M. M. Haque, S. Pervin, and Z. Begum, "Literature review of automatic single document text summarization using NLP," *International Journal of IAS*, vol. 3, no. 3, pp. 857-865, 2013.

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

[7] K. Sarkar, "An approach to summarizing Bengali news documents," in *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, Chennai, India, 2012, pp. 857-862.

[8] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159-165, 1958.

[9] H. P. Edmundson, "New methods in automatic extracting," *Journal of the ACM*, vol. 16, no. 2, pp. 264-285, 1969.

[10] M. I. Efat, M. Ibrahim, and H. Kayesh, "Automated Bangla text summarization by sentence scoring and ranking," in *Proceedings of International Conference on Informatics, Electronics & Vision (ICIEV)*, Dhaka, Bangladesh, 2013, pp. 1-5.

[11] Banglapedia: the National Encyclopedia of Bangladesh. Dhaka: Asiatic Society of Bangladesh, 2003.

[12] G. Miller, "WordNet: a lexical database for English," *Communications of the Association for Computing Machinery (CACM)*, vol. 38, no. 11, pp. 39-41, 1995.

[13] Bengali WordNet, "Indradhanush WordNet Development for the Bengali Language," Dept. of Information Technology, Ministry of Information and Communication Technology, Govt. of India, 2017, [Online]. Available: http://www.isical.ac.in/~lru/externalprojects.html.

[14] M. A. Karim, M. Kaykobad, M. Murshed, *Technical Challenges and Design Issues in Bangla Language Processing*. Hershey, PA: Information Science Reference, 2013.

[15] N. Uzzaman, "Bangla language and research on Bangla language processing: its motivation and impact!," 2008. [Online]. Available: https://sites.google.com/a/naushadzaman.com/www/BigPicture-URCS-NZ-Bangla.pdf?attredirects=0.

[16] M. M. Haque, S. Pervin, and Z. Begum, "Automatic Bengali news documents summarization by introducing sentence frequency and clustering," in *Proceedings of 18th International Conference on Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2015, pp. 156-160.

[17] 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, 2010.

[18] H. Saggion and T. Poibeau, "Automatic text summarization: past, present and future," in *Multi-source, Multilingual Information Extraction and Summarization*. Heidelberg: Springer, 2013, pp. 3-21.

[19] E. Canhasi and I. Kononenko, "Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization," *Expert Systems with Applications*, vol. 41, no. 2, pp. 535-543, 2014.

[20] A. M. Azmia and S. Al-Thanyyan, "A text summarizer for Arabic," *Journal of Computer Speech & Language*, vol. 26, no. 4, pp. 260–273, 2012.

[21] T. Islam and S. M. Masum, "Bhasa: a corpus-based information retrieval and summariser for Bengali text," in *Proceedings of the 7th International Conference on Computer and Information Technology*, Dhaka, Bangladesh, 2004.

[22] N. Uddin and S. A. Khan, "A study on text summarization techniques and implement few of them for Bangla language," in *Proceedings of 10th International conference on Computer and Information Technology*, Dhaka, Bangladesh, 2007, pp. 1-4.

[23] A. Das and S. Bandyopadhyay, "Topic-based Bengali opinion summarization," in *Proceedings of the 23rd International Conference on Computational Linguistics (COILING10)*, Beijing, China, 2010, pp. 232-240.

[24] K. Sarkar, "A keyphrase-based approach to text summarization for English and Bengali documents," *International Journal of Technology Diffusion (IJTD)*, vol. 5, no. 2, pp. 28-38, 2014.

[25] S. R. El-Beltagy and A. Rafea, "KP-Miner: a keyphrase extraction system for English and Arabic documents," *Journal Information Systems*, vol. 34, no. 1, pp. 132-144, 2009.

[26] ROUGE 2.0: a Java package for automatic summary evaluation [Online]. Available: http://www.rxnlp.com/rouge-2-0/.

[27] Indian Statistical Institute, "List of stop words for Bengali language," 2016 [Online]. Available: http://www.isical.ac.in/~fire/data/stopwords/.

[28] M. Islam, M. Uddin, and M. Khan, "A light weight stemmer for Bengali and its use in spelling checker," Center for Research on Bangla Language Processing, Dhaka, Bangladesh, 2007.

[29] Society for National Language Technology Research, "Bengali POS Tagger," [Online]. Available: http://nltr.org/snltr-software.

[30] A. Das and S. Bandyopadhyay, "SentiWordNet for Bangla," in *Knowledge Sharing Event-4: Task 2: Building Electronic Dictionary*. Mysore, India: Knowledge Sharing Event, 2010.

[31] M. Chowdhury, I. Khalil, and M. H. Chowdhury, *Bangla Vasar Byakaran*. Dhaka: Ideal Publishers, 2000.

[32] H. Mamud, *Vasa Shikkha, Bangla Vasar Byakaran O Rachanariti*. Dhaka: The Atlas Publishing House, 2011.

[33] Occupation in Bangladesh, "Name of occupation in largest job site," [Online]. Available: http://bdjobs.com.

[34] Gpedia [Online]. Available: http://www.gpedia.com/bn.

[35] A. Ekbal, R. Haque, and S. Bandyopadhyay, "Named entity recognition in Bengali: a conditional random field approach," in *Proceedings of International Joint Conference on Natural Language Processing*, Hyderabad, India, 2008, pp. 589-594.

[36] Z. R. Siddiqui, *English-Bangla Dictionary*, 2nd ed. Dhaka: Bangla Academy, 2011.

[37] G. M. Kiron, Ajker Bishaw (General Knowledge, Bangladesh and International Affairs). Dhaka: Premier Publications, 2014.

[38] Bengali names [Online]. Available: http://www.indiachildnames.com/regional/bengalinames.aspx.

[39] Post office of Bangladesh [Online]. Available: http://www.bangladeshpost.gov.bd/postcode.asp.

[40] L. Gravano, P. Ipeirotis, H. 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, no. 4, pp. 28-34, 2001.

[41] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 25, no. 5, pp. 513–523, 1988.

[42] A. Abuobieda, N. Salim, A. T. Albaham, A. H. Osman, and Y. J. Kumar, "Text summarization features selection method using pseudo genetic-based model," in *Proceedings of International Conference on Information Retrieval & Knowledge Management*, Kuala Lumpur, Malaysia, 2012, pp. 193–197.

[43] M. A. Fattah and F. Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization," *Computer Speech and Language*, vol. 23, no. 1, pp. 126–144, 2009.

[44] D. R. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," *Journal of Computational Linguistics*, vol. 28, no. 4, pp. 399-408, 2002.

[45] Rule based system [Online]. Available: http://www.j-paine.org/students/lectures/lect3/node5.html.

[46] Markov process [Online]. Available: digital.cs.usu.edu/~cy an/CS7960/Markov_Chains.ppt

[47] Bangla Natural Language Processing Community [Online]. Available: http://bnlpc.org/research.php.

[48] R. Ferreira, F. Freitas, L. de Souza Cabral, R. D. Lins, R. Lima, G. Franca, S. J. Simskez, and L. Favaro, "A four dimension graph model for automatic text summarization," in *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Atlanta, GA, 2013, pp. 389-396.

[49] J. Chen and H, Zhuge, "Summarization of scientific documents by detecting common facts in citations," *Future Generation Computer Systems*, vol. 32, pp. 246–252, 2014.

[50] C. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of the Human Technology Conference (HLT-NAACL-2003)*, Edmonton, Canada, 2003, pp. 71-78.

[51] S. Hariharan, T. Ramkumar, and R. Srinivasan, "Enhanced graph based approach for multi document summarization," *The International Arab Journal of Information Technology*, vol. 10, no. 4, pp. 334-341, 2013.

Md. Majharul Haque https://orcid.org/0000-0003-3144-1717

He completed Bachelor of C.S.E. from the University of Development Alternative, M.S. degree in IT from University of Dhaka, Bangladesh. Since 2011, he is with the C.S.E Department of University of Dhaka, Bangladesh as a PhD student.

Suraiya Pervin

She completed both her Bachelor of Science and Masters of Science degree from the University of Dhaka. She did her Ph.D. from IIT, Kharagpur, India. She is now a professor at the University of Dhaka at the Department of C.S.E.

Zerina Begum

She completed Bachelor of Science degree, Masters of Science degree and M.Phil. from the University of Dhaka. She did her Ph.D. from the University of Dhaka in Applied Physics. She is now a professor at the University of Dhaka at the institute of IT.