Thesis for the Degree of Doctor of Philosophy

A New Approach of Bangla News Document Summarization

Md. Majharul Haque

Registration No: 106/2015-16

Department of Computer Science and Engineering
University of Dhaka
Dhaka - 1000, Bangladesh

November, 2017

A New Approach of Bangla News Document Summarization

Md. Majharul Haque

Department of Computer Science and Engineering

University of Dhaka

Dhaka - 1000, Bangladesh

November, 2017

A New Approach of Bangla News Document Summarization

by

Md. Majharul Haque

Registration No: 106/2015-16

Supervised by

Prof. Suraiya Pervin, Ph. D.

Prof. Zerina Begum, Ph. D.

Submitted to the Deptartment of Computer Science & Engineering
of the Faculty of the Engineering and Technology in
University of Dhaka for partial fulfillment
of the requirements of the degree of
Doctor of Philosophy

As the candidate's supervisor, I have approved this dissertation for submission.

Name: Dr. Suraiya Pervin

Signed: _____

Name: Dr. Zerina Begum

Signed: _____

Declaration of Authorship

We declare that this thesis titled "A New Approach of Bangla News Document Summarization" and the works presented in it are our own. We confirm that:

- The full part of the work is done during Ph. D. research study in the University of Dhaka, Bangladesh.
- Any part of this thesis has not previously been submitted for a degree or any other qualification in this University or any other institution.
- We have consulted the published works of others with appropriate references.
- This thesis work is done entirely by us and our contributions and enhancements from other works are clearly stated.

Signed:

Candidate

Countersigned:

Supervisor: Dr. Suraiya Pervin

Co-supervisor: Dr. Zerina Begum

Abstract

The object of this research work is to propose a new method of automatic Bangla news document summarization. It is noticeable that the existing English text summarization systems may not be directly applicable for Bangla for the complexities of Bangla language in grammatical rules, structure of sentences, different placement of subject and object, etc. Again, the research work for Bangla language processing is difficult because there is hardly any automated tool to facilitate research work. In this challenging situation, a new approach for Bangla news document summarization has been presented here by introducing pronoun replacement and an improved version of sentence ranking. Major parts of this approach are (i) preprocessing the input document, (ii) word tagging, (iii) replacement of pronoun, and (iv) sentence ranking. Replacement of pronoun has been accomplished here for the first time to minimize the dangling pronoun in summary. After replacing pronoun, sentences are ranked by considering (i) term frequency, (ii) sentence frequency, (iii) numerical figures (presented in words and digits), and (iv) title words. If two sentences has at least 60% cosine similarity, frequency of larger sentence is increased and remove smaller sentence which eliminates redundancy. Moreover, the first sentence has been specially considered for containing any title word. Again, numerical figure has been

identified from words and digits to assess the importance of sentences despite the variety of forms for any numerical figure in Bangla. For achieving the target of this proposed method, 3000 news documents have been analyzed and some Bangla grammar books have been studied. The effect of each incorporated feature has been demonstrated with step by step performance analysis. From the evaluation results of the proposed method, the F-measure scores for ROUGE-1 and ROUGE-2 have been found as 0.6003 and 0.5708 respectively and the accuracy of pronoun replacement has been found as 71.80%. The proposed method has minimized the dangling pronoun in summary for 89.75% than the latest Bangla text summarization system. Again, the text summarization performance of the proposed method has been observed as 9.39% (based on ROUGE-1 F-measure score) and 12.52% (based on ROUGE-2 F-measure score) better than the latest existing method.

Acknowledgment

By the grace of Almighty Allah who is the most gracious and merciful, I have accomplished my Doctoral Thesis. Special innumerous thanks go to my supervisors Prof. Dr. Suraiya Pervin and Prof. Dr. Zerina Begum for their indispensable guidelines in my research work.

I wish to thanks Information and Communication Technology (ICT) Division, Ministry of ICT, Government of the Peoples Republic of Bangladesh for awarding me a fellowship scholarship for Ph.D. research work. Thanks goes to the Central Bank of Bangladesh for approving my deputation for Ph.D. study.

I would like to give thanks to my honorable teachers of the department CSE and IIT, University of Dhaka, for their invaluable comments and suggestions. It is noticeable that I got a comfortable place for research work in the Samsung Innovation Lab in dept. of CSE. I am also indebted to all the staff of the University of Dhaka for their help.

Thanks go to my fellow researchers, colleagues and friends who assisted me in computing, writing, and thinking. I also remember the encouragement of my senior colleague Md. Amir Hossain Pathan sir of Bangladesh Bank.

Md. Majharul Haque

November, 2017

Table of Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1
Introduction

1.1 Introduction

The objective of the present work is to develop an innovative method of automatic Bangla news document summarization. A sophisticated approach has been proposed here by introducing pronoun replacement and an improved version of sentence ranking. In this chapter, an overview of text summarization and its necessity in real life has been described. Explanation has been given on automatic text summarization with its' application, necessity of Bangla news document summarization, motivating scenario, challenges in the research work on Bangla language, and the goals of this research work with major contributions. Some terminologies have also been enlightened as natural language processing, expert system, semi supervised process, markov model, etc. The structure of this endeavor has been outlined at the end of this chapter.

1.2 Text Summarization

Text summarization is a process of distilling the most salient information from source(s) for any particular purpose, and outlines important aspects of the

document(s) in a precise way. It should be informative to indicate the document's relevance to the reader. It is also non-repetitive for being as brief as possible and surely providing the most significant content of the text. Simply, a summary text is a derivative of a source(s) text condensed by selection and/or generalization on important content [3]. Eduard Hovy [4] and Radev et al. [5] formally defined summary as:

A summary can be defined as a text that is produced from one or more texts, that contain a significant portion of the information of the original text(s), and that is no longer than half of the original text(s).

1.3 Application of Text Summarization

In our every sphere of daily living, a huge number of documents are found to get information which are quite difficult to read. Sometime we send E-mail, Fax and SMS etc. from one spot to another of the world where concise information is required. In this way, text summarization can be used in the following issues:

- Save reading time
- Summarize reports to SMS or WAP-format for small devices such as mobile phones/PDA
- View compressed figure of the search results in search engine
- Multi-document summarization for composing the state-of-the-art work of a topic
- Make a brief description of a book which will help us to decide whether we will buy the book or not

- Generate short description of television, radio, and entertainment programs
- Navigate and look for the information in the World Wide Web
- Create summary of books of digital libraries, news portals, journals, etc.

Several types of summaries can be inferred from a text as follows: a) indicative summaries (that provide an idea of the text without giving any content), b) informative summaries (that do provide a short version of the content) [6].

1.4 Automatic Text Summarization

Automatic text summarization is to generate summary by machine as computer, cell phone, PDA, etc. The primary goal of automatic text summarization is to condense the source text with preserving its significant content within a short time [7,8]. In the automatic text summarization system, the score is calculated for each sentence on the basis of location, cue words, length, etc. Finally, top scored sentences are selected from the source document(s) to generate summary. General process flow of automatic text summarization system is given in the figure 1.1 [1].

On the basis of summarization procedures, approaches can be divided into two broad categories - (i) extraction and (ii) abstraction [9]. Extraction techniques simply copy the most significant and representative sentences. But in abstractions, special forms of summaries are generated that are formed from the most imperative topics in the text. Reformulation of contents is done while abstraction. In most of the cases of extraction, importance of a sentence is measured based on its format and position in the text rather than its semantic information [9]. Extraction is domain independent and needs no background knowledge but abstraction is domain

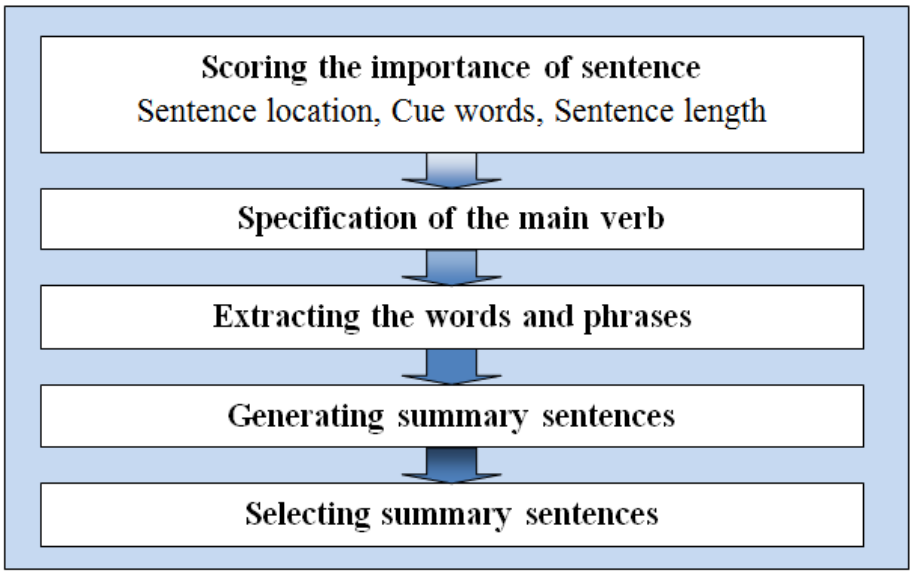


Figure 1.1: General process flow of automatic text summarization system [1]

dependent which requires background knowledge and is specific goal oriented [10]. Deep natural language processing is required in abstraction which is yet to reach a mature stage [11]. In abstraction, the selected sentences are further processed to restructure them [12] but extraction just settle on what is essential and what is unnecessary [10].

Some systems reduce the sentences to form shorter summary which is called sentence reduction [13]. There is another diversification in summaries such as generic or query-focused [14]. Query-focused summary presents the information that is most relevant to the given queries by users but a generic summary gives an overall sense of the document's content.

1.5 Importance of Automatic Text Summarization

The amount of available information increases rapidly with the development of information technology and wide use of Internet [15]. Nowadays, we share, store, write and publish text through the revolutionized advances of Internet, hardware and software. In this regard, a new era of information explosion is impending [15]. The estimated size of the web in February, 2015 was around 4.42 billion pages [16] and this number is growing everyday at a fast pace [17]. In this situation, it is really hard to come across the expected information from such a huge materials. Life is made up of some small amount of moments where we cannot go through all things' top to bottom. So, no one has time to read everything and we often have to make critical decisions based on what we are able to assimilate.

People use search engines (i.e. Google, Yahoo, etc.) as information retrieval tools to come across required information where outputs are abundant. Users often find each retrieved document very lengthy that is very tedious and time consuming to read. Therefore the automatic text summarization is needed to process the huge amount of Internet Data efficiently, scavenging useful information from it [17]. Summaries can aid users to evaluate the relevance of a document without thoroughly reading the whole text.

It is noticeable that research work on automatic text summarization was commenced with English text in 1958 [18]. Lots of proficient researchers presented their methods on English text summarization using various types of automated tools and techniques [3,4,19–26]. Some helpful tools (i.e. MS AutoSum, Summarist, etc.) are also available to generate automatic summary for English text but the

output is still not in 100% expected level [21]. So, the research works to find a sophisticated solution for automatic text summarization is still an ongoing process especially in this age of information overload. It is a matter of fact that unlike English which has seen a large number of systems developed to cater to it, other languages are less fortunate [27]. So, the development of text summarization has no mentionable progress for other languages specifically Bangla.

Point to be stated that we are Bangladeshi people and our National language is Bangla [28]. The overwhelming majority of Bangla speaking people lives in the eastern flank of South Asia that surrounds the Bay of Bengal where most of them are living in Bangladesh [29]. Despite significant progress in Information and Communication Technology (ICT) and the availability of a huge, enriched English knowledge database around the globe, the potential ICT benefit continues to elude a large majority of the Bangla-speaking population who are not equipped with English [29]. A very few research works have been conducted on automatic text summarization in Bangla language processing (BLP) [2, 30–36]. In these circumstances, we are focusing on Bangla text summarization. Text summarization can be needed for scientific documents, literature, news documents, books, etc. Today, online contents of Bangla news documents are growing very rapidly and mass people are reading this regularly. So, we are going to propose a method for Bangla news document summarization to make something helpful for mass people. Again, based on our observation, Bangla news documents have some common structures of text than other documents for which we have found news document summarization comparatively easier than other types of document. Moreover, the type of document which is comparatively easy to summarize has been selected because research work on Bangla has some noticeable challenges still (discussed in

section 1.7). The necessity for Bangla news document summarization is given in more details in the next section.

1.6 Importance of Text Summarization in BLP

Bangla is the national language of Bangladesh. Around 98% of the total population of Bangladesh speak in Bangla as their native language [37]. Based on the economic survey - 2015, there are 62.30% literate people in Bangladesh and most of them are used to Bangla language only. It is the official language of the state of West Bengal and the co-official language of the state of Tripura, Assam and the union territory of Andaman and Nicobar Islands [37]. Mother tongue of around 8.11% people of India [38] is Bangla. Moreover, it is the seventh most spoken language among around 3500 languages all over the world [28]. Around 250 million of people are using Bangla in their daily living [28]. In this situation, Bangla text summarization (BTS) specially for Bangla news document summarization is important for the followings:

- Many computerized contents are being developed in Bangla all over the world.
- Electronic Bangla text is increasing without any bounds in the cyber world and people are overloaded with huge volume of texts.
- Online version of Bangla newspaper is also growing rapidly.
- Mass people of Bangladesh are spending lots of time in reading Bangla newspaper regularly.

Therefore, an efficient Bangla text summarization technique is essential for researchers, international news agencies and individuals.

1.7 Motivation

We have already discussed that the volume of online Bangla text is growing rapidly. But, very few research works have been conducted for Bangla text summarization. However, the existing methods [2, 30–36] have some limitations as:

- Numerical figure, which is presented in words, can't be identified in the existing methods. So, some significant information (quantity of something special, specific date, amount of money), those are written in words, will not be considered in summary.
- Some sentences are selected by the existing methods with dangling pronoun where the corresponding noun is missing. It is noticeable that only one dangling pronoun in summary is enough to raise misunderstanding. In this regard, the summary can't deliver appropriate message. So, the user will be misguided with misinformation.
- The first sentence has not been significantly considered in the existing methods where it is positionally first and contains the full title often for Bangla news document.

We may consider the following figure 1.2 as a motivation example:

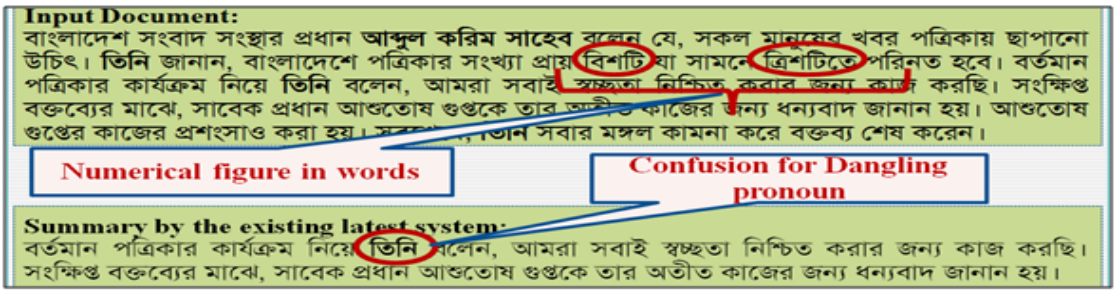


Figure 1.2: Sample input document and the summary generated by the latest existing method [2]

In the figure 1.2, a sample summary generated by the latest existing method [2] has been given with the indication of problems. The problems are:

i. Missing important information as numerical figure presented in words.

ii. The first sentence is not considered specially.

iii. Confusion for dangling pronoun which will misguide users.

iv. One sentence may cover information of multiple sentences. But such sentences are not identified.

So, the existing methods of Bangla text summarization [2,30–36] are not enough to serve the purpose for summarizing Bangla text document. Therefore, there is a pressing necessity to develop an efficient method for alleviating the burden of large community of Bangla speaking people. With this motivation, we have identified the major goals of this thesis. But there are also some challenges and difficulties for doing research work on Bangla language.

1.8 Challenges in the Research Work on Bangla Language

Automatic text summarization requires a model of human comprehension, production of language and semantic analysis, etc. [23]. The performance of baselines is very close to artificial intelligence system because summarizing requires deep understanding of texts [39]. The contents of document are not well defined always which throw a challenge in generating the desired precis. It can be large documents like journals, novels and books or it can be short documents such as emails, news articles and dialogue' scripts. Text from various sources has its' different structures and difficulties. Even all the journals or novels of same genre have no similar format. So the summarization procedure should be independent from structure of document. There are also some significant aspects that should be followed while generating summary [40]:

- Summaries may be formed from single or multiple documents
- Summaries must preserve significant information
- Summaries will be short with entire specification

To alleviate the burden of large volume of text, very few research works have been conducted for Bangla [2, 30–36]. So, for this large community of Bangla-speaking people, more research work is very much necessary especially for Bangla information retrieval. But research work for Bangla language is difficult for the following issues:

i. Based on our study, automatic tools are hardly available for Bangla language to facilitate research work.

ii. For Bangla language, there is no lexical database like WordNet [41]. Though a similar tool is ongoing to be developed it has limited features [42].

iii. There is no database of ontological meanings for Bangla words that can be used programmatically.

iv. Subject and object of all sentences need to be identified for proper recognition of structures of sentences which is complex in Bangla than that of English. Because, the placement of subject in English sentence is generally before the verb phrase, auxiliary verb or it may appear after the word 'by' in passive voice but subject may be existed in several places in Bangla sentence.

v. The scope of knowledge sharing is also limited as there are a few researchers in this arena.

vi. Works have been done by researchers in scattered and there is very little unification [43].

vii. Scarcity of free and open source software [43].

viii. Bangla language has been derived from Sanskrit and mostly maintained the rules of inconsistencies [43].

1.9 Terminologies

1.9.1 Natural Language Processing (NLP)

Natural Language Processing is a field of computer science, artificial intelligence and linguistics. It deals with the interactions between machines and human languages which accomplish task on analyzing, understanding and generating language [44]. Naturally, NLP is used by humans in order to interact with computers in both oral and written contexts instead of computer languages [44]. It is an interdisciplinary field based on versatile grounds as follows [45]:

i. Computer engineering which provides methods for model illustration, algorithm devise and accomplishment

ii. Linguistics which categorizes linguistic forms and practices

iii. Mathematics where formal models and methods are analyzed

iv. Psychology which studies models and theories of human behavior

v. Statistics which offers procedures for predicting measures based on sample records

There are many challenges involved in NLP such as natural language understanding which enables computers to derive meaning from human language. Another major issue is natural language production to generate language on human readable format. Some general applications of NLP are given below:

i. Word Sense Disambiguation (WSD) [46, 47]: Identify the sense of a given input word based on its context.

ii. Information Retrieval (IR) [48]: Information retrieval has itself a large ground of research by which necessary information or knowledge can be discovered. This area overtakes traditional database searching. Nowadays, hundreds of millions of people are using IR systems when they use web search engine or find documents from archive.

iii. Machine Translation [49]: It is helpful for translating from one language to other. This application is very useful for bossiness and scientific purposes because the international collaboration grows exponentially.

iv. Question Answering (QA) [50]: In this task, NLP, IR and machine learning are integrated together and it is really complex. The principal aim of QA is to localize the correct answer to a question written in natural language from a non-structured collection of documents.

1.9.2 Expert System

In artificial intelligence (AI), an expert system is a computer system that emulates the decision-making ability of a human expert. In the simplest sense, AI is the study of developing computer programs which exhibit human-like intelligence. In AI, two things are needed for intelligent behavior as (i) capability of reasoning and (ii) a knowledge-base with which to reason. Because of the nature of these intelligent computer programs, they were aptly called expert systems [51]. An expert system is a computer program designed to model the problem-solving ability of a human expert. The program models consists of four characteristics: knowledge, reasoning, conclusions, and explanations respectively.

The expert system models the knowledge of the human expert, both in terms

of content and structure. Reasoning is modeled by using procedures and control structures which process the knowledge. Conclusions given by the system must be consistent with the findings of the human expert. The system can explain "why" various questions are being asked, and "how" a given conclusion was obtained. One of the principal attractions of expert systems is to extend the application of computers beyond the conventional mathematical processes where the computer can carry on a somewhat natural conversation.

1.9.3 Rule Based System

In computer science, rule-based system is utilized to store and manipulate knowledge in a useful way. They are often used in artificial intelligence applications and research. An example of rule-based system is the domain-specific expert system that uses rules to make deductions of choices. It can be used to perform lexical analysis to compile or interpret computer programs, or in NLP [52].

Rules typically take the form of an IF:THEN expression, or as a more specific example, IF 'red' AND 'octagon' THEN 'stop-sign'.

A typical rule-based system has four basic components [52]:

i. A rule base which is a specific type of knowledge base.

ii. An inference engine which takes action based on the interaction of input and the rule base. The interpreter executes a program by performing the following match-resolve-act cycle:

- Match: In this first phase, the left-hand sides of all productions are matched against the contents of working memory. As a result, a

conflict set is obtained, which consists of instantiation of all satisfied productions.

- Conflict-resolution: In this phase, one of the production instantiations in the conflict set is chosen for execution. If no productions are satisfied, the interpreter halts.
- Act: In this third phase, the actions of the production selected in the conflict-resolution phase are executed.

iii. Temporary working memory.

iv. A user interface through which input and output signals are received and sent respectively.

1.9.4 Markov Model

In probability theory, a Markov model is used to model randomly changing systems where it is assumed that future states depend only on the current state [53]. There are four common Markov models (given in table 1.1) used in different situations, depending on whether every sequential state is observable or not [53]:

Table 1.1: Markov models

	System state is fully observable	System state is partially observable
System is autonomous	Markov chain	Hidden Markov
System is controlled	Markov decision process	Partially observable Markov decision process

From the four Markov models, we have used Markov chain and Hidden Markov model that have been discussed below:

1.9.4.1 Markov chain model

A Markov chain is a type of Markov process that has either discrete state space or discrete index set (often representing time). In this model, all the states are known and the rule is applied based on the known state [54].

1.9.4.2 Hidden Markov model

A Hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. In this model, some states are known and some states are hidden that need to be discovered. After that, specific rule is applied based on new discovered state [55].

1.9.5 Semi Supervised learning

Semi-supervised learning is a class of supervised learning techniques that makes use of unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data) [56]. The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, semi-supervised learning can have a great practical value.

1.10 Goals of This Research Work

We are Bangladeshi people and our mother tongue is Bangla [28]. It has already been mentioned that a few research works have been reported for Bangla language [2, 30–36]. Today, online Bangla contents (especially Bangla news documents) are growing very rapidly and mass people are reading this regularly. So, we are going to propose a method for Bangla news document summarization which will be beneficial for the large community of Bangla speaking people.
The specific goal of this research work is to ensure human' relieve from the burden of huge volume of text by introducing a new approach of Bangla news document summarization. In this regard, the following issues have been considered:

i. Replacing pronoun by corresponding noun so that the number of dangling pronoun will be minimized in summary. This feature will also help in better sentence ranking because a noun may have higher term-frequency score but this noun can be presented in pronoun form in any sentence. Then, the sentence, which is containing the pronoun of the noun, will not get any score for that noun.

ii. Introducing better sentence ranking with sentence frequency calculation. Here, a sentence with high frequency means it is covering the information of more sentences. Sentence frequency is calculated based on 60% or more cosine similarity between two sentences where smaller sentence is deleted and frequency of larger sentence is increased. So, redundancy is eliminated which makes the summary free from repeated content.

iii. Identifying numerical figure, presented in words and digits, has been proposed

here for sentence ranking. Numerical figure (presented in digits) has been considered in several existing methods as an indication of sentence' importance [31, 57, 58]. But, it can be presented in words which is difficult to identify. In this method, we are going to propose a way for detecting numerical figure (presented in words and digits) and incorporated this in sentence ranking.

To accomplish this research work, we have scrutinized 3000 Bangla news documents (news documents of approximately 30 days from the Daily Prothom-alo which is the most popular newspaper of Bangladesh). Seven domain experts of Bangla language, who have completed four years graduation on Bangla language (their mother tongue is Bangla and they read Bangla newspaper regularly), helped us in this research work. After a detail discussion with those persons regarding the structures of sentences of Bangla language and analysis of news documents, we have utilized some rules for special tagging, dependency parsing, subject and object recognizing and pronoun replacing.

1.11 Contributions of the Thesis

Text summarization can be divided into two broad categories as (i) single document text summarization and (ii) multi-document text summarization [33]. Single document text summarization accepts only one document as input and multiple documents text summarization accepts several documents at a time to generate summary. The research work presented in this thesis is to generate summary for single Bangla news document. The following contributions are made in this research work:

i. List of suffixes are utilized to identify more verbs than the existing Part of speech tagger [59]

ii. A special tagging process has been introduced here to identify the followings:

- Acronym
- Repeated words
- Numerical figure from words with and without any suffix
- Numerical figure from digits with and without any suffix
- Initial of name
- Occupation
- Name of people (part of name)
- Name of places

iii. Dependency parsing has been accomplished for the first time in Bangla language. For dependency parsing we have identified 23 rules and defined the priority of these rules. We have used Rule-Based semi supervised technique to identify more words using the known surrounding words.

iv. Identifying full human name including the first, middle, last name, initial and designation. The existing method of Named Entity (NE) Recognition [60] has not been followed here because: (a) there is no verification process for identified NE, (b) using only pre-defined list, (c) no recall method of full name from part of name.

v. Categorizing the full name as subject or object

vi. Recalling full human name from the part of name

vii. Identifying corresponding noun of pronoun

viii. Replacing pronoun in suitable format with the corresponding noun to minimize the dangling pronoun (where corresponding noun is missing) in summary.

ix. Better sentence ranking by introducing the following:

- Calculating sentence frequency
- The score of TF-IDF and Sentence frequency has been boosted up for the replacement of pronoun by corresponding noun
- Counting numerical figure from words and digits with and without any suffix
- Considering the first sentence significantly

1.12 Organization of This Thesis

The dissertation is organized in nine chapters. The first chapter introduces the work done in this research work with the discussion on automatic text summarization, necessity of automatic text summarization especially for Bangla news document, motivating scenario, challenges in the research work on Bangla language, natural language processing, expert systems, rule based systems, semi supervised learning, and markov model. Goals of this research works and contributions are also explained.

The second chapter illustrates the literature review in the arena of automatic Bangla text summarization. The aspects that affect the process of text summarization, extraction and abstraction based text summarization are discussed

in detail for both English and Bangla text. The discussion about all the methods is given with their strength and weakness.

The third chapter describes the proposed method at a glance. All the steps of this approach as- a) preprocessing, b) word tagging, c) replacing pronoun by corresponding noun, and d) sentence ranking and summary generation, are briefly presented in this chapter.

The fourth chapter elaborately explains the user input and preprocessing of text. The inputs are Bangla news document and the required ratio between summary to source document. While preprocessing of text document, some major tasks are incorporated here such as segmenting document, removing stop words and word stemming.

The fifth chapter describes the word tagging as general and special tagging. In general tagging, parts of speech are tagged and some list of suffixes are considered to tag more verbs as very much inflectional forms of verbs are used in Bangla text. In special tagging the nature of words are identified as acronym, initial, repeated words, occupation, part of human name, name of organization and places. After general and special tagging, some words are found untagged and then dependency parsing is utilized to tag more words and verify the tagging based on the surrounding words.

Chapter six goes through a series of advances for the replacement of pronoun by corresponding noun so that the number of dangling pronoun can be minimized from summary. In this regard, full human named entities are identified and categorized them as subject and object. A mechanism is given so that full name can be recalled from the part of name and finally the replacement process of pronoun by corresponding noun in suitable format is explained.

Chapter seven addresses sentence ranking and selection of top ranked sentences for summary generation. The followings are considered here for sentence ranking: (i) term frequency inverse document frequency, (ii) sentence frequency, (iii) numerical figure from words and digits, (iv) title words, and (v) the first sentence. The values of all the attributes are tuned for better summarization performance. After sentence ranking, one third top ranked sentences are selected as the final summary.

Chapter eight is dedicated on evaluation and discussion on results of this proposed method. Some key factors are discussed initially as intrinsic and extrinsic method of evaluation, evaluation measures, and data set. Evaluation results are depicted for each step as general tagging, special tagging, full human name identifying, pronoun replacing, etc. Improvement of performance is shown with generating subversion after including each feature for summary generation. The proposed method is compared with the four latest existing methods based on ROUGE-1 and ROUGE-2 evaluation scores. After all, the discussion on evaluation results is also illustrated with the percentage of improvement of performance and the minimization rate of dangling pronouns from the existing methods.

This dissertation work is concluded in chapter nine where a short review of this proposed method is presented with mentioning some significant evaluation results. Finally, upcoming research direction is sketched out with a clear vision.

Chapter 2
Literature Review

2.1 Introduction

The necessity of automatic text summarization is expanding simultaneously with the increasing of online information without any bounds. The demand of the automatic creation of text summaries has appeared in many areas. For example, summaries of news documents, summaries of emails, summaries of information (for government officials, businessmen, researches and others), summaries of web pages to grasp the information at a glance. The field of text summarization has witnessed continuous involvement of many researchers in the attempt to look for different strategies [61]. A range of procedures, helpful for document shortening named neural networks, semantic graphs, decision trees, regression models, fuzzy logic and swarm intelligence, etc., have been incorporated to find significant portion of the text.

The objective of this chapter is to provide a comprehensive literature review on various strategies of automatic text summarization and to explore the trends. The categorization of text summarization system as (i) extraction and (ii) abstraction [9] has been explained. The remarkable point is that the research works on text

summarization was started with English text [18]. So, most of the features of existing Bangla text summarization methods have been taken from the methods of English text summarization [2, 30–36]. In this regard, most of the significant techniques for English and then Bangla in the field of automatic text summarization have been enlightened here. Detail description has also been given about some recently proposed methods.

The rest of the chapter is organized as follows: an overview on extraction and abstraction based summarization is presented in section 2.2 and 2.3 respectively. Description on text summarization in English language is given in section 2.4 with mentioning the name of significant features. Divergence between English and Bangla language is discussed in section 2.5. The review study on Bangla text summarization is portrayed in section 2.6 and finally this chapter is concluded in section 2.7.

2.2 Extraction Based Text Summarization

In the research work of automatic text summarization, sentence extraction techniques have been extensively employed. This is a shallow approach compared to knowledge-intensive deeper approaches which require additional knowledge-base such as ontology or linguistic knowledge. In this superficial and low-cost approach, statistical heuristics are utilized to identify the salient sentences of a text.

The approach to extract the most important sentences from the given documents needs to have some concrete features. The basic factor of extraction based text summarization is not to omit any important text unit but on the other hand unimportant ones should be ignored in summary preparation. Extractive methods are usually performed in three steps [62, 63]:

i. Creating an intermediate representation of the original text

ii. Sentence scoring

iii. Selecting top scored sentences as the summary

There are various approaches for extraction based text summarization such as:

- Surface level approaches: The earlier methods of text summarization [18, 62] followed the surface level approaches as considering term frequency (TF), length and position of sentences, cue-phrases, title words, etc.

- Corpus-based approaches: In this approach, bunch of documents are considered. Here, knowledgebase is utilized [23] and common terms in most of the documents are taken as insignificant by calculating term frequency inverse document frequency (TF-IDF) [64].

- Cohesion-based approaches: Sentences and words can be coupled in various ways as well as repetition, co-reference, synonymy and semantic association as expressed in thesauri. This coupling is used in this approach to generate summary [27, 65].

- Rhetorical Structural Theory (RST) based approaches: The theory of Rhetorical Structure of document was proposed by William C. Mann et al. [66]. In this approach, text coherence and relational attributes are considered [67, 68].

- Graph-based approaches: Generally, a graph can be denoted in the form of G = (V, E) where V represents the graph's vertex or node and E is the edge between any two vertices. In the context of text documents, vertex

represents sentence and edge is the link between two sentences. Using this approach, documents can therefore be represented as a graph where each sentence becomes the vertex and the edge between each vertex corresponds to the similarity between the two sentences [69].

2.3 Abstraction Based Text Summarization

Abstractive summarization approaches use information extraction, fusion and compression using ontological knowledge [11]. In automatically generated abstracts, summarization technique has been moved from the use of purely extractive methods to the generation of language. Abstraction involves recognizing a set of extracted passages together to constitute something new, something that is not explicitly mentioned in the source, and then replacing them in the summary with the (ideally more concise) core concept [11]. The abstraction based techniques have been proposed in various ways as follows:

- Cut and Paste of text: An analysis from abstract written by human, Hongyan Jing et al. [70] presented "Cut and Paste" based text abstraction. This technique has also been used in [71] where a set of words have been extracted and then arranged them to form sentences.

- Sentence lessening: Sentence reduction strategy has been followed here by deleting the extraneous phrases [72].

- Topic signature identifying: The goal of this technique is to filter the input to retain only the most important, central, topics [23]. Topic identification has been accomplished by using various complementary techniques based on

text structure, cue words, high-frequency indicator phrases, and discourse structure. Finally, the extracted materials are reformulated into a coherent, densely phrased, new text.

- Supervised training: In this technique, the system will be given a set of training data [73] or a model of relationship between the appearance of some features in a document and the appearance of corresponding features in the summary [74].

It has been found after a lot of research works that to date, no parser has been developed to produce the knowledge structures from text and to construct language from the results [4]. Eduard Hovy [4] demonstrated that no system can perform interpretation without aforementioned domain knowledge. But acquiring enough prior domain knowledge is so difficult that summarizers have only attempted it in a small way. In this regard, the interpretation remains blocked for the problem of domain knowledge acquisition. So, before producing abstracts by summarization systems, this problem needs to be resolved.

2.4 Text Summarization in English Language

The state-of-the-art works focused on text summarization in various languages which was started with English text [18]. The voice-over automatic i.e. computerized abstraction began around five decades ago by using surface level indicators for sentence extraction by H. P. Luhn [18] in 1958. Here, the "significance" factor of a sentence is derived from an analysis of its containing words. It was proposed that the frequency of word in an article establishes a useful measurement of words' impact. The idea was that when writing about a given

topic, a writer repeats some words as the text is constructed. Thus, relevance was considered proportional to its frequency and the term frequencies were used to score and to select sentences for the summary.

The method of H. P. Luhn [18] was first extended by incorporating position of sentences and cue-phrases by P. B. Baxendale [75]. It was declared that sentence can be important based on its' position and containing certain cue-words (i.e., words like "important" or "relevant") or the words of heading. G. J. Rath et al. [76] in 1961 stated that purely statistical method of producing extracts were suspected of being inadequate, and hence other methods were sought.

H. P. Edmundson [62] in 1969 accomplished a notable progress after around ten years from the beginning of research on text recapitulation. He incorporated three additional methods with the standard keyword method and the word frequency to determine the sentences' weight. These are:

- Cue method: The hypothesis of this technique is that the presence or absence of certain cue words will compute the significance of a sentence.
- Title method: The weight of a sentence is calculated as a sum of all the content words materializing in the title, headings and sub-headings of a text.
- Location method: Sentences that are taking place in the initial position of paragraphs have a higher probability of being pertinent.

Sometime, specific locations of the text (headlines, titles, first paragraphs, etc.) tend to hold significant information. The simple method of taking the lead (the first paragraph) as summary often outperforms other methods. Especially with newspaper articles, Brandow et al. [22] claimed to achieve very good results by selecting the first sentences as summary. Kupiec et al. [24] and Teufel et al. [77]

conducted experiment with similar algorithms and reported that this single method gives the best results for news, scientific, and technical articles.

Extensive research for deriving the optimum position policy was done by Lin et al. in 1997 and they stated that different text genres have different focuses on position [78]. They considered techniques of tailoring the position method towards optimality over a variety of summarization. With the number of 2097 documents in this method, the result illustrated that the first and the last paragraph fully cover majority of the significant text. Though their summarization technique is position based they predicted that discourse structure significantly varies over domains for which position method is a bit tough.

Today, various research works are available in the arena of English text summarization [19, 20]. Text correlation based summary was introduced by M. A. K. Halliday et al. in 1976 [65]. Gerard Salton offered term frequency inverse document frequency method for multiple documents text summarization in 1989 [64]. Edward Hovy et al. [23] in 1999 used symbolic world knowledge with Information Retrieval and statistical method. Naive-bayes classifier was used by Julian Kupiec et al. [24] and Bayesian classifier was introduced by Aone et al. [79] for selecting important sentences. In a hybrid machine learning model in 2014 [25], several features were utilized such as the similarity of words, text format, cue-phrases, term frequency, title, sentence' location, etc.

M. A. K. Halliday et al. [65] in 1976 performed the first research to explore the degree of subjectivity of two aspects of the lexical cohesion: (i) the word cluster (lexical chains) that are formed and (ii) the lexical semantics relations that are perceived between the words. The linguistic study was emphasized here and tried to form inter sentence groups of related words to make the summary cohesive.

Jane Morris et al. [80] in 1991 utilized cohesion chains and Regina Barzilay et al. [81] in 1997 exploited lexical chains in their method. They proposed to launch from a set of title words of the document to construct lexical chains, adjoining of words that have similar meaning or related to the title. WordNet thesaurus was used by them for determining cohesive relations between terms (i.e., repetition, synonymy, antonymy, hypernymy and homonymy). Tiedan Zhu et al. [82] emphasized on logical-closeness rather than topical-closeness. It was based on synonymy and not strong enough to compute the coherence of sentences. The cohesion problem was also addressed by Aqil M. Azmi et al. [27] in 2012 and stated that automatically generated summary which lacks cohesion should be considered as a poor summary. Moreover, they claimed that cohesive summary generation can be an immense research area.

Rhetorical Structure Theory (RST) has also been used to represent text coherence [19]. RST has been utilized for multiple documents text summarization [83–85] and showed that RST can represent interrelationship between text units. RST was introduced by William C. Mann et al. [66] in 1988 to serve as a discourse structure in the field of computational linguistics. They established a definitional foundation of RST and identified three points: a) the predominance of nucleus/satellite' structural patterns, b) the functional basis of hierarchy and c) the communicative role of text structure. In each rhetorical relation, the nucleus indicates the most important information in the relation while the satellite provides secondary information to the nucleus. The satellite in turn may be another nucleus (multi nucleus) [27]. The criteria for deciding which satellites to keep or eliminate from the RST marks the different summarization schemes [86].

Li Chengcheng [87] in 2010 presented an effective method using RST for

successful automatic text summarization which was based on natural language generation. The scheme of this procedure: (a) analyzing the candidate sentences, (b) discovering the rhetorical relations and (c) forming the essential part of sentences constructive for ultimate recapitulation.

Atkinson and Munoz [85] claimed that by employing rhetorical knowledge one may obtain better quality summaries for multiple documents text summarization. Similar to RST, two basic data structures were proposed by Dragomir R. Radev titled cube and graph data structure by describing cross-document structure theory (CST) [88].

Mani et al. [89] in 1997 presented a graph based methods where concepts denoted by words, phrases, and proper names in the document were considered as nodes. Graph-based methods like LexRank [69] and TextRank [90] represent document(s) as a text similarity graph constructed by taking sentences as vertices and the similarity between sentences as edges. A hub-authority base framework was introduced similar to graph based model that unites the text content with the following, "cue phrase", "sentence length" and "first sentence" [91].

S. Hariharan et al. [92] in 2013 proposed two enhancements on two graph based methods namely- LexRank (threshold) and LexRank (continuous) [69] by reducing redundancy and applying position weight mechanism. Using graph based method in 2014 Canhasi et al. [93] improved coverage in query based shortening system and Rafael Ferreira et al. [17] dealt with redundancy and information diversity problem.

Tai Liao et al. [94] presented extraction based multiple topics document summarization by combining statistics and document relationship map. Several features have been utilized in a hybrid machine learning model such as the similarity of words, text format, cue-phrases, term frequency, title, sentence' location, etc.

[25].

Some researchers have given their valuable efforts in specialized sides. For example, Jingqiang Chen et al. [95] proposed summarization of multiple scientific documents by detecting common facts in citations, Sarah Rastkar et al. [96] invented automatic summarizer of bug reports for software developers, Jaya Kumar et al. [97] formulated the method of multi-document news summarization particularly for disaster issue.

Besides all of these, still there is a qualitative dissimilarity between synopsis generated by existing automated summarizer and the abstract written by human [75]. In spite of some shortcomings, a number of methods have been started to emerge lately with either sentence compressing capability [98] or re-producing technique [80]. Sentence reduction is also a crucial task that could move automatic summarization very close to what human construct. In the information mining, a long sentence is likely to be favored because it has better chance to contain core topic. On the other hand, long sentence tends to contain some clauses that are unimportant for the summary for which removing unnecessary phrase(s) is required to concise the output.

The research of text shortening was limited in the beginning stage for the lack of powerful devices and the complexity of deep natural language processing [18, 62]. Today, interest in automated text summarization has been increased for the growing presence of on-line text especially on the web. In this arena, artificial intelligence was incorporated during 1970s [99, 100] and information retrieval (IR) was introduced during 1990s [23]. The recent direction of the summarization field is going for abstraction where proper language understanding has been enforced.

Deriving from an analysis of human written abstract, Hongyan Jing et al. [70]

presented a "Cut and Paste" strategy for abstraction based text summarization. Six operations were defined to transform chosen sentences from an article into the corresponding summary sentences resemblance to human written abstracts: (i) sentence reduction, (ii) sentence combination, (iii) syntactic transformation, (iv) lexical paraphrasing, (v) generalization and specification and (vi) reordering. Observation from 300 human written abstracts of newspaper articles, Jing and McKeown illustrated that only 19 percent of summary sentences do not have matching sentences in the document [72]. It was also found from the analysis that 78% of summary sentences in abstract produced by humans are based on cut-and-paste where cut-and-paste indicates vocabulary agreement through direct reuse [72]. Cut and paste strategy has been also utilized by Witbrock and Mittal [71] to extract a set of words from the input document and then arrange the words into sentences.

It can be said that summaries produced by the way of abstraction resemble to human summarization process more than extraction does. However, if large quantities of text need to be summarized, sentence extraction is more efficient method [11]. Comparatively extraction based method is found robust towards all kinds of input because deep natural language processing is not needed here [11]. It is noticeable that efforts of performing proper abstraction have not been very successful so far and an approximation called extraction is more feasible today [11].

Review studies have also been published for single document text summarization [19] and multi-document text summarization [20] for English text. Another research work has been accomplished to assess the 15 sentence scoring techniques that have been used in the last 10 years for text summarization [26]. Based on the study in [19, 20, 26], the following features have been frequently used

in English text summarization systems:

- Term frequency
- Cue words
- Important words
- Unimportant words
- Document title
- Numerical figure
- Word tagging
- Upper case lettered word
- Proper noun
- Sentence voting
- Sentence location
- Lexical indicator
- Lexical cohesion
- Lexical chain
- Rhetorical structure theory
- Neural Network
- Key-phrase
- Ontological knowledge

2.5 Divergence between Bangla and English Languages

There are a lot of research works for English text summarization [19,20]. But these may not be directly applicable for Bangla text because of the complexities of Bangla language in the structure of sentences, grammatical rules, inflection of words, etc. Details explanation is given below for which separate Bangla text summarization method is needed:

i. The words (especially verb) in Bangla text are very much inflectional [2]. In reality, the identification of verb is quite difficult because the verb may have a lot of suffixes in Bangla. The English word “say” can be “saying”, “said” and “says” on the basis of tense and person but this word can have various forms in Bangla. For example, the word "বল" (bol - say) can have three basic forms based on the first, second and third person in the present continuous tense only. Such as, 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 meaning of the word “you” in Bangla as "আপনি" (apni - you), "তুমি" (tumi- you) and "তুই" (tui - you) in respected, general and trivial form respectively. For all of these meaning of “you” in Bangla, the forms of 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 second person. In this way the word "বল" (bol - say) can have the following forms: "বলে" (bole - say), "বলেন" (bolen - say), "বলিস" (bolish - say), "বলি" (boli - say), "বলছে" (bolchhe - saying), "বলছেন" (bolchhen -

saying), "বলছ" (bolchho - saying), "বলছিস" (bolchhis - saying), "বলছি" (bolchhi - saying), "বলেছে" (bolechhe - said), "বলেছেন" (bolechhen - said), "বলেছ" (bolechho - said), "বলেছিস" (bolechhis - said), "বলেছি" (bolechhi - said), "বলুক" (boluk - say), "বলুন" (bolun - say), "বলল" (bollo - said), "বললেন" (bollen - said), "বললে" (bolle - said), "বললি" (bolli - said), "বললাম" (bollam - said), "বলত" (bolto - say), "বলতেন" (bolten - said), "বলতে" (bolte - said), "বলতিস" (boltis - said), "বলতাম" (boltam - said), "বলতেছি" (boltechhi - saying), "বলতেছ" (boltechho - saying), "বলতেছেন" (boltechhen - saying), "বলছিল" (bolchhilo - saying), "বলছিলেন" (bolchhilen - saying), "বলছিলে" (bolchhile - saying), "বলছিলি" (bolchhili - saying), "বলছিলাম" (bolchhilam - saying), "বলেছিল" (bolechhilo - saying), "বলেছিলেন" (bolechhilen - saying), "বলেছিলে" (bolechhile - saying), "বলেছিলি" (bolechhili - saying), "বলেছিলাম" (bolechhilam - saying), "বলবে" (bolbe - say), "বলবেন" (bolben - say), "বলবি" (bolbi - say), "বলব" (bolbo - say), "বলো" (bolo - say) [28, 101]. So the complexity of verb recognition in Bangla can't be compared with English.

ii. Numerical figure has been considered as significant for text summarization [57]. But numerical figure has variety of forms in Bangla. For example, the numerical figure "10" can be written as "10" or "ten" in English but in Bangla it can be written as "১০" , "১০টি", "১০টা", "১০খানা", "দশ", "দশটি", etc.

iii. Upper case lettered words are important in English text summarization [57] but in Bangle there is no upper case or lower case letter.

iv. Proper noun can be identified from English text by considering the upper case letter in the beginning of any word which is not possible in Bangla.

v. Subject and object identification is very much difficult in Bangla than that of English because subject may appear (in English text) in the beginning of

sentence, before auxiliary verb or main verb, or it may appear after the word 'by' in passive voice. But in Bangla, subject may appear in any place of the sentences.

vi. Hypernyms and hyponyms, lexical cohesion and lexical similarity can be easily identified from English text by using WordNet [41]. But there is no such kind of resource in Bangla. Though such a tool is ongoing to be developed, it has limited features [42].

So, it can be said that different text summarization procedure is needed for Bangla.

2.6 Text Summarization in Bangla Language

We have already discussed that there are a lot of electronic Bangla text and the volume of these text is growing rapidly. But, very few attempts have been reported for Bangla text summarization [33]. In this section, these attempts have been depicted with their strength and weakness.

In 2004, Islam and Masum [30] presented "Bhasa", a corpus oriented search engine and summarizer. It performs document indexing and information retrieval based on key words using vector space retrieval model [102] for Unicode Bangla text. Corpus files can be ranked and documents can be summarized by this method on the basis of frequent appearance of query terms. Here, document and query terms are treated as two vectors to get the similarity between them. A tokenizer has been used here that can determine different terms, abbreviations, tags, sentence' boundary, headings, and titles. This method has the following modules: (i) TF-IDF (term frequency inverse document frequency) calculation module, (ii) keyword

search module, and (iii) summary generation module. It has utilized the concept of useful, unimportant and important words list while ranking sentences. The problem of dangling pronoun from summary sentences has also been addressed.

This [30] is the first approach for Bangla text summarization along with search engine based on our study. In this method, the solution for the problem of dangling pronoun has been claimed without giving any explanation. Even, it is not specified which modules/sub-modules of this method is for text summarization or search engine. As per the TF-IDF calculation and similarity measurement of each sentence with a given query, it is exposed that the method is effective as search engine but not for summarization.

A few years later, some techniques from the investigation of English text summarization systems were applied to summarize Bangla text by Nizam Uddin et al. in 2007 [31]. They have proposed a technique by incorporating some existing methods of English as follows: (i) location method, (ii) cue method, (iii) title method, (iv) term frequency, and (v) numerical data. They have taken 40% higher ranked sentences from the input document as summary. It has been found that 40% extraction by this system has got the point 8.4 from human professional in the range of 0 to 10. The remarkable point of this paper [31] is to show that some features of English text summarization can also be applicable for Bangla. But this method didn't specify the exact contribution of each feature for sentence ranking. Here, numerical data has been counted for sentence scoring but numerical data can be presented in words instead of digits which can be considered for improvement. While evaluating this method, the score for each system generated summary has been calculated but the comparison with any model summary has not been shown. Moreover, the evaluation method is directly based on human professional.

In 2012, Kamal Sarkar [32] proposed an easy-to-implement approach like the method of Edmandson [62]. It has three major steps: (i) preprocessing, (ii) sentence ranking, and (iii) summary generation. In this method, thematic term has been utilized which is related to the main theme of a document. The term which has TF-IDF (Term Frequency Inverse Document Frequency) values greater than a predefined threshold value is taken as thematic term. The impact of thematic term has been investigated and the features like word-frequency, length and position of sentences have been utilized for sentence ranking. It was claimed that the system performs batter than LEAD baseline method (the first n words of an input article are considered as summary in LEAD baseline method). Average unigram based recall score was found as 0.4122. This method [32] is fully based on almost four decades old English text summarization method [62]. This can be upgraded by incorporating modern natural language processing technique like sentence clustering, redundancy removal, etc. Moreover, in the evaluation, only one model summary has been used for each of the test document. But more model summaries can be utilized for sophisticated evaluation result.

Again, in 2012, Kamal Sarkar proposed another method [33] by tuning each feature of his previous method [32] for better summarization performance. This approach has four major steps (i) preprocessing, (ii) extraction of candidate summary sentences, (iii) ranking the candidate summary sentences, and (iv) summary generation. This is also based on word-frequency, sentence position and sentence length that is similar to [32]. In this approach, some threshold points have been adjusted for position of sentences, TF-IDF values and minimum length of sentences. Too short sentence are discarded at the time of candidate summary sentence selection. Again, the words whose TF-IDF value is less than a predefined

threshold value are removed. All the parameters for sentence ranking have been tuned for better summarization performance. This method [28] has surpassed the LEAD baseline method and the methods described in [32]. All the features have been tuned here for better performance. But, this method is also based on an old English text summarization procedure [62]. This system can be upgraded by incorporating modern natural language processing techniques as discussed for the previous method. Moreover, the evaluation has been turned here against only one model summary.

In 2013, Md. Iftekharul Alam Efat et al. [34] introduced a method for Bangla text summarization by sentence scoring and ranking. Their system has three segments: (i) pre-processing the test document, (ii) sentence scoring, and (iii) generating summary. Sentence scoring is depended on term frequency, position, cue words and skeleton of the document. Here, skeleton of the document consists of the words in title and headers. The cue words can be as "মোটকথা" (motkotha - in short), "অবশেষে" (oboshese - at last), "ইতিমধ্যে" (itimoddhe - already), "যেহেতু" (jeheto - since), etc. The final score is a linear combination of all these features. The average accuracy of this proposed method has been claimed 83.57% against human generated summary. An experiment has also been turned to measure the contribution of each feature while sentence ranking. It is noticeable that evaulation has been accomplished using only 10 documents and standard evaluation has not been turned here to calculate precision, recall and F-measure [103, 104].

It was stated in their paper [34] that the system performs well if the document completely depends on a particular theme. So, the system can be enhanced by eliminating this dependency. Again, the evaluation result is also for a particular theme only which may not be comparable with other generic text summarization

methods.

Abstraction based Bangla text summarization system was proposed for the first time in 2014 by Jagadish Kallimani et al. [36]. They focused on attribute based Information Extraction (IE) rules and class based templates. Rule based classifier was used to categorize the document and to determine the applicable classes. In this system, classes are blueprints that indicate the multiple attributes to be identified for a given topic. And, attributes are primary pieces of information as - NAME, PLACE, DOB (date of birth), DOD (date of demise), and AWARDS. The most significant part of this system is the template based sentence generation where templates are generic structures of sentences. The extracted attributes are mapped with template to generate summary sentences. The authors claimed for adaptation of the system over four Indian languages as Kannada, Hindi, Bangla and Telugu. In the evaluation, the system achieved an average 86.24% precision, 78.93% recall, and 81.50% F-measure. The evaluation results claimed here are for attributes selection only and not for the generated summary.

It is noticeable that abstraction based English text summarization is yet in an immature stage [11] though the research work on English text summarization was begun in 1958 [18]. But this method [36] has reported for Bangla abstractive summarization. Attribute extraction of this method is remarkable which is required for informative sentence generation. But the utilized template is creating same structure of sentences always which is monotonous. It is also questionable that using template is enough or not for all types of sentences while abstraction. So, there is a scope of improvement for generating refined sentences with the identified attributes.

A single research work has been accomplished for multiple documents text

summarization for Bangla language in 2014 by Ashraf Uddin et al. [35]. In this paper, a primary summary is generated at first by sentence scoring on the basis of term-frequency. It has been reported that the words are replaced with their common synonym so that different words with same meaning will be treated as same word. Cosine similarity for each sentence to every other sentences of primary summary has been calculated to get the relevance between them. A graph based model is then applied with the A* [105] searching algorithm on the primary summary for creating final gist. It has been claimed that the selection of starting point of summary is effective by this method. Unigram based recall score was found 56% and the similarity between manual and system generated summary was shown 86.60%. Here, the evaluation result of simialrity measurement of 86.60% is not a standard process of evaluation [103, 104]. Again, they have taken only 8 sets of documents manually for evaluation purpose which is also not enough comparing to others [32, 33].

Bangla multiple documents text summarization has been introduced in this research work [35]. This method has selected the most relevant sentence as the starting point of summary but no reason has been stated behind this. Even the source of getting synonym for each word before term-frequency calculation has not been mentioned. And, there is no direction for ordering of the sentences of different sources which is very much necessary for multiple documents text summarization.

Apart from the previous approaches, keyphrase based summarization method outperforms for both Bangla and English text, which was proposed by Kamal Sarkar in 2014 [2]. Here, keyphrases are extracted as a sequence of words from any sentence that contains no punctuation mark and stop words. If any keyphrase consists of more than 5 words, it is not considered. Keyphrases with multiple words

are segmented to single words, double words and so on to get more keyphrases. All the keyphrases are ranked using phrase frequency inverse document frequency (PF-IDF) and sentences are scored based on their position and term frequency. There are two phases for summary generation. In the first phase (phase-1), candidate summary sentences are selected which contain top ranked keyphrases. Phase-1 considers the sentences for selection which appear early (within position 5) in the document as per the authors experiment. From these candidate sentences, top scored sentences are selected as final summary sentences. If phase-1 fails to generate summary of user desired length, phase-2 is activated. In the second phase (phase-2), more summary sentences are selected based on the sentences score from the rest of the sentences. In the evaluation, F-measure score has been claimed as 0.4242.

This [2] is the first task on keyphrase-based sentence extraction for Bangla text summarization. However, there are some limitations in sentence selection based on the frequency of keyphrases, term frequency and position of the sentences. Here, keyphrases with multiple words are breakdown to single words, double words, etc. so that there will be more keyphrases. Again, keyphrases with single word (got from the breakdown of a keyphrase of multiple words) can have chance to appear again as single word and in the breakdown of other keyphrases. So, keyphrases with single word have more chance to be high frequent. As the keyphrases are ranked based on their frequency, keyphrases with length 1 (single word) are getting higher rank. Besides, this method [33] does not set the minimum length of keyphrases and that is why single word keyphrases may mislead the result as they always get higher rank in analysis. Additionally, general positional score of sentences cant differentiate the importance of the first sentence which can be very significant for Bangla news

document. In the evaluation, this method outperforms all the existing methods of Bangla text summarization. But same type of method has already been introduced for English in [106]. For sentence scoring, only position and term frequency have been considered that have already been introduced around four decades ago [62]. Today, it can be seen that a lot of significant features have been invented by various researchers for text summarization [19, 20]. So, the performance of this research work can be enhanced by adding more features for sentence scoring.

2.7 Conclusion

Numerous types of research works have been accomplished by various researchers for summary generation from single and multiple documents. To explore the trends and analysis of research work in the arena of automatic text summarization, a comprehensive literature review has been depicted in this chapter for both English and Bangla . Strength and weakness of each technique have been discussed along with the scope of improvement. Similarity and differences among Bangle text summarization techniques have been shown by drawing a comparison table. It is expected that this literature review will help the next generation to know the basement of research works in Bangla text summarization and to get the direction for future works.

Chapter 3

Proposed Method

3.1 Introduction

This chapter includes the brief description of the proposed method. All the steps are discussed here in such a way so that anyone may get a general perception of the proposed procedure. Usually, the pattern of thinking for a specific problem is varied all over the world and sometime individual idea makes a great difference so as in text summarization. These innovative ideas come to the light as per the necessity. The necessity of text summarization also has a range of grounds. For example, summarization for news document, scientific publication, literature, crime report, etc.

In this regard, to meet the requirements of various types of documents shortening, efforts have been given by a lot of researchers to develop text summarization procedure in a variety of patterns. As the requirements are different, processes of way out are also different. Even, there are various types of research works for summarization of same type of document. Our proposed procedure is

for summarizing Bangla news document by introducing pronoun replacement and better sentence ranking. At a glance view of this procedure and the implementation model have been discussed in this chapter.

The rest of this chapter is organized as follows: all the steps of the proposed method are briefly explained in section 3.2. The implementation model is discussed in section 3.3. Finally, this chapter is concluded in section 3.4.

3.2 Proposed News Document Summarization Method

The proposed Bangla news document summarization approach consists of four main modules: preprocessing, word tagging, replacing pronoun, and sentence ranking with summarization generation respectively. The entire process flow of the proposed method is illustrated in the figure 3.1.

3.2.1 Preprocessing

Preprocessing of the input document is the starting point of this proposed method. Different types of user inputs are needed from one method to other as per the structural design and functional procedure. Single-document text summarization will take only one document but multi-documents text summarization will accept several documents. In our proposed method, a Bangla news document is taken as user input and then preprocessing will be started. There is an option to specify the ratio between the volume of input document and summary.

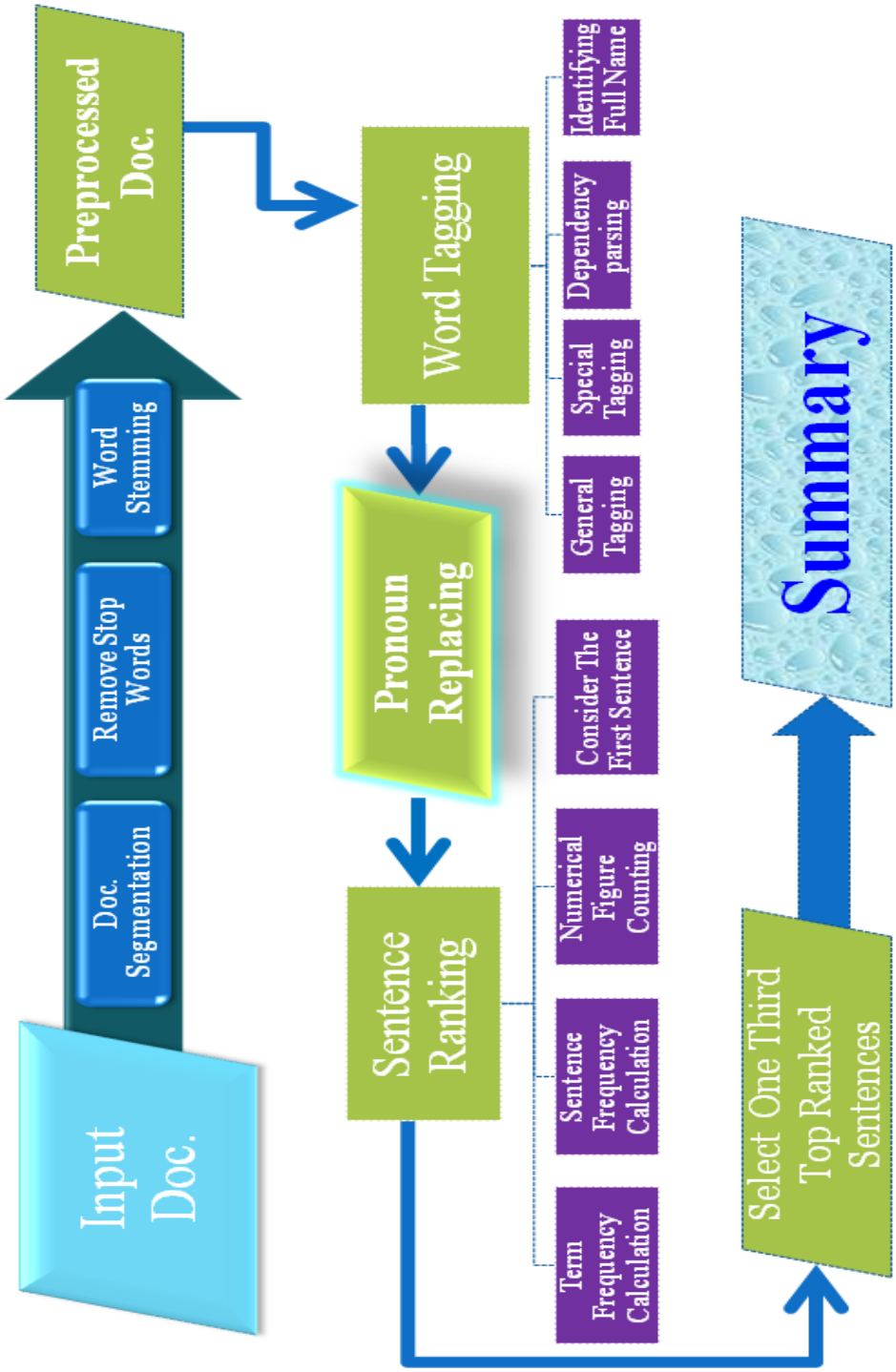


Figure 3.1: Process flow of the proposed automatic Bangla news document summarization system

After getting user input, preprocessing is presented in almost all of the existing methods for text summarization. The earliest works offered only the removal of stop words as preprocessing [18] but after some decades, researchers proposed extensive preprocessing for better summarization [98, 107]. This proposed method has three operations in this step as follows:

i. Segmenting document

ii. Removing stop words

iii. Word stemming

Input document is segmented to sentences based on the punctuation marks "।" or "?" as the end point of sentence. All the sentences are then tokenized to words on the basis of space among them. From the entire words, stop words such as "ও" (o - and), "এবং" (ebong - and), "আর" (are - and) are removed as per the list of 363 Bangla stop words [108]. These stop words are deleted so that they will have no effect in selecting important sentences. Generally, a word can be existed in many places with various suffixes for which same words are considered as different words. In this regard, word stemming is accomplished to map the words with different endings to a single word.

3.2.2 Word tagging

All the words are tried to be tagged with (i) general tagging, (ii) special tagging and (iii) dependency parsing. In the general tagging, words are tagged as noun, pronoun, adjective, verb, etc. by using a lexicon database [59] and SentiWordNet [109]. After general tagging, special tagging has been introduced to identify

words as acronym, numerical figure, initial, repeated words, name of occupation, organization, places and name of people. But, the nature of words in sentences can be varied due to the effect of surrounding words. Finally, dependency parsing has been incorporated so that any given tag can be updated (if needed) and untagged words can be tagged with the help of previously tagged words in a semi supervised way.

3.2.3 Replacing pronoun by corresponding noun

In summarization system, some sentences may be extracted with dangling pronouns where the corresponding nouns are missing. For these pronouns, user will not get any appropriate message from summary and there are chances to misunderstand the text. So, these dangling pronouns need to be replaced by corresponding nouns.

In this step, the following 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), viii) "তাহার" (tahar - his/her). Based on the analysis of Bangla news documents, it has been observed that the corresponding noun of any pronoun is existed in the immediate previous sentence or in the second immediate previous sentence for 88.63% times. This observation has been found by studying 3000 documents and doing experiment with 200 documents. The corresponding noun may be subject or object of the previous sentence in these cases. For recognizing subject and object of any sentence, nature of each word has been identified in the previous step.

Subject and object of any sentence can be the full human named entity for which full human names are identified. Point to be mentioned that only single word has

been considered in the previous step but full human name is consisted of multiple words (the first, last and middle name). For this reason, multiple words are taken into account to identify the full human named entities. After all, these full names are categorized as subject and object of the sentences. Sometime, part of name can be the subject or object in sentences instead of full name. In this situation, a mechanism has been introduced here so that the full name can be recalled from the part of name.

3.2.4 Sentence ranking and summary generation

For sentence ranking, values of some attributes are calculated for all the sentences and then sum-up all the attributes values to compute the final score of each sentence. Top scored sentences are assumed as top ranked sentences and vice versa. The following points are utilized in this method for sentence ranking:

i. Term frequency inverse document frequency (TF-IDF) score calculation: For TF-IDF calculation, all the words are stemmed (in preprocessing step) so that words with different endings can be treated as same word. Replacement of pronoun is also helpful for TF-IDF calculation because a pronoun does not have any TF-IDF score but after replacement by corresponding noun it will have TF-IDF score.

ii. Sentence frequency (SF) score calculation and redundancy elimination: In sentence frequency score calculation, if one sentence covers the 60% content of another sentence then smaller sentence will be removed and the frequency of bigger sentence will be increased. The 60% similarity has been taken because two sentences can be considered as same if they have 60% cosine

similarity [110]. Here, a sentence with high frequency means it is covering the information of more sentences.

iii. Counting score for the existence of numerical figure presented in words and digits: In this proposed method, numerical figures (presented in words and digits both) are identified to compute the score of sentences. Here, the technique of numerical figure identification will help to recognize something special such as the amount of money, specific date, etc. that are presented in words.

iv. Computation of score for title words: For sentence ranking, number of title words is counted as like some existing methods [32,33]. We have also observed from the analysis of 3000 news documents that title words convey the theme of the news documents for most of the times.

v. Special consideration of the first sentence: The first sentence is considered very significantly because it is in the first position and it contains the full title of the document often in Bangla news document. So, it is proposed here to select the first sentence always in summary if it contains any title word.

The ultimate rank of sentences is measured by summing up the calculated scores discussed in the above points individually for all sentences. All the calculated values are also tuned for better summarization performance. After sentence ranking, one third top ranked sentences are extracted as summary in this method.

In our case, we have implemented the proposed method along with the four latest existing methods of Bangla text summarization using a server side scripting language named PHP (Hypertext Preprocessor). And, the proposed method has been evaluated against the four latest existing methods. For evaluation, 600

summaries of 200 news documents (three summaries for each document) have been created manually by human professionals. Then, evaluation results have been generated by using ROUGE [103, 104] automatic evaluation package. The evaluation of performance and discussion on results has been given in the chapter 8.

3.3 Implementation Model of the Proposed Method

The proposed method has been implemented with Rule-Based technique where we have utilized four components as like Rule-Based system [52] as follows:

i. A list of rules as rule base. For example, "if the sentence is the first sentence, the sentence will be selected in summary".

ii. Inference engine to match rule(s) based on the current situation. If two or more rules are conflicted then resolve the conflict based on the priority of rules. For example, "if the sentence is the first sentence but it's length is less than 5, it will never be selected in summary".

iii. Temporary working memory to hold the selected rules.

iv. Interfaces for taking input from one step and giving output to the other step.

In some cases, we have used Hidden Markov Model and somewhere we have used Markov Chain Model (both models have been discussed in the chapter one). Because, the states are known for some cases (Markov Chain Model) and sometime, states are needed to be discovered (Hidden Markov Model) for applying specific rules.

3.4 Conclusion

In this chapter, the end to end process flow of the proposed method has been briefly described where Bangla news document is entered by user and expected summary is produced by summarizer. This chapter can be considered as an abstract outlook of this entire research work. Every step that has been illustrated here will have a detail explanation in the next chapters.

Chapter 4

Preprocessing

4.1 Introduction

This chapter addresses the first step of the proposed method which is started from user input and goes through preprocessing of the input document. Here, a Bangla news document is taken as input to generate summary. It will be primarily processed in this stage so that further steps can be accomplished.

There are so many types of activities to process input documents where some existing procedures furnish text, categorize document, remove stop words, and etc. In this regard, (i) stop words removal was done as preprocessing in the earliest work of text summarization [18], (ii) segmentation of text to different cluster was accomplished by Goldstein et al. as preprocessing by using supplementary available information about the documents set [111], (iii) term frequency matrix was generated by G. Salton [64], (iv) word hierarchical representation of text was proposed by Y. Ouyang et al. [107], and etc. The preprocessing of the input document in this research work includes (i) segmenting input document,

(ii) removing stop words, and (iii) word stemming. All of these tasks have been explained in this chapter.

The rest of the chapter is organized as follows: input from user is described in section 4.2, all the tasks for preprocessing of input document are explained in section 4.3. Finally, the chapter is concluded in section 4.4.

4.2 User Input

In this proposed method, the inputs are as follows:

i. A Bangla news document.

ii. Ratio between summary to source document. This option is included because one may expect the volume of summary will be forty percent where other may want thirty percent of the original text. If the user will not specify any ratio, the summary will be one third of the source document according to some existing methods [5, 8].

Sample input text from Bangla news document:

বাংলাদেশ সংবাদ সংস্থার পরিচালনা কমিটির প্রধান করিম সাহেব সাংবাদিকদের বলেন যে, সকল মানুষের খবর পত্রিকায় ছাপানো উচিৎ। তিনি জানান, ভবিষ্যতে পত্রিকা হবে গণমানুষের জন্য।

Figure 4.1: Sample input text

4.3 Preprocessing

After entering the input document, preprocessing is a significant step to formulate the text for further processing. From the systems point of view this is the first tread

on summarization. The following actions are accomplished on submitted document in this step:

4.3.1 Segmenting input document

Input document is segmented into sentences based on the punctuation marks "।" or "?" as the end point of sentence. An array of sentences is generated to hold the contents of the given document (presented in figure 4.1):

Sentence_array [1] = "বাংলাদেশ সংবাদ সংস্থার পরিচালনা কমিটির প্রধান করিম সাহেব সাংবাদিকদের বলেন যে, সকল মানুষের খবর পত্রিকায় ছাপানো উচিৎ"
Sentence_array [2] = "তিনি জানান, ভবিষ্যতে পত্রিকা হবে গণমানুষের জন্য"

After this segmentation, sentences are tokenized into words on the basis of space among them.

4.3.2 Removing stop words

Stop words are generally used in language to indicate the tense, adjective or for adapting grammatical structure. In computing, stop words are filtered out prior to or after in case of necessity while processing of natural language. It is thus because frequency is a measurement along with others criteria for any word to be selected as important. In this selection, some words should not be taken as important such as articles, auxiliary verbs, parts of speech, etc. The elimination of these is accomplished so that these words will have no effect in measuring the significance of any sentence.

For identifying stop words, a list of stop words is kept in the system with which

all the words of the input document are matched and removed the matched words. The list of 363 stop words for Bangla text has been collected from [108]. It is noticeable that this same list of stop words has been used in some existing Bangla text summarization methods [2, 32, 33]. Here, the list of stop words has pronoun also but these pronouns are not deleted because these (singular pronouns) will be replaced by corresponding nouns (discussed in chapter 6). The following figure 4.2 shows a sample text with selection of stop words in red color and sample text after removing stop words.

Original message for example:

বাংলাদেশ সংবাদ সংস্থার পরিচালনা কমিটির প্রধান করিম সাহেব সাংবাদিকদের বলেন যে, সকল মানুষের খবর পত্রিকায় ছাপানো উচিৎ। তিনি জানান, ভবিষ্যতে পত্রিকা হবে গণমানুষের জন্য

Message after stop words removal:

বাংলাদেশ সংবাদ সংস্থার পরিচালনা কমিটির প্রধান করিম সাহেব সাংবাদিকদের বলেন, মানুষের খবর পত্রিকায় ছাপানো । তিনি জানান, ভবিষ্যতে পত্রিকা গণমানুষের

Figure 4.2: Sample text for the example of stop words removal

4.3.3 Word stemming

Word stemming is the process of reducing inflected (sometimes derivative from another) words to their stem, base or root form. It is a process of linguistic normalization, in which the variant forms of a word are reduced to root form by a stemming algorithm [112]. The words in Bangla language are very much inflectional [33] for which word stemming is applied to map the words with different endings to a single word as in the figure 4.3:

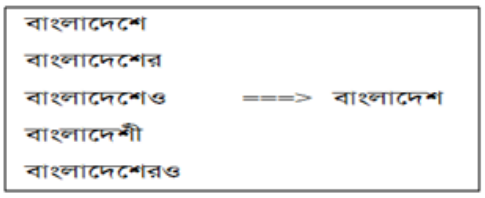


Figure 4.3: Example of word stemming

This step is very much essential for term frequency calculation because only for different endings same words will be treated as different words if stemming will not be applied. Word stemming is also required for sentence similarity measurement on the basis of existence of same words. If word stemming is not introduced, same words may be existed in two or more sentences but in different forms for which similarity between these sentences may not be identified. In this way, this will negatively impact on selection of important sentences which may lead to generate summary with trivial sentences.

There are a variety of algorithms to find the stem of a given word. Lookup algorithm uses a lookup table with all derived words from the root words to find the exact stem of any given word. In the lookup algorithm, an obvious disadvantage is to use a very large lookup table. On the other hand, suffix stripping algorithms do not rely on lookup table. If the user has enough knowledge in the challenges of linguistics and morphology and suffix stripping rules, suffix stripping approach can be a better way than lookup table algorithm. Rather than using lookup table, this algorithm uses some rules to find out the root form of words that make stemming operation time comparatively low. In addition to deal with suffixes only, several approaches also attempt to remove common prefixes. Some algorithms use two approaches in unison because single criterion is unable to deduce the stem always [112]. Besides using the light weight stemmer, complex scenario arises when the

irregular words are needed to be stemmed. For example, the light weight stemmer stems "গ্রামের" to "গ্রাম", but does not stem "পৌঁছেছিস" to "পৌঁছা" and "দিলেন" to "দেওয়া". In this regard, lookup table algorithm has also been applied to get the stem. For this reason, a list of irregular words with their different forms is used in this system as a lookup table [113]. The steps of word stemming is given below:

(i) Get the list of tokenized words after removing stop words.

(ii) Irregular words have been collected from [113] and kept in lookup table. In one column of lookup table, it has the irregular words and in another column, it has the stemmed form of the words.

(iii) Match each irregular word (collected from [113]) in the lookup table and get the corresponding stem of the word.

(iv) List of suffixes for Bangla words have been collected from [28] as predefined suffix list.

(v) Lightweight stemmer has been used to strip the suffixes using a predefined suffix list on a "longest match" basis [112].

(vi) Get the list of stemmed words.

Words in different forms	Words after stemming
"গ্রামের" (gramer), "গ্রামে" (grame)	"গ্রাম" (gram). In English: "village"
"হাটছেন" (hatchhen), "হাঁটিতেছেন" (hatitechhen)	"হাটা" (hata). In English: "walk"

Figure 4.4: More example of word stemming

Markov chain model [54] has been followed in the process of word stemming because all the states are known. The algorithm of word stemming is given below:

Algorithm 1: Word stemming

Input: W1: List of words from the segmentation all the sentences
Output: W2: List of stemmed words

Begin
`/* Initialization of some variables */`
SuffixList ← *List of suffixes for words collected from [28]*
LookupTable ← *List of irregular words*
`/* Irregular words have been collected from [113] that can't be stemmed using SuffixList. This lookup table has two columns. In one column, it has the irregular words and in another column, it has the stemmed form of the words. */`
n ← *number of sentences in the input document*
Wt ← *ϕ //to hold a word temporarily*
for *i* ← 1 ***to*** *n* **do**
    Wt ← *W*1[*i*]
    if *Wt is existed in LookupTable* **then**
        *W*1[*i*] ← *Stemmed form of the word Wt from Lookup Table*
    else
        Check that Wt has any suffix from the SuffixList
        if *Wt has any suffix from SuffixList* **then**
            Remove the suffix from Wt
            *W*1[*i*] ← *Wt*
        end
    end
end
*W*1 ← *W*2
return *W*2
End

4.4 Conclusion

Starting point as well as the user input of the proposed method and the preprocessing of text has been illustrated in this chapter. Nowadays, texts are widely available where texts from different sources look quite different. So, the preprocessing is an essential way to make the documents adoptable to the system. In this chapter, necessity of the accomplished processes as like removing stop words and word stemming have also been explained.

Chapter 5

Word Tagging

5.1 Introduction

The nature of word identification is word tagging. It is the process of marking up a word in text (corpus) as corresponding to a particular part of speech (POS), based on both its definition and its context. It is also depended on relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught in primary school level to identify words as nouns, pronouns, verbs, adjectives, adverbs, etc. Once performed by hand, POS tagging is now done in the context of computational linguistics by using algorithms. It is a very familiar issue for the natural language processing for almost all the languages where the rules of tagging can be different. And, beyond noun, pronoun, verb, etc. tagging, the word can be numerical figure, it can be named entity, acronym and etc. which is also essential to be identified.

In this chapter, the word tagging process is accomplished in the following three phases: (i) general tagging where parts of speech are tagged, (ii) special tagging where words are tagged as acronym, named entity of human, name of places, occupation, etc., (iii) dependency parsing to verify each given tag and identify

more words based on the effects of surrounding words.

The rest of the chapter is organized as follows: general tagging is explained in section 5.2, special tagging is presented in section 5.3, dependency parsing is given in section 5.4, and finally this chapter is concluded in section 5.5.

5.2 General Tagging

All the words are tried to tag as noun, pronoun, adjective, verb, preposition, etc. in this step by using a lexicon database [59] and SentiWordNet [109]. The lexicon database and SentiWordNet has limited number of predefined words. Using lexicon database, the words can be tagged as 'JJ' (Adjective), 'NP' (Proper noun), 'VM' (Verb), 'NC' (Common Noun), 'PPR' (Pronoun), etc. On the other hand, SentiWordNet has list of words with tag as 'a' (Adjective), 'n' (Noun), 'r' (Adverb), 'v' (Verb), 'u' (Unknown). Based on these predefined lists of words, we have experimented on 200 Bangla news documents and found that 65.13% words can be tagged.

The words (especially verb) in Bangla language are very much inflectional [2]. So many verbs are left untagged as lexicon database and SentiWordNet have not covered the entire inflection. Though word stemming has been introduced (in the preprocessing) here to identify root form of word, 100% inflectional forms of verb can't be stemmed [112]. Even, the recognition of verb is quite difficult because there are a lot of suffixes for verb in Bangla. But, identifying verb is very important for language processing task as verb is the chief word for any sentence [101]. So, if there is any word left untagged after using lexicon database [59] and SentiWordNet [109], we need to check the word if it is verb or not once again. In this regard, a list of suffixes [28] are taken into account for ultimate checking such as "ইতেছি" (itechhi),

"তেছিলেন" (techhilen), "লেন" (len), "সেন" (sen) , etc. Now, if the considered word has any of these suffixes [28], it is tagged as verb. List of suffixes [28] is given for all tenses in the table 5.1:

Table 5.1: List of suffixes for Bangla words

Tense	**Suffixes of verb**
Present tense	"এ" (a), "এন" (en), "অ" (o), "ইস" (ish), "ই" (e), "ইতেছে" (itechho), "ছে" (chhe), "চ্ছে" (cchhe), "ইতেছেন" (itechhen), "ছেন" (chhen), "চ্ছেন" (cchhen), "ইতেছ" (itechho), "ছ" (chho), "চ্ছ", (cchho), "ইতেছিস", (itechhis), "ছিস" (chhis), "চ্ছিস" (cchhis), "ইতেছি" (itechhi), "ছি" (chhi), "চ্ছি" (cchhi), "ইয়াছে" (iyachhe), "এছে" (echhe), "ইয়াছেন" (iyachhen), "এছেন" (echhen), "ইয়াছ" (iyachho), "এছ" (echho), "ইয়াছিস" (iyachhis), "এছিস" (echhis), "ইয়াছি" (iyachhi), "এছি" (echhi), "উক" (uk), "উন" (un)
Past tense	"ইল" (el), "ল" (lo), "ইলেন" (ilen), "লেন" (len), "ইলে" (ile), "লে" (le), "ইলি" (ili), "লি" (li), "ইলাম" (ilam), "লাম" (lam), "লুম" (lum), "ইত" (ito), "তে" (te), "তো" (tO), "ইতেন" (iten), "তেন" (ten), "ইতে" (ite), "তে" (te), "ইতিস" (itis), "তিস" (tis), "ইতাম" (itam), "তাম" (tam), "তুম" (tum), "ইতেছিল" (itechhilo), "ছিল" (chhilo), "ইতেছিলেন" (itechhilen), "ছিলেন" (chhilen), "ইতেছিলে" (itechhile), "চ্ছিলে" (cchhile), "ছিলে" (chhile), "ইতেছিলি" (itechhili), "ছিলি" (chhili), "চ্ছিলি" (cchhili), "ইতেছিলাম" (itechhilam), "ছিলাম" (chhilam), "ইয়াছিল" (iyachhilo), "এছিল" (echhilo), "ইয়াছিলেন" (iyachhilen), "এছিলেন" (echhilen), "ইয়াছিলে" (iyachhile), "এছিলে" (echhile), "ইয়াছিলি" (iyachhili), "এছিলি" (echhili), "ইয়াছিলাম" (iyachhilam), "এছিলুম" (echhilum), "ইয়াছি" (iyachhi), "এছিলাম" (echhilam)
Future tense	"ইবে" (ibe), "বে" (be), "ইবেন" (iben), "বেন" (ben), "ইবি" (ibi), "বি" (bi), "বো" (bO)

After using the list of suffixes, the percentage of word tagging has been increased from 65.13% (result of word tagging before considering the list of suffixes [28]) to 66.73%. The tagging in this step is a preliminary tagging and some tags may be updated in the next steps. Again, some words will be specifically tagged as acronym, named entity, occupation, etc. in the next step.

5.3 Special Tagging

Word is the principal ingredient of a language [28] and hence it is difficult to detect subject and object of any sentence without detecting the nature of words. A procedure is available for Bangla parts of speech tagging [114]. But, no procedure has been found for identifying nature of words as acronym, initial, repeated words, numerical figure from digits and words, occupation, etc. In this situation, nature of each word has been identified as follows:

5.3.1 Checking for English acronym

In Bangla news documents, there may have acronym that means the word is consisted of some English letters that are written in Bangla. For example: "ইউএনডিপি" (UNDP), "আইএলও" (ILO), "ইউএসএ" (USA), etc. For checking this type of words, all the English letters are written in Bangla as: "এ" (A), "বি" (B), "সি" (C) , "ডি" (D), "ই" (E), "এফ" (F), "জি" (G), "এইচ" (H), "আই" (I), "জে" (J), "কে" (K), "এল" (L), "এম" (M), "এন" (N), "ও" (O), "পি" (P), "কিউ" (Q), "আর" (R), "এস" (S), "টি" (T), "ইউ" (U), "ভি (V), "ডব্লিউ" (W) , "এক্স" (X), "ওয়াই" (Y), "জেড" (Z). Then, sorting them in descending order based on their string length where "ডব্লিউ" (W) will be in the first place and "এ" (A) will be in the last place. Now, match each letter of the word, for example: "ইউএনডিপি" (UNDP) will be matched with "ইউ" (U), "এন" (N), "ডি" (D), "পি" (P). Significant point is that sorting in descending order is done for ensuring the longest matching always. For example, "এন" (N) will not be matched with "এ" (A) at first time rather it will be fully matched with "এন" (N). In this way, a word is tagged as an acronym or not. Experiment shows 100% success rate for detecting acronym.

It is noticeable that the acronym identification can be done by using a list of acronym in a data file and search the data file to check a word is acronym or not. This procedure will be depended on list of predefined words as like lookup table algorithm. But we have utilized a dynamic process for checking each word without any predefined set of data.

5.3.2 Checking for Bangla initial

As like English acronym mentioned in the above sub section 5.3.1, there can be Bangla letters with spaces such as "আ স ম" (A S M), "আ ক ম" (A K M), etc. These letters will be tagged as Bangla initial. Based on experiment, the correctness of finding initial is 100%.

5.3.3 Checking for repeated words

Repeated words are special form of word combination where same word can be placed for two times consecutively [28]. For example, "ঠান্ডা ঠান্ডা" (thanda thanda - cold cold), "বড় বড়" (boro boro - big big), "ছোট ছোট" (choto choto - small small), etc. Some words are there, those are repeated partially such as "খাওয়া দাওয়া" (khawa dawa - eat drink). List of some other words have been collected from [101] where some irregular words are mentioned as repeated words as like "দেনা পাওনা" (dena paona - payable receivable), "ছোট বড়" (choto boro - small big), etc. If any word is matched with these words or reiterated for two times (fully or partially), they are tagged as repeated words. We have applied this technique on 200 Bangla news documents and found 100% accuracy on identifying repeated words. If the repeated words are fully repeated, it is an adjective which can be placed before noun, pronoun or verb but if the words are partially repeated, it is treated as noun [28].

5.3.4 Checking for numerical figure

In Bangla news document, numerical figure can be presented in words and digits. There can be suffix with the numerical figure for which any numerical figure can have variety forms in Bangla. In this regard, the identification of numerical figure is difficult. But, numerical figure is very significant to measure the importance of any sentence. Even, if we miss this, we may lose the consideration of any required items such as any amount of money, specific date, and quantity of something special. For recognizing numerical figure presented in words and digits, three conditions are checked as follows:

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

(ii) In the next part (if any) it has the followings: "শত" (shoto - hundred), "হাজার" (hazar - thousand), etc.

(iii) At last, it may have suffix "টি" (ti - this), "টা" (ta - this), etc.

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

5.3.5 Checking for occupation

Occupation is itself a significant word and for the human named entity identification, occupation is very much helpful by which named entity can be

recognized. If we get any word as occupation, we may consider the immediate next some words to find out named entity. There is a table with 80 entries (collected from [115, 116]) for the title of Bangladeshi occupation such as "মন্ত্রী" (montri - minister), "কৃষক" (krishok - farmer), "ছাত্র" (chhatro - student) etc. Each word 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 occupations or fully matched with the listed occupations, the word is tagged as occupation. It has been observed that the proposed technique can identify occupation for 91% times.

5.3.6 Checking for the name of organization

Name of organization is an important factor where any type of word can be the part of organization name. It has been detected from our analysis that name of an organization can be mentioned as follows:

(i) The full name of organization is given which follows the acronym of the name enclosed in parentheses. For example, "দুর্নীতি দমন কমিশন (দুদক)" - "Durniti Domon Commission (DUDOK) - Anti Corruption Commission (ACC)".

(ii) The last part of the organization name may have some specific words. Such as, "লিমিটেড" (limited - limited), "বিশ্ববিদ্যালয়" (bishawbiddaloy - university), "মন্ত্রণালয়" (montronaloy - ministry), "কোং" (kong - kong), etc. [60].

If there is any acronym according to the above point (i), enclosed with parentheses, count the number of letters in the acronym. Then, same number of words (immediately before the acronym) is tagged as a name of organization. In

this situation, the acronym can be consisted with the initial letters of the words immediately before the acronym otherwise this mechanism will not be applicable. Experiment shows that 95.60% organization names can be found which has acronym in parentheses after name. For this experiment, we have collected 650 acronyms from [117, 118].

According to the point (ii), if any of such words is presented in the text, check three words immediately before the specific word. Here, three words are considered as it has been observed in our analysis that organization is constituted with three words for most of the time. If the organization 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 word, consider them as the name of an organization. Name of organizations can be recognized for 87% times based on the point (ii).

5.3.7 Checking for probable human named entity

Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities as the name of persons, organizations, locations, etc. Statistical NER systems typically require a large amount of manually annotated training data. Here, a data file (manually annotated data as the first, middle or last name) has been used for human named entity identification with 7500 entries [119]. If any word is matched with these listed names, it is primarily tagged as name of human. Somewhere middle name may be used as first name and first name can be used as last name. So, it is not fixed for any part of name that is the first, middle or last name. Part of name is identified in this point for

around 80% times which will be re-checked and full name will be identified from these parts of names (discussed later in the section of full name identification in the next chapter).

5.3.8 Checking for the name of place

Name of places identification is similar to the process of human name identification. A statistical procedure has been followed with manually annotated training data. A table has been maintained with 700 entries for the list of division, district, upozila and municipality as name of places of Bangladesh [120]. Here division is the top level, district is second level and upozila or municipality is third level in regional segmentation. Further, we have collected 230 names of countries and their capital [117]. If any word is matched with these listed names of places, it is tagged as place. In this way, around 82% names of places can be detected in our experiment.

Moreover, the tagging in general and special tagging is static but the words can be dynamic in nature depending on the surrounding words in sentences. In this regard, dependency parsing has been introduced in the next step. Further, the general and special tagging of each word will be reconsidered in dependency parsing.

5.4 Dependency Parsing

A dependency parser analyzes the grammatical structure of a sentence, establishing relationship between “head” words and other words. The nature of words in

sentences can be varied due to the effect of surrounding words. So, it may be needed to update the tag of any word which is accomplished in general and special tagging process. For this reason, dependency parsing has been incorporated here so that any given tag can be updated (if needed) and untagged words can be tagged with the help of previously tagged words.

So far, there is no dependency parser for Bangla text by which dynamic nature of any word can be measured. Though the dependency parser is available for English text [121], it can't be used for Bangla text because there is a huge difference between the structure of sentences of Bangla and English [28]. Here, a dependency parsing technique for Bangla has been explained by which the dynamic nature of each word can be identified based on the adjacent surrounding words. The parsing mechanism is a rule based technique where the following rules have been utilized as follows:

(1) List of adjectives have been collected from [28] and fully repeated words (mentioned in the special tagging process) are treated as adjective. Adjectives are placed as neighboring words of noun or verb [28]. If any adjective (not tagged as repeated word in special tagging) has any suffix, it is treated as noun. There can be consecutive adjectives where noun or verb is placed after these consecutive adjectives.

(2) If the repeated word (as in special tagging) is fully repeated, it is an adjective otherwise the repeated word is noun. For example, "খাওয়া দাওয়া" (khawa dawa - eat drink), "দেনা পাওনা" (dena paona - payable receivable).

Rule for the above points (1) and (2):

ListAdjective = List of Adjective collected from [28]

IF (Any word is repeated for two times) THEN

TAG the words as Adjective

ListAdjective = ListAdjective U Repeated words

ELSE IF (The repeated words are partially repeated) THEN

TAG the words as Noun

END IF

IF (The Adjective has any suffix) THEN

TAG the words as Noun

END IF

IF (Previous word is Adjective for any word) THEN

IF (The word has any suffix which is related to verb as "ইতেছি" (itechhi), "তেছিলেন" (techhilen), "লেন" (len), etc.) THEN

TAG the word as Verb

ELSE

TAG the word as Noun

END IF

END IF

(3) Some words are generally placed before adjective, for example, "অপেক্ষা" (opekkha - than), "চেয়ে" (cheye - than), "অধিক" (odhik - more), etc. [28]. So, if any word (immediately after these words) is wrongly tagged as noun or verb or any other tag (without adjective), update the tag of the word as adjective.

Rule:

ComparativeWords = "অপেক্ষা" (opekkha - than), "চেয়ে" (cheye - than), "অধিক" (odhik - more), etc. [28]

IF (The previous word of any word is like ComparativeWords) THEN

TAG the word as Adjective

END IF

(4) List of words are used as prefix of another words in Bangla language [101]. Individually, this list of prefixes has no meaning but it can change the meaning of other words. The prefixes are "প্র" (pro), "পরা" (pora), "অপ" (opo), "সম" (somo), "নি" (ni), "পাতি" (pati), "কদ" (kod), "কু" (ku), "বি" (bi), "ভর" (bhor), "রাম" (ram), "স" (so), "সা" (sa), "সু" (su), "হা" (ha), etc. [28]. The words with these prefixes are generally noun or adjective [28, 101]. If the word is not existed in the list of adjective (as mentioned in the above points 1, 2, or 3), this is treated as noun.

Rule:

PrefixList = List of prefixes [101]
ListAdjective = List of adjective based on the above points (1), (2), or (3)
IF(Any word has prefix from PrefixList) THEN
    IF(The word is not existed in ListAdjective) THEN
        TAG the word as Noun
    END IF
END IF

(5) Some words are there as verb like "কর" (kor - do), "দেয়" (dey - give), "যায়" (zay - go), etc. These words may have suffix as "ইতেছি" (itechhi), "তেছিলেন" (techhilen), "লেন" (len), "সেন" (sen), etc. [28].

Rule:

VerbList = List of words as the root form of verb [28]
IF (The word is existed in VerbList) THEN

TAG the word as Verb

END IF

(6) If the previous word has been tagged in the special tagging as occupation, word with article (except occupation with article), fully repeated words or numerical figure, it can't be verb.

Rule:

IF(Previous word is occupation or fully repeated words or numerical figure) THEN

IF(The word has been tagged as Verb) THEN

TAG the word as Unknown

END IF

END IF

(7) There is a list of article as "টি" (ti - this), "টা" (ta - this), "খানা" (khana - that), "খানি" (khani - that), etc. which can be placed as suffix with noun or pronoun [101]. The list of pronoun is collected from [28, 101]. So, if any word (except pronoun) has articles, this will be considered as noun.

Rule:

ArticleList = List of articles [101]

PronounList = List of pronouns [28]

IF(Any word has suffix from ArticleList AND it is not existed in PronounList) THEN

TAG the word as Noun

END IF

(8) The word "গোটা" (gota - whole) can be placed before numerical figure and "খানা"

(khana - this), "খানি" (khani - this) can be placed after numerical figure [28].

Rule:

IF(The previous word of any word is "গোটা" (gota - whole)) THEN

TAG the word as Numerical figure

ELSE IF(The next word of any word is "খানা" (khana - this), "খানি" (khani - this)) THEN

TAG the word as Numerical figure

END IF

(9) There can be article along with numerical figure, occupation, organization and name of places as they are one kind of noun. So, a new tag will be given for each of them as numerical figure with article, occupation with article, etc. if they contain article.

Rule:

ArticleList = List of articles [101]

IF(Any word has suffix from ArticleList) THEN

IF (Word is tagged as Numerical figure) THEN

TAG the word as Numerical figure with article

ELSE IF (Word is tagged as Occupation) THEN

TAG the word as Occupation with article

ELSE IF (Word is tagged as Organization) THEN

TAG the word as Organization with article

END IF

END IF

(10) If there is a numerical figure anywhere in the sentence, the next word of

numerical figure is a noun. If the numerical figure is the last word of sentence, the noun is placed immediately before that. This noun is direct object [101]. The direct object is material and indirect object is personal.

There are two types of objects: direct and indirect [122]. Indirect objects are nouns or pronouns that identify to whom or for whom the action of the verb is performed. In order to have an indirect object, there is a direct object. The indirect object typically precedes the direct object. Direct objects are nouns, pronouns, clauses and phrases. Direct objects follow transitive verbs (action verbs that require something or someone to receive the action).

Rule:

IF(Previous word is Numerical figure) THEN

    TAG the word as Noun

ELSE IF(There is no word after Numerical figure) THEN

    TAG the previous word of Numerical figure as Noun

END IF

(11) There may have comma separated words where last word is separated by "ও" (o - and), "এবং" (ebong - and), "আর" (ar - also). In these cases, all the comma separated words are same in nature. For example, "ঢাকা, দিল্লি এবং কুয়ালালামপুরের মাঝে একটি সমঝোতা স্মারক হবে" (There will be a memorandum of understanding among Dhaka, Delhi and Kuala Lumpur). In this case, if the word "ঢাকা" (Dhaka) is known as name of place in the section of special tagging, we may identify "দিল্লি" (Delhi) and "কুয়ালালামপুর" (Kuala Lumpur) with same tagging as name of places.

Rule:

IF(One known word is comma separated with some other words) THEN

TAG all the comma separated words as the same tag of the known word

ELSE IF(One known word is separated from other words by "ও" (o - and), "এবং" (ebong - and), "আর" (ar - also)) THEN

TAG both the words as the same tag of the known word

END IF

(12) List of words are there as preposition/conjunction/interjection and they are treated as "অব্যয়" (Obboy) in Bangla language. The parts of speech which meaning can't be changed anywhere in the sentence is called "অব্যয়" (Obboy) [28]. These are tagged as stop words. A list of 363 stop words has been collected from [108] for Bangla language. These words can't have other tagging and have no dependency on surrounding words [28].

Rule:

StopWordList = List of stop words [108]

IF(Any word is existed in StopWordList) THEN

TAG the word as Stop word

END IF

(13) The word that is presented immediately before the words "দ্বারা" (dara - with), "দিয়া" (diya - with), etc. is a noun which is object [28]. If there are two words and both are noun before these listed words, the first one is indirect object (personal) and second one is direct object (material). The definition of indirect object and direct object has been given in the above point (10).

Rule:

IF(Any word is placed before the words "দ্বারা" (dara - with), "দিয়া" (diya - with),

etc.) THEN

TAG the word as noun

END IF

(14) The word which is placed after "দ্বারা" (dara - with), "দিয়া" (diya - with), etc. is verb.

Rule:

IF(Any word is placed after the words "দ্বারা" (dara - with), "দিয়া" (diya - with), etc.) THEN

TAG the word as verb

END IF

(15) General structures of sentences can be as: (a) subject + object (personal object) + object (material object) + adjective of verb + verb, or (b) subject + time related word + place related word + indirect object + direct object + adjective of verb + verb [28, 101]. We may identify subject and object by following these structures. The two basic structures have been given in this point which can be utilized to find out not only subject and object but also the nature of object as direct or indirect and verb, time or place related words, etc.

(16) If there is a noun with suffix "র" (r) or "এর" (er), there will be another noun after that. Again if the second one has similar suffix, this will follow another noun and so on. The last noun can be either subject or object of the sentence.

Rule:

IF (The word is Noun AND The word has suffix like "র" (r) or "এর" (er)) THEN

TAG the next word as Noun

END IF

(17) The words "ওহে" (ohe - hi), "হে" (he - hi) follows a human named entity. It is noticeable that some words have already been tagged as name of human in the section of special tagging by using a list of predefined names. But all the named entities can't be tagged for which this rule can be utilized.

Rule:

IF(The previous word of any word is "ওহে" (ohe - hi) OR "হে" (he - hi)) THEN

TAG the word as Name of human

END IF

(18) The suffix "কার" (kar) and "কের" (ker) are placed with the word which indicates time.

Rule:

IF(The word has suffix as "কার" (kar) OR "কের" (ker)) THEN

TAG the word as Time related word

END IF

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

(20) If the words "কখন" (kokhon - when), "কোথায়" (kothay - where), "কবে" (kobe - when), "কিভাবে" (kivabe - how) are existed in the middle position of sentence,

there will be two parts of sentence. In that case, the former part is primary part and later part is secondary part of sentence where primary part contains the main subject and main verb.

The above points (19) and (20) are very much needed to identify the principal verb and principal subject of sentence.

(21) The word immediately before "সমাহার" (somahar - combination) is a noun where the previous word of the noun is a numerical figure. For example, "তিন মাথার সমাহার" (tin mathar somahar - a combination of three heads).

Rule:

IF(The next word of any word is "সমাহার" (somahar - combination)) THEN

TAG the word as Noun

TAG the previous word as Numerical figure

END IF

(22) There are pair of words "যে-সে" (je-she – who-he), "যা-তা" (ja-ta – which-that), "যিনি-তিনি" (jini-tini – who-he), "যাকে-তাকে" (jake-take – whom-he), "যেই-সেই" (jei-shei – when-then), "যাহাকে-তাহাকে" (zahake-tahake – whom-him). If the first word is existed, the second word is also existed [28]. For example, "যিনি রাজা তিনি ঋষি" (jini raja tini rishi - he is the king of the sage).

(23) 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) or "যিনি x তিনি y" (zini x tini y - who 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 word and vice versa. For example, "যিনি রাজা তিনি ঋষি" (jini raja tini rishi - he is the king of the sage).

Here, if we can identify any word from "রাজা" (raja - king) or "ঋষি" (rishi - sage) then we will also identify another word.

Rule:

PairWordsList = List of Pair of words as "যে-সে" (je-she – who-he), "যা-তা" (ja-ta – which-that), "যিনি-তিনি" (jini-tini – who-he), etc.

IF (There are pare of words from PairWordsList in sentences) THEN

IF(Any word is known immediately after any of the pare words) THEN

TAG the word immediately after the other pare words with same tag

END IF

END IF

After dependency parsing, the tagging of words has been improved from 72.53% (result of word tagging after special tagging) to 79.50% in our experiment. In the dependency parsing, Hidden Markov model has been followed because some states are known and some are needed to explore based on the previous states. The algorithm of dependency parsing is given below:

Algorithm 2: Dependency parsing

Input: W: List of words with the tag after general and special tagging for all the sentences

Output: W: List of words with the tag after dependency parsing for all sentences

Begin

`/* Initialization of some variables */`

S ← List of sentences from the segmentation of a Bangla news document

R ←

Rule base which contains list of rules with their priority for dependency parsing

Rt ← ϕ//Rt is to hold the selected rule(s) temporarily

WTags ← ϕ// contain some words with tags temporarily

n ← number of sentences in the input document

m ← ϕ// contain the number of words in a sentence temporarily

for *i* ← 1 ***to*** *n* **do**

    WTags ← Get words with corresponding tags of i^{th} sentence

    m ← The total number of words in WTags

    for *j* ← 2 ***to*** *m* **do**

        Rt ← Get rules for the j^{th} word from rule base R

        if *Count(Rt) >1* **then**

            `/* More than one rules are conflicted */`

            Rt ← Get the rule with top priority from Rt

        end

        Apply Rt to tag the j^{th} word from the list of words W

    end

end

return *W*

End

An example of dependency parsing is given in the following figure 5.1. In the figure, a sentence with some tags is taken as input where the words "খুব" and "ভাল" are taged as “Adj” (adjective) and the next word "লোক" is unknown. Now, as per the rules (1) and (2) of dependency parsing (mentioned previously), the word after single or consecutive adjectives can be ”N” (noun) or ”V” (verb). Since there is no suffix with the word "লোক" it is taged as ”N” (noun) based on the rule (2).

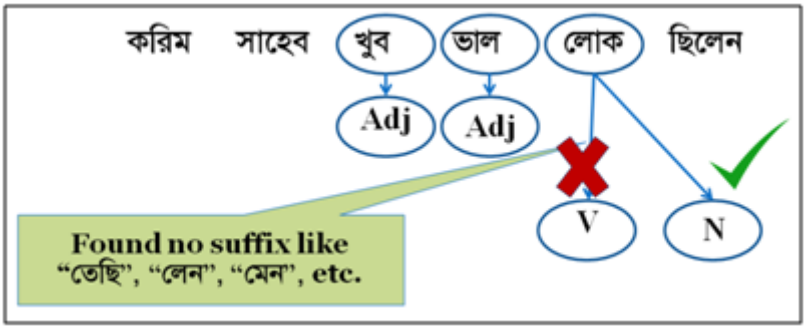


Figure 5.1: Example of dependency parsing

5.5 Conclusion

The process of identification of all the words have been described in this chapter in three steps as (i) general tagging, (ii) special tagging, and (iii) dependency parsing. The general tagging is a process to detect parts of speech in sentences with the help of list of predefined words. About 66.73% words have been tagged in general tagging. Special tagging has been introduced here to recognize the words as numerical figure, acronym, occupation, etc. and identified 72.53% words in total. After all, dependency parsing has been accomplished so that we may identify more words that can’t be tagged in general and special tagging process. Here, dependency parsing is a semi supervised way where some words are already

tagged and other words are tagged with the help of previously tagged words. The experimental results show that ultimately 79.50% words have been recognized after dependency parsing.

Chapter 6

Replacing Pronoun by Corresponding Noun

6.1 Introduction

Pronouns are used for representing nouns by which repetition of the same nouns entities can be minimized. In this situation, one can understand the pronouns by considering the corresponding nouns. In case of extraction based text summarization procedures, some sentences can be extracted where pronouns are available without corresponding nouns. These pronouns become dangling pronouns and make the information incoherent. So, the systems that have been developed for burden minimization from large volume of text, may deliver wrong message. Other than receiving a direction, the user will often be misguided with misinformation. In this regard, a technique is introduced here for replacing pronouns by corresponding nouns to generate dangling pronoun free summary.

Eight forms of pronouns are considered here for replacement as: (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). The corresponding nouns of these pronouns of any sentence are the subject or object of the immediate two previous sentences for 88.63% times (discussed in the chapter 3). So the subjects and objects need to be identified for all the sentences of the input document. But, identification of subject and object is complex in Bangla than that of English. Because, the placement of subject in English sentence is generally before the verb phrase, auxiliary verb or it may appear after the word by in passive voice. But, subject may be appeared in several places in Bangla sentence. Moreover, the subject and object can be full named entity. So, a technique is introduced here to find the full named entities by aggregating the parts of names (parts of names have been identified in the previous chapter). In some cases, parts of names are used in the sentences instead of full names. So, a mechanism is presented here so that the full names can be recalled from the parts of names and to make the replacement in suitable format without any inconsistency. The details explanation of the pronoun replacement procedure is depicted in this chapter.

The rest of the chapter is organized as follows: the process of full human named entity identification is presented in section 6.2, section 6.3 describes the mechanism for keeping the named entity in such a way so that full name can be recalled using part of name. Section 6.4 illustrates the way of replacement of pronoun and finally this chapter is concluded in section 6.5.

6.2 Finding the Full Name of Human

In case of general and special tagging (discussed in the previous chapter), most of the tags are depended on some lists of words. It is usual that whatever will be the range of lists, there will be limitation. Specially, some words have been tagged as human name in the step of special tagging where some other named entities can be available in the input text. Some words might be wrongly tagged as named entities. Even, identification of all the parts of a name is almost impossible on the basis of the list of predefined words. In this regard, more analysis is necessary for each sentence to get the full human named entities that are subjects or objects.

It is noticeable that the existing technique for named entity recognition [60] has not been utilized here. Because, primarily selected named entities may be ignored based on the impact of surrounding words which is very significant feature but not available in [60]. Some more words can be named entities in the document that cannot be indicated based on the predefined lists of words as like [60]. In these circumstances, the existing technique [60] is not suitable for us. For pronoun replacement, full name needs to be recalled using the part of name which is another distinguished feature of our approach.

Based on our observation on 3000 news documents, the name of human is written as full-name for the first time for around 95% times. Sometime full-name is existed with occupation. Then, part of the name may be used anywhere in the document. Parts of name may be there at the first time of a news document if the name has already been appeared in several news documents which make the name familiar to all. It is usual that after a series of news for a single event, the part of name for the people involved with the event become known to all. So, using part

of name may serve the purpose after using the full name of human.

By using the part of name it is quite difficult to find out the full name because any single word may be used for multiple meanings. For example, individually the word "সুরুজ" (Shuruz) may indicate for “sun” but "সুরুজ মিয়া" (Shuruz Miah) will indicate a name of person as there is a recognizable last name "মিয়া" (Miah). In this regard, the input document is checked thoroughly to find out the named entities where full names are existed as discussed in the earlier of this step. Multiple words are checked at a time in this step for getting all parts of names such as the first, last and middle names with or without any initial. Now, the following rules are brought into play based on our study of Bangladeshi news documents and Bangla grammar books [28, 101] to get named entities (full human names) from the entire document:

(1) Generally occupation exists before the name of human in any text document. So, if any word has occupation tag without any article, consider the immediate next four words. Four words are considered as there may have an initial also before the full name and a 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 words (at least one of the words should be tagged as part of name based on the step of special tagging in section 5.3).

Rule:

IF(The word is tagged as Occupation AND it has no article as suffix) THEN

Consider the next four words

IF(The next four words are untagged OR tagged as the first, middle, last name or initial) THEN

IF(At least one word has the first, middle, last name tagging) THEN

Select the words as one human named entity where the last word is tagged as the first, middle or last name tagging

ELSE

No named entity is found

END IF

END IF

END IF

(2) If there is any 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 adjacent two words also. Total three words are considered as there are generally three parts of a name (the first, middle and last part of name) [119]. From these three words, take the words as one named entity those are tagged as the first name, last name, middle name, initial or any untagged word. But, if no other words are there with the considered word to form the full name, ignore the word.

Rule:

IF(The word is tagged as the first, last or middle name or initial) THEN

Consider the adjacent two words

IF(The words are tagged as the first, middle, last name or initial) THEN

Select the words as one human named entity

ELSE

No named entity is found

END IF

END IF

(3) If there is any verb at the end of sentence, move from this verb to the beginning of sentence for collecting a named entity as like the above points (1) and (2) of this step.

(4) If there is a comma (punctuation mark) followed by a word with verb tag, there can be 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. move from this word to the beginning of the sentence for collecting named entity as like the above points (1) and (2) of this step.

(5) If any word is found as verb without an adjacent comma (punctuation mark) and the word is not at the end of sentence, move from this word to the end of sentence for collecting a named entity as like the above points (1) and (2).

(6) Based on our study, we have observed that some digits are enclosed with parentheses which indicate the age of a person immediately after name. For example: "আব্দুল বাতেন (২৪)" (Abdul Baten (24)), "আশুতোষ গুপ্ত (৩০)" (Ashutosh Gupto (30)), etc. So, look for named entity immediately before such digits enclosed with parentheses. In this regard, maximum three digits are considered as the indicator of age.

Rule:

IF(The word is tagged as Numerical Figure AND it is enclosed with parentheses) THEN

IF(The number of digits are not more than 3) THEN

Check three words immediately before the numerical figure and

apply the above rules (1) and (2) to get a full named entity

END IF

END IF

(7) There is named entity immediately after the word "নাম" (nam - name) and immediately before the word "নামে" (name - named) or "নামের" (namer - name').

Rule:

IF(The word is "নাম" (nam - name)) THEN

Check three words immediately after the word and apply rules (1) and (2) to get the full human named entity

ELSE IF(The word is "নামে" (name - named) OR "নামের" (namer name)) THEN

Check three words immediately before the word and apply rules (1) and (2) to get the full human named entity

END IF

(8) There may have wrongly selected human named entities in the previous points. For verifying every named entity, the immediate previous word for each named entity is considered. If the previous word is numerical figure, word with article or repeated words, the considered word is not taken as name of human. Some words are generally placed before adjective such as "অপেক্ষা" (opekkha - than), "চেয়ে" (cheye - than), "অধিক" (odhik - more), etc. [28]. So, if any of these words is existed immediately before the considered word, it can't be named entity. In this way, wrongly selected named entities will be removed.

Rule:

IF(The previous word of any named entity is repeated words OR word with

article OR numerical figure) THEN

The considered named entity will be deleted

ELSE IF(The previous word of any named entity is "অপেক্ষা" (opekkha - than) OR "চেয়ে" (cheye - than) OR "অধিক" (odhik - more), etc.) THEN

The considered named entity will be deleted

END IF

(9) It may be happened that all the named entities are selected properly but for replacing pronoun only singular pronouns are taken into account where the corresponding nouns (named entities) should also be singular. So, it is checked for each named entity that it is connected with another named entity with the words "ও" (o - and), "এবং" (ebong - and), "আর" (are - also), etc. Because these words are generally used for integrating two or more singular entities to make them plural. So, if the previous or next word of any named entity is one of these words, it will not be considered as corresponding noun.

Rule:

IF(The previous or next word of any named entity is "ও" (o - and), "এবং" (ebong - and), "আর" (are - also), etc.) THEN

Discard the named entity from consideration

END IF

Here, the procedure of full human named entity identification is based on Markov Chain model [54]. The algorithm of the named entity identification is given in the following page:

Algorithm 3: Identifying full human names

Input: W: List of words with the tags after word tagging
Output: ListNE: List of full human named entities

Begin
`/* Initialization of some variables */`
$S \leftarrow$ *List of sentences from the segmentation of a Bangla news document*
$R \leftarrow$ *List of Rules for full human name identification with priority of rules*
$ListNE \leftarrow \phi$*//ListNE will contain the list of full human named entity*
$Ln \leftarrow 0$*// number of found named entity*
$NE \leftarrow \phi$*// contains the parts of named entity*
$NE_Flag \leftarrow 0$*// A flag for verifying named entity*
$WTags \leftarrow \phi$*// for containing tags and words*
$n \leftarrow$ *number of sentences in the input document*
$m \leftarrow \phi$*// contain the number of words in a sentence temporarily*
for $i \leftarrow 1$ ***to*** n **do**
  $WTags \leftarrow$ *Get words and corresponding tags of* i^{th} *sentence*
  $m \leftarrow$ *The total number of words in WTags*
  for $j \leftarrow 1$ ***to*** m **do**
    if *tag(*j^{th} *word) = Occupation OR initial OR the first name of human* **then**
      $NE \leftarrow j^{th}$ *word*
      $NE_Flag \leftarrow 0$
      for $k \leftarrow j+1$ ***to*** $j+3$ **do**
        `/* Consider the next three words */`
        if *tag(*k^{th} *word) = first, middle or last name OR Untagged word* **then**
          $NE \leftarrow NE\ U\ k^{th}$ *word*
          if *tag(*k^{th} *word) = first, middle or last name* **then**
            $NE_Flag \leftarrow 1$*// verified as named entity*
          end
        end
      end
      if $Length(NE) \geq 2$ **then**
        `/* The length of name has more than one words */`
        if $NE_Flag = 1$ **then**
          Ln^{++} *// Increase the number of collected named entity*
          $ListNE[Ln] = NE$
          $j = k$*// forward the loop to kth word*
        end
      end
    end
  end
end
return $ListNE$
End

6.3 Keep the Named Entities in an Associative Array

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 maintained which means that the index of the array is word. For example, if a named entity is presented as "প্রধান শিক্ষক লতিফুর রহমান খান" (Prodhan Shikkhok Lotifur Rahman khan - Head Master Lotifur Rahman Khan), it will be placed in the array for five times based on the parts of name as in the figure 6.1.

Index		value
[প্রধান]	=>	প্রধান শিক্ষক লতিফুর রহমান খান
[Head]	=>	Head Master Lotifur Rahman Khan
[শিক্ষক]	=>	প্রধান শিক্ষক লতিফুর রহমান খান
[Master]	=>	Head Master Lotifur Rahman Khan
[লতিফুর]	=>	প্রধান শিক্ষক লতিফুর রহমান খান
[Lotifur]	=>	Head Master Lotifur Rahman Khan
[রহমান]	=>	প্রধান শিক্ষক লতিফুর রহমান খান
[Rahman]	=>	Head Master Lotifur Rahman Khan
[খান]	=>	প্রধান শিক্ষক লতিফুর রহমান খান
[Khan]	=>	Head Master Lotifur Rahman Khan

Figure 6.1: Structure of associative array for keeping named entities

This mechanism of associative array has been used 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 value of the resultant index "প্রধান শিক্ষক লতিফুর রহমান খান" (Prodhan Shikkhok Lotifur Rahman khan - Head Master Lotifur Rahman Khan). If the same part of name will be found in two names then both the index will be deleted and other parts of name will be considered.

The algorithm of keeping the list of named entities in associative array is given below:

Algorithm 4: Keep the list of named entities in an associative array

Input: ListNE: List of full human named entities
Output: NEA: An array which will contain the part of name as index and full name as value

```
1  Begin
   /* Initialization of some variables                              */
2  PNE ← φ// PNE is an array which will contain parts of named entity
3  Prt ← φ// contain a part of name
4  for i ← 1 to Count(ListNE) do
5  |  PNE ←
   |  explode(ListNE[i])// break a full named entity to parts of name
6  |  for j ← 1 to Count(PNE) do
7  |  |  Prt ← PNE[j]
8  |  |  if NEA[Prt] ≠ φ then
   |  |  |  /* set empty for duplicate part of name entry          */
9  |  |  |  NEA[Prt] ← φ
10 |  |  else
11 |  |  |  NEA[Prt] ← ListNE[i]
12 |  |  end
13 |  end
14 end
15 return NEA
16 End
```

6.4 Replacing Pronoun by Corresponding Noun

Though recognizing name of human is an important task, some other processes are also indispensable for replacing pronouns. For example, from the identified named entities, subject and object needs to be detected. In this step, some rules are applied for recognizing subject and object of each sentence, distinguishing the

corresponding nouns of pronouns and replacing pronouns as follows:

(1) For the replacement, eight forms of singular pronouns are taken into account from the input document 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). Considering the other forms of pronoun except these eight forms are left as future work. Because it will make the process very complex to handle plural forms of pronouns. Experiment with 600 summaries, it has been found that dangling pronouns are appeared for 98.50% times for the considered eight forms of pronouns only.

(2) Some special cases are there in the sentence of Bangla language where pronouns are available but they are not dangling pronouns. In these cases, pronouns are kept as they are. For example, "তাকে" (take - him/her) is followed by "যাকে" (zake - whom), "সে" (she - he/she) is followed by "যে" (ze - who), etc.

(3) The two immediate previous sentences are considered for getting the named entity as the corresponding noun of pronoun. It has been found in our experiment that the corresponding noun is available within the two immediate previous sentences for 88.63% times. So, get the named entity (corresponding noun) of immediate previous sentence as discussed in the previous step. If no named entity is available in the immediate previous sentence, look for the named entity in the second previous sentence. If only one named entity is presented, 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 subject or object plural which

is not considered here. If there are exactly two named entities, it is needed to decide which is subject and which is object. Special attention need to be placed on the replaceable pronoun that it will be replaced by subject or object of the previous sentence. In this regard, following rules are applied:

(i) Generally, it has been found that named entity with the following suffixes are object: "কে" (ke), "রে" (re), "এর" (er), etc. [28].

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

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

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

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

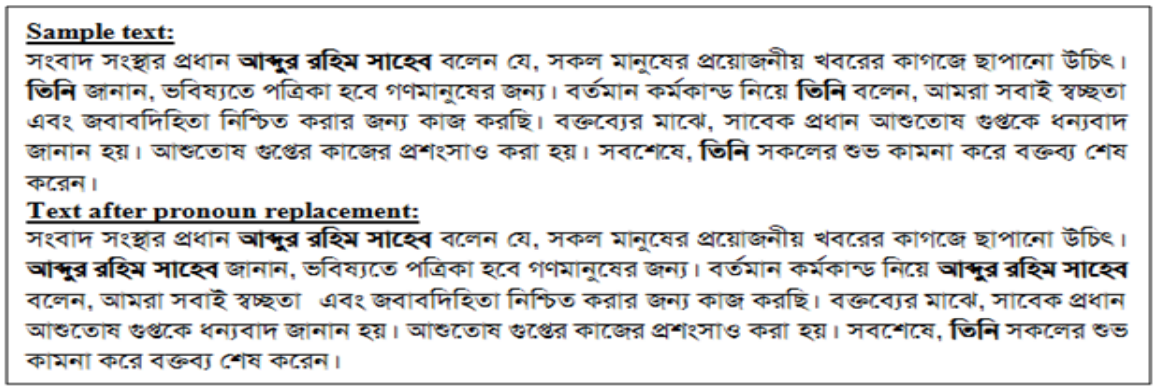
(vi) In most of the time, subject can be replaced by subject and object can be replaced by object while pronoun replacement. The pronouns such as "তাকে" (take - him/her), "তাহাকে" (tahake - him/her), "তার" (tar - his/her) and "তাহার" (tahar - his/her) will be replaced by object of previous sentence.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

(viii) For ensuring replacement of pronoun 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 with the noun, b) if the pronoun is "তার" (tar - his/her) or "তাহার" (tahar - his/her), a suffix "এর" (er) should be added with 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 our proposed technique, named entity may be existed as only one word where it is difficult to determine whether it is really a named entity or not. To overcome this situation, all the named entities of the input document have been kept in an associative array in word by word (discussed in the section 6.3). So, if the system needs to consider a single word is named entity or not, the word will be searched in the associative array. If the word is located as index in the array, the value of the resultant index is used to replace the pronoun otherwise it is left without replacement.

Point to be mentioned that 100% accurate result can't be produced by light weight stemmer [112]. So, both the stemmed form of the words and the actual words used in the text are considered for the general tagging, special tagging, dependency parsing, finding full name, and in replacing pronoun.


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।

Figure 6.2: Sample text for the example of replacement of pronoun

A sample text has been given in the figure 6.2 to illustrate the output of our proposed tehcnique where the sample event and names are imaginary. Here, the pronouns and nouns are marked with bold form. In the original message, there is one pronoun "তিনি" (tini - he) mentioned for three times. In the message after applying our technique, the pronoun "তিনি" (tini - he) has been replaced correctly for the first two times by corresponding noun "আব্দুর রহিম সাহেব" (Mr. Abdur Rahim). For the third times, it has not been replaced because the corresponding noun was not found within the two immediate previous sentences.

Markov Chain model [54] has been followed in the process of replacement of pronoun. The algorithom for replacement of pronoun by corresponding noun is given in the next page.

6.5 Conclusion

It has been discussed here that if there is any pronoun in the summary without the corresponding noun, the summary will deliver a confusing message to the user.

Algorithm 5: Replacement of pronoun by corresponding noun

Input: S1: List of sentences from the segmentation of a Bangla news document

Output: S2: List of sentences after replacement of pronoun

Begin

`/* Initialization of some variables */`

W ← List of words with the tag after dependency parsing for all sentences

R ← List of Rules for replacing pronoun with priority of rules

ListNE ← List of full human named entities

NEA ←
An array which will contain the parts of name as index and full name as value

PRO ← List of pronouns for replacement

TmpPRO ← ϕ// Contains the list of pronouns of a sentence

NEs ← ϕ// Contains the named entities of a sentence

TmpNEs ← ϕ// Contains some named entities temporarily

n ← number of sentences in S1

for *i ← 2* ***to*** *n* **do**

`/* This loop is starting from second position because the pronoun replacement is not required for the first sentence. */`

TmpPRO ←
Get the pronouns in 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 entity from NEs without any suffix as subject

if *Count(TmpNEs) = 1* **then**

`/* if the subject is singular named entity */`

Replace the TmpPRO by TmpNEs

end

else if *TmpPRO =* "তার" *OR* "তাকে" *OR* "তাহাকে" *OR* "তাহার" **then**

`/* for object */`

TmpNEs ←
Get the named entity from NEs with suffix as "কে, রে, এর, রের, র" *as object*

if *Count(TmpNEs) = 1* **then**

`/* if the object is singular named entity */`

Replace the TmpPRO by TmpNEs

end

end

end

S2 ← S1

return *S2*

End

In this situation, the procedure of replacement of pronoun by corresponding noun has been introduced in this chapter so that the summary will be unambiguous. The corresponding nouns of the pronouns can be subject or object of the previous sentence. So, subject and object have been identified and corresponding nouns of pronouns have been detected. Finally, a mechanism has been introduced to ensure the replacement in suitable format. It is also expected that this replacement process for solving the problem of dangling pronoun will help in further research work of Bangla information retrieval.

Chapter 7

Sentence Ranking & Summary Generation

7.1 Introduction

In any kind of summary generation system, a vital step is to select important sentences. Various strategies have been followed by researchers for significant sentence identification [8, 19, 20, 61]. Some researchers proposed query oriented strategy where sentences are selected that respond well to the users defined query. Sometime, the query is too inadequate to allow for meaningful measurement of similarity with sentences. One alternative is to follow the course set by information retrieval procedure and try to expand the query [111]. The maximum coverage formulation was introduced for summarization by Filatova et al. [123]. They have utilized a greedy approach and the main spotlight was on entity mining. On the other hand, You Ouyang et al. selected sentences using the concept of hierarchical representation [107]. Rada Mihalcea et al. [90] generated links between sentences

as per sentences similarity relation for summary sentences selection. In most of the cases of extraction based summary generation systems, sentences are selected based on their score or rank [19, 20].

In this chapter, 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. Sentence ranking is accomplished by doing the followings: (i) term frequency inverse document frequency (TF-IDF) score calculation, (ii) sentence frequency (SF) score calculation and redundancy elimination, (iii) counting score for the existence of numerical figure presented in words and digits, (iv) computation of score for title words, and (v) special consideration of the first sentence. Finally, one third (or the ratio between summary to source document set by the user) top ranked sentences are selected as the final summary.

The rest of the chapter is organized as follows: sentence ranking is explained in details in section 7.2, section 7.3 illustrates the procedure of summary generation. The algorithm of summary generation is given in section 7.4 and finally this chapter is concluded in section 7.5.

7.2 Sentence Ranking

The process of sentence ranking is accomplished by considering term frequency inverse document frequency (TF-IDF), sentence frequency (SF), numerical figure presented in words and digits, title words, and the first sentence. All these features are discussed below:

7.2.1 Calculation of TF-IDF

The TF-IDF or term frequency-inverse document frequency is used to measure the weight of terms as per their number of appearance. Typically, the TF-IDF weight is composed by the following two terms:

a) TF: Term Frequency, which measures how frequently a term exists in a document. Since the length of every document is usually varied, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is divided by the document length (total number of terms in the document) as a way of normalization [62]:

   TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

b) IDF: Inverse Document Frequency, which measures how significant a term is. While computing TF, all terms are considered equally important. However, it is known that some terms may appear a lot of times but have little magnitude. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following [62]:

   IDF(t) = log_e(Total number of documents / Number of documents containing term t)

In the information retrieval and text mining algorithm TF-IDF based weight has been frequently utilized [18, 26, 62]. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus [92]. The TF-IDF score is calculated with the following equations:

$$TF - IDF_{(t)} = TF * log(N/DF) \tag{7.1}$$

$$S_{TF-IDF_{(k)}} = \sum_{t=1}^{T} TF - IDF_{(t)} \tag{7.2}$$

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

7.2.2 Calculation of Sentence frequency (S_{SF}) and elimination of Redundancy

For sentence ranking, this proposed method has introduced second attribute as sentence frequency (S_{SF}) which is based on cosine similarity. Here, set of sentences S [s1, s2, s3, , sn] have been taken from the segmentation of input document. The sentence frequency of each sentence is set as 1 (one) at first. If one sentence has cosine similarity 60% or more with any other, smaller sentence is removed and the frequency of larger sentence will be the summation of the frequency of both of the sentences. As there is 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 ratio [110]. The sentence frequency (S_{SF}) score is calculated using the following equation:

$$\forall_{i\epsilon\{1,...,n\}} \begin{cases} \forall_{j\epsilon\{1,...,n\}} \\ \quad If(i=j) \text{ Then} \\ \quad\quad Continue \text{ //Same sentence} \\ \quad \text{Else If } Sim(S_i,S_j) \geq 60\% \text{ AND } len(S_i) > len(S_j) \text{ Then} \\ \quad\quad S_{SF(i)} = S_{SF(i)} + 1 \\ \quad\quad Remove S_j \text{ from the Sentence Set S} \\ \quad \text{Else If } Sim(S_i,S_j) \geq 60\% \text{ AND } len(S_j) > len(S_i) \text{ Then} \\ \quad\quad S_{SF(j)} = S_{SF(j)} + 1 \\ \quad\quad Remove S_i \text{ from the Sentence Set S} \end{cases} \tag{7.3}$$

where n is the number of sentences; $S_{SF(i)}$ is the sentence frequency of i^{th} sentence which is initially 1 (one) 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 measure between two sentences S_i = [wi1, wi2, .,wim] and S_j = [wj1, wj2, .,wjm] is computed as [124]:

$$Sim(S_i,S_j) = \frac{\sum_{k=1}^{m} W_{ik}W_{jk}}{\sum_{k=1}^{m} W_{ik}^2 . \sum_{k=1}^{m} W_{jk}^2}, i,j = 1,2,3,...,n \tag{7.4}$$

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

It is noticeable that the score of TF-IDF and SF has been boosted up for the replacement of pronoun (replacement has been discussed in chapter 6). For example, a noun can have better TF score but if the pronoun form of the noun is existed in any sentence, the sentence will not get any TF score for this pronoun. Moreover, pronoun and noun can't be matched but after replacement of

pronoun by corresponding noun they can be recognized as same entity for sentence frequency calculation.

In the following figure 7.1, the pronoun "তিনি" (tini-he) has no TF score but after replacing this by the corresponding noun "করিম সাহেব" (Korim Shaheb - Mr. Karim), TF score can be calculated.

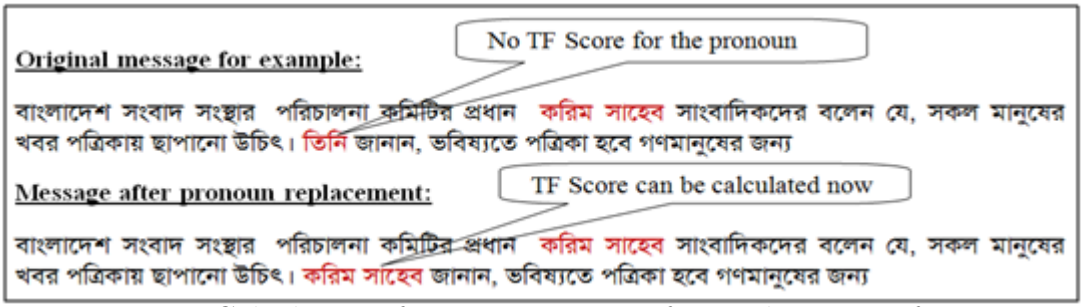


Figure 7.1: Calculation of Term Frequency after replacement of pronoun

7.2.3 Counting numerical figure presented in words and digits (S_N)

The third attribute is to count numerical figure for each sentence (S_N). The value of S_N for each sentence is set to 0 (zero) at first and for the existence of each numerical figure it will be incremented by 1 (one). In [17, 31, 57], 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. Even, numerical figure can have various suffixes in Bangla text. We may consider the following two sentences for example, "করিমের জন্ম সাল ২০০৬।তাহার বয়স দশ বছর।' (korimer jonmo shal 2006 ।tahar boyosh dosh bochhor - Karim's birth year is 2006. He is ten years old.). Existing procedure [17,31,57] can find a 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 here to recognize Bangla numerical figure from both words and digits mentioned in the special tagging part in chapter 5. 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 (in chapter 4) and count the numerical figure in digits and words based on the following equations:

$$\forall_{i\epsilon\{1,\ldots,n\}} N_{digits(i)} = Regexp(S_{(i)}, [0,1,2,3,4,5,6,7,8,9]) \tag{7.5}$$

$$\forall_{i\epsilon\{1,\ldots,n\}} N_{words(i)} = Regexp(S_{(i)}, [FormatOfNumInWords]) \tag{7.6}$$

$$\forall_{i\epsilon\{1,\ldots,n\}} S_{N(i)} = N_{digits(i)} + N_{words(i)} \tag{7.7}$$

where n is the number of sentences; N_{digits} and N_{words} are the number of numerical figure presented in digits and words respectively; *Regexp* function returns the number of matches between the corresponding sentence and the given pattern as the second argument of this function. The pattern for matching digits is 0 to 9 and for words is *FormatOfNumInWords* (explained in the section of special tagging in chapter 5). Finally both N_{digits} and N_{words} are summed up for each sentence individually to get S_N which is the score of sentence for numerical figure.

7.2.4 Computation of score for title words (S_T)

In several existing methods [17,34,62], title words have been considered for sentence scoring. We have also observed from the analysis of 3000 news documents that title words convey the theme of the news document in the most cases. The score of each

sentence for title word is set to 0 (zero) at first and incremented by 1 for the existence of each title word in the sentence. For computing the title words 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 the following equation:

$$\forall_{i\epsilon\{1,\ldots,n\}} S_{T(i)} = match(S_{W(i)}, TW) \tag{7.8}$$

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.

7.2.5 Special consideration of the first sentence

In some existing methods [2, 57, 75], the sentence score is depended on 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. But, in most of the time especially for Bangla news documents, the first sentence is much important than any other sentences as per our experiment which is explained in the lower part of this sub-section. So, general positional score (which is gradually decreasing as like [2, 57, 75]) is not applicable for the first sentence of news documents. Again, some existing summarization methods emphasized on sentences those contain any title word [17, 34, 62] and in Bangla news documents, the first sentence contains the full title often. So, an extra care is proposed here for the first sentence of the input document.

In the experiment with our training data set (discussed in chapter 8), it has been found that the first sentence is existed in the summary for 78% times.

So, if the first sentence is always kept in summary, there will be wrong selection for 22% (100 - 78) times. But, after scrutinizing one step ahead, it has been found that if the first sentence contains any title word, it is existed in summary for 88% times where error rate is 12% (100-88). So, it is proposed here to select the first sentence in summary if it contains any title word.

Point to be noted that this type of special care for the first sentence has been proposed here for the Bangla news documents only and it may not be suitable for others.

After measuring all the attributes value, the ultimate rank of each sentence is computed using the following equation where S_k is the rank of k^{th} sentence:

$$S_k = \begin{cases} w_1 \times S_{TF(k)} + w_2 \times S_{SF(k)} + w_3 \times S_{N(k)} + w_4 \times S_{T(k)}, \textit{if } k > 1 \\ max(S_k) + 1, \textit{if } k = 1 \textit{ and } S_1 \textit{ contains any title word} \end{cases} \tag{7.9}$$

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 rank of the first sentence will be set as the highest value + 1 if it contains any title word so that it will be selected always.

The values of the coefficients w_1, w_2, w_3, and w_4 in the above equation are obtained by tuning them for the best results of summary generation by proposed system. In order to select optimal value for each coefficient, experiment is done with a training data set of 200 documents and 600 model summaries (3 model summaries for each document) generated by human professionals. In the devised experiment, the summary is generated for each training document by setting a value of each coefficient (ranging from 0.00 to 1.00). 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 chapter 8). After generating all

the summaries from the training documents, the average F-measure for each value of coefficient is calculated by comparing the system generated summary and the corresponding model summary. In this way, the optimal values are identified as 0.87, 0.09, 0.02, and 0.21 for w_1, w_2, w_3, and w_4 respectively. The figures for showing the tuning graph for the adjustment of w_1, w_2, w_3, and w_4 are given in the next chapter.

7.3 Summary Generation

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

$$\forall_{i\epsilon\{1,\ldots,n/3\}} SumSen = SumSen \cup ExtTopScored(S) \tag{7.10}$$

where n is the number of sentences; ExtTopScored function extract top scored sentences from sentence 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 input document to summary based on [5] if it is not specified by user.

7.4 Algorithm of Sentence Ranking & Summary Generation

The procedure of sentence ranking and summary generation is based on the Markov Chain model [54]. The algorithm of this procedure is given below:

Algorithm 6: Sentence ranking & summary generation

Input: S: List of sentences after replacement of pronoun by corresponding noun

Output: SUMMARY: Summary sentences

Begin

`/* Initialization of some variables */`

$WT \leftarrow$ *List of title words of the input document*

$S_{TF-IDF} \leftarrow 0$*// TF-IDF score*

$S_{SF} \leftarrow 0$*// Score for sentence frequency*

$S_N \leftarrow 0$*// Score of numerical figure*

$S_{TITLE} \leftarrow 0$*// Score for title words*

$SCORES \leftarrow \phi$*//For containing the scores of all the sentences*

$n \leftarrow$ *number of sentences in S*

for $i \leftarrow 1$ ***to*** n **do**

  $S_{TF-IDF} \leftarrow$ *TF-IDF score based on equation 7.2*

  $S_{SF} \leftarrow$ *Sentence frequency score based on equation 7.3*

  $S_N \leftarrow$ *Score for numerical figure based on equation 7.7*

  $S_{TITLE} \leftarrow$ *Score for title words based on equation 7.8*

  $SCORES \leftarrow$

  Accumulated score of i^{th} *sentence based on equation by equation 7.9*

end

Sort SCORES in Descending order

for $i \leftarrow 1$ ***to*** n **do**

  if *SCORES[i]* $\geq$ *Top one third scores from the SCORES* **then**

    `/* Keep the` i^{th} `sentence in summary */`

    $SUMMARY = SUMMARY \cup S[i]$

  end

end

return $SUMMARY$

End

7.5 Conclusion

Important sentences selection is the part and parcel in the process of automatic text summarization. Beyond the surface level measures of sentence selection (such as position of sentences, matching with title or heading, the first or last sentence of each paragraph), something more have been introduced here for sentence ranking as (i) sentence frequency calculation, (ii) numerical figure identification presented in words and digits, and (iii) consideration of the first sentence specially. The value of each attribute of sentence ranking has been tuned for better summarization performance. Ultimate rank of each sentence is computed by aggregating the values of all the attributes. After all, one third top ranked sentences have been selected as summary. In the next chapter, performance evaluation of the proposed method is depicted.

Chapter 8
Results and Discussion

8.1 Introduction

Evaluating summarization method is a difficult task and the sophisticated way is yet to be achieved [5]. In this situation, several techniques have been applied to measure the quality of summary which is depended on the followings: a) importance of selected contents and b) presentation quality. Again, presentation quality can be assessed based on grammatical correctness and coherence. Considering all these aspects, evaluation procedures are divided into two main categories as: a) intrinsic mode and b) extrinsic mode [23].

Intrinsic evaluation of selected contents is usually done by comparing system generated summaries with model summaries written by human professionals. More specifically, evaluation is achieved by measuring the overlap between model summary and the automatically extracted summary as in ROUGE evaluation system [103]. In extrinsic evaluation method, the quality of summary is judged based on how it affects the completion of some other task. For example, human

annotators use the system generated summaries and human generated summaries to categorize documents and compare their accuracy in categorization to measure the quality of the summaries.

In this chapter, discussion is given about intrinsic evaluation of the proposed method. Here, the proposed procedure for text summarization is compared with the four latest existing methods and evaluated using ROUGE [103] against the summarization by human professionals.

The rest of the chapter is organized as follows: the description of data sets, evaluation procedure, basic system requirements and evaluation measures is given in section 8.2, all the experimental results and the discussion on results are illustrated in section 8.3 and 8.4 respectively. Finally, this chapter is concluded in section 8.5.

8.2 Experiments

8.2.1 Data sets

It has been discussed about the challenges for the research work in Bangla natural language processing in chapter 1. There is no automated tool for facilitating research work. Another problem that there is no benchmark data set which can be utilized for evaluating Bangla text summarization (BTS) system.

For training and evaluation of the proposed method, 3400 Bangla news documents (each document has 18 to 25 lines of Unicode text) have been collected from the most popular Bangladeshi newspaper the Daily Prothom-Alo (May 2015). These news documents contain variety of news that cover a wide range of topics like political, sports, crime, economy, environment, etc. We have analyzed 3000

documents to understand the structure of sentences in news document and identify the rules for replacing a pronoun by corresponding noun. For other 400 news documents, three human judges have generated summary for each document. Human generated summaries are considered as reference/model summaries. These 400 documents-summaries are divided into two data sets as (i) randomly selected 200 documents with corresponding model summaries are taken as training set for adjusting the value of w_1, w_2, w_3, and w_4 for sentence ranking (discussed in chapter 7) and (ii) other 200 documents with corresponding model summaries are treated as 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 Internet so that other researchers may evaluate their systems with this [125].

Point to be mentioned that the data set of 400 test documents is around ten times larger than the evaluation data set of some existing methods [2, 32, 33]. Moreover, some existing methods [2, 32, 33] were evaluated against one model summary only. But the evaluation is turned here with three model summaries of each test document. Here, someone may raise question for using human generated model summaries. But, the remarkable point is that human generated model summaries were also used for English text summarization methods despite the existence of benchmark data set [95, 126] and for other languages where there was no benchmark data set [27, 32, 33].

8.2.2 Procedure

In this research work, ROUGE [103] automatic evaluation package has been utilized, a widely used metric, to evaluate the automatically generated summaries of the proposed method. Updated ROUGE package has been applied here because it can be used for Bangla Unicode text [104]. In this package, there are two folders: (i) one folder will contain the model summaries and (ii) other folder will contain the system generateds summaries. After that, it will automatically generate the values of evaluation measures(discussed later in 8.2.4).

8.2.3 System Requirements

The proposed system of automatic text summarization is a web-based platform independent system. Followings are the basic requirements for this approach:

(i) Windows operation system version 98 or later. Or, Linux (Ubuntu) operating system version 13.04 or later

(ii) Apache web server version 2.2.11 or later

(iii) PHP (Personal Home Page) version 5.2.9 or later

(iv) MySQL Database version 5.1 or later

8.2.4 Evaluation Measures

Evaluating the quality of a summary is a difficult problem, principally because there is no obvious “ideal” summary [5]. Even, for relatively straightforward news articles, human summarizers tend to agree for around 60% content overlapping [5].

The quality of summary text is often assessed manually by human as there is no standard of text. It is known that there are two types of summary evaluation named: a) intrinsic, and b) extrinsic [127]. The main approach for quality of summary determination is the intrinsic content evaluation which is often done by comparing with an ideal summary (written by human professional) [23]. Some intrinsic method rated summaries as per readability, informativeness, fluency and coverage for one of the larger studies [22]. Some measures are given below as per the incorporated arrangements for evaluation of summary by various researchers:

(a) Text quality measures: There are several aspects of text quality assessment such as grammaticality, non-redundancy, text coherence and structure.

(b) Co-selection measures: In co-selection measures, the principal evaluation metrics are [128]:

(i) Precision (P): It is the number of sentences occurring in both system generated summary and ideal summary divided by the number of sentences in the system generated summary.

(ii) Recall (R): It is the number of sentences occurring in both system generated summary and ideal summary divided by the number of sentences in the ideal summary.

(iii) F-measure: The integrated measure that incorporated both Precision and Recall is F-measure.

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

equations:

$$Precision(P) = \frac{A \cap B}{A} \tag{8.1}$$

$$Recall(R) = \frac{A \cap B}{B} \tag{8.2}$$

$$F - measure = \frac{2 \times P \times R}{P + R} \tag{8.3}$$

(c) Content-based measures: Content-based measures can compute how much two sentences have matching with each other in terms of words.

(d) Task-based measures: Usability of summary for a specific task is measured here.

From the above four types of measurement, the Co-selection measures have been selected in this research work. It is noticeable that Co-selection measures have been considered in several text summarization systems for Bangla [2, 32, 33], English [25, 92], and other [27]. In the Co-selection based measurement of this proposed method, the system generated summary has been compared with three model summaries of 200 news documents each and the results of evaluation is the average results of the comparisons.

8.3 Experiments and Results

8.3.1 Results of word tagging

Experimental result of word tagging is given in the table 8.1 for each phase. The experiment has been conducted on 31525 words of 200 test documents.

Table 8.1: Results of word tagging of different phases

Phases of word tagging	**Number of tagged words**	**Percentage of tagged words**
General tagging	20532	65.13%
Considering suffixes for verb	21038	66.73%
Special tagging	22865	72.53%
Dependency parsing	25062	79.50%

Table 8.2: Experimental results of special tagging

#	**Nature of words**	**Success rate of identification**
1	English acronym	100.00%
2	Bangla initial of name	100.00%
3	Repeated words	100.00%
4	Numerical figure from digits	100.00%
5	Numerical figure from words	100.00%
6	Occupation	91.00%
7	Name of organization based on the number of letter in acronym which enclosed in parentheses	95.60%
8	Name of organization based on some specific last words	87.00%
9	Probable human named entity	80.00%
10	Name of places	82.00%

In the results of special tagging (shown in table 8.2), it has been found that some nature of words have been identified for 100% as acronym, initial, numerical figure from digits and words. The procedures have followed some patterns to identify these

(acronym, initial, numerical figure from digits and words) and not based on any limited number of predefined words. These specific patterns are the main reason for 100% success rate achievement. But some nature of words can't be identified completely. Here, occupation has been identified as 91%, name of organization by considering acronym and specific last words are 95.60% and 87% respectively, name of human and places have been identified as 80% and 82% correspondingly. The procedures have utilized some lists of predefined words to identify occupation, name of organization, name of human, and places. The noticeable point is that the utilized lists have limited number of predefined words for which 100% success rate is not possible for now.

8.3.2 Results on replacement of pronoun

From the evaluation data set (200 news documents), number of singular pronoun is counted. Now, input these documents to the system and the followings are counted for performance measurement for pronoun replacement: (i) correctly replaced pronouns, (ii) how many pronouns have been kept without replacing, and (iii) incorrectly replaced pronouns. The results of replacement of pronoun and number of pronouns have been given in the following tables 8.3 and 8.4 respectively.

Table 8.3: Result on pronoun replacement for 200 news documents

Number of pronouns	Kept unchanged	Replaced correctly	Replaced incorrectly
255	60	183	12

Table 8.4: Number of singular pronoun counting from 200 news documents

#	Singular Pronoun	Frequency	Percentage
1	"তিনি" (tini - he)	160	62.75%
2	"তার" (tar - his/her)	48	18.82%
3	"সে" (she - he)	33	12.94%
4	"তাকে" (take - him)	14	5.49%
5	"ইনি" (ini - he)	0	0.00%
6	"উনি" (uni - he)	0	0.00%
7	"তাহার" (tahar - his/her)	0	0.00%
8	"তাহাকে" (tahake - him)	0	0.00%
Total number of pronouns		255	100.00%

The results in the above table 8.3 illustrates that the system can replace 183 pronouns correctly from 255 pronouns in total, which implies the accuracy of the system is 71.80%. The reason, for which 100% accuracy could not be achieved, has been discussed in section 6.4. And, the incorrect replacement can be happened if the subject and object can't be identified properly while word tagging. More explanation has been given in the following section 8.3.8 regarding pronoun replacement.

8.3.3 Step by step improvements of performance for each feature

The proposed method has the following features:

(i) Replacing pronoun by corresponding noun which minimize the number of

dangling pronoun in summary.

(ii) Sentence ranking by accomplishing the followings: (a) term frequency inverse document frequency (TF-IDF) calculation, (b) sentence frequency (SF) 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 values of coefficients of all the attributes listed in the above point (ii).

From the above listed features, term frequency calculation, counting the numerical figure from digits and title word score have already been introduced in some existing methods which have been described in chapter 2. Now, to justify the significance of each feature of this proposed method, progress of performance has been given in the figure 8.1 and table 8.5 after inclusion of each feature. Here, Precision, Recall and F-measure have been calculated using the equations 8.1, 8.2, and 8.3 respectively upon training data set (discussed in the section 8.2.1). In the following figure 8.1, 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.

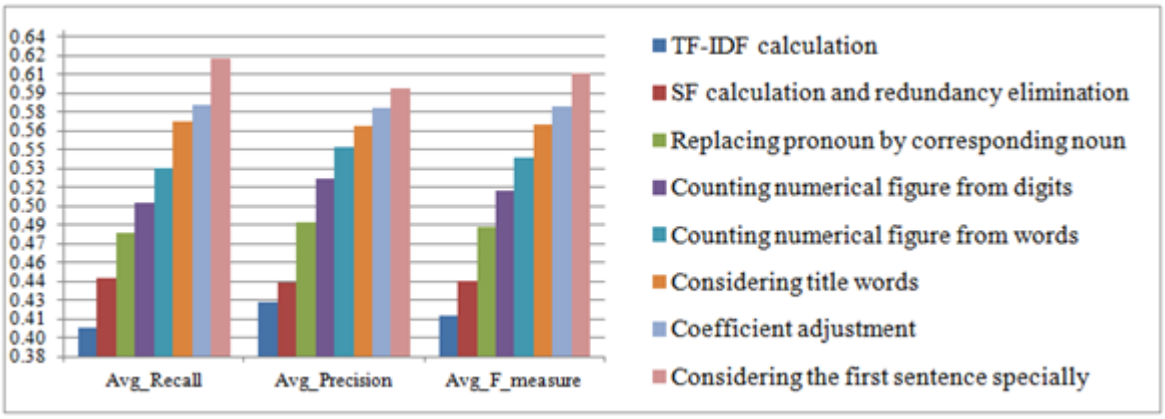


Figure 8.1: Step by step improvement of performance for including each feature

Table 8.5: Percentage of improvement for including each feature

#	**Features**	**F-Measure Score after including each feature**	**Improvement after including each feature**
1	TF-IDF	0.4125	N/A
2	Sentence frequency	0.4402	6.71%
3	Pronoun replacement	0.4827	9.66%
4	Count numerical figure from digits	0.5124	6.15%
5	Count numerical figure from words and digits	0.5386	5.11%
6	Title words	0.5653	4.96%
7	Coefficients adjustment	0.5793	2.47%
8	Considering the first sentence specially	0.6052	4.47%

The result in the above table has been generated using our training data set and for ROUGE-1 scores. Here, the improvement column shows the improvement of performance from the previous step. The first step (TF-IDF calculation) has no previous step for which the improvement column has value N/A (Not Applicable) for this.

8.3.4 Coefficients tuning for sentence ranking

The coefficients for the attributes of sentence ranking (mentioned in chapter 7, equation 7.9) have been tuned for the better performance as follows:

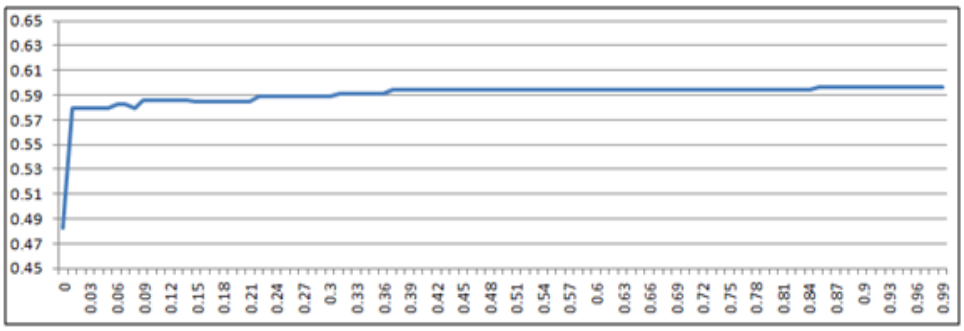


Figure 8.2: F-measure for various values of w_1

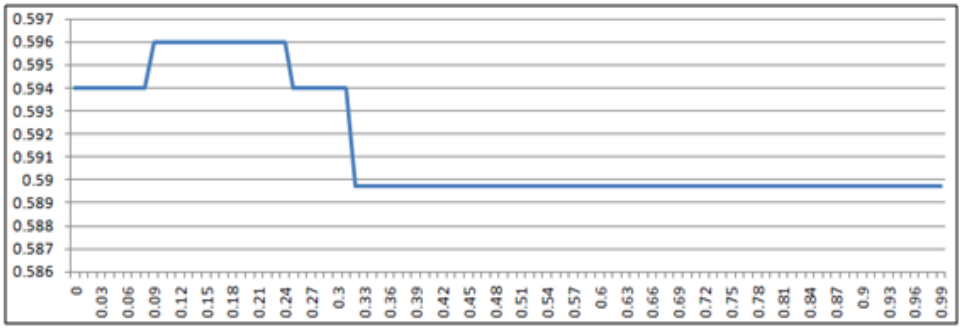


Figure 8.3: F-measure for various values of w_2

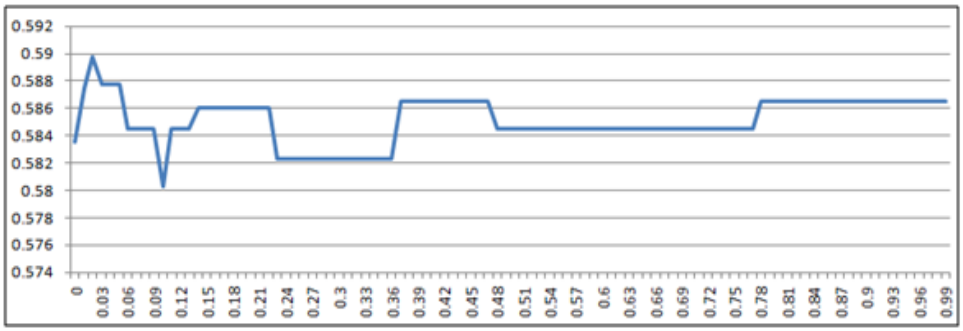


Figure 8.4: F-measure for various values of w_3

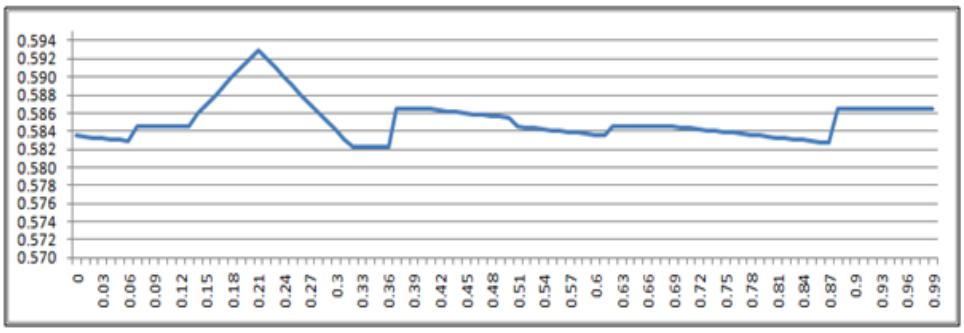


Figure 8.5: F-measure for various values of w_4

8.3.5 ROUGE Evaluation scores

To judge the efficiency of the proposed method, experiments have been conducted on 200 news documents. In each time, the system generated summary is compared with three model summaries of each document and computed the average value of Precision, Recall and F-measure with ROUGE automatic evaluation package [104]. The average results of ROUGE-1 and ROUGE-2 evaluation scores for 200 documents with 95% confidence interval is depicted in table 8.6.

Table 8.6: Average of ROUGE-1 and ROUGE-2 scores of the proposed system

	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

8.3.6 Comparison among the existing approaches of BTS

Eight approaches have been discussed in 2.6 with their pros and cons. The following table 8.7 turns the comparison among these approaches based on their incorporated features and evaluation results:

Table 8.7: Comparison among the existing approaches

Researcher(s), year	**Incorporated features**	**Remarks**	**Evaluation results**
Islam and Masum, 2004, [30]	i) Term frequency, and ii) Useful word list with important and unimportant word list	This method has keyword search module for summary generation where keywords are selected on the basis of TF-IDF and list of useful, important and unimportant words.	No evaluation has been drawn.
Md. Nizam Uddin and Shakil Akter Khan, 2007, [31]	i) Using location method, ii) Using Cue method, iii) Using Title method, and iv) Using Term frequency method	This research work has shown that some features of English text summarization can be used for Bangla text.	Got 8.4 from human professional in the range of 0 to 10 point. The standard evaulation process [103, 104] has not been followed here.

Kamal Sarkar, 2012, [32]	i) Term frequency, ii) Sentence length, and iii) Sentence position	The impact of thematic term has been investigated and some statistical measures have been incorporated for sentence scoring.	Unigram based recall score was found as 0.4122.
Kamal Sarkar, 2012, [33]	i) Term frequency, ii) Sentence length, and iii) Sentence position	It was claimed that the features used here in more effective way for news document summarization than in the previous method [32].	Precision, Recall and F-measure values have been claimed 0.3659, 0.5064 and 0.4169 respectively.
Md. Iftekharul Alam Efat, Mohammad Ibrahim and Humayun Kayesh, 2013, [34]	i) Term frequency, ii) Sentence position, and iii) Skeleton of document	Their system is alienated into three segments as pre-processing the test document, sentence scoring and summarization based on sentences' score.	The average accuracy of this proposed method has been found 83.57% against human generated summary. The evaulation has been accomplished using only 10 documents and the evaluation result is for a particular theme only which can't be taken as standard process of evaluation.

Jagadish Kallimani, Srinivasa and Eswara Reddy, 2014, [36]	i) Parts of speech tagging, ii) Named entity recognition, and iii) Utilizing template of sentence	Input document is categorized to apply specific classes. Some attributes are extracted from the document and mapped with the template of sentence for summary sentence generation.	The system achieved an average of 86.24% precision, 78.93% recall, and 81.50% F-measure. The evaluation results claimed here only for attributes selection. There is no comparison between any model summary and system generated summary.
Kamal Sarker, 2014, [2]	i) Keyphrase extraction, ii) Sentence position, and iii) Term frequency	This is a keyphrase-based sentence extraction method for both Bangla and English text document. Two phases approach have been applied here. In the first phase, sentences are selected on the basis of top ranked keyphrases and sentences' score. Second phase is activated if summary can't be created in the first phase and select more sentences based on sentences' score.	The F-measure score has been found 0.4242 in the evaluation.

Md. Ashraf Uddin, Kazi Zakia Sultana and Md. Akibul Alam, 2014, [35]	i) Term frequency, ii) Cosine similarity among sentences, and iii) Using A* [105] searching algorithm	This is a multi-document text summarization system. A primary summary is generated at first by sentence scoring. A graph based model is then applied with the A* [105] searching algorithm on the primary summary for creating the final gist.	Unigram based Recall Score was claimed as 56%. Here, only 8 sets of documents have been selected manually for evaluation purpose which is not enough comparing to others [32, 33].

8.3.7 Comparison with existing methods based on ROUGE score

The proposed procedure along with four existing methods [2, 32–34] (discussed in 2.6) on Bangla single document text summarization has been implemented with a server side scripting language named PHP (Hypertext Preprocessor). All of the methods have been evaluated with same data set (discussed in 8.2.1) for which the results have been varied from the respective results claimed by the corresponding authors of the existing methods. Comparison results based on the average ROUGE-1 and ROUGE-2 scores have been depicted in figure 8.6 and 8.7 respectively where method 1 is presented in [32], method 2 is in [33], method 3 is in [34], and method 4 is in [2]. In the four existing methods, K numbers of top ranked sentences are selected for final summary where K is specified by user. In our implementation for all the methods, one third sentences are selected as final summary for evaluation. Again, same list of stop words [108] have been used for

implementing all the methods.

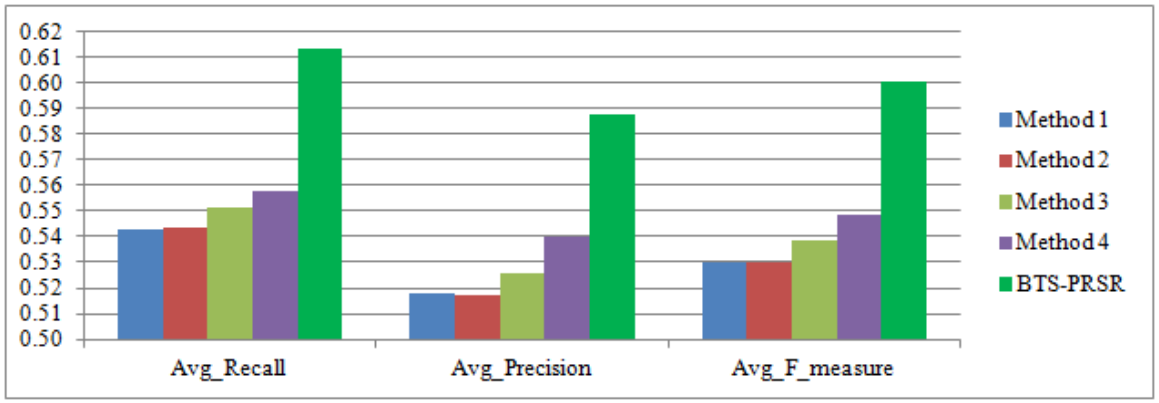


Figure 8.6: Results of comparison based on ROUGE-1 scores

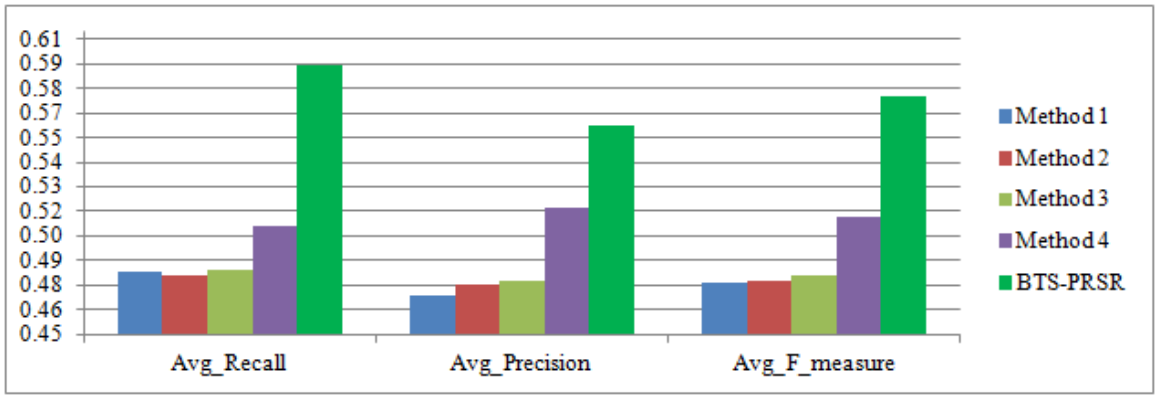


Figure 8.7: Results of comparison based on ROUGE-2 scores

In the above figures, the proposed method is indicated as BTS-PRSR (Bangla Text Summarization by introducing Pronoun Replacement and an improved version of Sentence Ranking). Point to be mentioned that the comparison results presented in figure 8.6 and 8.7 are generated with 95% confidence intervel. In the following table 8.8, improvement of the proposed method is displayed from the existing methods:

Table 8.8: Improvement of text summarization in the proposed method than the four latest existing methods

#	**Methods**	**Improvement based on ROUGE-1 score**	**Improvement based on ROUGE-2 score**
1	Method 1 [32]	13.28%	20.85%
2	Method 2 [33]	13.26%	20.45%
3	Method 3 [34]	11.48%	19.83%
4	Method 4 [2]	9.39%	12.52%

8.3.8 Number of dangling pronouns in summaries

An important aspect of the proposed method is that it has focused on the problem of dangling pronoun in summary for the first time. We have drawn an analysis about the number of dangling pronoun in the output of existing four methods along with the proposed method (BTS-PRSR) in the table 8.9. Another result is given in the table 8.10 to show the minimization rate of dangling pronoun.

Table 8.9: Number of dangling pronouns in summary

#	**Methods**	**Number of dangling pronoun in summary**
1	Method 1 [32]	80
2	Method 2 [33]	76
3	Method 3 [34]	81
4	Method 4 [2]	78
5	BTS-PRSR	8

Table 8.10: Minimization rate of dangling pronoun from the four latest existing methods

#	**Methods**	**Minimization rate of dangling pronoun**
1	Method 1 [32]	90.00%
2	Method 2 [33]	89.50%
3	Method 3 [34]	90.12%
4	Method 4 [2]	89.75%

The above results show that number of dangling pronoun can't be minimized for 100% for which 8 (eight) dangling pronouns are there in the output of the proposed method. The reason for this situation can be as follows:

i The corresponding noun of the pronoun may not be presented within the previous two sentences (explained in the section 6.4).

ii Nature of all the words could not be identified (discussed in the section 8.3.1) which can be a reason for not recognizing the corresponding noun of pronoun.

8.4 Discussion

In this proposed method, some innovative features have been introduced for getting better performance. In the figure 8.1, improvement of performance has been presented for incorporating each feature. By generating summary using only term-frequency (the first step in figure 8.1), the F-measure score has been found as 0.4124 (using the training data set). Then, after incorporating each feature the F-measure score has been raised (shown in table 8.5).

In the previous section, the average of Recall, Precision and F-measure scores of ROUGE-1 and ROUGE-2 have been shown for the proposed method. The comparison of the proposed method with the four latest Bangla text summarization methods has also been demonstrated for ROUGE-1 and ROUGE-2 scores respectively. It has been found that the proposed method outperforms all of them. The improvement of performance from the four latest existing methods has been given in the table 8.8.

The proposed method has another noteworthy feature to minimize the dangling pronoun in summary. Based on our analysis, only one dangling pronoun in the summary is enough to deliver wrong message to the user. The result in the table 8.10 shows that the proposed system has significantly minimized the number of dangling pronoun in the summary.

After all, it can be said that the proposed procedure is helpful enough for alleviating the problem of information gap for the existence of dangling pronoun. This will also be useful for the research work on Bangla language processing such as information mining, opinion mining, etc. So, the proposed method can add some value for Bangla information retrieval systems.

From the overall results, 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.

8.5 Conclusion

In this chapter, intrinsic evaluation technique has been followed to assess the performance of the proposed method. Evaluation has been done by measuring the

similarity of system generated summaries with human professionals' summaries using ROUGE evaluation package. Here, two data sets (each data set contains 200 news documents and 3 summaries of every document) have been utilized for training and evaluation of the system. The results of performance for all sub steps have been illustrated as result of word tagging in each phase, dependency parsing, pronoun replacing, etc. Subversions of the proposed method have been created to show the improvement of performance for including each feature. The comparisons have been drawn against the four latest existing methods based on the ROUGE-1 and ROUGE-2 scores where the proposed method has outperformed all. The F-measure scores for ROUGE-1 and ROUGE-2 have been found as 0.6003 and 0.5708 respectively. Moreover, the percentage of improvement has been depicted for ROUGE scores and number of dangling pronouns. The improvement from the latest existing method has been found as 9.39% and 12.52% for ROUGE-1 and ROUGE-2 F-measure scores respectively. The number of dangling pronoun has also been minimized in the summary of the proposed method from the latest existing method for 89.75%.

Chapter 9

Conclusion

9.1 Conclusion

A new approach has been illustrated here to summarize Bangla news document by introducing dangling pronoun replacement and an improved version of sentence ranking. Though there are a lot of research works for English text summarization which may not be directly applicable for Bangla because of the complexities of Bangla language in the structure of sentences, grammatical rules, inflection of words, etc. Again, the research work for Bangla is also difficult as there is hardly any automated tool to facilitate research work, no database for ontological knowledge of words and limited scope of knowledge sharing. The necessity and challenges of automatic text summarization as well as Bangla news document summarization have been explained in chapter 1. Despite of these difficulties and challenges, in this dissertation, an innovative method for summarizing Bangla news document has been introduced. A review study has been portrayed in chapter 2 to enumerate the basement of automatic Bangla text abridgement with their pros

and cons including their limitations and scope of improvement. Comparison has been turned to show the similarities and differences for the existing Bangla text summarization systems.

In the review study (in chapter 2), it has been indicated with reference that most of the incorporated features in various existing methods of Bangla have been taken from existing English text summarization procedures. But, the proposed method has some significant new features as (i) special tagging, (ii) dependency parsing, (iii) replacement of pronoun by corresponding noun, (iv) sentence frequency calculation, and (v) numerical figure identification from words. The proposed method has four steps as follows: (i) preprocessing, (ii) word tagging, (iii) replacing pronoun by corresponding noun, and (iv) sentence ranking and summary generation. All these steps have been briefly discussed in chapter 3 to show at glance view of the whole procedure. The first step of the proposed method is preprocessing of input document (explained in chapter 4) which includes (i) segmenting input document, (ii) stop words removing, and (ii) word stemming. After preprocessing, nature of each word has been identified in chapter 5 in three phases: (i) general tagging where parts of speech are tagged, (ii) special tagging where words are tagged as acronym, named entity of human, name of places, occupation, etc. and (iii) dependency parsing to verify each given tag and identify more words based on the effects of surrounding words. After these three phases of word tagging, accuracy has been found as 79.50%. It is a matter of fact that if there is any pronoun in summary sentence without the corresponding noun, the pronoun will be dangling pronoun and it will deliver a confusing message to user. To overcome this problem, replacement of pronoun by corresponding noun has been accomplished in chapter 6 with 71.80% accuracy. The proposed method has minimized the dangling pronoun

in summary for 89.75% than the latest Bangla text summarization system. It is expected that the process of replacement of pronoun will be helpful for any kind of Bangla information retrieval procedures.

For summary sentences selection, sentence ranking (discussed in chapter 7) has been accomplished by considering the followings: (i) term-frequency inverse document frequency (TF-IDF), (ii) sentence frequency (SF), (iii) numerical figure presented in words and digits, (iv) title words, and (v) the first sentence if it contains any title word. Here, sentence frequency is calculated based on the 60% or more cosine similarity between two sentences where the frequency of larger sentence is increased and the smaller sentence is removed. Again, replacement of pronoun (given in chapter 6) has boosted up the sentence scoring for TF-IDF and sentence frequency (SF) calculation. Because a pronoun can have no score but if it is replaced by noun then TF-IDF and SF scores can be calculated for this. Moreover, numerical figure has been identified from words as well as digits for 100% and the first sentence has been treated specially for better sentence ranking. In Bangla text, any numerical figure can have variety of forms than that of English which makes the numerical figure identification difficult and challenging. Finally, all the features for sentence ranking have been adjusted with tuning for better summarization performance. Chapter 8 has been dedicated for evaluation and discussion on results of the proposed method. For performance evaluation, the proposed method along with the four latest Bangla text summarization methods has been implemented using a server side scripting language named PHP (Hypertext Preprocessor). Efficiency has been measured with the ROUGE automatic evaluation tools against 200 documents and 600 model summaries (3 summaries for each document). As there is no benchmark data set for evaluating Bangla text summarization, in-house

data set (discussed in 8.2.1) has been used. To demonstrate the step by step improvement of performance for including each feature of the proposed method, seven sub-versions have been created and showed the progress. In the performance evaluation, the F-measure scores for ROUGE-1 and ROUGE-2 have been found as 0.6003 and 0.5708 respectively for the proposed method. As per the evaluation results, the proposed system outperforms four latest Bangla text summarization systems and the performance is 9.39% (based on ROUGE-1 F-measure score) and 12.52% (based on ROUGE-2 F-measure score) better than the latest existing method.

The proposed method has some limitations also as follows: (i) nature of all the words can't be identified for 100% (discussed in section 8.4.1), (ii) though the replacement of pronoun has been introduced, accuracy of dangling pronouns minimization is 71.80% and some pronouns are replaced incorrectly (explained in section 8.4.7), (iii) the words presented in different synonyms can't be treated as same word because we dont have any tool for synonym identification.

Overall, it can be said that the proposed system obtained potential outcome as per the results of evaluation not only for higher ROUGE evaluation scores but also for minimizing dangling pronouns from summary to deliver an unambiguous message. So, it is expected that the proposed system will bring serenity for human by mitigating the burden of huge volume of text and lessen the valuable time spending in getting precise information.

9.2 Future Works

We hope to consider the following tasks as future works:

- Overcome the limitations of the proposed method (discussed in the previous section) in accuracy in natura of words detection, pronoun replacement, and synonym identification.
- Multiple Bangla news documents summarization to generate a single summary for the news which has been published in different newspapers.
- Bangla single and multiple documents (other than news document) summarization where structure of texts are not so organized.

Bibliography

[1] H. Mamiko, M. Yosihiro, and S. Satoshi, "Summarizing newspaper articles using extracted informative and functional words," *Journal of NLP*, vol. 9, no. 4, pp. 55–73, 2002.

[2] K. Sarkar, "A keyphrase-based approach to text summarization for english and bengali documents," *International Journal of Technology Diffusion*, vol. 5, no. 2, pp. 28–38, April 2014.

[3] A. Hamou-Lhadj and T. Lethbridge, "Summarizing the content of large traces to facilitate the understanding of the behaviour of a software system," in *Proceedings of the 14th IEEE International Conference on Program Comprehension (ICPC)*. IEEE, 2006, pp. 181–190.

[4] E. Hovy, *Automated Text Summarization*. In: Mitkov, R. (Ed.), The Oxford Handbook of Computational Linguistics. Oxford University Press, 2005, chapter 32, pp. 583-598.

[5] 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, December 2002.

[6] K. S. Jones, "Automatic summarizing: factors and directions," *Advances in automatic text summarization*, pp. 1–12, 1999.

[7] J. K. Yogan and S. Naomie, "Automatic multi document summarization approaches," *Journal of Computer Science*, vol. 8, no. 1, pp. 133–140, 2012.

[8] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," *Journal of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, pp. 258–268, August 2010.

[9] I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim, "Summac: a text summarization evaluation," *Natural Language Engineering*, vol. 8, no. 1, pp. 43–68, March 2002.

[10] B. Cretu, Z. Chen, T. Uchimoto, and K. Miya, "Automatic summarization based on sentence extraction: A statistical approach," *International Journal of Applied Electromagnetics and Mechanics*, vol. 13, no. 1-4, pp. 19–23, 2002.

[11] Y. Shiren, T. S. Chua, M. Y. Kan, and L. Qiu, "Document concept lattice for text understanding and summarization," *Information Processing & Management*, vol. 43, no. 6, pp. 1643–1662, 2007.

[12] K. Ganesan, C. Zhai, and J. Han, "Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions," in *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, 2010, pp. 340–348.

[13] H. Jing, "Sentence reduction for automatic text summarization," in *Proceedings of the Sixth Conference on Applied Natural Language Processing*,

ser. ANLC '00. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000, pp. 310–315.

[14] Y. Ouyang, L. Wenjie, L. Sujian, and L. Qin, "Applying regression models to query-focused multi-document summarization," *Journal of Information Processing & Management*, vol. 47, no. 2, pp. 227–237, March 2011.

[15] A. Dongmei, Z. Yuchao, and Z. Dezheng, "Automatic text summarization based on latent semantic indexing," *Journal of Artificial Life and Robotics. Springer*, vol. 15, no. 1, pp. 25–29, August 2010.

[16] M. d. Kunder, "The size of the world wide web," February 2015, [Online]. Available: http://www.worldwidewebsize.com. [Accessed: 15- February-2015].

[17] R. Ferreira, L. de Souza Cabral, F. Freitas, R. D. Lins, G. de Frana Silva, S. J. Simske, and L. Favaro, "A multi-document summarization system based on statistics and linguistic treatment," *Expert Systems with Applications. Elsevier*, vol. 41, no. 13, pp. 5780–5787, October 2014.

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

[19] M. M. Haque, S. Pervin, and Z. Begum, "Literature review of automatic single document text summarization using NLP," *International Journal of Innovation and Applied Studies*, vol. 3, no. 3, pp. 857–865, July 2013.

[20] M. M. Haque, S. Pervin, and Z. Begum, "Literature review of automatic multiple documents text summarization," *International Journal of Innovation and Applied Studies*, vol. 3, no. 1, pp. 121–129, May 2013.

[21] O. M. Foong, A. Oxley, and S. Sulaiman, "Challenges and trends of automatic text summarization," *International Journal of Information and Telecommunication Technology*, vol. 1, no. 1, pp. 34–39, 2010.

[22] R. Brandow, K. Mitze, and L. F. Rau, "Automatic condensation of electronic publications by sentence selection," *Information Processing & Management*, vol. 31, no. 5, pp. 675–685, 1995.

[23] E. Hovy, and C. Lin, *Automated Text Summarization in SUMMARIST*. Cambridge: MIT Press, 1999, in Inderjeet Mani and Mark T. Maybury (editors), Advances in Automatic Text Summarization, pp. 81-94.

[24] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proceedings of the 18th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1995, pp. 68–73.

[25] M. A. Fattah, "A hybrid machine learning model for multi-document summarization," *Applied intelligence*, vol. 40, no. 4, pp. 592–600, 2014.

[26] R. Ferreira, L. D. S. Cabral, R. D. Lins, G. P. Silva, F. Freitas, G. Cavalcanti, R. Lima, S. J. Steven, and F. Luciano, "Assessing sentence scoring techniques for extractive text summarization," *Expert systems with applications*, vol. 40, no. 14, pp. 5755–5764, 2013.

[27] A. M. Azmi and S. Al-Thanyyan, "A text summarizer for arabic," *Journal of Computer Speech and Language*, vol. 26, no. 4, pp. 260–273, August 2012.

[28] C. Munir, K. Ibrahim, and H. C. Mofazzal, *Bangla Vasar Byakaran*. Ideal publication, Dhaka, November 2000.

[29] M. A. Karim, M. Kaykobad, and M. Murshed, *Technical Challenges and Design Issues in Bangla Language Processing.* Published in the United States of America by Information Science Reference (an imprint of IGI Global), June 2013.

[30] M. T. Islam and S. Masum, "Bhasa: A corpus based information retrieval and summarizer for bengali text," in *Proceedings of the 7th International Conference on Computer and Information Technology*, December 2004.

[31] M. N. Uddin and S. A. Khan, "A study on text summarization techniques and implement few of them for bangla language," in *Proceedings of the 10th International Conference on Computer and Information Technology (ICCIT-2012).* IEEE, 2007, pp. 1–4.

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

[33] K. Sarkar, "An approach to summarizing bengali news documents," in *proceedings of the International Conference on Advances in Computing, Communications and Informatics, ACM*, 2012, pp. 857–862.

[34] M. I. A. Efat, M. Ibrahim, and H. Kayesh, "Automated bangla text summarization by sentence scoring and ranking," in *Proceedings of the International Conference on Informatics, Electronics & Vision (ICIEV).* IEEE, 2013, pp. 1–5.

[35] M. U. Ashraf, K. Z. Sultana, and M. A. Alam, "A multi-document text summarization for bengali language," in *International Forum on Strategic*

Technology (IFOST). Chittagong University of Engineering & Technology (CUET), 2014.

[36] J. S. Kallimani, K. Srinivasa, and B. E. Reddy, "A comprehensive analysis of guided abstractive text summarization," *International Journal of Computer Science Issues (IJCSI)*, vol. 11, no. 6, pp. 115–121, 2014.

[37] S. Children, "History of bengali language," 2017, [Online]. Available: https://www.cs.mcgill.ca/~rwest/link-suggestion/wpcd_2008-09_augmented/wp/b/Bengali_language.htm. [Accessed: 05-May-2017].

[38] The Times of Inida, "Nearly 60% of indians speak a language other than hindi," 2017, [Online]. Available: http://timesofindia.indiatimes.com/india/Nearly-60-of-Indians-speak-a-language-other-than-Hindi/articleshow/36922157.cms. [Accessed: 05-May-2017].

[39] I. A. Laura Alonso, "Representing discourse for automatic text summarization via shallow nlp techniques," 2011, [Online]. Available: http://www.iula.upf.edu/materials/050304alonso.pdf. [Accessed: 17-March-2016].

[40] M. Juan and M. Torres, *Automatic Text Summarization*. John Wiley & Sons, November 2014, chapter 2, pp. 22-52.

[41] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[42] Indian Statistical Institute, "A lexical database for bengali," 2015, [Online]. Available: http://www.isical.ac.in/~lru/wordnetnew/index.php/site/aboutus. [Accessed: 28-October-2015].

[43] N. U. Zaman, "Big picture seminar series," 2008, [Online]. Available: http://www.cs.rochester.edu/u/naushad/survey/BigPicture-URCS-NZ-Bangla.pdf. [Accessed: 15-February-2016].

[44] E. Charniak and D. McDermott, *Introduction to Artificial Intelligence*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1985.

[45] B. Z. Manaris, "Natural language processing: A human-computer interaction perspective," *Advances in Computers*, vol. 47, pp. 1–66, December 1998.

[46] A. Gelbukh, G. Sidorov, and Y. H. Sang, "Evolutionary approach to natural language word sense disambiguation through global coherence optimization," *WSEAS Transactions on Communications*, vol. 2, no. 1, pp. 11–19, 2003.

[47] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[48] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[49] A. Feldman, "Computational linguistics: Models, resources, applications," *Comput. Linguist.*, vol. 32, no. 3, pp. 443–444, September 2006.

[50] S. Ferrández and A. Ferrández, "The negative effect of machine translation on cross—lingual question answering," in *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*, ser. CICLing '07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 494–505.

[51] E. A. Feigenbaum, "The art of artificial intelligence: I. themes and case studies of knowledge engineering," Stanford, CA, USA, Tech. Rep., 1977.

[52] Notes for students, "Rule based system," November 2000, [Online]. Available: http://www.j-paine.org/students/lectures/lect3/node5.html. [Accessed: 01-April-2017].

[53] C. Yan, "Markov process," 2008, [Online]. Available: http://digital.cs.usu.edu/~cyan/CS7960/Markov_Chains.ppt. [Accessed: 01-April-2017].

[54] M. Hazewinkel, *Markov chain.* New York, NY, USA: Encyclopedia of Mathematics, Springer, 2001.

[55] S. Fine, Y. Singer, and N. Tishby, "The hierarchical hidden markov model: Analysis and applications," *Mach. Learn.*, vol. 32, no. 1, pp. 41–62, Jul. 1998.

[56] O. Chapelle, B. Schlkopf, and A. Zien, *Semi-Supervised Learning*, 1st ed. The MIT Press, 2010.

[57] 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 the International Conference on Information Retrieval Knowledge Management.* Kuala Lumpur, Malaysia: IEEE, May 2012, pp. 193–197.

[58] M. A. Fattah and F. Ren, "GA, MR, FFNN, PNN and GMM Based Models for Automatic Text Summarization," *Comput. Speech Lang.*, vol. 23, no. 1, pp. 126–144, January 2009.

[59] Society for Natural Language Technology Research, "Bengali pos tagger," 2016, [Online]. Available: http://nltr.org/snltr-software. [Accessed: 01-April-2016].

[60] A. Ekbal and S. Bandyopadhyay, "Bengali named entity recognition using support vector machine," in *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008, pp. 51–58.

[61] H. Saggion and T. Poibeau, *Automatic Text Summarization: Past, Present and Future*, 1st ed. Springer-Verlag, Berlin, Heidelberg, July 2012, in Multi-source, Multilingual Information Extraction and Summarization, pp. 3-21.

[62] H. P. Edmundson, "New methods in automatic extracting," *Journal of the Association for Computing Machinery*, vol. 16, no. 2, pp. 264–285, April 1969.

[63] A. Nenkova and K. McKeown, *A survey of text summarization techniques*. Springer US, January 2012, in Mining Text Data.

[64] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

[65] M. A. Halliday and R. Hasan, "Cohesion in," *English*, 1976, Longman, London.

[66] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text-Interdisciplinary Journal for the Study of Discourse*, vol. 8, no. 3, pp. 243–281, 1988.

[67] D. Marcu, "Discourse trees are good indicators of importance in text," *Advances in automatic text summarization*, pp. 123–136, 1999.

[68] M. Taboada and M. Stede, "Introduction to rst (rhetorical structure theory)," 2009, [Online]. Available: http://www.sfu.ca/rst/pdfs/RST_Introduction.pdf. [Accessed: 09-June-2016].

[69] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp. 457–479, 2004.

[70] H. Jing and K. R. McKeown, "Cut and paste based text summarization," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. Association for Computational Linguistics, 2000, pp. 178–185.

[71] M. J. Witbrock and V. O. Mittal, "Ultra-summarization (poster abstract): a statistical approach to generating highly condensed non-extractive summaries," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 315–316.

[72] H. Jing and K. R. McKeown, "The decomposition of human-written summary sentences," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 129–136.

[73] R. Jin and A. G. Hauptmann, "Title generation for machine-translated documents," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 1229–1234. [Online]. Available: http://dl.acm.org/citation.cfm?id=1642194.1642258

[74] M. Banko, V. O. Mittal, and M. J. Witbrock, "Headline generation based on statistical translation," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2000, pp. 318–325.

[75] P. B. Baxendale, "Machine-made index for technical literature -an experiment," *IBM Journal of Research and Development*, vol. 2, no. 4, pp. 354–361, October 1958.

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

[77] S. Teufel and M. Moens, "Sentence extraction as a classification task," in *Proceedings of the ACL*, vol. 97, no. 1997, 1997, pp. 58–65.

[78] C. Lin, and E. Hovy, "Identifying topics by position," in *Proceedings of the 5th Conference on Applied Natural Language Processing*. Association for Computational Linguistics, 1997, pp. 283–290.

[79] C. Aone, M. E. Okurowski, J. Gorlinsky, and B. Larsen, "A scalable summarization system using robust nlp," *Journal of Intelligent Scalable Text Summarization*, pp. 66–73, 1997.

[80] J. Morris and G. Hirst, "Lexical cohesion computed by thesaural relations as an indicator of the structure of text," *Computational linguistics*, vol. 17, no. 1, pp. 21–48, 1991.

[81] R. Barzilay, "Lexical chains for summarization," Ph.D. dissertation, Citeseer, 1997.

[82] T. Zhu and X. Zhao, "An improved approach to sentence ordering for multi-document summarization," in *Proceedings of the 4th International Conference on Machine Learning and Computing*, vol. 25, 2012, pp. 29–33.

[83] X. Yong-dong, W. Xiao-long, L. Tao, and X. Zhi-ming, "Multi-document summarization based on rhetorical structure: Sentence extraction and evaluation," in *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on.* IEEE, 2007, pp. 3034–3039.

[84] P. C. Cardoso, M. L. Jorge, and A. P. Thiago, "Exploring the rhetorical structure theory for multi-document summarization," in *Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, XXXI.* Sociedad Española para el Procesamiento del Lenguaje Natural-SEPLN, 2015.

[85] J. Atkinson and R. Munoz, "Rhetorics-based multi-document summarization," *Expert Systems with Applications*, vol. 40, no. 11, pp. 4346–4352, 2013.

[86] V. R. Uzêda, T. A. S. Pardo, and M. d. G. V. Nunes, "Evaluation of automatic text summarization methods based on rhetorical structure theory," in *Intelligent Systems Design and Applications, 2008. ISDA'08. Eighth International Conference on*, vol. 2. IEEE, 2008, pp. 389–394.

[87] L. Chengcheng, "Automatic text summarization based on rhetorical structure theory," in *Computer Application and System Modeling (ICCASM), 2010 International Conference on*, vol. 13. IEEE, 2010, pp. 13–595.

[88] D. R. Radev, "A common theory of information fusion from multiple text sources step one: cross-document structure," in *Proceedings of the 1st SIGdial workshop on Discourse and dialogue-Volume 10*. Association for Computational Linguistics, 2000, pp. 74–83.

[89] I. Mani and E. Bloedorn, "Multi-document summarization by graph search and matching," in *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence*, Providence, Rhode Island, July 1997, pp. 622–628.

[90] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts." Stroudsburg, Pennsylvania: Association for Computational Linguistics, July 2004.

[91] J. Zhang, L. Sun, and Q. Zhou, "A cue-based hub-authority approach for multi-document text summarization," in *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*. IEEE, 2005, pp. 642–645.

[92] S. Hariharan, T. Ramkumar, and R. Srinivasan, "Enhanced graph based approach for multi document summarization." *Int. Arab J. Inf. Technol.*, vol. 10, no. 4, pp. 334–341, 2013.

[93] 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.

[94] T. Liao and Z. Liu, "Research of summarization extraction in multiple topics document," in *Dependable, Autonomic and Secure Computing, 2009.*

DASC'09. Eighth IEEE International Conference on. IEEE, 2009, pp. 859–860.

[95] 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.

[96] S. Rastkar, G. C. Murphy, and G. Murray, "Automatic summarization of bug reports," *IEEE Transactions on Software Engineering*, vol. 40, no. 4, pp. 366–380, 2014.

[97] Y. J. Kumar, N. Salim, A. Abuobieda, and A. T. Albaham, "Multi document summarization based on news components using fuzzy cross-document relations," *Applied Soft Computing*, vol. 21, pp. 265–279, 2014.

[98] N. Agarwal, K. Gvr, R. S. Reddy, and C. P. Rosé, "Towards multi-document summarization of scientific articles: making interesting comparisons with scisumm," in *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages.* Association for Computational Linguistics, 2011, pp. 8–15.

[99] K. McKeown and D. R. Radev, "Generating summaries of multiple news articles," in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 1995, pp. 74–82.

[100] S. R. Young and P. J. Hayes, "Automatic classification and summarization of banking telexes," in *Proceedings of CAIA*, 1985, pp. 402–409.

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

[102] S. M. Wong, W. Ziarko, and P. C. Wong, "Generalized vector spaces model in information retrieval," in *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 1985, pp. 18–25.

[103] C. Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1.* Association for Computational Linguistics, 2003, pp. 71–78.

[104] G. Kavita, "Java package for evaluation of summarization tasks with updated rouge measures," 2016, [Online]. Available: http://kavita-ganesan.com/content/rouge-2.0. [Accessed: 25-May-2016].

[105] A. Aker, T. Cohn, and R. Gaizauskas, "Multi-document summarization using a* search and discriminative training," in *Proceedings of the 2010 conference on empirical methods in natural language processing.* Association for Computational Linguistics, 2010, pp. 482–491.

[106] K. Sarkar, "Automatic single document text summarization using key concepts in documents," *Journal of information processing systems*, vol. 9, no. 4, pp. 602–620, 2013.

[107] Y. Ouyang, W. Li, and Q. Lu, "An integrated multi-document summarization approach based on word hierarchical representation," in *Proceedings of*

the ACL-IJCNLP 2009 Conference Short Papers. Association for Computational Linguistics, 2009, pp. 113–116.

[108] Indian Statistical Institute, "List of stop words for bengali language," 2016, [Online]. Available: http://www.isical.ac.in/~fire/data/stopwords_list_ben.txt. [Accessed: 18-February-2016].

[109] A. Das and S. Bandyopadhyay, "Sentiwordnet for bangla," *Knowledge Sharing Event-4: Task*, vol. 2, 2010.

[110] L. Gravano, P. G. Ipeirotis, H. V. Jagadish, N. Koudas, S. Muthukrishnan, L. Pietarinen, and D. Srivastava, "Using q-grams in a dbms for approximate string processing," *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 28–34, 2001.

[111] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell, "Summarizing text documents: sentence selection and evaluation metrics," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 121–128.

[112] M. Z. Islam, M. N. Uddin, and M. Khan, "A light weight stemmer for bengali and its use in spelling checker," in *Proceedings of the 1st Intl. Conf. on Digital Comm. and Computer Applications (DCCA 2007)*, Amman, Jordan, March 2007.

[113] M. R. Mahmud, M. Afrin, M. A. Razzaque, E. Miller, and J. Iwashige, "A rule based bengali stemmer," in *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*. IEEE, 2014, pp. 2750–2756.

[114] A. Ekbal, R. Haque, and S. Bandyopadhyay, "Bengali part of speech tagging using conditional random field," in *Proceedings of Seventh International Symposium on Natural Language Processing (SNLP2007)*, 2007, pp. 131–136.

[115] Gpedia, "Gpedia, your encyclopedia," 2016, [Online]. Available: http://www.gpedia.com/bn. [Accessed: 25-June-2016].

[116] BdJobs.com, "Occupation in bangladesh, name of occupation in largest job site in bangladesh," February 2016, [Online]. Available: http://bdjobs.com. [Accessed: 25-June-2016].

[117] G. M. Kiron, *Ajker Bishaw*. Premier publications, Dhaka, January 2014, general Knowledge, Bangladesh and International Affairs, Edition - 68.

[118] Z. R. Siddiqui, *English-Bangla Dictionary*. Bangla Academy, Dhaka, June 2011, second edition, Published by Shahida Khatun.

[119] IndiaChildNames.com, "Indian child names," 2016, [Online]. Available: http://www.indiachildnames.com/regional/bengalinames.aspx. [Accessed: 25-June-2016].

[120] Bangladeshpost, "Bangladesh post office online," 2016, [Online]. Available: http://www.bangladeshpost.gov.bd/postcode.asp. [Accessed: 10-February-2016].

[121] D. Chen and C. D. Manning, "A fast and accurate dependency parser using neural networks." in *EMNLP*, 2014, pp. 740–750.

[122] Write.com, "Direct and indirect objects," 2017, [Online]. Available: http://www.write.com/writing-guides/general-writing/grammar/direct-and-indirect-objects. [Accessed: 21-March-2017].

[123] E. Filatova and V. Hatzivassiloglou, "Event-based extractive summarization," in *Proceedings of ACL Workshop on Summarization*, vol. 111. Barcelona, Spain., 2004.

[124] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[125] Bangla Natural Language Processing Community, "Dataset for evaluating bangla text summarization system," 2016, [Online]. Available: http://bnlpc.org/research.php. [Accessed: 08-August-2016].

[126] R. Ferreira, F. Freitas, L. d. S. Cabral, R. D. Lins, R. Lima, G. França, 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)-Volume 01.* IEEE Computer Society, 2013, pp. 389–396.

[127] I. Mani and M. T. Maybury, *Advances in automatic text summarization.* MIT Press, 1999, vol. 293.

[128] D. R. Radev, H. Jing, and M. Budzikowska, "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user

studies," in *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*. Association for Computational Linguistics, 2000, pp. 21–30.

Appendix A

List of Acronyms

ATS	Automatic Text Summarization
BLP	Bangla Language Processing
BTS	Bangla Text Summarization
BTS-PRSR	Bangla Text Summarization by Introducing Pronoun Replacement & an improved version of Sentence Ranking
CST	Cross-document Structure Theory
Dc	Cosine Distance
DUC	Document Understanding Conference
HC	Highlighted Content
IDF	Inverse Document Frequency
IR	Information Retrieval
ISI	Indian Statistical Institute
KEA	Keyword Extraction Algorithm
MDTS	Multiple Documents Text Summarization
MMR	Maximum Marginal Relevance
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
PDA	Personal Digital Assistant

PHP	Hypertext Preprocessor
QA	Question Answering
ROUGE	Recall Oriented Understudy for Gisting Evaluation
RST	Rhetorical Structural Theory
SC	Summary Content
SMS	Short Message Service
TF	Term Frequency
WSD	Word Sense Disambiguation

Appendix B

List of Stopwords

The following list of Bangla stop words has been collected from the Indian Statistical Institute [108]:

অবশ্য	ওদের	তাঁদের	পারে	রকম
অনেক	ওঁদের	তাঁহারা	পরে	শুধু
অনেকে	ওখানে	তাঁরা	পরেই	সঙ্গে
অনেকেই	কত	তাঁর	পরেও	সঙ্গেও
অন্তত	কবে	তাঁকে	পর	সমস্ত
অথবা	করতে	তাই	পেয়ে	সব
অথচ	কয়েক	তেমন	প্রতি	সবার
অর্থাত	কয়েকটি	তাকে	প্রভৃতি	সহ
অন্য	করবে	তাহা	প্রায়	সুতরাং
আজ	করলেন	তাহাতে	ফের	সহিত
আছে	করার	তাহার	ফলে	সেই
আপনার	কারও	তাদের	ফিরে	সেটা
আপনি	করা	তারপর	ব্যবহার	সেটি
আবার	করি	তারা	বলতে	সেটাই
আমরা	করিয়ে	তারৈ	বললেন	সেটাও
আমাকে	করার	তার	বলেছেন	সম্প্রতি

আমাদের	করাই	তাহলে	বলল	সেখান
আমার	করলে	তিনি	বলা	সেখানে
আমি	করলেন	তা	বলেন	সে
আরও	করিতে	তাও	বলে	স্পষ্ট
আর	করিয়া	তাতে	বহু	স্বয়ং
আগে	করেছিলেন	তো	বসে	হইতে
আগেই	করছে	তত	বার	হইবে
আই	করছেন	তুমি	বা	হৈলে
অতএব	করেছেন	তোমার	বিনা	হইয়া
আগামী	করেছে	তথা	বরং	হচ্ছে
অবধি	করেন	থাকে	বদলে	হত
অনুযায়ী	করবেন	থাকা	বাদে	হতে
আদ্যভাগে	করায়	থাকায়	বার	হতেই
এই	করে	থেকে	বিশেষ	হবে
একই	করেই	থেকেও	বিভিন্ন	হবেন
একে	কাছ	থাকবে	বিষয়টি	হয়েছিল
একটি	কাছে	থাকেন	ব্যবহার	হয়েছে
এখন	কাজে	থাকবেন	ব্যাপারে	হয়েছেন
এখনও	কারণ	থেকেই	ভাবে	হয়ে
এখানে	কিছু	দিকে	ভাবেই	হয়নি
এখানেই	কিছুই	দিতে	মধ্যে	হয়
এটি	কিন্তু	দিয়ে	মধ্যেই	হয়েই
এটা	কিংবা	দিয়েছে	মধ্যেও	হয়তো
এটাই	কি	দিয়েছেন	মধ্যভাগে	হল
এতটাই	কী	দিলেন	মাধ্যমে	হলে

এবং	কেউ	দু	মাত্র	হলেই
একবার	কেউই	দুটি	মতো	হলেও
এবার	কাউকে	দুটো	মতোই	হলো
এদের	কেন	দেয়	মোটেই	হিসাবে
এঁদের	কে	দেওয়া	যখন	হওয়া
এমন	কোনও	দেওয়ার	যদি	হওয়ার
এমনকী	কোনো	দেখা	যদিও	হওয়ায়
এল	কোন	দেখে	যাবে	হন
এর	কখনও	দেখতে	যায়	হোক
এরা	ক্ষেত্রে	দ্বারা	যাকে	জন
এঁরা	খুব	ধরে	যাওয়া	জনকে
এস	গুলি	ধরা	যাওয়ার	জনের
এত	গিয়ে	নয়	যত	জানতে
এতে	গিয়েছে	নানা	যতটা	জানায়
এসে	গেছে	না	যা	জানিয়ে
একে	গেল	নাকি	যার	জানানো
এ	গেলে	নাগাদ	যারা	জানিয়েছে
ঐ	গোটা	নিতে	যাঁর	জন্য
ই	চলে	নিজে	যাঁরা	জন্যও
ইহা	ছাড়া	নিজেই	যাদের	বেশ
ইত্যাদি	ছাড়াও	নিজের	যান	দেন
উনি	ছিলেন	নিজেদের	যাচ্ছে	তুলে
উপর	ছিল	নিয়ে	যেতে	ছিলেন
উপরে	জন্য	নেওয়া	যাতে	চান
উচিত	জানা	নেওয়ার	যেন	চায়

ও	ঠিক	নেই	যেমন	চেয়ে
ওই	তিনি	নাই	যেখানে	মোট
ওর	তিনিই	পক্ষে	যিনি	যথেষ্ট
ওরা	তিনিও	পর্যন্ত	যে	টি
ওঁর	তখন	পাওয়া	রেখে	এসব
ওঁরা	তবে	পারেন	রাখা	
ওকে	তবু	পারি	রয়েছে	

Appendix C

Examples of Generated Summaries

Example 1:

Title of input document: বাজেটই দেশের আর্থ-সামাজিক উন্নয়নের ভিত্তি

Text of input document: বাজেটেই প্রতিফলিত হয় দেশের আর্থ-সামাজিক উন্নয়নের ভিত্তি।দারিদ্র্য দূরীকরণের জন্য জাতীয় বাজেটে যদি পর্যাপ্ত বাজেট রাখা না হয়, তাহলে দেশের সার্বিক উন্নয়ন আশা করা যায় না। বুধবার (২৪ জুন) রাজাধানীর জাতীয় প্রেসক্লাবে "বাজেট পরবর্তী অতি দারিদ্র্য নিরসন: জাতীয় বাজেটের ভূমিকা" শীর্ষক মতবিনিময় সভায় কৃষি ব্যাংকের সাবেক চেয়ারম্যান খোন্দকার ইব্রাহিম খালেদ এসব কথা বলেন।মতবিনিময় সভাটি আয়োজন করে "উন্নয়ন সমন্বয়" নামের একটি সংগঠন।তিনি বলেন, বর্তমান সরকার ভালো অবস্থানে রয়েছে।সারা দেশের উন্নয়ন ঘটাতে এ সরকারের সক্ষমতা রয়েছে।আলোচিত পদ্মা সেতু বিষয়ে ইব্রাহিম খালেদ জানান, এ সরকারই দেখিয়ে দিলো, বিশ্বব্যাংকের টাকা ছাড়াও পদ্মা সেতুর কাজ শুরু করা যায়।বায়াত্তর সালে ধনী-দরিদ্রের মধ্যে বেশি ফারাক ছিল না।এখন ধনী-দরিদ্রের মধ্যে বিশাল ফারাক দেখা দিয়েছে বলেও মন্তব্য করেন তিনি।মতবিনিময় সভায় ঢাকায় বিশ্ববিদ্যালয়ের শিক্ষক, সমাজকল্যাণ মন্ত্রণালয়ের স্থায়ী কমিটির সদস্যসহ সংগঠনটির নেতাকর্মীরা উপস্থিত ছিলেন।

Human expert' summary: দারিদ্র্য দূরীকরণের জন্য জাতীয় বাজেটে যদি পর্যাপ্ত বাজেট রাখা না হয়, তাহলে দেশের সার্বিক উন্নয়ন আশা করা যায় না।বুধবার (২৪ জুন) রাজাধানীর জাতীয় প্রেসক্লাবে "বাজেট পরবর্তী অতি দারিদ্র্য নিরসন: জাতীয় বাজেটের ভূমিকা" শীর্ষক মতবিনিময় সভায় কৃষি ব্যাংকের সাবেক চেয়ারম্যান খোন্দকার ইব্রাহিম খালেদ এসব কথা বলেন।মতবিনিময় সভাটি আয়োজন করে "উন্নয়ন সমন্বয়" নামের একটি সংগঠন।

System generated summary: বাজেটেই প্রতিফলিত হয় দেশের আর্থ-সামাজিক উন্নয়নের

ভিত্তি। দারিদ্র্য দূরীকরণের জন্য জাতীয় বাজেটে যদি পর্যাপ্ত বাজেট রাখা না হয়, তাহলে দেশের সার্বিক উন্নয়ন আশা করা যায় না।বুধবার (২৪ জুন) রাজাধানীর জাতীয় প্রেসক্লাবে "বাজেট পরবর্তী অতি দারিদ্র্য নিরসন: জাতীয় বাজেটের ভূমিকা" শীর্ষক মতবিনিময় সভায় কৃষি ব্যাংকের সাবেক চেয়ারম্যান খোন্দকার ইব্রাহিম খালেদ এসব কথা বলেন।

Example 2:

Title of input document: লেনদেন আবারও ৫০০ কোটির ঘরে

Text of input document: দেশের প্রধান শেয়ারবাজার ঢাকা স্টক এক্সচেঞ্জে (ডিএসই) লেনদেন আবারও ৫০০ কোটির ঘরে নেমে এসেছে।সপ্তাহের প্রথম কার্যদিবসে গতকাল রোববার ডিএসইতে লেনদেনের পরিমাণ ছিল প্রায় ৫১৩ কোটি টাকা।অন্য শেয়ারবাজার চট্টগ্রাম স্টক এক্সচেঞ্জের (সিএসই) লেনদেন ৫০ কোটি থেকে কমে গতকাল দিন শেষে নেমে এসেছে ৩৮ কোটিতে।লেনদেন কমার পাশাপাশি দুই বাজারে সপ্তাহের প্রথম কার্যদিবসে সূচকও কমেছে। বাজার-সংশ্লিষ্ট ব্যক্তি ও প্রতিষ্ঠানের কর্মকর্তাদের মতে, গতকালের বাজারে বিনিয়োগকারীদের মধ্যে মুনাফা তুলে নেওয়ার প্রবণতা বেশি ছিল।এ কারণে ক্রয়াদেশের চেয়ে বিক্রয়াদেশ ছিল বেশি।ফলে সূচক কমেছে।মার্চেন্ট ব্যাংক আইডিএলসি ইনভেস্টমেন্টসের পর্যালোচনা প্রতিবেদন অনুযায়ী, গত কয়েক দিনে নির্দিষ্ট কিছু কোম্পানির শেয়ারের দাম বেড়েছে।সেসব শেয়ার বিক্রি করেই মূলত মুনাফা তুলে নিয়েছেন বিনিয়োগকারীরা।গত কয়েক দিনের বাজার পরিস্থিতি পর্যালোচনা করে দেখা গেছে, শেয়ারবাজারে তালিকাভুক্ত কম মূলধনী কিছু কোম্পানির শেয়ারের দাম বেশ বেড়েছে।নতুন তালিকাভুক্ত ইনফরমেশন টেকনোলজি কনসালট্যান্টসের (আইটিসি) ১০ টাকার প্রতিটি শেয়ারের বাজারমূল্য মাত্র ছয় কার্যদিবসে ৭০ টাকা ছাড়িয়েছে।

Human expert' summary: দেশের প্রধান শেয়ারবাজার ঢাকা স্টক এক্সচেঞ্জে (ডিএসই) লেনদেন আবারও ৫০০ কোটির ঘরে নেমে এসেছে। গত কয়েক দিনের বাজার পরিস্থিতি পর্যালোচনা করে দেখা গেছে, শেয়ারবাজারে তালিকাভুক্ত কম মূলধনী কিছু কোম্পানির শেয়ারের দাম বেশ বেড়েছে। নতুন তালিকাভুক্ত ইনফরমেশন টেকনোলজি কনসালট্যান্টসের (আইটিসি) ১০ টাকার প্রতিটি শেয়ারের বাজারমূল্য মাত্র ছয় কার্যদিবসে ৭০ টাকা ছাড়িয়েছে।

System generated summary: দেশের প্রধান শেয়ারবাজার ঢাকা স্টক এক্সচেঞ্জে (ডিএসই) লেনদেন আবারও ৫০০ কোটির ঘরে নেমে এসেছে।অন্য শেয়ারবাজার চট্টগ্রাম স্টক এক্সচেঞ্জের (সিএসই) লেনদেন ৫০ কোটি থেকে কমে গতকাল দিন শেষে নেমে এসেছে ৩৮ কোটিতে।নতুন তালিকাভুক্ত ইনফরমেশন টেকনোলজি কনসালট্যান্টসের (আইটিসি) ১০ টাকার প্রতিটি শেয়ারের বাজারমূল্য মাত্র ছয় কার্যদিবসে ৭০ টাকা ছাড়িয়েছে।

Appendix D

List of Publications

International Journal Papers

1. M. M. Haque, S. Pervin, and Z. Begum, "An Innovative Approach of Bangla Text Summarization by Introducing Pronoun Replacement and Improved Sentence Ranking," *International Journal of Information Processing Systems (JIPS)*, vol. 13, no. 4, pp. 752-777, August 2017. *(Indexed in SCI, SCOPUS, DBLP, etc.).*

2. M. M. Haque, S. Pervin, and Z. Begum, "Enhancement of Keyphrase-Based Approach for Automatic Bangla Single Document Text Summarization," *Under second round review in the International Journal of Asian and Low-Resource Language Information Processing (TALLIP)*, May 2017. *(Indexed in SCI, SCOPUS, DBLP, etc.).*

3. M. M. Haque, S. Pervin, and Z. Begum, "Rule based replacement of pronoun by corresponding noun for bangla news documents," *International Journal of Technology Diusion (IJTD)*, vol. 8, no. 2, pp. 26-42, April 2017. *(Indexed in ACM Digital Library, DBLP, etc.).*

4. M. M. Haque, S. Pervin, and Z. Begum, "Literature review of automatic single document text summarization using NLP," *International Journal of*

Innovation and Applied Studies, vol. 3, no. 3, pp. 857-865, July 2013. *(Indexed in ResearchGate, Google Scholar, etc.).*

5. M. M. Haque, S. Pervin, and Z. Begum, "Literature review of automatic multiple documents text summarization," International *Journal of Innovation and Applied Studies*, vol. 3, no. 1, pp. 121-129, May 2013. *(Indexed in ResearchGate, Google Scholar, etc.).*

International Conference Papers

1. M. M. Haque, S. Pervin, and Z. Begum, "Enhancement of keyphrase-based approach of automatic bangla text summarization," in *Proceedings of the Region 10 Conference (TENCON)*. IEEE, 2016, pp. 42-46.

2. M. M. Haque, S. Pervin, and Z. Begum, "Automatic bengali news documents summarization by introducing sentence frequency and clustering," in *Proceedings of the 18th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2015, pp. 156-160.