International Journal of Technology Diffusion
Volume 8 • Issue 2 • April-June 2017

Rule Based Replacement of Pronoun by Corresponding Noun for Bangla News Documents

Md. Majharul Haque, University of Dhaka, Dhaka, Bangladesh

Suraiya Pervin, University of Dhaka, Dhaka, Bangladesh

Zerina Begum, University of Dhaka, Dhaka, Bangladesh

ABSTRACT

The object of this research work is to replace pronoun by corresponding noun for Bangla news documents. To the best of our knowledge, this is the first initiative to solve the problem of dangling pronoun where corresponding noun is not available. If the information retrieval procedures extract any sentence with dangling pronoun, it may raise confusion to the user. To mitigate this problem, a method has been proposed here by using general and special tagging, dependency parsing, full name identifying and finally pronoun replacing. For achieving the target of this method, 3000 Bangla news documents have been analyzed and some grammar books have been studied. Seven knowledgeable persons in the arena of Bangla language also helped us in this research work. Finally, the proposed method shows 71.80% accuracy in the evaluation for replacing pronoun.

KEYWORDS

Bangla News Documents, Corresponding Noun, Dangling Pronoun, Dependency Parsing, Information Retrieval

INTRODUCTION

Bangla is the 7th most spoken language in the world from 3500 languages and around 250 millions of people are using Bangla (Chowdhury, Khalil, & Mofazzal, 2000). It is the mother language of Bangladesh (Islam, 2003) and the second most spoken language in India (Olivet, 2015). Based on the economic survey - 2015, there are 62.30% literate people in Bangladesh where most of them are used to Bangla language only. Nowadays, many computerized contents are being developed in Bangla and online version of Bangla newspaper is also growing rapidly. So, electronic version of Bangla text is increasing without any bounds in the cyber world and people are overloaded with huge volume of texts. To alleviate this burden of large volume of text, very few research works have been conducted for Bangla (Islam & Masum, 2004; Sarkar, 2014; Haque, Pervin, & Begum, 2016). In this situation, the Bangla-speaking people have been deprived from the advantage of information technology. So, for this large community of Bangla-speaking people, more research work is very much necessary especially for Bangla information retrieval. But research works for Bangla language is difficult for the following issues:

DOI: 10.4018/IJTD.2017040102

1. Based on our study, automatic procedures are hardly available for Bangla language to facilitate research work.
2. For Bangla language, there is no lexical database like WordNet (Miller, 1995). Though a similar tool is ongoing to be developed it has limited features (Indian Statistical Institute, 2015).
3. There is no database of ontological meanings for Bangla words that can be used programmatically.
4. 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.

Some other problems have been discussed in (Karim, Kaykobad, & Murshed, 2013; Zaman, 2015) about the research work on Bangla. Even, the scope of knowledge sharing is also limited as there are a few researchers in this arena. Despite these difficulties, a method has been presented here which focuses upon a problem in the output of Bangla information retrieval procedures.

In the output of information retrieval, some sentences may be available with dangling pronoun(s) where the corresponding noun is missing. These pronouns will make the information incoherent. So, the systems, which 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.

The objective of this research work is to make the output of Bangla information retrieval procedures free from dangling pronoun. Otherwise, there is a huge probability to misunderstand the text by user because only a single dangling pronoun is enough to deliver wrong message. In these circumstances, a method has been proposed here to resolve this problem with the following major contributions:

1. Detecting the nature of each word in the sentence of Bangla news document as noun, pronoun, verb, subject, object, numerical figure, acronym, name of people and places, etc. In this regard, words are tagged in two phases as general and special tagging.
2. Dependency parsing of words to verify the nature of each word because a word may have dynamic nature for the effect of surrounding words. Some untagged words are also tagged using dependency parsing.
3. Locating pronouns and distinguishing each of them as subject or object.
4. Identifying the corresponding noun of the pronoun and replacing the pronoun in suitable format.

Based on the authors' findings, this is the first initiative for the replacement of pronoun in Bangla. For English language, one system has been developed by Stanford NLP (Stanford CoreNLP, 2015) for noun-pronoun matching but this can't be used for Bangla as the structure of Bangla sentence is much different from English (Chowdhury et al., 2000).

To accomplish this research work, the authors have scrutinized 3000 Bangla news documents (news documents of around one month from the Daily Prothom-alo which is the most popular newspaper of Bangladesh). Seven knowledgeable persons of Bangla language, who have completed four years' graduation on Bangla language (their mother tongue is Bangla and they read Bangla newspapers regularly), helped in this research work. After a detail discussion with those persons regarding the structure of sentences of Bangla language and analysis of news documents, some rules have been utilized here. Based on these rules, special tagging, dependency parsing, subject and object recognizing and after all pronoun replacing have been accomplished._

PROPOSED METHOD

In this proposed method, eight forms of pronouns are considered for replacement by corresponding noun as follows: 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). Only singular forms of pronouns that can be used for human are considered for replacement. So, the corresponding noun should also be singular human named entity. Based on the authors’ analysis, the corresponding noun of any pronoun is existed as subject or object of the immediate previous sentence or the second immediate previous sentence for 88.63% times. The details explanation of the proposed method has been given as follows:

Preprocessing

Document Segmentation

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.

Word Stemming

Word stemming is a procedure by which words with different endings will be mapped to a single word. For example, “খেলছে” (khelchhe-playing), “খেলেছে” (khelechhe-played) will be “খেলা” (khela-play) and “করেছে” (korechhe-did), “করছেন” (korchhen-doing) will be “করা” (kora-do). In this step, the light weight stemmer (Islam, Uddin, & Khan, 2007) has been incorporated.

General Tagging

All the words are tried to tag as noun, pronoun, adjective, verb, preposition, conjunction and interjection in this step by using a lexicon database (Society for Natural Language Technology Research [SNLTR], 2015) and SentiWordNet (Das & Bandyopadhyay, 2010). Here, 65.13% words are tagged as per the experiment with 200 test documents because lexicon database (SNLTR, 2015) and SentiWordNet (Das & Bandyopadhyay, 2010) have limited number of predefined words.

The words (especially verb) in Bangla language are very much inflectional (Sarkar, 2014). 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 previous step) here to identify root form of word, 100% inflectional forms of verb can’t be stemmed (Islam et al., 2007). In reality, the identification of verb is quite difficult because the verb may have a lot of suffixes in Bangla. For example, 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 is 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), "বলতছে" (boltechho - saying), "বলতছেনে" (boltechhen - saying), "বলছলি" (bolchhilo - saying), "বলছলিনে" (bolchhilen - saying), "বলছলিে" (bolchhile - saying), "বলছলিি" (bolchhili - saying), "বলছলিাম" (bolchhilam - saying), "বলছেলি" (bolechhilo - saying), "বলছেলিনে" (bolechhilen - saying), বলছেলিে (bolechhile - saying), "বলছেলিি" (bolechhili - saying), "বলছেলিাম" (bolechhilam - saying), "বলবে" (bolbe - say), "বলবনে" (bolben - say), "বলবি" (bolbi - say), "বলব" (bolbo - say), "বল◌ো" (bolo - say) (Chowdhury et al., 2000; Mamud, 2011). So the complexity of verb recognition in Bangla can't be compared with English.

But, identifying verb is very important for language processing task as verb is the chief word for any sentence (Mamud, 2011). In this regard, if there is any word left untagged after using lexicon database (SNLTR, 2015) and SentiWordNet (Das & Bandyopadhyay, 2010), we need to check the word if it is verb or not. Finally, a list of suffixes (Chowdhury et al., 2000) are taken into account for ultimate checking such as "ইতছেি" (itechhi), "তছেলিনে" (techhilen), "লনে" (len), "সনে" (sen), etc. Now, if the considered word has any of these suffixes (Chowdhury et al., 2000), it is tagged as verb.

After using the list of suffixes, the percentage of words tagging has been increased from 65.13% (result of word tagging before considering the list of suffixes (Chowdhury et al., 2000)) to 66.73%. The tagging in this step is a preliminary tagging and some tags may be updated in the next steps.

Special Tagging

It is remarkable here that subject and object identification is necessary to replace pronoun because the corresponding noun of replaceable pronoun is subject or object of the previous sentence. It is well known that word is the principal ingredient of a language (Mamud, 2011) and hence it is difficult to detect subject and object of any sentence without recognizing the nature of words. A procedure is available for Bangla parts-of-speech tagging (Ekbal, Haque, & Bandyopadhyay, 2008a) but there is no procedure 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:

Checking for English Acronym

In Bangla news documents, there may have acronym that means the word is consists of some English letters that are written in Bangla. For example: "ইউএনডপিি" (UNDP), "আইএলও" (ILO), etc. For checking this type of words, all the English letters are written in Bangla such as: "এ" (A), "বি" (B), "সি" (C) …. …. …. "ডব্লউি" (W), "এক্স" (X), "ওয়াই" (Y), "জডে" (Z) and sorting them in descending order based on their string length where "ডব্লউি" (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: "ইউএনডপিি" (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 English acronym or not. Experiment shows 100% success rate for detecting English acronym.

Checking for Bangla Initial

As like English acronym mentioned in the point (1), there can be Bangla letters with spaces such as "আ স ম" (A S M). These letters will be tagged as Bangla initial. Based on experiment the correctness of finding initial is 100%.

Checking for Repeated Words

In Bangla language, same words can be written for two times (Chowdhury et al., 2000). Such as "ঠান্ডা ঠান্ডা" (thanda thanda - cold cold). 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 (Mamud, 2011) where some irregular words are mentioned as repeated words such as "দনো পাওনা" (dena paona – payable receivable). If any word is matched with these words or repeated for two times (fully or partially), they are tagged as repeated words. In the most of the cases, such words are used as adjective and placed before noun, pronoun or verb (Chowdhury et al., 2000). We have applied this technique on 200 news documents and found 98% accuracy on identifying repeated words.

Checking for Numerical Figure

For recognizing numerical figure presented in words and digits, three conditions are checked as follows:

1. 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), "তনি" (tin - three) to "নরিানব্বই" (niranobboi – ninety-nine). While checking numerical figure from digits, decimal point (.) is also considered.
2. In the next part (if any) it has the followings: "শত" (shoto - hundred), "হাজার" (hazar - thousand), etc.
3. 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 number.

We have experimented on 200 test documents and observed that 100% numerical figure can be identified from both digits and words.

Checking for Occupation

There is a table with 80 entries (collected from (Gpedia, 2016; BdJobs, 2016)) 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), "শকি্ষামন্ত্রী" (shikkhamontri – Education minister) and so on. In this way, if any word has suffix from the list of occupation or fully matched with the listed occupation, the word is tagged as occupation. It has been observed that the proposed system can identify occupation for 91% times.

Checking for the Name of Organization

It has been detected from our analysis that name of an organization can be mentioned as follows:

1. The full name of organization is given which follows the acronym of the name enclosed in parentheses. For example, "দুর্নীতি দমন কমশিন (দুদক)" - "Durniti Domon Commission (DUDOK) – Anti Corruption Commission (ACC)".
2. The last part of the organization name may have some specific words such as "লমিটিডে" (limited - limited), "বশি্ববদি্যালয়" (bishawbiddaloy - university), "মন্ত্রণালয়" (montronaloy - ministry), "কে◌াং" (kong - kong), etc (Ekbal, Haque, & Bandyopadhyay, 2008b).

If there is any acronym according to the point (1), enclosed with parentheses, count the number of letters in the acronym and then same number of words (immediately before the acronym) are tagged as a name of organization. Experiment 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 Bangla Academy dictionary (Siddiqui, 2011) and a book of general knowledge (Kiron, 2014).

According to the point (2), 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 is 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 point (ii).

Checking for Probable Human Named Entity

A data file has been used here for first name, middle name and last name with 7500 entries where most of them are collected from (Indian child names, 2015). 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 first, middle or last name. Part of name is identified in this point for more than 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).

Checking for the Name of Place

A table has been maintained with 700 entries for the list of division, district, upozila and municipality as name of places of Bangladesh (Bangladesh Post Office, 2016). Here division is the first 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 (Kiron, 2014). 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.

From the 31525 words (from 200 test documents), 5.80% untagged words are identified in special tagging which raise the words tagging from 66.73% (result of general tagging) to 72.53%. Some experimental results on the special tagging process are given in Table 1.

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.

Dependency Parsing of Each Word

The nature of words in sentences can be varied due to the effect of surrounding words. So, it may need to update the tag of any word which is accomplished in general and special tagging process. 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 as follows:

Table 1. Experimental results of special tagging using 200 test documents

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

1. List of adjectives has been collected from (Chowdhury et al., 2000) and fully repeated words (mentioned in the special tagging process) are treated as adjective. Adjectives are placed as neighboring words of noun or verb (Chowdhury et al., 2000). If any adjective (not tagged as repeated word in special tagging) has any suffix, it is treated as noun. There may have 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.
3. Some words are generally placed before adjective for example "অপকে্ষা" (opekkha - than), "চয়ে" (cheye - than), "অধকি" (odhik - more), etc. (Chowdhury et al., 2000).
4. List of words are used as prefix of another words in Bangla language (Mamud, 2011). Individually, this list of prefix has no meaning but it can change the meaning of other words. The words with these prefixes are generally noun or adjective (Chowdhury et al., 2000; Mamud, 2011). 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.
5. The word which is presented after adjective is noun or verb. If the adjacent words of adjective have suffix as "ইতছেি" (itechhi), "তছেলিনে" (techhilen), "লনে" (len), etc. (Chowdhury et al., 2000), it will be treated as verb otherwise noun (Mamud, 2011).
6. If the previous word has been tagged in the special tagging as occupation, word with article (except occupation with article), repeated words or numerical figure, it can't be verb.
7. Some words are there as verb like "কর" (kor - do), "দয়ে" (dey - give), "যায়" (zay - go), etc. These words may have suffix as "ইতছেি" (itechhi), "তছেলিনে" (techhilen), "লনে" (len), "সনে" (sen), etc. (Chowdhury et al., 2000).
8. A name is presented before verb where the name is treated as subject if this is not an adjective.
9. There is a list of article as "টি" (ti - this), "টা" (ta - this), etc. which can be placed as suffix with noun or pronoun (Mamud, 2011). The list of pronoun is collected from (Chowdhury et al., 2000; Mamud, 2011). So, if any word (except pronoun) has articles, this will be considered as noun.
10. There can be article along with number, 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 number with article, occupation with article, etc. if they contain article.
11. The word "গ◌োটা" (gota - whole) can be placed before numerical figure and "খানা" (khana - this), "খানি" (khani - this) can be placed after numerical figure (Chowdhury et al., 2000).
12. 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 (Mamud, 2011). Here, direct object is material and indirect object is personal.
13. 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.
14. List of words are there as preposition/conjunction/interjection and they are treated as "অব্যয়" (Obboy) in Bangla language. They are tagged as stop words. A list of 363 stop words has been collected from (Indian Statistical Institute, 2016) for Bangla language. These words can't have other tagging and have no dependency on surrounding words (Chowdhury et al., 2000).
15. The word that is presented immediately before the words "দ্বারা" (dara - with), "দয়িা" (diya - with), etc. is a noun which is object (Chowdhury et al., 2000). 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).
16. The word which is placed after "দ্বারা" (dara - with), "দয়িা" (diya - with), "দয়িে" (diye - with), etc. is verb.
17. General structures of sentence 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 (Chowdhury et al., 2000; Mamud, 2011). We may identify subject and object by following these structures.

18. If there is a noun with suffix “র” (r), “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.
19. The words “ওহে” (ohe - hi), “হে” (he - hi) follows a human named entity.
20. The suffixes “কার” (kar) and “করে” (ker) are placed with the word which indicates time.
21. 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.
22. 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.
23. The word immediately before “সমাহার” (somahar - combination) is a noun where the previous word of the noun is a numerical figure.
24. 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 (Chowdhury et al., 2000).
25. 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.

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. See Table 2.

Finding the Full Name of Human for the Identification of Subject and Object

In case of general and special tagging, all the tags are depended on some lists of words. But it is apparent that whatever the range of lists will be, there is a limitation. Specially, some words have been tagged as human name in the step of special tagging but there may be some other named entities available in the entire input text. Some words might be wrongly tagged as named entity. 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 thoroughly to get the human named entity properly which is subject or object.

Table 2. Experimental results of word tagging from 31525 words of 200 documents

Word tagging in different phases	Number of tagged words	Percentage of words
Tagging by list of words from (SNLTR, 2015; Das & Bandyopadhyay, 2010)	20532	65.13%
Tagging after utilizing list of suffixes for verb	21038	66.73%
After special tagging	22865	72.53%
After dependency parsing	25062	79.50%

It is noticeable that the existing technique for named entity recognition (Ekbal et al., 2008b) has not been utilized here. Because, primarily selected named entity may be ignored based on the impact of surrounding words which is very significant feature but not available in (Ekbal et al., 2008b). Some more words can be named entity in the document that cannot be indicated based on the predefined lists of words as like (Ekbal et al., 2008b). In these circumstances, the existing technique (Ekbal et al., 2008b) 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, 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. Part of name may be there at the first time of a news document if the name has already come 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 functions. 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 entity where full name is existed as discussed in the earlier of this step. Multiple words are checked at a time in this step for getting all parts of name such as first, last and middle name 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 (Chowdhury et al., 2000; Mamud, 2011) to get named entity (full name) 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 full name has three parts usually (the first, last and middle name). From these four words, take the words as named entity that are tagged as the first name, middle name, last name, noun or any untagged word (at least one of the words should be tagged as part of name based on the step of special tagging).
2. If there is any 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) (Indian child names, 2015). From these three words, take the words as named entity those are tagged as name, noun, Bangla initial or any untagged word. But, if no other words are there with the considered word to form the full name, ignore the word.
3. If there is a comma (punctuation mark) followed by a word with verb tag, there may have 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 sentence for collecting a named entity as like the first two points of this step.
4. 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 first two points 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.
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.

7. There is named entity immediately after the word “নাম” (nam - name) and immediately before the word “নামে” (name - named) or “নামের” (namer – name’). Experiment shows that this rule is correct for 98% scenarios.
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 also considered. If the previous word is number, 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. (Chowdhury et al., 2000). 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 entity will be removed from the list of collected names.
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 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 similar 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.

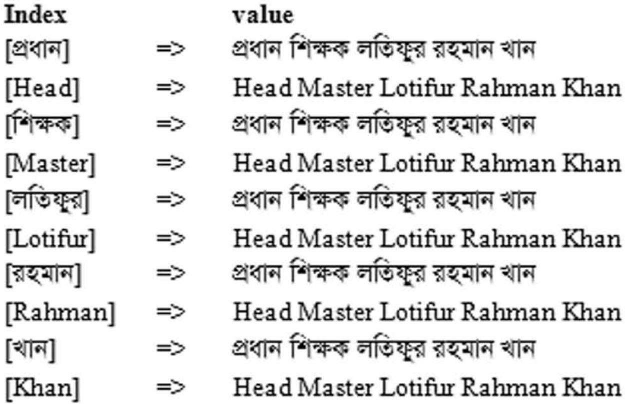
After finding all the named entities, a simple and well organized mechanism has been incorporated here to keep them easily accessible. An associative array has been maintained which means that the index of the array will be word. For example, if a named entity is presented as “প্রধান শিক্ষক লতিফুর রহমান খান” (Head Master Lotifur Rahman Khan), it will be placed in the array for five times based on the parts of name as in Figure 1.

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” (Head Master Lotifur Rahman Khan).

Replacing Pronoun

Though recognizing name of human (singular subject and object) is an important step, some other things are also indispensable for replacing pronoun. Because, from the identified named entities,

Figure 1. Structure of associative array for keeping named entities



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

subject and object needs to be detected. In this step, some rules are applied for recognizing subject and object of each sentence and to distinguish the corresponding noun of a pronoun 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.
2. Some special cases are there in the sentence of Bangla language where pronoun is available but it is not dangling pronoun. In these cases, pronouns are kept as it is. 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 replaceable pronoun. It has been found in the authors’ 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:
    a. Generally, it has been found that named entity with the following suffixes are object: “কে” (ke), “রে” (re), “এর” (er), etc. (Chowdhury et al., 2000).
    b. 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.
    c. 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.
    d. 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.
    e. 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.
    f. 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.
    g. 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 used as subject of any sentence.
    h. For ensuring replacement of pronoun in suitable format, the following things 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.

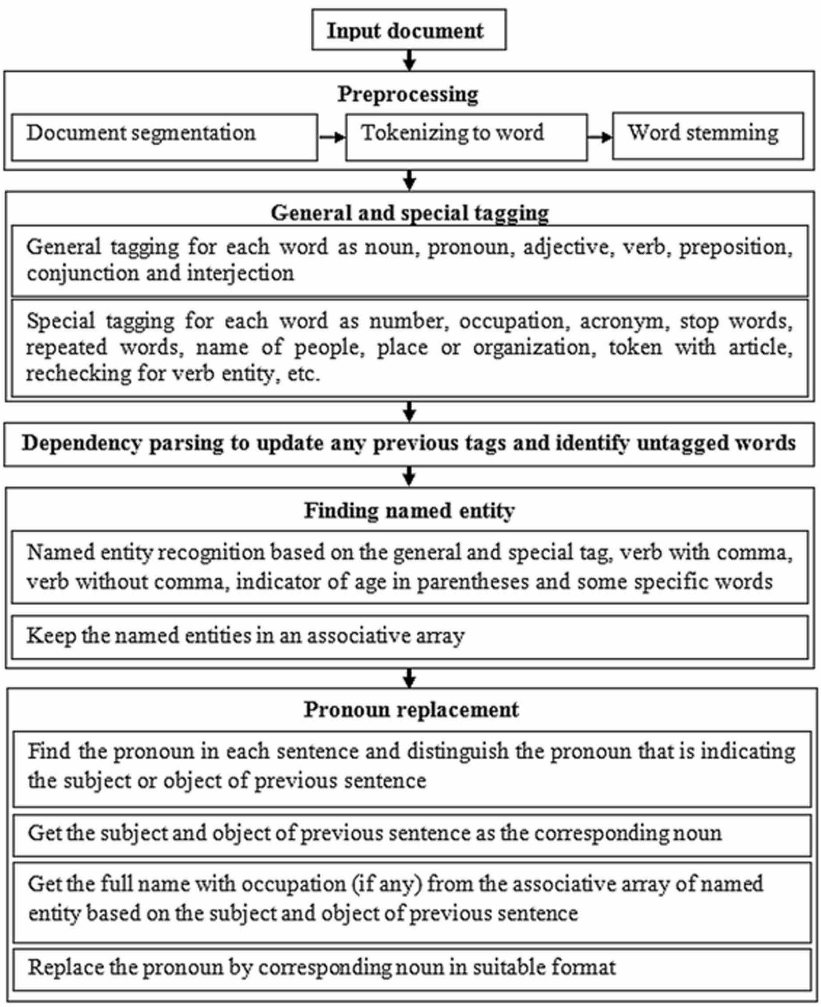
In our proposed technique (Figure 2), named entity may be existed as only one word where it is difficult to determine whether it is really a named entity or not. To overcome this situation, all the named entities of input document have been kept in an associative array in word by word (discussed in the previous step). So, if the system needs to consider a single word is named entity or not, it 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 (Islam et al., 2007). 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.

EVALUATION AND DISCUSSION ON RESULTS

Based on the authors' findings, there is no benchmark dataset to evaluate this method. For training and evaluation, 3200 Bangla news documents (with 15 to 25 lines of text each) have been collected from

Figure 2. Process flow of the proposed method


Input document
Preprocessing
Document segmentation
Tokenizing to word
Word stemming
General and special tagging
General tagging for each word as noun, pronoun, adjective, verb, preposition, conjunction and interjection
Special tagging for each word as number, occupation, acronym, stop words, repeated words, name of people, place or organization, token with article, rechecking for verb entity, etc.
Dependency parsing to update any previous tags and identify untagged words
Finding named entity
Named entity recognition based on the general and special tag, verb with comma, verb without comma, indicator of age in parentheses and some specific words
Keep the named entities in an associative array
Pronoun replacement
Find the pronoun in each sentence and distinguish the pronoun that is indicating the subject or object of previous sentence
Get the subject and object of previous sentence as the corresponding noun
Get the full name with occupation (if any) from the associative array of named entity based on the subject and object of previous sentence
Replace the pronoun by corresponding noun in suitable format

the most popular Bangladeshi newspaper. Randomly selected 3000 documents have been analyzed to train the system and other 200 documents have utilized for performance evaluation. From these 200 documents, number of singular pronoun is counted. Now, input these documents to the system and the followings are counted for performance measurement: (i) correctly replaced pronouns, (ii) how many pronouns have been kept without replacing, and (iii) incorrectly replaced pronouns. The numbers of pronouns, results of finding named entities and results of replacement of pronoun have been given in the Table 3, Table 4 and Table 5, respectively for 200 news documents.

The results in Table 5 illustrates that the system can replace 183 pronouns correctly from 255 pronouns in total, which implies the accuracy of the system is 71.80%. To the best of authors' knowledge, there is no existing system for pronoun replacement with which this proposed method can be compared.

A sample text has been given in the Figure 3 to illustrate the output of our proposed technique 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 replacing pronoun, it has been replaced correctly for the first two times by corresponding noun "আব্দুল করিম সাহেব" (Mr. Abdul Karim). For the third times, "তিনি" (tini - he) has not been replaced because the corresponding noun was not found within the two immediate previous sentences.

Table 3. Number of singular pronoun counting from 200 Bangladeshi news documents

#	Singular Pronoun	Frequency
1	তিনি (tini - he)	160
2	তার (tar – his/her)	48
3	সে (she - he)	33
4	তাকে (take - him)	14
5	ইনি (ini - he)	0
6	উনি (uni - he)	0
7	তাহার (tahar – his/her)	0
8	তাহাকে (tahake - him)	0
Total Number of Pronouns		255

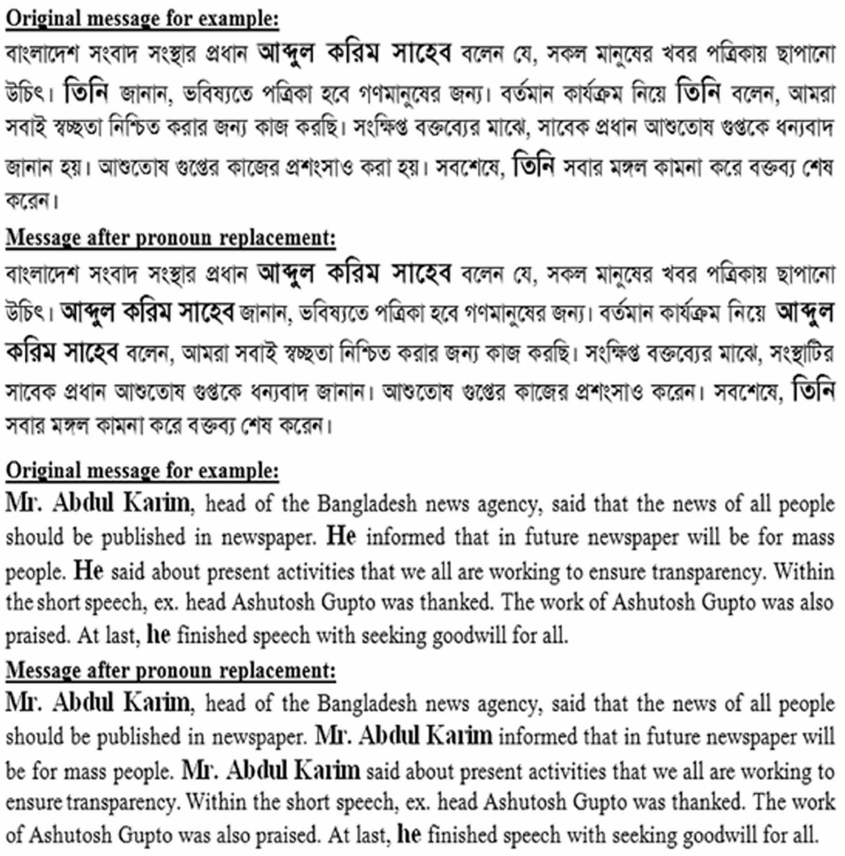
Table 4. Success rate of finding named entities, subject and object

#	Activities	Success rate
1	Part of name identification	80%
2	Full name identification	76.50%
3	Recall the full name from the part of the name	74.50%
4	Categorize the named entities as subject and object	73%

Table 5. Result on pronoun replacement for 200 Bangladeshi news documents

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

Figure 3. Sample text for the example of replacement of pronoun



Original message for example:

বাংলাদেশ সংবাদ সংস্থার প্রধান **আব্দুল করিম সাহেব** বলেন যে, সকল মানুষের খবর পত্রিকায় ছাপানো উচিৎ। **তিনি** জানান, ভবিষ্যতে পত্রিকা হবে গণমানুষের জন্য। বর্তমান কার্যক্রম নিয়ে **তিনি** বলেন, আমরা সবাই স্বচ্ছতা নিশ্চিত করার জন্য কাজ করছি। সংক্ষিপ্ত বক্তব্যের মাঝে, সাবেক প্রধান আশুতোষ গুপ্তকে ধন্যবাদ জানান হয়। আশুতোষ গুপ্তের কাজের প্রশংসাও করা হয়। সবশেষে, **তিনি** সবার মঙ্গল কামনা করে বক্তব্য শেষ করেন।

Message after pronoun replacement:

বাংলাদেশ সংবাদ সংস্থার প্রধান **আব্দুল করিম সাহেব** বলেন যে, সকল মানুষের খবর পত্রিকায় ছাপানো উচিৎ। **আব্দুল করিম সাহেব** জানান, ভবিষ্যতে পত্রিকা হবে গণমানুষের জন্য। বর্তমান কার্যক্রম নিয়ে **আব্দুল করিম সাহেব** বলেন, আমরা সবাই স্বচ্ছতা নিশ্চিত করার জন্য কাজ করছি। সংক্ষিপ্ত বক্তব্যের মাঝে, সংস্থাটির সাবেক প্রধান আশুতোষ গুপ্তকে ধন্যবাদ জানান। আশুতোষ গুপ্তের কাজের প্রশংসাও করেন। সবশেষে, **তিনি** সবার মঙ্গল কামনা করে বক্তব্য শেষ করেন।

Original message for example:

Mr. Abdul Karim, head of the Bangladesh news agency, said that the news of all people should be published in newspaper. **He** informed that in future newspaper will be for mass people. **He** said about present activities that we all are working to ensure transparency. Within the short speech, ex. head Ashutosh Gupto was thanked. The work of Ashutosh Gupto was also praised. At last, **he** finished speech with seeking goodwill for all.

Message after pronoun replacement:

Mr. Abdul Karim, head of the Bangladesh news agency, said that the news of all people should be published in newspaper. **Mr. Abdul Karim** informed that in future newspaper will be for mass people. **Mr. Abdul Karim** said about present activities that we all are working to ensure transparency. Within the short speech, ex. head Ashutosh Gupto was thanked. The work of Ashutosh Gupto was also praised. At last, **he** finished speech with seeking goodwill for all.

After all, it can be said that the system 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, automatic text summarization, opinion mining, etc. So, the proposed method can add some value for Bangla information retrieval systems.

CONCLUSION

A method has been proposed here to resolve the problem of dangling pronoun in the output of Bangla information retrieval procedures. For establishing this technique, general and special tagging of each word have been accomplished. In general tagging, 65.13% words are tagged using lexicon database and SentiWordNet. After that, by considering the inflection of verb, a list of suffixes is utilized to identify more verbs and 66.73% words have been tagged. Then, special tagging has been introduced and 72.53% words are identified. Here, the general and special tagging are static but the words can have dynamic nature due to the affect of surrounding words. To overcome this problem, dependency parsing has been accomplished to verify any given tag and identify more untagged words by which total 79.50% words have been recognized. Finally, corresponding noun of any dangling pronoun has been detected and replace the pronoun appropriately in suitable format. The system has been trained and evaluated using 3000 and 200 Bangla news documents respectively. In the performance evaluation, the proposed method shows 71.80% accuracy. We believe that 71.80% accuracy can be acceptable as this is the first initiative for the replacement of dangling pronoun in Bangla. It is also expected that the process for the detection of nature of words will also be helpful for other types of research work on Bangla language processing.

In this research work, only singular pronouns are being replaced and some pronouns are replaced incorrectly. There is also some dependency upon the perspective of Bangladesh for using the list of local names of people and places. In future, we hope to extend this method so that it will be more robust and generalized.

REFERENCES

Bangladesh Post Office. (2016). *Post office of Bangladesh*. Retrieved http://www.bangladeshpost.gov.bd/postcode.asp

BdJobs. (2016). *Name of occupation in largest job site in Bangladesh*. Retrieved from http://bdjobs.com

Chowdhury, M., Khalil, I., & Mofazzal, H. C. (2000). Bangla Vasar Byakaran. Dhaka: Ideal publication.

Das, A., & Bandyopadhyay, S. (2010). SentiWordNet for Bangla. In Knowledge Sharing Event-4: Task 2: Building Electronic Dictionary. Mysore.

Ekbal, A., Haque, R., & Bandyopadhyay, S. (2008a). Maximum Entropy Based Bengali Part of Speech Tagging. *Advances in Natural Language Processing and Applications Research in Computing Science*, *33*, 67–78.

Ekbal, A., Haque, R., & Bandyopadhyay, S. (2008b). Named Entity Recognition in Bengali: A Conditional Random Field Approach. *Proceedings of the International Joint Conference on Natural Language Processing*.

Gpedia. (2016). *Your Encyclopedia*. Retrieved from http://www.gpedia.com/bn

Haque, M. M., Pervin, S., & Begum, Z. (2016). Enhancement of Keyphrase-Based Approach of Automatic Bangla Text Summarization. Proceedings of the International IEEE TENCON conference, Singapore (pp. 42-46).

Indian child names. (2015). *Bengla Names*. Retrieved from http://www.indiachildnames.com/regional/bengalinames.aspx

Indian Statistical Institute. (2015). *Bengali WordNet, A Lexical Database for Bengali*. Retrieved from http://www.isical.ac.in/~lru/wordnetnew/index.php/site/aboutus

Indian Statistical Institute. (2016). *List of stop words for Bengali language*. Retrieved from http://www.isical.ac.in/~fire/data/stopwords/

Islam, M. T., & Masum, S. M. A. (2004). Bhasa: A Corpus-Based Information Retrieval and Summariser for Bengali Text. *Proceedings of the 7th International Conference on Computer and Information Technology*.

Islam, M. Z., Uddin, M. N., & Khan, M. (2007). *A light weight stemmer for Bengali and its Use in spelling Checker. Center for research on Bangla language processing*. CRBLP.

Islam, S. (Ed.). (2003). *Banglapedia, The national Encyclopedia of Bangladesh*. Dhaka: Asiatic Society of Bangladesh.

Karim, M. A., Kaykobad, M., & Murshed, M. (2013). *Technical Challenges and Design Issues in Bangla Language Processing*. United States of America: Information Science Reference (an imprint of IGI Global). doi:10.4018/978-1-4666-3970-6

Kiron, G. M. (2014). *Ajker Bishaw, General Knowledge, Bangladesh and International Affairs* (68th ed.). Dhaka: Premier publications.

Mamud, D. H. (2011). *Vasa Shikkha, Bangla Vasar Byakaran O Rachanariti*. Dhaka: The Atlas Publishing House.

Miller, G. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, *38*(11), 39–41. doi:10.1145/219717.219748

Olivet. (2015). Second most spoken languages around the world. Retrieved from http://graduate.olivet.edu/news-events/news/second-most-spoken-languages-around-world

Sarkar, K. (2012). An approach to summarizing Bengali news documents. *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* ACM (pp. 857-862). doi:10.1145/2345396.2345535

Sarkar, K. (2014). A Keyphrase-Based Approach to Text Summarization for English and Bengali Documents. *International Journal of Technology Diffusion*, *5*(2), 28–38. doi:10.4018/ijtd.2014040103

Siddiqui, Z. R. (Ed.). (2011). *English-Bangla Dictionary. Bangla Academy* (2nd ed.). Shahida Khatun.

Society for natural language technology research. (2015). *Bengali POS Tagger*. Retrieved September 13, 2015, from http://nltr.org/snltr-software

Stanford CoreNLP. (2015). *Co-reference*. Retrieved from http://nlp.stanford.edu:8080/corenlp/process

United Nations. (2015). *International Mother Language Day, Inclusive Education through and with Language - Language Matters*. Retrieved from http://www.un.org/en/events/motherlanguageday

Zaman, N. U. (2008). *Big Picture Seminar Series*. Retrieved from http://www.cs.rochester.edu/u/naushad/survey/BigPicture-URCS-NZ-Bangla.pdf