Joint 2019 8th International Conference on Informatics, Electronics & Vision (ICIEV) &
3rd International Conference on Imaging, Vision & Pattern Recognition (IVPR)

Bengali Stop Phrase Detection Mechanism using Corpus Based Method

Rakib Ul Haque
Dept. of CSE
University of Asia Pacific
Dhaka, Bangladesh
rakibulhaqueraj@gmail.com

Parisa Mehera
Dept. of CSE
University of Asia Pacific
Dhaka, Bangladesh
parisamehera@gmail.com

M. F. Mridha
Dept. of CSE
University of Asia Pacific
Dhaka, Bangladesh
firoz@uap-bd.edu

Md. Abdul Hamid
Dept. of CSE
University of Asia Pacific
Dhaka, Bangladesh
ahamid@uap-bd.edu

***Abstract*—This paper discusses a corpus-based method for the detection of the stop phrase. These phrases must be detected and eliminated during NLP in the quest for attaining efficient indexing in modern Information Retrieval (IR) systems. A complete set of stop phrases for the Bengali language has not been developed yet. In this paper, a corpus-based approach is introduced for recognizing and extracting Bengali stop phrases. This proposed technique indicates that an input paragraph will be tokenized in several required manners and after that identification of stop phrases will be obtained by checking through the corpus. Accepted stop phrases will be sent for uniqueness. Outcomes of this proposed approach for stop phrases detection are notable where accuracy, precision, and recall results are observable. Eliminating these stop phrases will further reduce the time complexity of those algorithms, which were used in case of text summarizing and IR system.**

***Index Terms*—Bengali Language Processing (BLP); stop phrase; Natural Language Processing (NLP); NLTK; Corpus Based**

I. Introduction

In NLP, Stop phrases are indicated as frequent phrases for a language which have a slight influence on the actual meaning of a particular language. They are very alike to stop words [1] but here instead of an explicit word it indicates to choose an entire phrase of a block. These phrases are not important for Bengali text summarizing and also for information retrieval system. These phrases must be detected and eliminated during NLP in the quest for attaining efficient indexing in modern Information Retrieval (IR) systems. For example, at the time of writing a Bengali book of fairy tales for kids, the phrase 'once upon a time' (একদা এক সময়) appears frequently. This phrase is a stop phrase and this should be added to the list of stop phrases to stop them from being indexed by search engines. Example:

আমাকে লন্ডন হোটেল দেখান।
(Show me London hotels.)

Here, "show me" (আমাকে দেখান) is a stop phrase. Bengali Stop Phrases are found in Bengali Novels, Bengali fairy-tales, Bengali storybooks etc. A Bengali fairy-tale book is a hypothetical scenario and there is no work done with a book of fairy tales in real life. However, in Bengali business documents there are some regularly recurrent phrases which are better off gridlocked. Expressions like, 'the price of this article' (এই নিবন্ধনের মূল্য), 'consumer opinion of this product' (এই পণ্যের ভোক্তা মতামত), etc which occur often in the document could be added to the list of stop phrases. There is no exhaustive list of stop phrases in the Bengali language. Only Google has developed a list stop phrase for English language [2]. We have developed a list of Bengali stop phrase for this research. Sometimes Bengali stop phrases do not behave like stop phrase but used as content phrases, which might be meaningful for that sentence. That kind of sentences are like the following:

আমাকে টাকা দেখান
(Show me money.)

This phrase could be considered a stop phrase but it might be referred to a film, in which this Bengali phrase could be used frequently.

A. Types of Stop Phrase

There is no Introduction and types of stop phrase are available. In Bengali stop phrases can be of different types. Since there is no proper classification of stop phrase, a new sentence (বাক্য) wise classification of stop phrases are proposed and classified below. There are three types of sentences according to their structure [3]. They are,

- Stop phrase in Simple sentence (সরল বাক)
- Stop phrase in Complex sentence (জটিল বাক্য)
- Stop phrase in Compound Sentence (যৌগিক বাক্য)

Fig. 1. shows a diagram for the classification of Stop phrases. In the following, Table I there are some examples of Bengali stop phrases for different kind of sentences (বাক্য) structure wise.

978-1-7281-0788-2/19/1.00 ©2019 IEEE

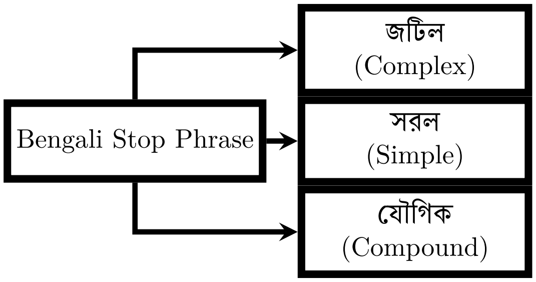

Fig. 1. Classification of Bengali Stop phrases (Sentence structure wise).

TABLE I
STOP PHRASES WITH EXAMPLE

Sentence (বাক্য)	**Example** (উদাহরণ)
Simple (সরল)	Cox's Bazaar is the largest sea-beach in the world. (কক্সবাজার বিশ্বের বৃহত্তম সমুদ্র সৈকত।)
Complex (জটিল)	Though he lives in America, he fluently speaks Bengali. (যদিও তিনি আমেরিকাতে থাকেন, তবুও তিনি অনর্গল বাংলা বলেন।)
Compound (যৌগিক)	Mou shouted, and all the people started to clap. (মৌ চিল্লাচ্ছিল, এবং সবাই হাততালি দিচ্ছিল।)

Phrases containing underline in Table-I are indicated as stop phrases. Again, there are five categories of stop phrases in sentences according to their meaning and functions [3]. These are:

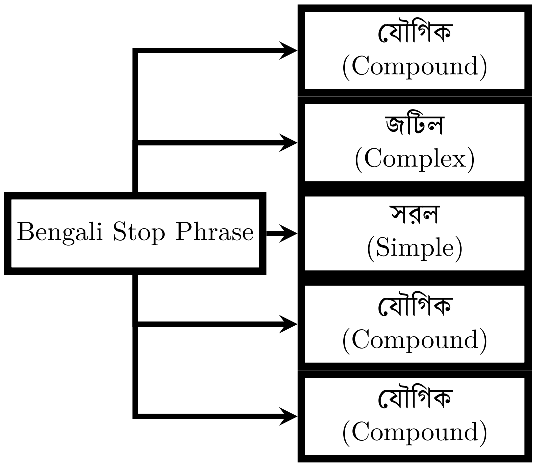

Fig. 2. Classification of Bengali Stop-phrases (Sentence meaning and function wise).

- Stop phrase in Assertive sentence (বর্ণনামূলক বাক্য)
- Stop phrase in Interrogative Sentence (প্রশ্নবোধক বাক্য)
- Stop phrase in Imperative Sentence (অনুজ্ঞাসূচক বাক্য)
- Stop phrase in Optative Sentence (প্রার্থনাসূচক বাক্য)
- Stop phrase in Exclamatory Sentence (বিস্ময়সূচক বাক্য)

Fig. 2. is the diagram for the classification of Stop phrases. In the following TABLE - II there are some examples of Bengali stop phrases for different kind of sentences (বাক্য) meaning and function wise.

TABLE II
STOP PHRASES WITH EXAMPLE

Sentence (বাক্য)	**Example** (উদাহরণ)
Assertive (বর্ণনামূলক)	Everyone could read this poem. (সবাই এই কবিতা পড়তে পারে।)
Interrogative (প্রশ্নবোধক)	Do you need some money? (আপনার কী কিছু টাকা প্রয়োজন?)
Imperative (অনুজ্ঞাসূচক)	Please, give me a pen. (দয়া করে আমাকে একটি কলম দিন।)
Optative (প্রার্থনাসূচক)	May the patient come round early. (তারাতারি যেন রোগীটি সুস্থ হয়)
Exclamatory (বিস্ময়সূচক)	Hurrah! We have won the game. (হুররে! আমরা এই খেলা জিতেছি।)

Phrases containing underline in Table-II are indicated as stop phrases. Rest of this paper is organized as follows. In Section II there will be brief discussion about some works related to the proposal in this paper. Proposed model will be explained in Section III. Performance evaluation and result analysis is carried out in Section IV. Finally, we conclude our paper in Section V.

II. Related Literature

To remove stop word and stop phrase Google, at first classify the stop word and stop phrase form the corpus list then decide to execute the search on that queries eliminating or without eliminating the stop-phrases. If the outcomes or the results are similar, then stop word and stop phrases elimination may not affect the search. If the resulted lists are not similar, then stop word and stop phrases elimination may affect the search. So they should not be eliminated [2]. To make a list of stop word [4] a completely new tactic, which is known as term-based indiscriminate testing which has been presented depending on Kullback - Leibler divergence measurement. This method determines how informative a term is and hence permits us to create a stop word list. In [5] author argue that removing corpus-specific stop words after representation of boundary, which might be much clear and the outcome is more alike. In [6] authors tried to identify the possibility and risk caused by interdisciplinary ways appropriate to automatic stop words list training. This technique fits in the period of pragmatic representations. In [7], authors have proposed a semantic way to automatically recognize and eliminate stop words based upon the data from Twitter. In [8], authors have explained an approach in order to distinguish copied ways in terms of text groups is obtainable. Authors showed that stop word 'n-grams' uncover significant data for piracy exposure. They are capable to seize synthetic and syntactic comparisons based on doubtful and also real data. In [9] authors suggested a machinated combined approach depends on the mathematical and informational model. The aim is from Chinese language identification and detection of stop-word and forming a list. For [10] Arabic language authors planned and apply an effective stop word eliminating method depends on (FSM) Finite State Machine. In [11] author's main goal is to identify the significance of different elemental compression methods in SVM (Support Vector Machines). Matching the increment of performance at the time of applying to the standard REUTERS-21578 benchmark. In [12] authors investigate the effect of the different word such as OR, AND for connecting sentence wise relationship. Developing a conceptual program to recognize synonymous connection out of hyponymic connection in tags including multiple words. In [13] Hindi words like noun containing data set were practically created by doing the measurement. A Hindi Word Net developed at IIT Bombay used to take the definition of the recognition.

From the above discussions, it is clear that there is no work done on stop phrase detection in any language except English.

III. Implementation

A. Environment and Data-set

NLTK (Natural Language Toolkit) is the built-in library used for Bengali language handling. We have used Jupyter notebook from Anaconda. Environment and size of the data set are listed below.

TABLE III
Environment & Data-set-size

Library	NLTK
Language	Python
Development Environment	Jupyter Note-book
Stop phrase No.	554
No. of Text Files	150

B. Algorithm & Flow-chart

Fig 3 is the primary process for the detection of the stop phrase. In Fig. 3, Bengali input texts are sent for the detection process. The outcomes of the detection process are sent for uniqueness.

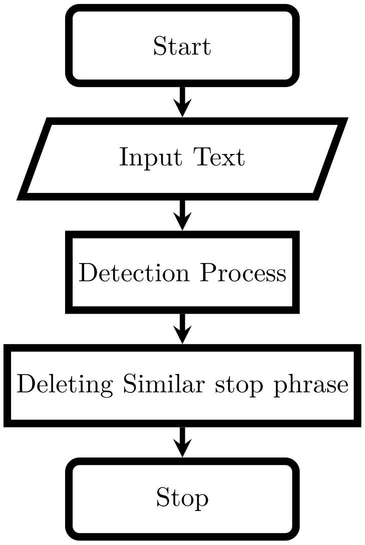

Fig. 3. Stop phrase Detection Process.

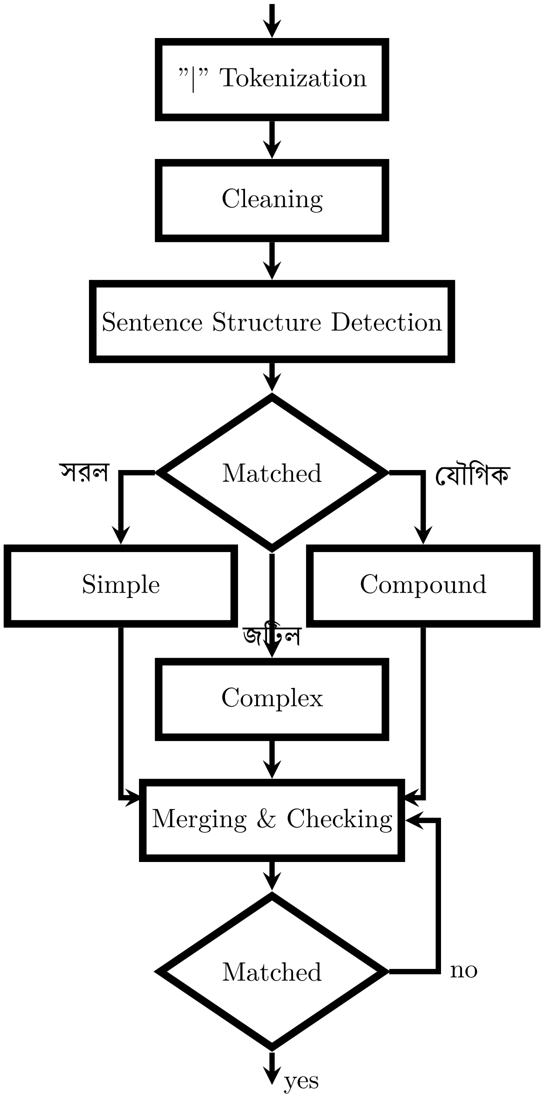

Fig. 4. Flow chart of the detection process for stop phrase (sentence structure wise).

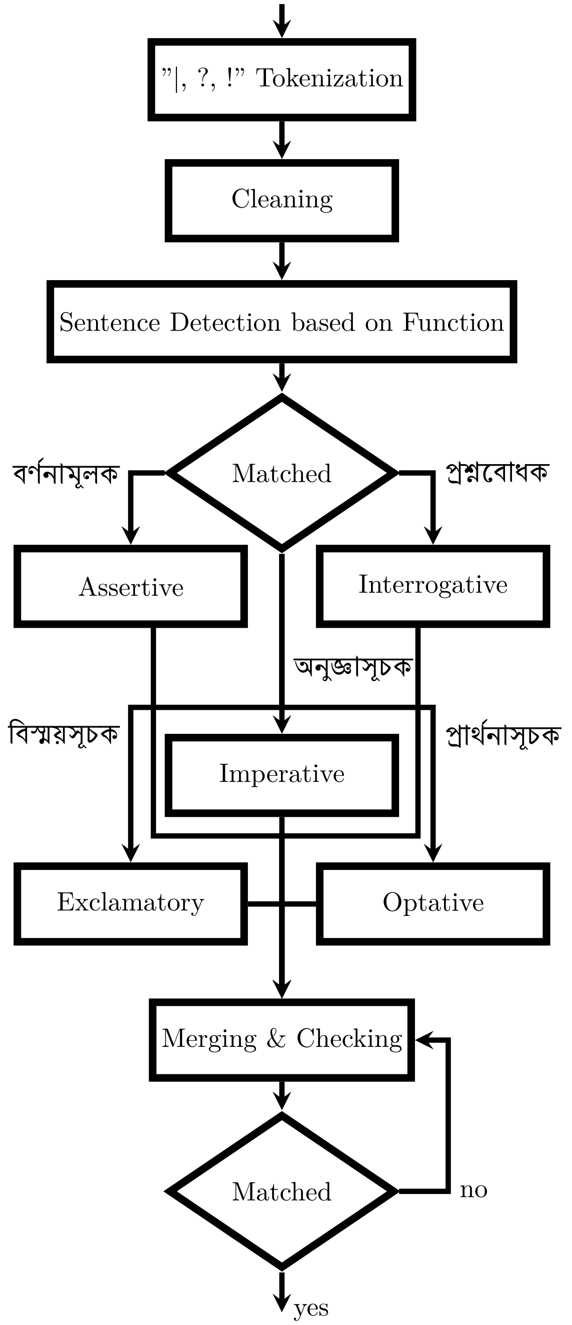

Fig. 5. Flow chart of the detection process for stop phrase (sentence meaning and function wise).

Two proposed schemes are explained to detect stop phrase. They are

1) Corpus-based approach depending on sentence structure wise. (Fig. 4)
2) Corpus-based approach depending on sentence meaning and function wise. (Fig. 5)

Fig. 4 shows the detection process of stop phrase based on sentence structure wise categorization. A complex sentence is detected, when a sentence has যদিও, যদি, যেহেতু (Though, if, since) etc. A Compound sentence is detected, when a sentence has এবং, কিন্তু, অথবা (and, but, or) etc. Except for the compound and complex sentence, all other sentences are detected as a simple sentence. Tokenization is done after the detection of sentences. From the list of tokenized words, two words are combined to create a phrase and checked with the corpus. If matches that phrase is sent for uniqueness. Example

Though he lives in America,he fluently speaks Bengali.
(যদিও তিনি আমেরিকাতে থাকেন, তবুও তিনি অনর্গল বাংলা বলেন।)

This sentence is recognized as a complex sentence. Here, "Though he", "he fluently" "যদিও তিনি", "তবুও তিনি" phrases are recognized as a stop phrase. After removing them, there is no change occurred in the meaning of the sentence.

Fig. 5 shows the detection process of stop phrase based on sentence meaning and function wise categorization. Interrogative and exclamatory sentences are detected when any sentences have ('?','!') in them. An imperative sentence doesn't have a subject. If a sentence has ("may", "wish") etc in them, that sentence is detected as an optative sentence. Sentences which don't belong to the above-mentioned category are detected as Assertive sentences. Tokenization is done after the detection of sentences. From the list of tokenized words, two words are combined to create a phrase and checked with the corpus. If matches that phrase is sent for uniqueness. Example

Hurrah! We have won the game.
(হুররে! আমরা এই খেলা জিতেছি।)

This sentence is recognized as a exclamatory sentence. Here, "Hurrah! We""হুররে! আমরা" phrases is recognized as a stop phrase. After removing it, there is no change occurred in the meaning of the sentence.

Following algorithm is for stop phrase removal. Only the special case where stop phrase act as a main phrase is ignored in our system.

Data: Bengali text with stop phrase
Result: stop phrase list of that document

```
Procedure check( list_of_words)
    for each word in list_of_words do
        phrase =current_word+' '+next_word;
        value = call stop-phrase_function (phrase1);
        if value equals to 1 then
            send(phrase1);
        else
            Ignored;
    end
    return;
```

Algorithm 1: Corpus based algorithm for stop phrase

IV. Result analysis

Finally, we measure the performance of the proposed method. We have considered F-score, precision, accuracy and recall. Before calculating them, there are some formulas and variables that requires brief explanation. These are: False Negative (FN), False Positive (FP), True Positive (TP), True Negative (TN), PC = Predicted Class, and AC = Acute Class.

TABLE IV
PREDICTED and ACUTE CLASS

AC	**PC**	
	Positive	***Negative***
Positive	FP	TN
Negative	TP	FN

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F - Score = 2 \times \frac{PPV \times TPR}{PPV + TPR} \quad (4)$$

TABLE V
Performance Evaluation

Sentence **(বাক্য)**	**Accuracy**	**Recall**	**Precision**	**F-Score**
সরল (Simple)	75.00%	85.71%	80.00%	82.75%
জটিল (Complex)	75.00%	91.67%	73.33%	81.48%
(যৌগিক) (Compound)	85.00%	87.5%	93.33%	90.32%

TABLE VI
STOP PHRASES WITH EXAMPLE

Sentence **(বাক্য)**	**Accuracy**	**Recall**	**Precision**	**F-Score**
Assertive (বর্ণনামূলক)	85.00%	93.33%	87.50%	90.32%
Interrogative (প্রশ্নবোধক)	90.00%	94.44%	94.44%	94.44%
Imperative (অনুজ্ঞাসূচক)	75.00%	85.71%	80.00%	82.75%
Optative (প্রার্থনাসূচক)	75.00%	76.47%	92.86%	83.87%
Exclamatory (বিস্ময়সূচক)	90.00%	94.12%	94.12%	94.12%

V. CONCLUSION

The Bengali language is the 5th most spoken language having 228 million speakers [12]. There must be a complete set of stop phrase list for this language which is imperative for NLP and information retrieval of Bengali language. This paper proposes a corpus based methodology for the detection of stop phrase from a given context and eliminating those in the quest of achieving simpler and faster data processing of Bengali language. The Experimental result indicates that the application of proposed algorithm yields numerous conveniences for faster processing of Bengali language. We strongly believe that our effort will help advancing the Bengali Language processing significantly. Our future plan is to implement a deep learning based stop phrase detection method.

Acknowledgment

This work is supported by The Institute for Energy, Environment, Research and Development (IEERD), University of Asia Pacific (UAP).

References

[1] Senem Kumova Metïn, Bahar Karaoğlan, “Stop Word Detection As a Binary Classification Problem,” Anadolu Üniversitesi Bilim ve Teknoloji Dergisi A- Uygulamalı Bilimler ve Mühendislik, Anadolu University Journal of Science and Technology A- Applied Sciences and Engineering, 2017 - Volume: 18, Number: 2, Page: 346 – 359, DOI: 10.18038/aubtda.322136.

[2] Google,“Locating meaningful stopwords or stop-phrases in keyword-based retrieval systems”, United States Patent, 7,409,383Tong, August 5, 2008.

[3] Shree Jagadish Chanda Ghosh, ”Adhunik Bengali Bakoron”, pp. 1-200, Presedency Library, Dhaka (1934).

[4] Rachel Tsz-Wai Lo, Ben He, Iadh Ounis, “Automatically Building a Stop word List for an Information Retrieval System,” 5th Dutch-Belgium Information Retrieval Workshop (DIR) ’05 Utrecht, the Netherlands .

[5] Alexandra Schofield, Mans Magnnusson, David Mimno, “Pulling Out the Stopword Removal for Topic Models,” Proceedings of the 15th conference of the european chapter of the association for the computational linguistics: Volume:2, EACL, Valencia, Spain, April-2017.

[6] S. Popova, L. Kovriguina, D. Mouromtsev, I.Khodyrev, “Stopwords in Keyphrase Extraction Problem,” proceeding of the 14th conference of fruct association, Espo, Finland. 11-15.Nov.2013

[7] Hassan Saif, Miriam Fernandez and Harith Alani,“Automatic Stopword Generation using Contextual Semantics for Sentiment Analysis of Twitter,” ISWC-PD’14 proceding of the 2014 International Conference on Posters & demonstrations track – volume 1272.

[8] Efstathios Stamatatos, “Plagiarism Detection Using Stopword n-grams,” This is a preprint of an article published in the Journal of the American Society for Information Science and Technology, 62(12), pp. 2512-2527, Wiley, 2011.

[9] Feng Zou, Fu Lee Wang, Xiaotie Deng, Song Han, Lu Sheng Wang, “Automatic Construction of Chinese Stop Word List”, Proceedings of the 5th WSEAS International Conference on Applied Computer Science, Hangzhou, China, April 16-18, 2006 (pp1010-1015)

[10] Riyad Al-Shalabi, Ghassan G Kanan, J.M.Jaam, A. Hasnah, Eyad Hailat, “Stop-word removal algorithm for Arabic language”, Conference: Information and Communication Technologies: From Theory to Applications, 2004. Proceedings. 2004 International Conference on.

[11] Catarina Silvatt, Bemardete Ribeirot, “The Importance of Stop Word Removal on Recall Values in Text Categorization”, Conference: Neural Networks, 2003. Proceedings of the International Joint Conference on, Volume: 3.

[12] Eduard Dragut, Fang Fang, Prasad Sistla, Clement Yu,Weiyi Meng, “Stop Word and Related Problems in Web Interface Integration”, Proceedings of the VLDB Endowment 2(1):349-360, Auigust, 20009. DOI:10.14778/1687627.168766.

[13] Satyendr Singh and Tanveer J. Siddiqui, “Evaluating Effect of Context Window Size, Stemming and Stop Word Removal on Hindi Word Sense Disambiguation”, International Conference on Information Retrieval & Knowledge Management (CAMP), March 2012, DOI: 10.1109/InfRKM.2012.6204972.

[14] SIL International, ”ethnologue (2019, 22nd edition)”, US, 22nd december 2019.