Implementation and Performance Evaluation of Semantic Features Analysis System for Bangla Assertive, Imperative and Interrogative Sentences

Article · November 2018

CITATION
1

READS
32

4 authors, including:

Fm Rahat Hasan Robi
Bangabandhu Sheikh Mujibur Rahman Science & Technology University, Gopalganj
5 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Md Monowar Hossain
Rajshahi University of Engineering and Technology
4 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Implementation and Performance Evaluation of Semantic Features Analysis System for Bangla Assertive, Imperative and Interrogative Sentences

Md. Sajidul Karim
Dept. of Computer Science & Engineering
Bangabandhu Sheikh Mujibur Rahman Science & Technology University, Gopalganj, Bangladesh
biplobsajidul@gmail.com

Nur-E-Alam Jony
Dept. of Computer Science & Engineering
Bangabandhu Sheikh Mujibur Rahman Science & Technology University, Gopalganj, Bangladesh
nuralam.cse.bsmrstu@gmail.com

F.M.Rahat Hasan Robi
Dept. of Computer Science & Engineering
Bangabandhu Sheikh Mujibur Rahman Science & Technology University, Gopalganj, Bangladesh
rahatcse10@gmail.com

Md.Monowar Hossain
Dept. of Computer Science & Engineering
Bangabandhu Sheikh Mujibur Rahman Science & Technology University, Gopalganj, Bangladesh
murad0904045@gmail.com

Md. Torikur Rahman
Dept. of Computer Science & Engineering
Uttara University, Dhaka, Bangladesh
torikurrahman@gmail.com

***Abstract*— For natural language processing of Bangla sentences, semantic analysis is needed. Sentence structure and validity are two very important parts of semantic analysis. To understand the meaning of a sentence by a machine, it requires a systematical way with a rule-based approach. In this paper, a rule-based approach is applied which includes grammatical rules and structure of Bangla language, parts of speech, relationship of each word with others for determining sentence types and validation.**

Keywords— semantic; rule; validation; relationship; word-id

I. INTRODUCTION

As a medium of communication human uses natural language for expressing thoughts with others. Bangla language has a lot of diversity. It contains infinite number of different words. Sometimes while communicating with each other in Bangla, we can still understand the meaning of the sentences which even not follow the rules of grammar and syntax. People can do this because they know about the language, know the grammar and they have previous experience. But for a machine, it is very much difficult to understand the real meaning of a sentence. So, a rule-based approach is needed for machine to understand the meaning of a sentence. According to Bangla grammar, we have identified some common structures of every types of sentences. We have worked with three types of sentences i.e. assertive, interrogative and imperative. We have also worked with the relationship of words with each other for determining sentence validation. Some possible outcomes of this paper are mentioned in the following: -

- Proposed a methodology for determining Bangla sentence types and validation.
- Proposed some semantic features for different types of part speech of Bangla words.
- Generate rule-based structure of the given sentence according to Bangla grammar.
- Classifying three types of sentences i.e. assertive, interrogative and imperative.
- Identifying whether a sentence is semantically valid, partially valid or invalid.

II. RELATED WORKS

The Bangla language parsing is one of the challenging tasks of natural language processing. Many researchers worked on Bangla Language in different ways. Some works have been done on dictionary constriction [1]. The Bangla sentence parsing including assertive, interrogative and imperative by a set of context sensitive grammars (CSG's) is shown in [2]. Verify Bangla sentence structures using N-Gram is shown in [3]. A comprehensive approach for CFG rules to parse all types of sentences including complex, compound, exclamatory and optative sentences was shown in [4]. A set of Context sensitive phrase structure rules for structural representation of Bangla natural language sentences is implemented in [5]. An automatic classification of Bangla sentence by sensing the definition present in Bangla wordnet is in [6]. In [8], a structural analysis for Bangla sentence of different sentence is proposed for machine translation. In [9], for getting the root word from any given Bengali word, a rule based Bengali stemmer is proposed.

III. SEMANTIC APPROACHES

There are two kinds of semantic approaches, one is lexical semantic and the other is compositional semantic.

A. Lexical semantic approach

Lexical semantics is a subfield of linguistic semantics. The units of analysis in lexical semantics are lexical units which include not only words but also sub-words or sub-units such as affixes and even compound words and phrases. Lexical units make up the catalogue of words in a language, the lexicon. Lexical semantics looks at how the meaning of the lexical units correlates with the structure of the language or syntax.

978-1-5386-8207-4/18/$31.00 ©2018 IEEE

B. *Compositional Semantic Approaches*

Semantic compositionality allows the construction of complex meanings from simpler elements based on the principle that the meaning of a whole is a function of the meaning of the parts. The target knowledge structures constructed in this approach are typically logic expressions. The structures created through this approach are usually categorized as logical formulae in some calculus.

IV. PROPOSED METHODOLOGY

The basic block diagram of the overall process depicted in Figure 1

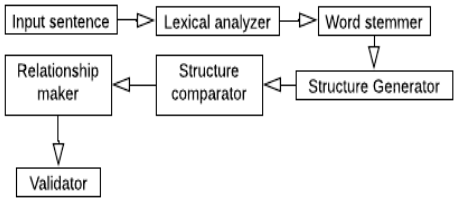

Fig. 1. The components of the proposed analyzer

A sentence is given as an input to the system. The lexical analyzer breaks the input sentence into lexemes. The lexemes are given as the input of word stemmer. The word stemmer produces stem words from the lexemes. From the stem words, structure generator produces the structure of the input sentence according to their parts of speech. After that, the structure comparator compares the generated structure with the common sentence structure. If it matches, the relationship maker makes relation between each word with others. From the relations of words, the validator calculates the relationship percentage and produces a result. Here is the Algorithm of the Proposed Method-

Step 1 - First we have to input a Bangla sentence as a string, like "Goru Akashe ore"
Step 2 - Tokenize the sentence, like "Goru + akashe+ ore"
Step 3 - Extract stem word from the input sentence, like Goru, akash, ora
Step 4 - Identify parts of speech of each stem word of the sentence, like Goru(Bishessho), akash(Bishessho), ora(kria)
Step 5 - Then assign id to each stem word by matching stored information in database, like Goru(3(b)), akash(10(b)), ora(4(k))
Step 6 - Generate sentence structure with the parts of speech of each word of the sentence, like Bishessho(Jati) + Bishessho(Nam) + Kria (Somapika)
Step 7 - Match with the common sentence structure with the generated sentence structure of input sentence.
Step 8 - (Match with valid common sentence structure)? X : Y. Here X= Match and Y= Not Match.

If Condition is true then,

- Generate all possible relation with each word to other
- Assign relationship percentage according to the daily usages of each relationship of words.
- Averaging the relationship percentage from the number of relationships and the sum of relationship percentage.
- If the average percentage is between 65 to 100 then, declare the input sentence is valid.
- If the average percentage is between 40 to 64 then, declare the input sentence is partially valid.
- If the average percentage is between 0 to 39 then, declare the input sentence is invalid.

If Condition is false then,

- Return Invalid sentence format.

The activity diagram of the proposed system depicted as Figure 2

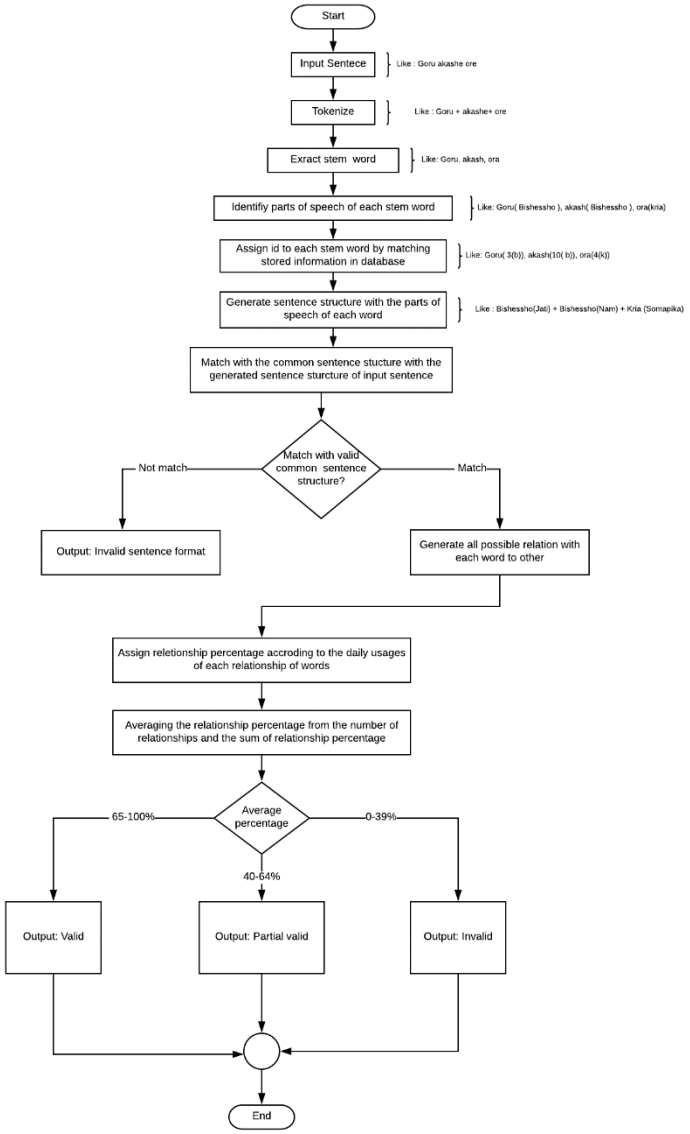

Fig. 2. Activity diagram of the proposed system

V. SEMANTIC FEATURES OF DIFFERENT PARTS OF SPEECH OF BANGLA

The features or characteristics that help to recognize a part of speech are known as semantic features. In semantic analysis many semantic features should be collected to easily identify the tokens.

A. *Semantic Features of Bishessho*

Bishessho is generally used as subject in simple assertive sentence. Generally, Bishessho in Bangla sentence contains the following semantic features: [Nam, Jati, Bostu, Somosti, Vab, Gun].

For some sentences, few features become more dominant than the other for example "Karim sudorshon chele". Some semantic features of common Bishessho have been shown in the following Table I.

TABLE I. SEMANTIC FEATURES OF BISHESSHO

id	Word (Bishessho)	Nam	Jati	Bostu	Somoshti	Vab	Gun
1	Karim	1	0	0	0	0	0
2	sudorshon	0	0	0	0	0	1
3	chele	0	1	0	0	0	0

B. Semantic Features of Sorbonam

In Bangla sentences, Sorbonam is used instead of Bishessho. When it is used, it is considered as subject in the sentence. So, it plays a vital role in semantic analysis. Generally semantic features of Sorbonam are similar to that of Bishessho. There are also some extra features which should be counted for identifying Sorbonam. In Bangla sentence, some Kria mainly correspond to this feature. As for example ["ke asbe","ki nibo"]. Semantic features of common Sorbonam have been shown in the following Table II.

TABLE II. SEMANTIC FEATURES OF SORBONAM

id	Word (Sorbonam)	Bakti	Proshno	Songjog
1	Ami	1	0	0
2	amra	1	0	0
3	ke	0	1	0
4	ki	0	1	0
5	je	0	0	1

C. Semantic Features of Bisheshon

Bisheshon is the quality of the subject. Sometimes, in sentence subject comes along with Bisheshon, in which the Bisheshon becomes a main part for the semantic analysis. In general, Bisheshon has two following semantic features [nam, vab]. Some of these features are "Gunoban", "Buddiman", "bhalo" etc. Some commonly used Bisheshon within their semantic features are shown in the following Table III.

TABLE III. SEMANTIC FEATURES OF BISHESHON

id	Word (Bisheshon)	Nam	Vab
1	Gunoban	1	0
2	Buddiman	0	0
3	Bhalo	0	1

D. Semantic Features of Kria

Kria is related to the semantic features of a subject. Kria is the most dominant factor of a sentence, because the event that is occurred represents the situations and also the meaning. For example, in the sentence "gachti hette jabe", "gachti khai" to verify the Kria walking whether it matches in connection with "gachti". Some commonly used Kria within their semantic features are shown in the following Table IV.

TABLE IV. SEMANTIC FEATURES OF KRIA

id	Word (Kria)	Somapika	Osomapika
1	khai	1	0
2	Thakbe	0	1
3	Jabe	0	1

VI. RULE GENERATORS FOR SENTENCE STRUCTURE

For each type of sentence there are many structures. We have worked with three types of sentence. From our dataset, we have figured out15 types of assertive, 8 types of interrogative and 7 types of imperative sentence's structure. It may be increased if more sentences are added to the dataset. Some of structures are shown in the following Table V.

TABLE V. STRUCTURES OF VARIOUS TYPES OF SENTENCES

Sentence		Rules
Assertive	1.	Bishessho(Nam) + Bishessho(Bostu) + Kria (Somapika)
	2.	Bishessho(Bakti) + Bisheshon(Vab) + Bishessho (Somoshti)
	3.	Bishessho(Bakti) + Obboi+Bishessho (Bakti) + Bisheshon(Gun) + Bishessho(Somoshti)
Interrogative	1.	Sorbonam(Bakti) + Proshno + Bishessho(Bostu) + Kria(Somapika)
	2.	Sorbonam(Bakti) + Proshno + Bishessho(Bakti)+ Bishessho(Somoshti) + Kria(Osomapika)
	3.	Bishessho (Bakti) + Proshno + Kria(Osomapika)
Imperative	1.	Bishessho(Bostu) + Bivokti + Kria(Somapika)
	2.	Kria(Osomapika) + Bishessho(Bostu) + Kria (Somapika)
	3.	Bishessho(Bakti) + Bochon + Bishessho (Bostu) + Bishessho(Bostu) + Kria(Somapika)

VII. VALIDATION CHECKING

We have assigned id for each word. These are shown in the below table. In each column, there is one Bangla parts of speech. According to the parts of speech each word is kept in its corresponding column. We denote Bishessho as(b), Sorbonm as(s), Kria **as(k),** Bisheshon as(bh) and Obboi as(ob).

TABLE VI. WORD ID FOR DIFFERENT PARTS OF SPEECH

id	Bishessho (b)	Sorbonam (s)	Kria (k)	Bisheshon (bh)	Obboi (ob)
1.	Sajidul	Se	khai	valo	Na
2.	goru	Ami	othe	kharap	Ha
3.	akashe	Tumi	jai	mota	ebong
4.	boi	eti	pore	choto	othoba
5.	bati	Apni	nivao	boro	
6.	vat		Ore	lomba	
7.	surjo		vashe	khato	

According to the daily usages in our real world, words are categorized into some category for showing relationship with each other. We have made a dataset of 23,000 sentences

which includes about 90 thousand of words. In a specific passage of that dataset, all the sentences are interrelated of that passage. Then we try to find out the relationship between every single word with others and assign a relationship percentage manually and store this relation and percentage in the database. Some examples are given below how we consider it. A passage contains the following sentences: -

"উপমহাদেশের মানুষের প্রধান খাদ্য ভাত। অন্যদিকে মধ্যপ্রাচ্যের লোকজন শুঁকনো জাতীয় খাবার বেশি খায়। তারা প্রধানত রুটি খেয়ে থাকে। আবার ইউরোপের দিকে আমিষ জাতীয় খাবার বেশি খায়। বিভিন্ন পশুর মাংস এদের পছন্দের তালিকার শীর্ষে যেমন- মুরগি, ঘোড়া, গরু ইত্যাদি।"

From the above information, it seems that there is a strong relationship between 'manush' and 'khaoya'. So, this relationship is a universal truth. So, we assign a 91-100% relationship percentage. On the other hand, the human takes various types of food in various regions. Hence, a relationship like between 'vat' and 'manush is usual, not universal. So, we assign a relationship percentage of 71-90% of these kinds of relationships. Another passage contains the following sentences: -

"আমাদের দেশে অন্যতম গৃহপালিত পশু হচ্ছে গরু। এরা সাধারণত নিরীহ ও শান্ত প্রকৃতির হয়ে থাকে। খাদ্য হিসেবে ঘাস, ভাতের মাড়, কুড়া ইত্যাদি খেয়ে থাকে। গ্রামের লোকজন নিজেদের বাড়িতেই গরু পালন করে থাকে। সকাল বেলা মাঠে নিয়ে যায় আরসন্ধ্যায় বাড়িতে নিয়ে আসে। খোলা আকাশের নিচে সারাদিন মাঠে চড়ে ঘাস খায়। চতুষ্পদ জন্তু এই গরু, গ্রামের মানুষের এক প্রকারের বন্ধু। গরুর দুধে রয়েছে প্রচুর পরিমাণে পুষ্টি। এর কিছু শারীরিক বৈশিষ্ট্য নিম্নরুপঃ চার পা, দুই কান, দুই শিং, দুই চোখ, একলেজ আরএকনাক বিশিষ্ট বিশালাকার দেহের অধিকারি হয়ে থাকে গরু।"

From the above information, we have learned some characteristics of cow. Now for a sentence like 'goru akashe ore', there is a relationship between 'goru' and 'akash'. But it is sometimes. So, we assign a 41-70% relationship percentage of these kind of relations. On the other hand, there is no relation between 'goru and 'ore' because we know that cow has no wing for flying. Hence, a relationship like between 'goru' and 'ore' is false (very rare). So, we assign a relationship percentage of 0-20% to this kind of relation. From these examples a mathematical formula can be formulated: -

f (a,b) € (1,0)

$$f(a,b) = \begin{cases} 1, & a \text{ and } b \text{ are related} \\ 0, & a \text{ and } b \text{ are not related} \end{cases}$$

Where f is the relationship function between two words a and b. If there exists any the relationship between a and b, we assign 1 otherwise 0. Then we count how many times this relation comes in a sentence in an interrelated passage. Then we calculate the average relationship of that specific relation in percentage.

$$p = \frac{\sum_{i=1}^{n} f_i(a,b)}{TS} \times 100\%$$

Where, the relationship percentage and TS is the total number of sentences in which words 'a' and 'b' exist. The relationship is given by a percentage of range. We have categorized words in 5 categories. This is shown in Table VII.

TABLE VII. RELATIONSHIP PERCENTAGE RANGE

No	Usages	Relationship Percentage Range
1.	Universal Truth	91- 100%
2.	Usually	71 - 90%
3.	Sometimes	41 – 70%
4.	Rare	21 – 40%
5.	Very Rare	0 – 20%

Then we have assigned a percentage with one word to another according to their daily usages. In each row, there is one word. The following cell of each row, there is an id for a word and the relationship percentage. We give the relation with each word to other. An example, the word 'Sajidul' is a Bishessho (b) which has an id 1 from table 6, column 1, row 1 and denoted by 1(b). The relationship of this word with other word like khai (in column 3, row 1, id 1(k)) is 95% percent according to the relationship and daily a usage of these words. These are shown in Table VIII.

TABLE VIII. RELATIONSHIP PERCENTAGE

id	Word	Relation				
1	Sajidul	1(k)	2(k)	3(k)	4(k)	6(b)
		95%	45%	88%	88%	98%
2	goru	1(k)	6(k)	3(bh)	4(bh)	3(b)
		92%	0%	40%	5%	10%
3	boi	1(k)	4(k)	1(bh)	2(bh)	4(b)
		0%	100%	80%	80%	85%
4	ghash	7(k)	3(bh)	4(bh)	5(bh)	7(b)
		0%	25%	80%	80%	65%
5	bati	1(bh)	2(bh)	5(k)	5(bh)	7(b)
		60%	60%	98%	80%	82%
6	vat	1(k)	4(k)	7(k)	1(bh)	4(b)
		95%	0%	2%	75%	70%
7	surjo	1(k)	2(k)	3(k)	5(k)	2(b)
		0%	96%	70%	0%	0%
8	akashe	1(k)	2(k)	3(b)	6(k)	1(b)
		80%	40%	94%	95%	80%
9	Ami	1(k)	1(bh)	2(bh)	3(bh)	6(b)
		80%	80%	80%	85%	85%
10	Tumi	1(bh)	2(bh)	3(bh)	4(k)	4(b)
		82%	79%	80%	83%	68%
11	eti	1(bh)	2(bh)	3(bh)	4(bh)	6(b)
		80%	80%	80%	80%	80%
12	Apni	1(bh)	2(bh)	3(bh)	4(bh)	6(b)
		80%	80%	80%	80%	80%

When a sentence is given as an input, the system first divides the sentence into single words and extract the stem word [9]. According to the Bangla parts of speech, the words are categorized by checking the word id table. Then it matches the sentence structure with the Bangla grammar sentence structure [8]. If it matches with any valid Bangla sentence structure, it shows the type of the given sentence. Otherwise it shows that the given sentence is not given in a structured way. It is not a correct sentence. Example of some sentence structure:

1) *Assertive Sentence*

i) Sajidul vat khai

Sajidul -------> vat = 98%

Sajidul------->khai = 95% Average = 96%
Vat-------->khai = 95%

ii) goru akashe ore
goru--------->akashe = 10%
goru---------> ore = 0% Average = 35%
akashe--------->ore = 95%

2) Imperative Sentence

i) batiti nivao
bati ------> nivao = 98% Average = 98%

3) Interrogative Sentence

i) Tumi ki boi poro?
Tumi ------> poro = 83%
Tumi--------->boi = 68% Average = 83.66%
boi---------->poro = 100%

For evaluating validation of the given sentence, it finds out the relationship with each word to others from the relationship table as percentage. This technique refers to the identification of actual meaning of a word based on its distinct contextual environments [5]. After getting all relationship percentage of all combination, it calculates the average relationship percentage. We categorized the average percentage into three categories. These are shown in Table IX.

TABLE IX. VALIDATION PERCENTAGE

No	Validation	Average Percentage Range
1.	Valid	65 - 100%
2.	Partial Valid	40 - 64%
3.	Invalid	0 – 39%

According to the average relationship percentage, the system will produce a decision, is the given sentence is valid, partial valid or invalid.

VIII. Experimental Results

We have tested many sentences of various types with various lengths. Our system has produced a satisfactory level of accuracy of 85.4%. The total experimental result is shown in Table X.

TABLE X. ACCURACY RATE

Sentence Type	Sentence length	Input sentence	Correct sentence	Accuracy rate
Assertive	3	95	87	86.3%
	5	60	49	
	7	35	28	
Interrogative	4	85	72	82.8%
	5	65	53	
	6	30	24	
Imperative	2	90	81	87.2%
	3	80	70	
	4	25	19	
Total		565	483	85.4%

IX. Discussion and Conclusion

It is easy for human to understand the meaning of sentences while communicating with others using language. But for machine it is hard by using natural language. Hence, we proposed a ‘rule’ and ‘relation’ based method for the determining sentence’s type and validity for Bangla language. We generate many rules/sentence structures and find out the relationship between each single word with others. Here, a comparison study of our work is given below in table XI with others. We focused on four factors-methodology, accuracy, validity determining and sentence types classifying.

TABLE XI. COMPARISON STUDIES

Research work	Methodology	Accuracy (%)	Determine validity	Classify sentence type
Our work	Rule and relation-based	85.4%	Yes	Yes
Nur Hossain et. al[3]	N-gram	93.1%	Yes	No
Arefin et. al[2]	Parse tree	83.09%	No	Yes
Pal et. al[6]	Sensing definition	91%	Yes	No

References

[1] Ismail, Sabir, M. Shahidur Rahman, and Md Abdullah Al Mumin. "Developing an automated Bangla parts of speech tagged dictionary." Computer and Information Technology (ICCIT), 2013 16th International Conference on. IEEE, 2014.

[2] Arefin, Mohammed Safayet, et al. "An empirical framework for parsing Bangla assertive, interrogative and imperative sentences." Computer and Information Engineering (ICCIE), 2015 1st International Conference on. IEEE, 2015.

[3] Nur Hossain Khan, Md, et al. "Verification of Bangla Sentence Structure using N-Gram." Global Journal of Computer Science and Technology (2014).

[4] Mehedy, Lenin, SM Niaz Arifin, and M. Kaykobad. "Bangla syntax analysis: A comprehensive approach." Proceedings of International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh. 2003.

[5] Hoque, Mohammed Moshiul, and Muhammad Masroor Ali. "Context-sensitive phrase structure rule for structural representation of Bangla natural language sentences." Proceedings of ICCIT, Dhaka (2004): 615-620.

[6] Pal, Alok Ranjan, Diganta Saha, and Niladri Sekhar Dash. "Automatic classification of bengali sentences based on sense definitions present in bengali Wordnet." arXiv preprint arXiv:1508.01349 (2015).

[7] Anwar, Md Musfique, Nasrin Sultana Shume, and M. A. A. Bhuiyan. "Structural analysis of Bangla sentences of different tenses for automatic Bangla machine translator." International Journal of Computer Science and Information Security 8.9 (2010).

[8] Mahmud, Md Redowan, et al. "A rule based bengali stemmer." Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on. IEEE, 2014.