2018 21st International Conference of Computer and Information Technology (ICCIT), 21-23 December, 2018

A Comprehensive Parts of Speech Tagger for Automatically Checked Valid Bengali Sentences

Nahid Hossain
Computer Science and Engineering
United International University
Dhaka, Bangladesh
nahid@cse.uiu.ac.bd

Mohammad Nurul Huda
Computer Science and Engineering
United International University
Dhaka, Bangladesh
mnh@cse.uiu.ac.bd

Abstract— **This paper constructs a method for corroborating Bengali sentences whether they are equitable syntactically, semantically and pragmatically; then a method is devised for detecting parts of speech (POS) from the valid sentences. In our approach, we have analyzed several techniques to check whether construction of inputted Bengali sentence is valid, and finally, we have chosen the best technique among them. Moreover, after analyzing the construction of several Bengali sentences we have designed a rule-based algorithm for detecting POS with a significant accuracy. The idea of ignoring sentences with grammatical mistakes helped significantly to achieve higher accuracy and to reduce execution time. Moreover, our projected method achieved an accuracy of 91.45% which is the highest among similar POS tagger.**

Keywords—parts of speech; Bengali Sentence; Syntax; Semantics; Pragmatics

I. INTRODUCTION

Bengali is an Indo-Aryan language and is the native language of Bangladesh and West Bengal (India). Bengali is the seventh most spoken language by native speakers around the world. There are 200 million native speakers and about 300 million speakers in total [1]. It is written in Bengali alphabets with a Unicode range from U+0980 to U+09FF. Nowadays Bengali is used vastly in different web platforms and software. Thus, natural language processing of Bengali language is essential for superior machine learning. It is very fundamental to check a sentence whether it is valid syntactically, semantically and pragmatically before any further processing. In addition, determination of different parts of speech (POS) in a valid sentence is another integral part of natural language processing.

Previously, a considerable amount of work had been done in POS tagger in English and other languages. However, a few efficient studies were done on the identification of Bengali POS. S. Dandapat, et al. [2] proposed two stochastic based taggers in 2007 where manually annotated corpus was used. Ekbal, et al. [3] proposed a support vector machine(SVM) based tagger in the year 2008. The POS tagger has an accuracy of 86.84%. In the year 2010, an unsupervised POS tagger for the Bengali language based on a Baum-Welch trained Hidden Markov Model(HMM) approach was proposed by H. Ali[4]. Sarkar and Gayan [9] proposed a supervised trigram POS tagger for Bangla using second-order HMM in 2012. They have also developed a bigram POS tagger to which their trigram approach had been compared. In 2015, Hoque and Seddiqui [10], projected a Bangla POS tagger using Bangla stemmer and rule-based analyzer.

However, most of these approaches were done only for research purposes with negligible consideration of the real-life applications. Therefore, most of these approaches showed several imperfections in real-life applications. They have detected POS of individual words without detecting POS from complete Bengali sentences which is more practical. Thus, our proposed method is designed to work on valid complete Bangla sentences. Moreover, these researchers did not check whether a Bengali word/sentence is valid before POS detection. Although highly accurate POS tagger is very essential for NLP applications, we found no such POS tagger available for Bangla. These are some reasons that motivated us to do a research on Bangla POS tagger that would audit the validity of an inputted Bengali sentence and determine the parts of speech from the valid Bengali sentence considering all real-life applications with the highest accuracy possible.

In our research, at first, we have designed an approach to analyzing whether an inputted Bengali sentence is authentic. To validate a sentence we ensured that all the three distinct aspects (syntax, semantics, and pragmatics) of sentences are properly satisfied. After that, we have designed a rule-based approach [proposed method] to detect diverse POS from the valid sentences. The originality of the paper is to incorporate validation testing before the POS detection and to incorporate all the parts of speech tagging for Bengali sentences. Moreover, we have achieved significant accuracy using our proposed method.

The paper is organized as follows: In Section II, our paper describes the proposed method with proper examples, algorithms, and step by step demonstration of the algorithms. The paper demonstrates experiments and experimental set up in Section III. Section IV interprets result from the experiments, while Section V concludes the paper with future remarks.

II. PROPOSED METHOD

In a sentence written in English, the subject comes first, the verb second, and the object comes third, which is a subject-verb-object (SVO) sentence structure [5]. A simple sentence structure in English is given in Figure 1.

978-1-5386-9242-4/18/$31.00 ©2018 IEEE

Figure 1: Simple English sentence ("He eats rice") structure.

However, in a sentence written in Bengali, the subject comes first, then the object and finally the verb and this is called subject-object-verb (SOV) sentence structure[6]. Figure 2 exhibits a simple Bengali sentence structure. Bengali words in the sentence "সে ভাত খায়", corresponding IPA (International Phonetic Alphabet) and English meaning are given in Table 1.

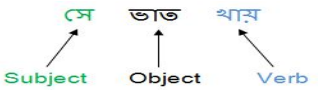

Figure 2: Simple Bengali sentence("সে ভাত খায়") structure

Table 1: Bengali sentence "সে ভাত খায়" with its IPA representation and English meaning

Bengali Words	IPA Representation	English Meaning
সে	/s/ /e/	He
ভাত	/b^h/ /a/ /t̪/	rice
খায়	/k^h/ /a/ /e̯/	eats

In our approach, we have branched our work into two dominant tasks. At the first step, we checked whether the inputted Bengali sentence is authentic obeying three aspects(namely Syntax, Semantics, and Pragmatics). Finally, we detect the POS from the legitimate sentences. Our first step comprises of three stages: i) checking whether an inputted Bengali sentence follows the SOV or SV or SO structure correctly to ensure the syntax property, ii) Checking whether the inputted sentence is semantically valid using our algorithm, and iii) finally we checked whether the sentence is valid pragmatically using the N-gram model [7]. For our final step, we have developed CFG (Context Free Grammar) and implemented it to identify all POS meticulously.

A. *Algorithm for Syntax Checking*

SE ← Input a Bengali sentence.
IF no repeated word in SE THEN
  S ← Extract subject-phrase from SE
  O ← Extract object-phrase from SE
  V ← Extract verb-phrase from SE
  IF SE follows SOV or SV or SO
    Print "SE is syntactically valid"
  END IF
END IF

B. *Algorithm for Semantics Checking*

SE ← Input a Bengali sentence
S ← Extract subject-phrase from SE
O ← Extract object-phrase from SE
V ← Extract verb-phrase from SE
P_SO ← Find relative frequency probability [8] *between S & O*
IF P_SO is greater than 0 THEN
  P_SV ← Find relative frequency probability [8] *between S & V*
  IF P_SV is greater than 0 THEN
    Print "SE is semantically valid"
  END IF
END IF

C. *Algorithm for Pragmatics Checking*

SE ← Input a Bengali sentence
w^n ← Break SE into n words // w^n = set of n words
Find the bigram sentence probability [7] *using,*

$$P(w^n)=\prod_{k=1}^{n} P(w_k \mid w_{k-1}) \text{ where } P(w_k \mid w_{k-1}) = \frac{Count(w_{k-1}w_k)}{Count(w_{k-1})}$$

IF $P(w^n)$ *is greater than 0 THEN*
  Print "The SE is pragmatically valid".
END IF

D. *Algorithm for POS Detection*

SE ← Input a Bengali sentence.
Check whether the syntax of SE is valid.[Syntax checking algo.]
IF syntactically valid THEN
  Check whether the semantics of SE is valid.[Semantics checking algo.]
  IF semantically valid THEN
    Check whether pragmatics of SE is valid.[pragmatics checking algo]
    IF pragmatically valid THEN
      w^n ← Break SE into n words.
      k ← 1 //Deal with first word
      WHILE k != n //Whether n word complete
        Pass each word, w^k from w^n through the given grammar.
        Display POS result.
        k ← k+1 //Go to next word
      END WHILE
    END IF
  ELSE IF
END IF

E. *Algorithm for Testing*

N ← Calculate number of the total sentence from the test set.
Total_Words ← 0
C ← 0
FOR i=1 TO N
  SE ← Take a sentence from the test set.
  W^n ← Break SE into words //w^n set of n words
  FOR k=1 to n
    POS[W^k] ← Detect_POS(W^k)
    Total_Words ← Total_Words+1
    IF(POS[W^k] is correct)
      Count[POS[W^k]] ← Count[POS[W^k]]+1
      C ← C+1

END IF
END FOR
END FOR

III. EXPERIMENTS

A. *Data Corpus*

We have utilized the following two corpora for our experiments.

i) **Training Corpus**

It consists of 4090 Bengali sentences (approximately 23,000 words). Table 2 demonstrates the different types of sentences used in training.

Table 2: Different types of sentences used in training

Sentence Type	Number of sentences
Simple	2740
Compound	694
Complex	506
Unorthodox	150

ii) Testing Corpus

For test corpus, we have used 2550 Bengali sentences (approximately 14,000 words). Test Sentences are of the aforementioned type of the training sentences in Table 2.

B. *Experimental Set up*

For finding the total recognition rate, A and error rate, E we have used equation (1) and (2), where C and T represent the total number of correctly recognized words and total words respectively.

$$A = \frac{C}{T} * 100\ \% \quad (1)$$

$$E = 100\% - A \quad (2)$$

We have implemented the following two methods for POS detection in our experiments

i) Probability-based method
ii) Rule-based method [Proposed Method]

For POS detection in Rule-based method, we have used the following abstract CFG for valid Bengali sentences,

sentence→ sub-phrase + obj-phrase + ver-phrase
sub-phrase→ prn | nou | prn+nou | nou+nou | prn+prn | prn+nou+nou | nou+nou+nou
obj-phrase → ε | adj | prep | art+nou | num+prep | prep+adj | prep+nou | adj+nou | adj+adj | art+adj+nou | adj+adj+nou | num+nou+nou | prep+num+nou
ver-phrase → ε | ver | adv | ver+adv | ver+ver | adv+v | ver+ver+adv | adv+ver+adv
prn → সে | তুমি | আমি | আমরা | তোমার | আমার | …
nou → ভাত | মানুষ | ডিম | মুরগি | ডিম | অংক | ফুটবল |…
adj → বেয়াদব | লাজুক | দ্রুত | কষ্ট | সৎ | নিয়মিত | নষ্ট | অসৎ |…
art→ রোজ | একজন | একটি | …
prep→ থেকে | কাছে | পর্যন্ত | দিয়ে | …
num→ এক | দুই | তিন | চার | পাঁচ | ছয় |…
ver→ খাই | যাব | পাড়ে | বেড়াতে | পড়ছে | উড়ে |…
adv→না | নই | কালকে | কোন | সারাজীবন |…

In the probability-based method, we have conceived a database of 4090 (Training Corpus) Bengali sentences where all the POS of words were manually tagged beside each sentence to determine the probability for new Bengali sentences. When a Bengali sentence is inputted for determining the POS, it is segregated into words and obtained the maximum probability of POS for each word.
In the probability-based method, the procedure of sentence validation and testing is the same as our proposed Rule-based method. Algorithm for the method is given below.

SE ← Input a Bengali sentence
w^n ← Break the sentence into n words.
FOR k=1 TO n
  FOR p=1 TO size of ArrayOfPOS
    // Where ArrayOfPOS contains all the different POS
    //names(such as noun, pronoun, adjective etc)
    PK ← Find the number of occurrence of p^{th} POS in k^{th} location of every sentence.
    PKRate ← (PK/ Total Sentence).
    NUM ← Number of how many sentences have w_k word at k^{th} position.
    POSNUM ← Number of how many sentences contains w_k word and of p^{th} POS and at k^{th} position.
    NUMRate ← (NUM/POSNUM).
    IF(PKRate + NUMRate) greater than previous iteration THEN
      Save this current result as maximum probability, because w^k word at k^{th} position of sentence can be p^{th} POS.
    END IF
  END FOR
  Print the POS with maximum probability for k^{th} word.
END FOR

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Parsing of the Bengali sentence "আমরা ভালো ফুটবল খেলি" (IPA: *amra b^halo futbl k^heli*), English meaning: '*We play good football*') is demonstrated in Figure 3 based on our grammar.

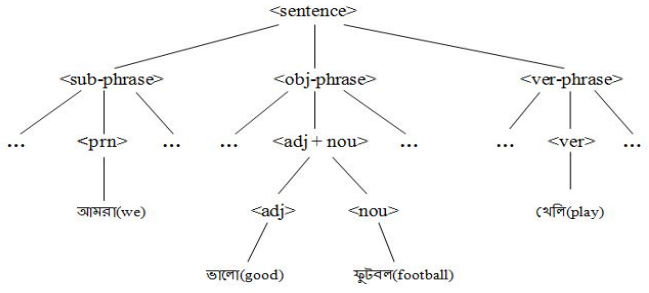

Figure 3: Parsing of the Bengali sentence "আমরা ভালো ফুটবল খেলি"

Figure 4 demonstrates the impact of inputting a syntactically and semantically invalid Bengali sentence "ভালো আমরা খেলে ফুটবল" (IPA: *b^halo amra k^hele futbl*, English meaning: *good we play*

football). Before detecting the POS, the result shows that the sentence is not valid.

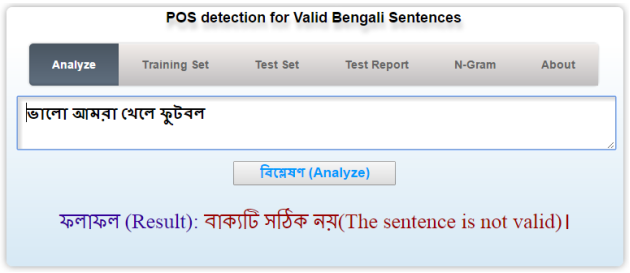

Figure 4: Analyzing syntactically and semantically incorrect Bengali sentence "ভালো আমরা খেলে ফুটবল".

The impact of pragmatically invalid Bengali sentence "পাখি ভালো ফুটবল খেলে" (IPA: *pak^{h}i b^{h}alo futbl k^{h}ele*, English meaning: '*Bird plays good football*') is shown in Figure 5.

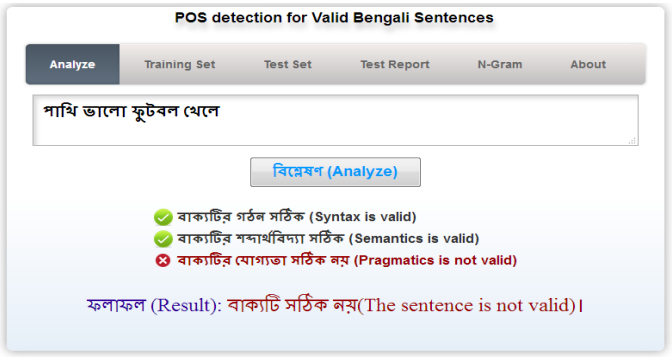

Figure 5: Analyzing pragmatically incorrect Bengali sentence "পাখি ভালো ফুটবল খেলে".

Figure 5 shows that Bengali sentence "পাখি ভালো ফুটবল খেলে" is syntactically and semantically valid, but the sentence is not pragmatically valid. This is due to the fact that, N-gram model returns zero probability for the sentence.

For observing the consequence of a proper sentence in all aspects we put a valid Bengali sentence "আমরা ভালো ফুটবল খেলি" (IPA: *amra b^{h}alo futbl k^{h}eli* , English meaning: '*We play good football*'). The status of the sentence is demonstrated in Figure 6 and the detected POS is shown in Figure 7.

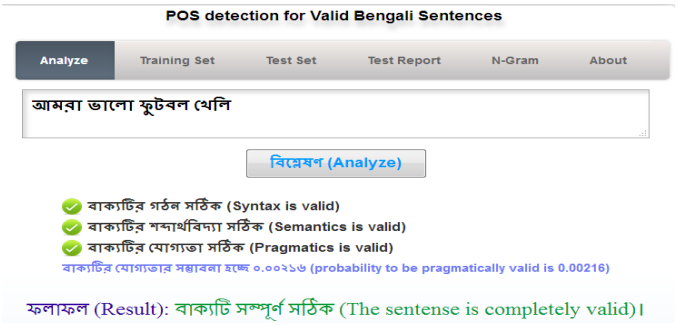

Figure 6: Analyzing with the valid Bengali sentence "আমরা ভালো ফুটবল খেলি"

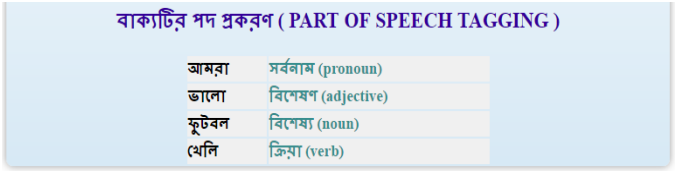

Figure 7: POS detection result from valid Bengali sentence "আমরা ভালো ফুটবল খেলি"

The test result of our proposed method for individual POS is demonstrated in Figure 8 and the result summary is shown in Figure 9. For example, in Figure 8 total বিশেষ্য(nouns(are 4047 and correctly recognized বিশেষ্য(nouns) are 3731. Therefore incorrect recognition of বিশেষ্য (nouns(are 316. Thus the error rate of বিশেষ্য (nouns) detection is 7.81%.

Figure 9 shows that the total experimental words of all types (বিশেষ্য-noun, সর্বনাম-pronoun,বিশেষণ-adjective,ক্রিয়া-verb,পদান্বয়ী অব্যয়-preposition, আরটিকেল-article,সংখ্যা-number, বিশেষণের বিশেষণ-adverb), total recognized words, total incorrectly recognized words, average error rate, and accuracy rate.The figures are 13853, 12668, 1185, 8.55% and 91.45% respectively.

পদসমূহ(Part of Speech)	মোট(Total)	চিহ্নিত(recognized)	চিহ্নিত হয়নি(not recognized)	ভুলের হার (error rate)
বিশেষ্য (noun)	4047	3731	316	7.81%
সর্বনাম(pronouns)	2851	2654	197	6.91%
বিশেষণ(adjectives)	1487	1391	96	6.46%
ক্রিয়া(verbs)	3813	3370	443	11.62%
পদান্বয়ী অব্যয়(propositions)	194	185	9	4.64
আর্টিকেল(articles)	67	67	0	0%
সংখ্যা(numbers)	80	80	0	0%
বিশেষণের বিশেষণ(adverbs)	1314	1190	124	9.44%

Figure 8: POS detection result for the test set using our proposed method.

SUMMERY	
মোট শব্দ(Total words)	: 13853
মোট চিহ্নিত(Total recognized words)	: 12668
ভুল চিহ্নিত(Wrongly recognized words)	: 1185
গড় ভুলের হার (Average error rate)	: 8.55%
সূক্ষ্মতা(Accuracy)	: 91.45%

Figure 9: Summary of POS detection result for the test set using the our proposed method.

On the other hand, the accuracy rate of the probability-based method is low in comparison with our proposed method. The accuracy rate of this method is 82.32%. Summary of the result of our probability based POS detection method is shown in Figure 10.

SUMMERY	
মোট শব্দ(Total words)	: 13853
মোট চিহ্নিত(Total recognized words)	: 11404
ভুল চিহ্নিত(Wrongly recognized words)	: 2449
গড় ভুলের হার (Average error rate)	: 17.68%
সূক্ষ্মতা(Accuracy)	: 82.32%

Figure 10: Summary of POS detection result for test set using probability based method.

V. CONCLUSION AND FUTURE WORK

Our proposed approach exhibits a rule-based POS detection method. The paper compares the rule-based POS detection with a probability based counterpart. Rule-based method [proposed method] shows 91.45% of accuracy with only 8.55% of average error rate, while 82.32% accuracy with 17.68% of average error rate is observed from the probability-based method. Most of the inaccuracies are perceived among verb and adverb detection.

Although our research has reached its goal, there were negligible limitations. One is our proposed method works solely for the Bengali language. Moreover, our proposed method is feeble to identify POS in unorthodox sentences. Authors are working on it and are expecting to elucidate the limitations in near future. Authors are planning to upsurge the size of training and testing set. In addition, authors would like to deal with more complex sentence structures of Bengali language for POS detection and make the grammar more precise in near future to increase the accuracy rate.

REFERENCES

[1] Wikipedia contributors. "Bengali language." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 11 Apr. 2016. Web. 17 Apr. 2016

[2] Sandipan Dandapat, Sudeshna Sarkar, Anupam Basu. 2004. "A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali". In Proceedings of International Conference on Computational Intelligence-2004

[3] Ekbal, A. Bandyopadhyay, S., "Part of Speech Tagging in Bengali Using Support Vector Machine", ICIT- 08, IEEE International Conference on Information Technology, pp. 106-111, 2008Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[4] Hammad Ali (2010), "An Unsupervised Parts-of-Speech Tagger for the Bangla language", Department of Computer
Science, University of British Columbia. 2010.

[5] Wikipedia contributors. "Subject–verb–object." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 8 Apr. 2016. Web. 21 Apr. 2016.

[6] Wikipedia contributors. "Subject–object–verb." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 16 Apr. 2016. Web. 21 Apr. 2016.

[7] Sidorov, Grigori (2013). "Syntactic Dependency-Based *n*-grams in Rule Based Automatic English as Second Language Grammar Correction". *International Journal of Computational Linguistics and Applications*: 169–188.

[8] Merialdo, B. 1994. Tagging English Text with Probabilistics Model. Computational Linguistics, 20(2): 155-172.

[9] Sarkar, Kamal & Gayen, Vivek. (2012). A practical part-of-speech tagger for Bengali. 10.1109/EAIT.2012.6407856.

[10] Nesarul Hoque, Md & Seddiqui, Hanif. (2015). Bangla Parts-of-Speech tagging using Bangla stemmer and rule based analyzer. 440-444. 10.1109/ICCITechn.2015.7488111.