

Syntactic and Semantic analysis of Bangla language for developing grammar checker system

Bibekananda Kundu

CDAC Kolkata

bibekananda.kundu@cdackolkata.in

Abstract

This paper describe about types of error may occur in the time of writing a text and methodology of development of grammar checker to rectify such error and alert the user of every error it detects. In this paper the grammatical and ungrammatical error are described on the basis of Bangla Language.

1. Introduction

A computer conversational system can potentially help a foreign language student to improve his/her fluency. One of its potential roles could be to correct ungrammatical sentences. An important component of such a system is to provide corrections of the students' mistakes and will alert the student of every error it detects. This system may also help professional writer as a proof checker by converting ungrammatical structures to grammatical forms. The system can detect and suggest rectifications for a number of grammatical errors, resulting from the lack of agreement, order of words in various phrases etc., in literary style Bangla texts. The purpose of the system is to find various grammatical mistakes in the formal texts written in Bangla language. While detecting grammatical mistakes, the focus is on keeping the false alarms to minimum. For every detected error, system provides enough information for the user to understand why the error is being marked. This system as a whole can be used as a post editor for a number of applications for Bangla like machine translation, optical character recognition etc., where the output needs to be corrected grammatically before providing the end results. Grammar checking is one of the most widely used tools within natural language engineering applications. Bangla being spoken by more than 200 million people [The Summer Institute

for Linguistics (SIL) Eth-nologue Survey (1999).], no significant work is done on grammar checking of Bangla text. Three methods are widely used for grammar checking in a language; syntax-based checking, statistics-based checking and rule-based checking. Our system will based on hybrid approach, i.e. both rule-based and statistical based checking.

2. Types of Error

Text Error can be categorized in to two main parts:

- a) Language Learning Error: error occurs due to lack of knowledge of the language.
- b) Performance Error: error occurs due to over sight of the user.

In linguistic point of view text error can be categorized in to two main parts:

- a) Structural Error: error occurs for inserting, deleting or moving one or more words.
- b) Non Structural Error: error occurs for replacing an existing word with a different one.

In the time of writing a text error can be occur due to mainly four operations. These are as follows:

- 1) Replacement
- 2) Addition
- 3) Deletion
- 4) Transposition

When error occurs for more than one operation this type of error is know as Composite Error.

A) Morphological replacement:

i) Verb Suffix Replacement: A verb in a sentence may be replaced by same verb with a different form. For example: in the sentence রাম ভাত খাচ্ছিলাম (Itrans: raama bhaata khaachchhilaama) the verb খাচ্ছিলাম (Itrans: khaachchhilaama) is changed from খাচ্ছিলেন (Itrans: khaachchhilen) due to replacement operation.

ii) Case Replacement: case marker associated with pronoun and noun may be replaced. For example: in the sentence এটা কাকারা বই (Itrans: eTaa kaakaaraa ba_i) the suffix রা (Itrans: raa) of the noun কাকা (Itrans:kaakaa) is changed from genitive case র (Itrans: ra) due to replacement operation.

iii) Adjectival suffix Replacement: In the sentence দমাময়ী শিক্ষক আসছেন । (Itrans: daYaamaYii shikShaka aasachhena) the female suffix ময়ী (Itrans: maYii) of the word দয়া (Itrans: daYaa) is changed from male suffix ময় ((Itrans: maYa) due to replacement operation.

B) Similar word or Cohort replacement:

Incorrect Sentence: বলে বাঘ থাকে । (Itrans: bale baagha thaake)

Correct Sentence: বনে বাঘ থাকে । (Itrans: bane baagha thaake)

here বলে (Itrans: bale) and বনে (Itrans: bane) are cohort of each other but বলে (Itrans: bale) is verb and বনে (Itrans: bane) is noun.

Addition Operation:

A) Repeated words:

Bangla: আমি ভাল ভাল ছেলে ।

Itrans: aami bhaala bhaala Chele

English: I am good good boy.

B) Unnecessary words:

Bangla: পরমাণু অনু অপেক্ষা অধিক ক্ষুদ্রতর ।

Itrans: paramaaNu anu apekShaa adhika kShudratara

English: atom is more smaller than molecule

Deletion Operation:

A) Implicit Subject:

Bangla : তোমার মঙ্গল করুন। (Subject : ঈশ্বর is missing here)

Itrans: tomaara ma~Ngala karuna (Subject : iishbara is missing here)

English: May bless you. (Subject: God is missing here)

B) Implicit Verb:

Bangla: তুমি কি এবার মাধ্যমিক পরীক্ষা ? (Verb: দেবে is missing here)

Itrans: tumi ki maadhyamika pariikShaa ? (Verb: debe is missing here)

English: Will you matriculation exam? (Verb: give is missing here)

Transposition Operation:

Incorrect Sentence: থেকে গাছ ফল পড়ে । (Itrans: theke gaachha phala pa.De) Here the Post position থেকে (theke) is placed before noun গাছ (gaachha).

Correct Sentence: গাছ থেকে ফল পড়ে । (Itrans: gaachha theke phala pa.De)

Types of Grammatical Errors:

a) Tense Error:

Example 1:

Bangla: আমি প্রশ্নপত্র পড়ব ও উত্তর দিয়েছিলাম

Itrans: aami prashnapatra pa.Daba o uttara diYechhilaama

English: I will read the question paper and I gave the answer.

Example 2:

Bangla: আমি দরজা খুলছিলাম তখন সে ঘরে ঢুকে পড়েছিল

Itrans: Jakhana aami darajaa khulachhilaama takhana se ghare Dhuke pa.Dechhila

English: When I was opening the door then he entered the room.

Example 3:

Bangla: গতকাল আমি সিনেমা যাব

Itrans: gatakaala aami sinemaa Jaaba

English: Yesterday I will go to Cinema.

b) Person Error:

Bangla: ছাত্ররা নিশ্চয় বিদ্যালয় যদি সে পরীক্ষা দিতে চায় ।

Itrans: chhaatraraa nishchaYa bidyaalaYa Jaabe Jadi se pariikShaa dite chaaya

English: student must goes to school if he wants to appear in the exam.

c) Improper pronoun reference:

Bangla: রবি রামকে সমস্যাটা বলল এবং সে সমস্যাটা সমাধান করল

Itrans: rabi raamake samasyaaTaa balala eba.n se samasyaaTaa samaadhaana karala

English: Rabi tells the problem to Ram and he solves the problem.

Here the reference of pronoun সে (se) is not clear whether it is referring to রবি (rabi) or রাম (raama)

d) Improper use of punctuation:

Example 1:

Bangla: তোমার নাম কি ।

Itrans: tomaara naama ki ।

English: What is your name .

Here the punctuation '।' is used instead of '?' symbol.

Example 2:

Bangla: আমি, দেখলাম সে আসছে । (incorrect use of comma)

Itrans: aami, dekhalaama se aasachhe .

English: I, see he is coming.

e) Sentence Fragment:

Example 1:

Bangla: আমি গান গাইব | যদি তুমি নাচ |

Itrans: aami gaana gaa_iba | Jadi tumi naacha |

English: I will sing, if you dance.

Here single sentence incorrectly fragmented into two single part.

f) Invalid Subject-Verb agreement:

Subject and Verb have to agree with respect to number and person. আমি ভাত খাবেন (*Itrans:* aami bhaata khaabena) is an incorrect sentence because the subject আমি (*Itrans:* aami) is first person non honorific but the person information of the verb খাবেন (*Itrans:* khaabena) is third person honorific. Some time subject and verb agreement differ due to their semantic relation. That is expectation of the verb is not satisfied by the subject. For example in the sentence মানুষ ফুটবল খায় (*Itrans:* maanuSha phuTabala khaaYa) is an incorrect sentence because human (মানুষ *Itrans:* maanuSha) can not eat (খায় *Itrans:* khaaYa) football (ফুটবল *Itrans:* phuTabala).

3. Methodology

Statistical approach can be used on Bangla corpora. The system will work on Transformation-Based Error-Driven learning algorithm [4]. This algorithm has been used in many areas in the Natural Language Processing applications including POS (part of speech tagging), prepositional phrase attachment and word classification. The natively annotated text is compared to the true annotation as indicated by a small manually annotated corpus, and transformations are learned that can be applied to the output of the initial state annotator to make it better resemble the truth. The phrase structure learning algorithm is trained on a corpus of partially bracketed text which is also annotated with part of speech (POS) information. The learning begin in a native initial state, knowing very little about the phrase structure of the corpus. Transformations are learned automatically which transformed the output of the native parser into output which better resembles the phrase structure found in the training corpus. Once a set of transformation has been learned, the system will be capable of changing unstructured sentence to a most probable structured sentence. For checking the correctness of the grammar we simply measure the probability of a sentence using n-gram analysis. We will use part of speech tags rather than individual words for calculating n-gram probability.

The threshold value determines the success and the failure of the system. The determination of the threshold value depends on the empirical test and has to be finalized depending on initial test results.

Rule Based approach is required for checking linguistic grammatical construction as follows:

- Subject-Verb Agreement
- Checking of implicit verb in the sentence

In rule-based checking linguistic rule will be help full for mainly two parts

- 1) Insertion of postposition, conjunct, modals in the valid position of the sentence.
- 2) Insertion/modification of noun/verb inflection: The nouns and verbs, appearing in their uninflected forms in the reduced input, can be substituted by any of their inflected forms. Improper inflection of noun/verb can be modified depending on linguistic rules.

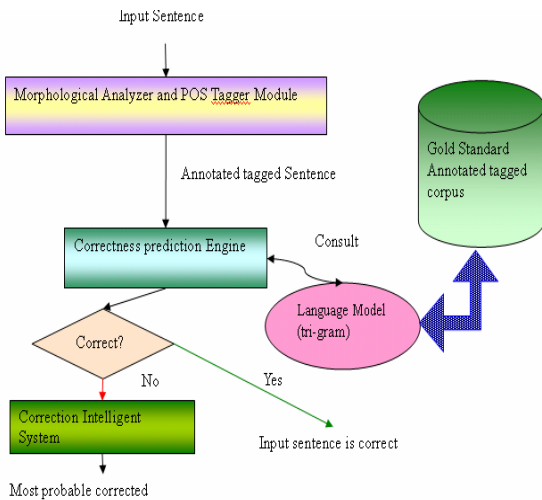
Using the Language Model (tri gram) if the probability of correctness of the given sentence found less than a threshold value then following statistical and grammatical procedure will be applied on that sentence to correct it:

- 1) Transposition of POS Tag for syntactically correction of the given sentence. That means “Noun Adjective” combination can be rewritten as “Adjective Noun” in a given noun phrase.
- 2) Full Cohort Generation:
If a word is correctly spelled by wrongly placed in a sentence then generate the cohort of the word to correct it. For example in the sentence “Taj Mahal is bigger then Kutub Minar”, the word “then” is correctly spelled by wrongly placed in the sentence. So we will generate cohort of “then” and will find out that “than” is the most probable cohort for the given sentence.
- 3) Deletion of repeated similar words:
- 4) Insertion of implicit Determiners, Verb, Preposition/Post Position and Pronoun in the valid syntactic position of the sentence.
- 5) Checking Subject-Verb agreement and modify the verb depending the gender, number and person of the subject.

Checking the Semantic expectation and valency of the verb with the subject and object of the sentence.

For example “I eat water” the verb “eat” expect some solid edible thing such as bread but the ontological

category of object “water” is liquid which does not satisfy the expectation of verb in the sentence. Again if we replace the “water” with “grass” then it may not be expectation of the subject “I” because vegetarian animal such as “Cow”, “Goat” has more expectation of the object “grass” than human. We can use semantic analysis /Ontology for transferring ungrammatical sentence to a grammatical sentence. Overall diagram of the system given below:



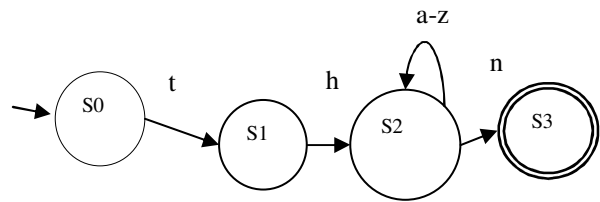
4. Calculation of Cohort Weight

Cohort List Generation: For generating cohort list we can use regular expression.

Say for the incorrect grammatical sentence “Taj Mahal is bigger then Kutub Minar.”, the word “then” is correctly spelled but wrongly placed in the sentence so we can correct the sentence by generating all the cohort of “then” and using maximum probable cohort for this sentence. Now using regular expression we can generate the cohort as follows:

- a) Cohort List by modification:
[a-z]+hen, t[a-z]+en, th[a-z]+n, the[a-z]+
- b) Cohort List by addition:
[a-z]+then, t[a-z]+hen, th[a-z]+en, the[a-z]+n, then[a-z]+
- c) Cohort List by Transposition:
then, hten, tehn, thne,
- d) Cohort List by Deletion:
ten, thn, hen, the

Consider only the valid correct spelled words from the generated cohort list. From the regular expression th[a-z]+n



We get a valid spelled word “than” which will be most appropriate for the sentence. So the correct sentence will be “Taj Mahal is bigger than Kutub Minar.”

For selecting the most appropriate cohort from the cohort list we will consider the following steps:

- i) Calculate degree of similarity to the word actually typed: if cohort is same as actual type word then cohort weight will be 1 else weight will be less than 1 say 0.888
- ii) Calculate degree of fit in the given sentence context: measurement of likelihood bound between the tags of each cohort member and the tags of the words before and after.
- iii) Frequency of usage in General Language.
- iv) If the cohort member occurs in a grammatical idioms or preferred collocation with surrounding words, then relative weight is increased.
- v) Domain dependent lexical preference of cohort

Cohort weight = multiplication of the weight from generated by step i) to step v).

Most appropriate cohort will be the cohort with maximum weight.

5. Rules and Exception

Rule for detecting simple question type sentence where '?' punctuation not present:

Rule: if in a simple sentence S, there exist one or zero finite verb VBF and only one question word WH (Such as के, कि, कब, कितना) then place '?' punctuation at the end of the sentence.

Exception:

- 1) কে ঐ সব চিঠির লেখিকা তাও সে জানত ।
Itrans: ke aara paatra dekhabe ।
English: who is the writer of these letters that he also knew.

Rule for tense checking:

Rule for simple sentence: if time adverbial ADV and finite verb VBF belong to sentence S then time of the adverb and the finite verb will be same.

$ADV(wi), VBF(wj)_S \Rightarrow$
 $Time(ADV(wi)) = Time(VBF(wj))$

Example:

গতকাল আমি বাজারে গিয়েছিলাম।
Itrans: gatakaala aami baajaare giYechhilaama
English: Yesterday I went to market.

Here $Time(ADV(গতকাল(Itrans: gatakaala)))$ and $Time(VBF(গিয়েছিলাম(Itrans: giYachhilaama)))$ are same. Both indicate past tense.

Rule for complex sentence: If in a complex sentence two individual sentences S1 and S2 is connected by correlative conjunct CRC the tense of the finite verb of S1 will be same as tense of the finite verb S2.

Example:

যখন তুমি যাবে তখন আমি যাব ।
Itrans: Jakhana tumi Jaabe takhana aami Jaaba
English: When you will go then I will go.

Here $S1 =$ তুমি যাবে (*Itrans: tumi Jaabe*) and $S2 =$ আমি যাব (*Itrans: aami Jaaba*) and $CRC =$ { যখন (*Itrans: Jakhana*), তখন (*Itrans: takhana*) } $Tense(VBF_S1(যাবে (Itrans: Jaabe))) = Tense(VBF_S2(যাব (Itrans: Jaaba)))$

Both indicate future tense.

6. Conclusion

In this paper, author have described common types of error that may occur in the time of writing a text and a methodology to solve some of the errors using hybrid approach (both rule-base and statistical approach). Some rules and exception are also described in this paper for better understanding of scope and limitation of this system.

7. Reference

[1] Md. Jahangir Alam, Naushad Uzzaman, and Mumit Khan. 2006. N-gram based Statistical Grammar Checker for Bangla and English. In Proc. of ninth International Conference on Computer and Information Technology (ICCIT 2006), Dhaka, Bangladesh.

[2] Eric Atwell and Stephen Elliott, Dealing with ill-formed English text, The Computational Analysis of English, Longman, 1987.

[3] Daniel Naber. 2003. A Rule-Based Style and Grammar Diplomarbeit Technische Fakultät, Checker. University Bielefeld, Germany. (Available at: http://www.danielnaber.de/language/tool/download/style_and_grammar_checker.pdf (1/10/2007))

[4] Brill, Eric .1993. Automatic grammar induction and parsing free text: A transformation-based approach. In proceedings, 31st Meeting of the Association of Computational Linguistics, Columbus, OH.

[5] Atwell, Eric Steven (forthcoming b) "Transforming a Parsed Corpus into a Corpus Parser", to appear in Proceedings of the 1987 ICAME 8th International Conference on English Language Research on Computerised Corpora, Helsinki, Finland