Parsing Bengali Text - an Intelligent Approach

Goutam Kumar Saha

Scientist-F, Centre for Development of Advanced Computing, (CDAC), Kolkata
Mail to: CA- 2 / 4 B, Baguiati, Deshbandhu Nagar, Kolkata 700059, INDIA sahagk@gmail.com,

Abstract

This paper demonstrates how a proposed parser identifies part of speech (POS) of Bengali lexicons in Bengali (or Bangla) sentences along with annotating semantics information. A parser is used in machine translation (MT) for identifying the part of speech in a sentence. We call it an intelligent parser as it handles semantics as well as POS identification. The aim of such rule-based intelligent Bangla parser is to ease the task of handling semantic issues in the subsequent stages in machine translation.

Keywords: Bangla Text Parsing, Semantics Parsing and Disambiguation.

1. Introduction

Syntax deals with the structure of natural language sentences. A parser produces a structural description on applying a given computational grammar on the input sentence. In general, a parser is expected to identify structural and thematic relations in the sentence. The assignment of structural description to words depends upon the grammar that is, a description language and a set of structural constraints. A parser attempts to analyze the sequence of symbols presented to it based on the grammar [1,2,3,4]. Intelligent parsing technique through semantic analysis using tagged lexicons and word's major class (on ontology) proves to be helpful for more accurate machine translation. Human language like Bangla is very rich in inflections, vibhakties (suffix) and karakas, and often they are ambiguous also. That is why Bangla parsing task becomes very difficult. At the same time, it is not easy to provide necessary semantic, pragmatic and world knowledge that we humans often use while we parse and understand various Bangla sentences. Bangla consists of total eighty-nine part-of-speech tags. Bangla grammatical structure generally follows the structure: subject-object-verb (S-O-V) structure. We also get useful POS information from various inflections at morphological parsing. We describe here an intelligent parsing algorithm for Bangla sentences that relies on semantic analysis (through Ontology and rules) of a Bangla sentence during the task of parsing. Bangla consists of eighty-nine parts-of-speech tags [6]. Intelligent parsing is very important for machine translation [5,7] also.

2. Intelligent Approach to Bengali Parsing

In Bangla we do not have concept of small or capital letters. Unlike English, every letter in a Bangla word is capital only. Thus we find difficulties in understanding whether a word is a proper noun or not. For example, the Bangla word "BISWAS" can be a proper noun (i.e., a family name of a person) as well as an abstract noun (with the meaning of *faith* in English). For example, in order to understand the following Bangla sentence, we must need an intelligent parser. A parser takes the Bangla sentence as input and parses every sentence according to various rules of parsing like:

Sentence (S) ⟶ **<Noun Phrase (NP)> < Verb Phrase (VP) >**

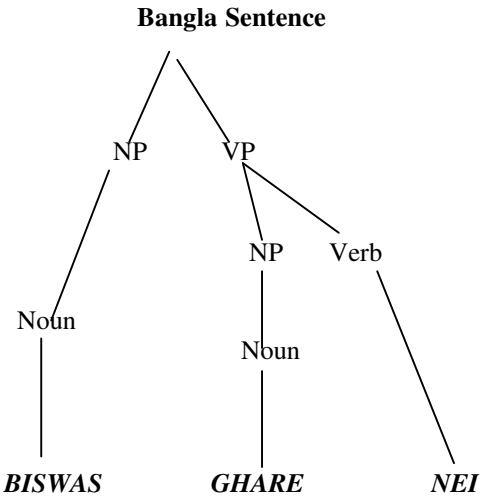
NP ⟶ **<Article> < Noun >**

VP ⟶ **< NP> <Verb>**

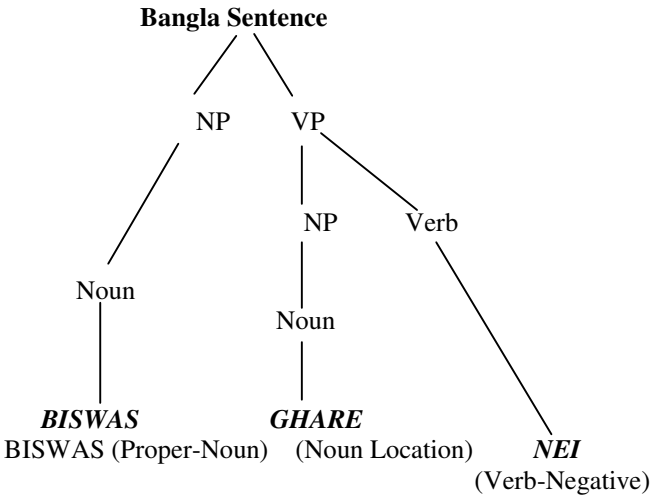
NP ⟶ **<Noun / Pronoun>**

The value of log- likelihood- ratio also gives us information on disambiguation for unsupervised one. *BISWAS GHARE NEI*. (in English: *Biswas is not at room*) is parsed in both the conventional as well as in intelligent ways.

The most frequently used Bangla word "BHALO" has also POS and Word Sense disambiguation like: *good* (adjective), *well* (adverb) and *goodness/benefaction* (abstract noun). Ram Ekta (a) Bhalo (good-adjective) Chhele (boy) / Ram Bhalo (well-adverb) Khele (i.e., Ram plays well) / Ramer Bhalo (benefaction-abstract noun) Karo (i.e., Do for Ram's goodness).

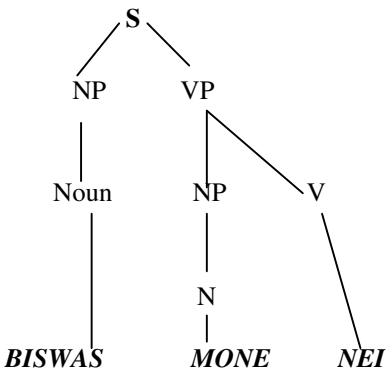


(A Conventional Parsing)

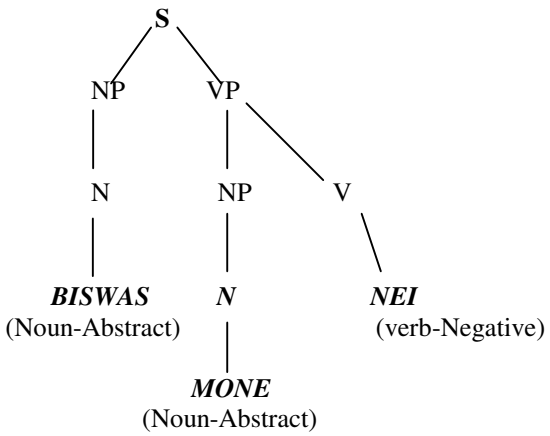


(An Intelligent Parsing)

Similarly, for the Bangla sentence "***MONE BISWAS NEI"*** *(*in English: *No faith is in mind)* both the ways of parsing are shown below.

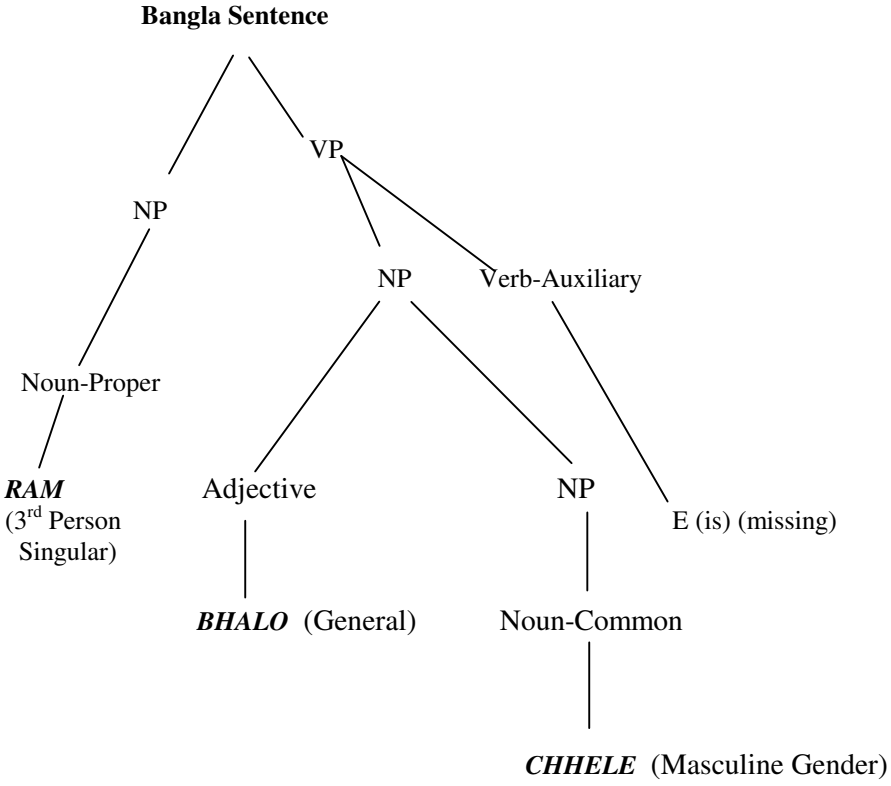


(A Conventional Parsing)



(An Intelligent Parsing)

Again in Bangla language often we do not use verb. For an another example, the parsing of the Bangla sentence (with missing verb say denoted by *E*) *RAM BHALO (good) CHHELE* (boy) (in English, Ram is a good boy) is shown below.



3. Intelligent Bangla Parsing Algorithm

The following steps explain how the proposed intelligent parser works.

Step 1. Start Parsing
Step 2. While $i <=$ Number of words in the Bangla Sentence
Step 3. Start annotating POS to a Bangla word
Step 4. If a word is not in Lexicon, assign it as Noun-Proper.
Step 5. If POS-Ambiguity, resolve it by Bangla grammar rules.
Step 6. If word-sense-ambiguity, apply N-gram model to assign correct word-sense. (Resolving word-sense ambiguity using Bi or Tri gram i.e., scanning 2-3 surrounding words for knowing an ambiguous word by its surrounding words.)

Step 7. End While
Step 8. Repeat step 2 through 7 until the end of sentences in text.
Step 9. End

4. Conclusion

The proposed intelligent Bangla parser concentrates not only on the syntactic structure (according to Bangla grammar using SOV structure) but also on the semantics issues. While parsing, we believe semantic issues are more important than syntactic ones in human languages (for example, Bangla), which are not rigidly structured. We used vibhakties (e.g. *ke* for noun/pronoun, *chhe* for verb etc.) or suffix for understanding a word's parts of speech also. The proposed approach is equally applicable for other human languages also that are loosely structured.

References

1. D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," Proc. 33rd Annual meeting of the ACL, Cambridge, MA, USA, pp 189-196, 1995.
2. D. Yarowsky, "Decision list for lexical ambiguity resolution: Application to accent restoration in Spanish & French," Proc. 32nd Annual meeting of the ACL, Las Cruce, NM, 1994.
3. Robert C. Berwick, et al., *Principle Based Parsing: Computation and Psycholinguistics,* Kluwer Academic, Boston, 1991.
4. D. Jurafsky and J. Martin, *Speech and Language Processing,* Pearson, 2003.
5. Goutam Kumar Saha, "English to Bangla Translator: The BANGANUBAD," International Journal-CPOL, Vol. 18 (4), pp.281-290, USA, 2005.
6. Goutam Kumar Saha, et al, "Computer Assisted Bangla Words POS Tagging," Proc. International Symposium on Machine Translation NLP and TSS (iSTRANS-2004), New Delhi, pp 248-251, 2004.
7. Goutam Kumar Saha, "The E2B Machine Translation: A New Approach to HLT," ACM Ubiguity, Vol. 6(32), 2005, ACM Press, USA.

Author's Biography

Goutam Kumar Saha has been working for last eighteen years as a computer scientist in renowned R&D organisations. He is presently working as a Scientist-F in the Centre for Development of Advanced Computing, CDAC, Kolkata, India. His past employer is LRDE, Defence R&D Organisation, Bangalore, India. He has authored many international research papers on Fault Tolerant Computing and NLP. He is a reviewer for IEEE Magazine, AMSE J (France), IJCPOL and CSI Journal. He is a fellow in IETE, Senior member in CSI, IEEE. He is an associate editor of the ACM Ubiquity. He can be reached also via gsaha@acm.org, sahagk@gmail.com.