# A Bangla Semantic Parser Using Context-Free-Grammar

**Conference Paper** · September 2017

DOI: 10.1109/CTCEEC.2017.8455103



**4 authors:**

M. Firoz Mridha Ph. D.
Bangladesh University of Business and Technology (BUBT)
**58** PUBLICATIONS **107** CITATIONS
SEE PROFILE

Hanif Bhuiyan
The Commonwealth Scientific and Industrial Research Organisation
**11** PUBLICATIONS **25** CITATIONS
SEE PROFILE

Shammi akhtar Shammi
University of Asia Pacific
**4** PUBLICATIONS **11** CITATIONS
SEE PROFILE

Dr. Aloke Kumar Saha
University of Asia Pacific
**27** PUBLICATIONS **54** CITATIONS
SEE PROFILE

Proceedings of International Conference on Current Trends in Computer, Electrical, Electronics and Communication" (ICCTCEEC)
08-09 September, 2017, Karnataka, India

# A Bangla Semantic Parser Using Context-Free-Grammar

M. F. Mridha, Hanif Bhuiyan, Shammi Akhtar and Aloke Kumar Saha
Dept. Of Computer Science and Enginerring
University of Asia Pacific, Dhaka, Banglaesh
mdfirozm@yahoo.com

*Abstract*—This research work describes a computer system for understanding the parsing of Bangla sentences. It draws on recent developments in Natural Language Processing (NLP) research to look at the past, present, and future of NLP technology in a new light. The research work of Bangla Language Processing (BLP) was started in late1980s in Bangladesh and it already produced some substantive results. NLP is a theory-motivated range of computational techniques for the automatic analysis and representation of human language. Here the researchers present a semantic parser for analyzing Bangla language semantics applying Bangla Lexicon. Semantic parsing augments the stratum of comprehension of NLP than the syntactic parsing, that is primarily indulges in untangling the syntactic ambiguities of the words. In this work, the researchers have tried to develop a Context-Free-Grammar (CFG) for semantic parser. It is the grammar that consist rules with a single symbol on the left-hand side of the rewrite rules. The researchers have tried to make a rule based grammar using CFG. Tokenize of sentence is extensive deed of this research. According to every token, the researchers present their pronunciation and morphological attachment. Finally, the researchers have analyzed Bangla language semantics applying Bangla lexicon.

**Keywords—Semantic Parser; Context-Free-Grammar; NLP; Morphological Analysis; Bangla Language Processing.**

## I. INTRODUCTION

Natural Language Processing is a field of computer science, artificial intelligence and linguistics concerned with the interaction between computer and human. It is an ambit of experiment and application that explores how computers can be used to fathom and manipulate natural language text or speech. However, in order to peruse this language verbally or literally in machine a media is necessary, which performs as an interpreter and complier of a machine. This media between human and machine can be named semantic parser. Semantic parsing is a process, which represents the natural language into a specific format with keeping the original meaning. In order to do such semantic operation parser performs the semantic analyzing on the text. The analyzing is performed by a bunch of tokens that determines the grammatical structure of the sentence as well as the appropriate meaning of the sentence through a vast database. Usually, a semantic parser deals between the grammatical structure and relations of the words in a sentence. Because the meaning of any given corpus is actually the relation and interaction between the entities. In addition, it uses the knowledge sources and grammatical rules to determine the responsibilities of the words in the text Sentence. It is very important for unambiguous representation of a sentence in stochastic parsing. For English language a lots of works have done [1] and some showed impressive tasks. However, there have been very few research endeavors regarding Bangla language due to lack of language processing mechanism.

In terms of speaking in language, Bangla language is the seventh. Almost 210 million people in the world speak in this language. Therefore, such a mechanism like semantic parsing which can used to semantically map the Bangla language would be so worthy. Few researchers had been done regarding this task [2, 3] but still have some of untapped issues to solve. For an example, "সে খাট খায়" which means in English "He eat bed". But in a syntactic manner it has no fruitful meaning although it is grammatically correct. Therefore, to overcome such issues a Bangla semantic parser which seems conceivable to determine the meaningful sentence as well as the grammatical structure also. To perform such semantic mapping of Bangla language at first we need a Bangla semantic parser. As English is the communication language worldwide. In world business, science and government-related matters and entertainment all site is comfortable in English language. There is high communication gap for non-English speakers. The gap which is exists between English speakers and the information and culture in other people. The Bangla semantic parser can make the bridge between native and others speaking people can get most profit out of it. Moreover, the impact of Bangla computing is not restricted to socioeconomic in future. It will also essay the Bangla language unity on the global period. At present most of the researcher are trying to develop a machine translation system for Bangla language which is in the developing stages but they still contain limitations. The researchers also try to develop a Semantic Parser for Bangla Language Processing (BLP). For this reason, the researchers try to build a most powerful semantic parser in Bangla language. In this paper, the researchers present a Bangla Semantic parser which consists of analyzing process for language sentence input both

**978-1-5386-3243-7/17/$31.00 ©2017 IEEE.**

syntactically and semantically. A set of semantic rules are used to ensure the grammatical structure and semantic meaning of the sentence and a Bangla WordNet is used to untangle the words' ambiguity. The approach works in three steps, first determine the meaning of the words in the sentence, second confirm the grammatical structure of the sentence through the grammatical rule and finally justify the meaningful relation between the words to verify the meaningful sentence. The goal of Bangla semantic parser is to make a process that will analyze, understand, and create languages which humans use naturally, so that in the last it will be able to address the computer as if it were addressing another person. Since text can hold knowledge at many different granularities, from simple token –based representations, to rich high –level logical representations across document collections, this research seeks to work at the right level of analysis for the applications concerned. This research work, implements a technique to parse Bangla sentence in a new approach.

The organization of this paper is as follows: In Section 2, we describe the Literature Review, Section 3 has the short description about Semantic Parser, Section 4 depicts the proposed framework; Section 5 demonstrates our Results. Finally, Section 6 draws the curtains by concluding amid some heeds towards our future work.

## II. LITERATURE REVIEW

A significant number of research works have been done to analyze the Bangla Language (one of the widely used language in the world) in order to understand the exact meaning and information. In order to decipher the meaning of a Bangla sentence an effective and efficient Bangla semantic parser is required. Some extensive and comprehensive works regarding this Bangla parse has been done where syntactic parsing, Bangla Wordnet, lexicon anaysis, ontology mapping and so on were studied as these are the prime factors for generating the parser of Bangla language [4,5,6]. Moreover, some works have reported significant progress on Morphological analysis on Bangla language based on semantic concept [7,8,9]. However, in terms of other language several works have done, Haniewicz et al. developed an algorithm for sentence based on semantic structure on polish language [10]. Furthermore, a throughout inspection was accomplished over the hybrid mechanism of merging rule-based and statistical technique into Hybrid Machine Translation (HMT) approach [11] which is suitable for automatically translating morphologically rich and syntactically different languages that abide the Subject Object Verb (SOV) order. Context Free Grammar (CFG) helps to generate sentence structure by performing some statistical methods using Parts Of Speech (POS) tagger. Additionally, predictive Parser and construction of the parse table [12] to recognize Bangla Grammar were summed up. Using these references, ideas about Bangla Grammar for morphological and semantic analysis were applied in order to prepare Bangla WordNet and morphological rules.

## III. BANGLA SEMANTIC PARSING

Semantic parsing is a way by which a sentence is tokenize and assigned with a suitable structure according to the meaning. The analyses of the Bangla syntactic and semantic structure of an input sentence use the Bangla grammar analysis rules and the parsing technique. Bangla sentence is the input of our semantic parser, which produces an annotated parse tree by applying semantic relationship among the words. A parser split the token into smaller elements using a set of rules, which describe its structure, and sequence of tokens to identify its grammatical structure which is defined by native formal grammar. Semantic representation provides simple information about the grammatical relationships, which we can easily understand and effectively implement by people without any specific language proficiency, those who want to gain textual relations. The relationships among the words in the sentence are represented uniquely as semantic relations [13, 14, 15]. We have proposed our own parser for obtaining the tag sets and context-free format grammar representation for the source structure. The pronouns, adverbs, singular, nouns, plural, persons, verb, tenses, adjectives etc are stored in a database.

## IV. PROPOSED FRAMEWORK

The proposed framework consists of three steps. The input of the system is a Bangla sentence. The preprocessing module tokenizes the sentence. After that, the parsing algorithm is applied on all the tokens to determine the semantic meaning of the words using word dictionary. Then reconstruct the sentence in several forms and send it to the parser. The processed sentences are passed through the parser using the stated CFG rules. The details description of the module is given in fig. 1.

### *A. Preprocessing*

The input sentence should be in Bangla. As, we are concerning about Bangla language so any other language will not be accepted. After getting the input the sentence the following changes are performed:

- Except Bangla language all kinds of language (e.g. Hindi, English, Portugese etc.), date (e.g. ৩/৫/১৬, ২/৫/১৬ etc.), numbers (২, ৩, ৫ etc.), special characters (e.g. @,&,% etc.) বাংলাযুক্তবর্ণ (Complex symbol) will be removed from the sentence.

- All kinds of punctuation will be remove from the sentence like commas (,), periods (এ. বি. এম. তাজুল থেকে এ বি এম তাজুল etc.), and honorific (মিঃ থেকেমি, ডাঃ থেকেমি etc.).

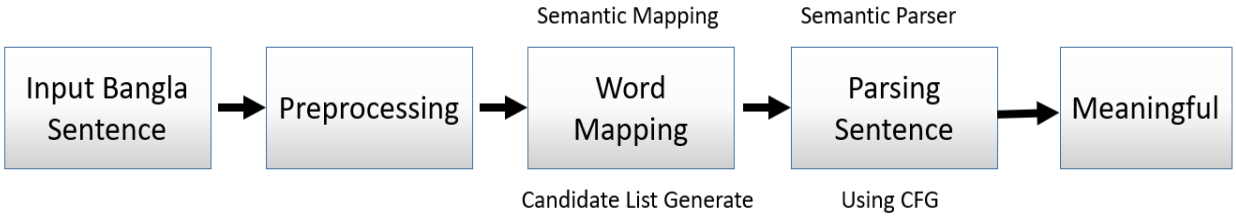


Fig. 1. Proposed System Workflow

### B. *Word Mapping*

This section outline in details to determine the semantic meaning of the sentence using word dictionary and the proposed CGF rules. In order to understand the word meaning a robust linguistic database (Table 1) is devised where each word dictionary represents Bangla language of lexical identifiers to denote word information related to each word or word sense. Although any natural language processing of a particular language is difficult and may be quite far-reaching, often the bulkiness of the experience about the sense of word rather than the token constraints of the word. Therefore, word analysis tends to be very small, and lots of words, although it may be different significantly in sense, will have the similar lexical description.

TABLE I. BANGLA WORD DICTIONARY

| Word | Root | Synonyms |
|---|---|---|
| খাওয়া(eat) | খা | খাইতেছে, খাইতেছি, খাই |
| যাওয়া(go) | যা | যাইতেছে, যাইতেছি, যাই |
| পাওয়া(get) | পা | পাইতেছে, পাইতেছি, পাই |
| দেওয়া(give) | দে | দিতেছে, দিতেছি, দেই |
| নেওয়া(take) | নে | নিতেছে, নিতেছি, নেই |

Once the preprocessing of the text is done then the proposed Bangla_parsing_algorithm is applied. The detail description of the algorithm is given below.

**Bangla_Parsing_Algorithm**
Input: Text, **x (a Bangla sentence)**
Word_databse, Ü = { U1 …. Un}
Output : string, meaningful sentence or not meaningful

```
1   Semantic_word_mapping (x)
2   Generate candidate set1: ∂1={w1,w2,.....wn}
3   Identify_word(w)
4           root(w) and If w € Ü then return w
6           else return Ø
7    end if
8   Generate candidate set1: ∂2={w1,w2,.....wn}
9   synonym(w)
10  root (w) and ifw € Ü thenreturn all synonyms of w
11  else return Ø
12  end if
13  Generate all possible sentences {s1,s2.....sn}
14  Semantic_parsing(s)
15  if all sentences parse coreectlty
16          print meaningful sentence
17    Else
18          print not meaningful sentence
19   end if
```

For mapping the word according to the stated dictionary first tokenize the sentence and thus make the first candidate set $\partial_1$={w1,w2,.....wn}. All the words are converted to its root form and then check in the database and if available then keep that word. Afterward, for making second candidate list apply the Bangla post tagging on the sentence and determine the verbs. Before marking the verbs all the verb words are converted into its root form. Then make the second candidate list$\partial_2$= { w1,w2……wn}. After that we look for the similar word of the verb word in the dictionary. Because, usually in Bangla sentence verb is a word for what actually the meaning of the sentence can be changed. Therefore we mark out all the possible synonyms of the verbs and then reconstruct the sentence. For example, someone type "আমি ভাত খাইতেছে" so in here the root is খাওয়া and synonyms could be খাইতেছে, খাইতেছি, খাই etc. So according to the proposed process the new possible sentences are "আমি ভাত খাইতেছে/ আমি ভাত খাই" etc.

## V. RESULT ANALYSIS

In the proposed method, the researchers have taken Bangla sentence as input and when end character is found then the method compares every word of sentence with the word dictionary entries. The corpora were automatically tagged with our selected word tag. the researchers have tested more than 10000 sentences and our tested output contain approximate 40000 token words, among the tested words the frequency of noun (Bisseo in Bangla) is 16250, pronoun (Shorbonum in Bangla) is 2420, verb (Kria in Bangla) is 12750, adjective (Bisason in Bangla) is 8630 and adverb (Abboy in Bangla) is 1400. The output generated by our system is given in table 2 below:

TABLE II. NUMBERS OF CATEGORIZED WORDS

| Podh (Part of Speech) | Total Words |
|---|---|
| Bisseo (Noun) | 16250 |
| Shorbonum (Pronoun) | 2420 |
| Kria (Verb) | 12750 |
| Abboy (Adverb) | 1400 |
| Bisason (Adjective) | 8630 |

When more sentences are tested and more rules will be added, then percentage accuracy will be increased. In the proposed system Bangla sentence is taken as input and when terminator symbol is found then the system compares each word of sentence with the word dictionary entries. If sentence is matched with the defined rule then information will display parsing output of the given sentence. If the word found having multiple tags then the system always try to search the most appropriate rule for that word. By using sets of rule, system detects the most appropriate word according to the meaning of the sentence. If the word is not available in the database, then the parser will not display any output. The outputs of tested sentences are given in fig. 2 below:

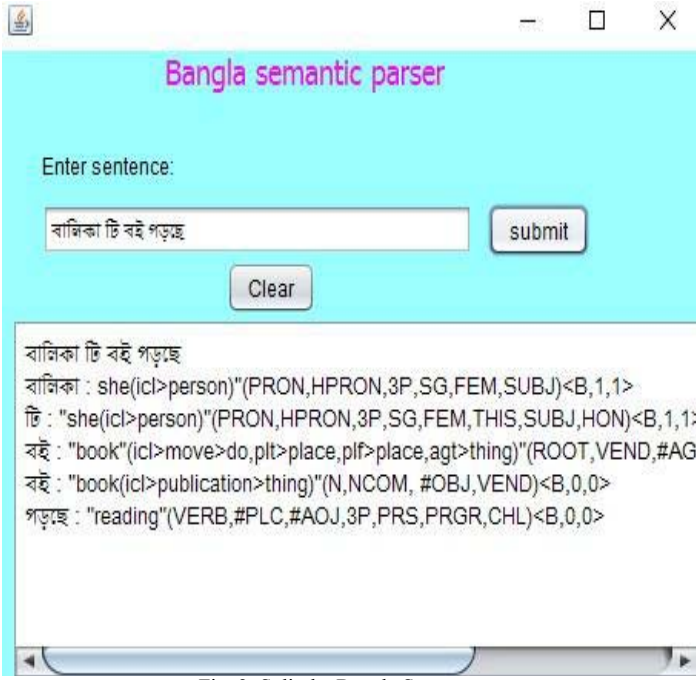

Fig. 2. Split the Bangla Sentence

TABLE III. ACCURACY CALCULATION OF SENTENCES

| No of Tested Sentences | No of Correct matched Sentences | Accuracy (%) |
|---|---|---|
| 10000 | 9157 | 91.57% |

Testing is performed with 10000 sentences and the accuracy of different stage of output is calculated. In our implementation we have resolved the different ambiguity and tag to each token in a Sentence was achieved an accuracy approximately 91.57% which is shown in table 3.

## VI. CONCLUSION

In this work, CFG based Bangla semantic parser has been developed that can be used as a lexical resource for NLP tasks for Bangla language. This Bangla semantic parser is generated from our database. In this research the researchers interrupt the Bangla sentence among the tokens is find out by using semantic representation and grammar and also generate a parse tree. The principal goal was to design a parser that is capable of accepting all types of Bangla sentences. This research works are as follows:

i) We have designed a parser module for Bangla Language Processing.
ii) Entry of different sentences of Bangla language in a lexicon.
iii) We have applied context-free grammar to parse the sentence of Bangla language.
iv) Output of the parser module represented as a list that can be represented and manipulated very easily.

Modifying Context-free grammar rule, the researchers have designed a more powerful parser for Bangla language. The main challenge of Bangla language parsing is the unambiguous representation of a sentence. Moreover, the researchers have solved the ambiguity of words by using the semantic rule. The accuracy of 91.57% was achieved from our proposed parser. During the experiment, it has been scrutinized that the accuracy was getting diminutive when the researchers have tested ambiguous words and sentences furthermore. In future this accuracy will be increased when more problems of ambiguities will be solved.

## *References*

[1] Sangeetha, J., S. Jothilakshmi, and Devendra Kumar. "An Efficient Machine Translation System for English to Indian Languages Using Hybrid Mechanism." International Journal of Engineering & Technology (0975-4024) 6.4 (2014).

[2] Hasan, K. M., Al-Mahmud, Amit Mondal, and Amit Saha. "Recognizing Bangla Grammar using Predictive Parser." International Journal of Computer Science & Information Technology (IJCSIT), Vol. 3, No. 6, Dec 2011.

[3] M. F. Mridha, Molla Rashied Hussein, Md. Musfiqur Rahaman, Jugal Krishna Das "A Proficient Autonomous Bangla Semantic Parser for Natural Language Processing", ARPN Journal of Engineering and Applied Sciences, VOL. 10, NO. 15, AUGUST 2015,ISSN 1819-6608, pp 6398-6403.

[4]Vijay Kumar, Pankaj K. Sengar(2010), "Segmentation of Printed Text in Devanagari Script and Gurmukhi Script", International Journal of Computer Applications Volume 3 – No.8, pp. 0975 –8887.

[5] Javed Ahmad Mahar, Ghulam Qadir MEMON(2010), "Rule Based Part of Speech Tagging of Sindhi Language", International conference on Signal Acquisition and processing.,pp.101-106.

[6] Pawan Goyal, Vipu Arora, Laxmidhar Behera (2009), "Analysis of Sanskrit text: Parsing and Semantic Relation", Springer-Verlag Berlin Heidaelberg., pp. 200-218.

[7] Ms Vaishali M. Barkadeet. al. (2010), "English to Sanskrit Machine Translation Semantic Mapper", International Journal of Engineering Science and Technology vol.2(10).

[8] Ms Vaishali M. Barkade. Prof. Prakash R. Devale, Dr.Suhas H. Patil(2010) , "English to Sanskrit Machine Translator Lexical Parser and Semantic Mapper", National Conference On" Information and Communication Technology"(NCICT-10).

[9] Khaled Shaalan (2010), "Rule-based Approach in Arabic Natural Language Processing", International Journal on Information and Communication Technologies, Vol. 3, No. 3,.

[10] Haniewicz K., Kaczmarek M., Adamczyk M., Rutkowski W.: Polarity lexiconfor the polish language: Design and extension with random walk algorithm. In:Advances in Systems Science, pp. 173–182, Springer, 2014.

[11] Kędzia P., Piasecki M., Orlińska M.: Word Sense Disambiguation Based on Large Scale Polish CLARIN Heterogeneous Lexical Resources. Cognitive Studies, (15), pp. 269–292, 2015, http://dx.doi.org/10.11649/cs.2015.019.

[12] Kobus C., Yvon F., Damnati G.: Normalizing SMS: are two metaphors better than one? Proceedings of the 22nd International Conference on Computational Linguistics,vol. 1, pp. 441–448, Association for Computational Linguistics, 2008. 41

[13] Haniewicz K., Kaczmarek M., Adamczyk M., Rutkowski W.: Polarity lexicon for the polish language: Design and extension with random walk algorithm. In: Advances in Systems Science, pp. 173–182, Springer, 2014.

[14] Beaufort R., Roekhaut S., Cougnon L.A., Fairon C.: A hybrid rule/model-based Finite-state framework for normalizing SMS messages. In: Proceedings of the 48thAnnual Meeting of the Association for Computational Linguistics, pp. 770–779, Association for Computational Linguistics, 2010.

[15] Buczynski A., Wawer A.: Shallow parsing in sentiment analysis of product reviews. In: Proceedings of the Partial Parsing workshop at LREC, vol. 2008, pp. 14–18, 2008.