Probabilistic Approach of Parsing Bengali Sentences

Conference Paper · December 2018

DOI: 10.1109/WIECON-ECE.2018.8783000

2 authors:

Ayesha Khatun
Chittagong University of Engineering & Technology
20 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Moshiul Hoque
Chittagong University of Engineering & Technology
72 PUBLICATIONS **210** CITATIONS
SEE PROFILE

Probabilistic Approach of Parsing Bengali Sentences

Ayesha Khatun, Mohammed Moshiul Hoque
Dept. of Computer Science & Engineering
Chittagong University of Engineering & Technology
Chittagong, Bangladesh
Email: ayeshankhatun@gmail.com, moshiulh@yahoo.com

Abstract— **Statistical parsing denotes the technique of syntactical analysis that is based on probabilistic inference from the corpus. In order to remove the structural ambiguity of sentences, a probabilistic technique of parsing Bengali sentences is proposed in this paper. In this paper, we use the CYK algorithm for parsing and apply binarization technique to improve the parsing efficiency. For this purpose, a set of probabilistic context-free grammars (PCFGs) is proposed based on intonation of the sentences and structures which are used for Bayesian inference. The proposed system is tested with 3500 sentences (10648 words) and achieved 85% overall accuracy for all kinds of Bengali sentences.**

Keywords— Bangla language processing, probabilistic context-free grammar, binarization, Bayesian inference.

I. INTRODUCTION

People speak lots of distinct things, what the way people are used to saying things also have some regularity and structure but in the parsing of sentences, we face many irregularities as well as ambiguities. The main goal of parsing within linguistics tries to construct the most likely parse of a sentence to avoid ambiguity which is called statistical parsing [1]. The crucial use of it is not only to solve the disambiguation problem but modern parser also help for understanding many natural linguistic tasks including thematic role labeling, question answering, text summarization and importantly used in machine translation [2]. The connection to grammatical rules with the probability which is learning from a large corpus, this is the main idea of probabilistic parser. In among all the chart parser algorithm, CKY is very popular for parsing efficiency, as a result in this model we used the probabilistic version of the CKY algorithm. To provide accurate probability to the grammar Bayesian Interface is used in PCFG, with the help of those PCFGs as well as dictionary, CKY algorithm help to find the most accurate parse tree. In this proposed model. Having a huge variety of sentence structure and ambiguities, language processing of Bangla is a challenging task. A statistical model can play an important role in it. For language modeling also in Bangla machine translation system, the statistical parser can be used.

In natural language processing, the parser can be categorized into three types; rule-based, generalized and statistical based parser. For parsing sentences when the grammatical rules are recursively applied then it defines as ruled based parser. And in this case, many ambiguities can arise. To resolve this ambiguity we need to write large and complex grammar rule which is really a very difficult task, on the other hand, the statistical parser can easily detect the ambiguity by training a large corpus. The traditional parser help to construct the syntax trees, where finding the highest probable parse is the main task of the statistical parser. A propabilistic model for parsing all kinds of Bengali sentences automatically is proposed in this paper. And for achieving this purpose PCFG is introduced to parse sentences according to intonation as well as to assign probability effectively, Bayesian algorithms are used. We applied left binarization technique to increase parsing efficiency. Golden standard method is applied to calculate the efficiency of the proposed system and it is successfully shown that this dynamic model using the CYK algorithm can parse different types of Bengali sentences accurately.

II. RELATED WORK

A tremendous amount of research work has been done in the area of statistical natural language processing. Michael Collins developed three types of lexicalized and generative structure for statistical parsing [3] which play an extremely powerful role in Statistical NLP. In Bangla Language processing, a large number of works has been done on the rule-based technique [4]. While an approach of Bengali grammar pattern by using shift-reduce parse is described in [5]. An architecture for translation the text between two types of Bangla local language is designed here [6], and CYK algorithm has been presented in [7]. Many research is done on ruled-based approach, however, only a few research has been found in a statistical-based approach. The algorithm of pobabilistic left corner is used for parsing Bengali sentences in [8, 9], where probabilistic CYK parsing algorithm is applied for parsing Bengali sentences in [10]. This method only can parse Bangla sentences according to the structure and the parsing evaluation is traditional. Current research activities are more intensive on parsing with traditional way, not in probabilistic way. The statistical NLP has very high research activities in other languages, for that reason this paper proposed a probabilistic approach of parsing Bengali sentences which help to parse sentences according to the structure as well as intonation. To get the more accurate probability to generate PCFG, this paper used Bayesian Inference and also evaluated the parser with F-score measurement technique which helped to give the quite satisfactory result.

III. PROPOSED FRAMEWORK

There are five modules in our proposed system with input and output representation and they are the Syntax analyzer, Rule generator, Statistical parser, Error Handler, and Lexicon. The schematic diagram of statistical parsing of Bangla sentences is presented in Fig. 1.

A. Syntax Analyzer:

The syntax analyzer is a program module which takes a sentence as split the input into distinct words, this simple which produce small element is called Token [2]. Firstly, we take an example of simple sentence, "একটি ছোট ছেলে মাছ ধরছে (ekti choto chele mas dhorche)" and the output of syntax analyzer will be presented as TOKEN= ekti (একটি), choto (ছোট), chele (ছেলে), mas (মাছ), dhorche(ধরছে).

B. Dictionary:

The dictionary is simply a database of strings with their corresponding parts of speech(POS) tag and probabilities [11]. It will check every word before placing it in the parse tree and if the grammatical structure is unknown or the input word does not contain in the database then it will create an error message. According to the input sentence, the lexicon will be shown as

ekti (একটি) → APR [0.4]; choto (ছোট) → AP [0.14]; chele (ছেলে) → N [0.4]; dhorche (ধরছে) → V [0.06]

here *ekti (একটি)* is a token and this token has POS tag which is a specifier (SPR). Dictionary or lexicon of Bangla sentences is shortly illustrated in Table I.

C. Rule Generator:

The important module is the Rule generator which is generates grammar rules of Bengali sentences shown as the set of a context-free grammar (CFG) [12]. Statistical parser needs grammar rules which is probabilistic, as a result we developed sophisticated probabilistic context-free grammar (PCFG) rules. Bangla PCFG rules go to the Statistical Parser module as input.

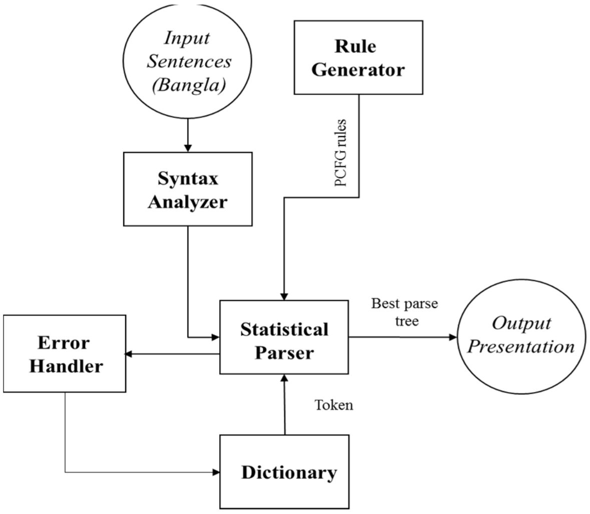


Fig. 1. A simple schematic diagram of parsing Bengali sentences with propabilistic approach.

1) A probabilistic context-free grammar (PCFG): The probabilistic context-free grammar is the technique of assigning the propability to each and every grammatical rules. Let G= (T, N, S, P) be the context free grammar, where terminal symbols is presented as T, which is a finite set. Where, finite nonterminal symbol is N, the start symbol shown as S and set of productions is P which is finite. The production A→ BC or A → ω, where A, B, C∈N and ω∈ T. The PCFG (G, θ), G and theta is a pair where G represent CFG and theta is a real-valued vector of length |R| and θ > 0 where all nonterminals

$$A \in N, \sum_{A \to \beta \epsilon P} \theta_{A \to \beta} = 1$$

Where β used as a variable ranging on $(N \times N) \cup T$ [13].

2) Bayesian inference for PCFG: Given a sequence of string terminals $w = (w_1, w_2, w_3 \ldots\ldots\ldots w_n)$ which is produced by a known context-free grammar G infer the production probability theta and offer to apply Bay's rule. The probability will be, $P(\theta|w) \propto P_G(w|\theta)P(\theta)$, *where*

$$P_G(w|\theta) = \prod_{i=1}^{n} P_G(w_i|\theta)$$

The joint posterior distribution over t and θ can be calculated and marginalized over t with

$$P(t, \theta|w) \propto P(w|t)P(t|\theta)P(\theta)$$
$$= P(\theta)(\prod_{i=1}^{n} P(w_i|t_i)\, P(t_i|\theta))$$

Where t is denoted a sequence of parse trees for w, here a small part of Bengali PCFGs and lexicon according to intonation like assertive, interrogative, imperative etc [14, 15] are represented in Table I. The majority part or the real view of the table has very low probability.

TABLE I. A SMALE PART OF PCFG AND LEXICONS OF THE BENGALI LANGUAGE

Rank	**Probabilistic CFG**	
	PCFGs and Lexicon	***Probability***
1	S → AS | IRS | IS | ES	0.65 | 0.2 | 0.25
2	AS → NP VP	1.00
3	IRS → NP VP	0.50
4	IS → NP VP	0.06
5	ES → NP VP	0.18
6	ES → INJ NP	0.06
7	INTJ → INTJ PN	0.2
8	NP → N	0.32
9	NP→ PRO	0.55
10	NP → NP NP	0.13
11	NP → SPR AP N	0.85
12	NP → WH AP	0.05
13	NP → NP WH	0.10
14	NP → AP NP	0.93
15	NP → NP PN	0.07
16	VP → V	0.92
17	VP → NP VP	0.08
18	VP → AP V	0.84
19	VP → V Ind	0.16
20	VP → VP PN	0.44 | 0.19 |0.37
21	VP → V Ind PN	0.6 | 0.4
22	N → hoimonti (হৈমন্তী) | korim (করিম) | polish (পুলিশ) | rajniti (রাজনীতি) | shasti (শাস্তি) | nam (নাম)	0.0023|0.273|0.078|0.082|0.045 |0.033

Rank	Probabilistic CFG	
	PCFGs and Lexicon	***Probability***
23	PRO → ami (আমি) | tumi (তুমি) | she (সে) | tara (তারা) | uni (উনি) | tui (তুই)	0.231|0.252|0.237|0.091|0.052|0.123
24	AP → sobuj (সবুজ) | valo (ভালো)	0.089 | 0.071
25	WH → ke (কে)| ki (কি) | kothay (কথায়) | kivabe (কিভাবে) | keno (কেন)	0.294 | 0.171 | 0.155 | 0.142 | 0.238
26	Ind → na (না) | ni (নি)	0.599 | 0.401
27	PN → ? | ! | ,	0.198 | 0.092 | 0.710
28	V → khay (খায়) | lekhe (লিখে)| jay (যায়)	0.0977 | 0.0743 | 0.0421
29	INJ→ ah (আহ) | aha (আহা)	0.000006 | 0.000034

D. Statistical Parser:

The statistical parsing mechanism of the proposed model is to use rule generator module for PCFG and with the help of lexicon and lexical analyzer, it produces the most probable parse tree.

1) Binarization: Conversion of an n-ary grammatical rules into binary grammar which is an equivalent and this is the most important part in parsing for bringing out an o(n3) time complexity [16]. For all kinds of chart parser such as CYK, Early parser etc are need to convert CFG into biranry form. The binarization technique not only optimizes the parser computational cost of tabular parser but also help to speed up the parsing action. Here we binarize the CFG explicitly into CNF form which is required by the CKY algorithm. An example of simple sentence [17] is "একটি ছোট ছেলে মাছ ধরছে (ekti choto chele mas dhorche)" and the PCFG of the following sentence is *NP → SPR AP N [0.85].*
Where specifier, adjective and Noun are represented as SPR, AP and N. After applying left binarization the grammatical rule will be

NP→ @SPR_AP N [0.85]
@SPR_AP → SPR AP [1.00]

Here we use left binarization technique which is select the left two pair and this technique will not affect the probability of CFG.

2) Probabilistic CYK algorithm: The CYK algorithm [2] is a polynomial bottom-up chart parser, where a table ($n \times n$) used to record analysis for substrings of a sentence $s = (w_1, w_2, w_3 \ldots\ldots\ldots w_n)$. The complexity of CYK is $O(n^3)$ where n represents the length of the string to be parsed. This paper used CYK parsing algorithm which is the probabilistic version of CYK and this algorithm can parse the sentences dynamically, help to produce the highest probable parse of a sentence.

E. Output:

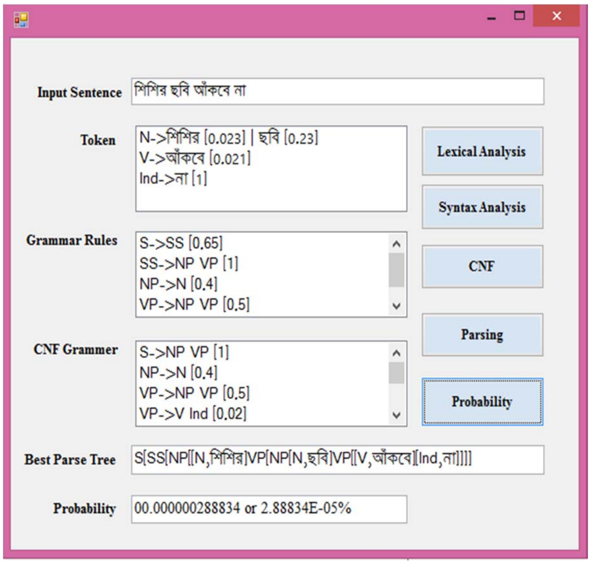
Finally, the output of the proposed system is come out and it is simply the highest probable parse tree which is represented as labeled based structure. If the left binarization technique is used in any grammatical rule, the output will also define that in GUI for the following input. the highest probability is *1.65E-8%* and the labeled based structure of the parse tree is.

$$S[NP[[SPR,\text{একটি}][@AP\_N\begin{bmatrix}AP,\\ \text{ছোট}\end{bmatrix}\begin{bmatrix}N,\\ \text{ছেলে}\end{bmatrix}VP[NP\begin{bmatrix}N,\\ \text{মাছ}\end{bmatrix}VP\begin{bmatrix}V,\\ \text{ধরছে}\end{bmatrix}]]$$

IV. Experimental results

A. Implementations

For implementing purpose the system is used MS SQL server to save the corpus, MS visual studio held to develop GUI and C# is the programming language. Siyam Rupali and Shonar Bengali fonts are used. Our proposed model is designed to parse assertive, interrogative, imperative and exclamatory. the corpus is collected from textbooks, blogs, the popular newspaper, novels and literatures of Bangladesh and proposed system is tested with different types of Bengali sentences.



Input Sentence: শিশির ছবি আঁকবে না

Token:
N->শিশির [0.023] | ছবি [0.23]
V->আঁকবে [0.021]
Ind->না [1]

Grammar Rules:
S->SS [0.65]
SS->NP VP [1]
NP->N [0.4]
VP->NP VP [0.5]

CNF Grammer:
S->NP VP [1]
NP->N [0.4]
VP->NP VP [0.5]
VP->V Ind [0.02]

Best Parse Tree: S[SS[NP[[N,শিশির]VP[NP[N,ছবি]VP[[V,আঁকবে][Ind,না]]]]

Probability: 00.000000288834 or 2.88834E-05%

Fig. 2. Probabilistic parsing of Bangla assertive sentences.

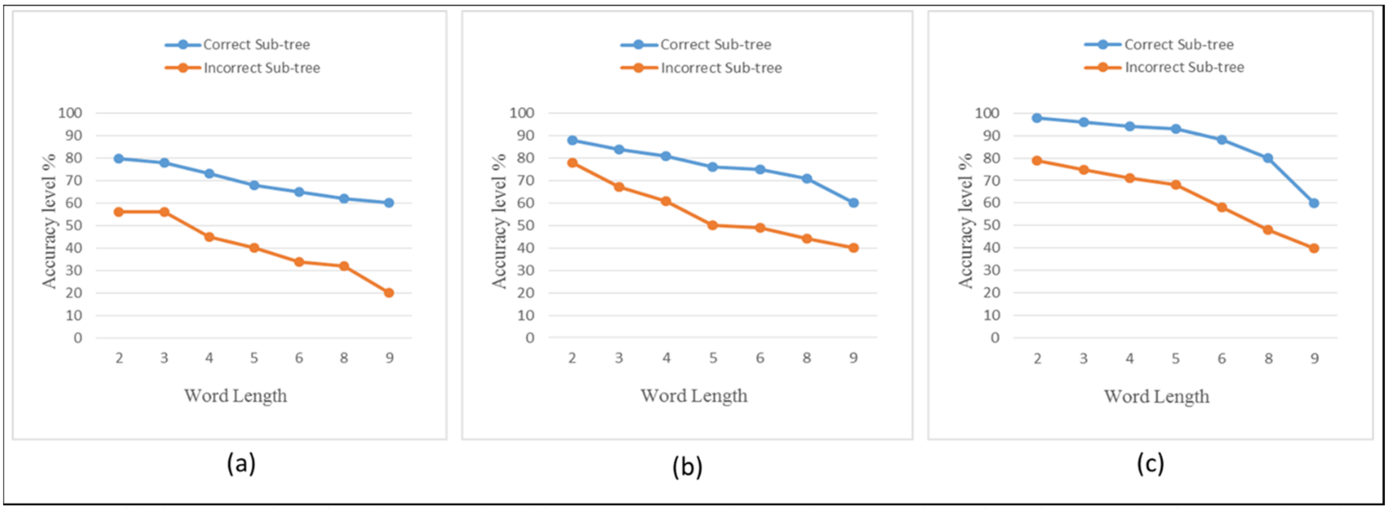
In following figure we shown assertive sentences and for doing this we use the special symbol in PCFG. With the help of algorithm, it produced the best output. The input sentence "shishir chobi akbe na (শিশির ছবি আঁকবে না)". Firstly it shows the tokens with related probability, then the PCFG of following sentences. After that, it will show the CNF form of grammar and finally, the best parse tree with the probability of input sentence will be shown as a output.

$$S[SS[NP[[N,\text{শিশির}]VP[NP\begin{bmatrix}N,\\ \text{ছবি}\end{bmatrix}VP\left[\begin{bmatrix}V,\\ \text{আঁকবে}\end{bmatrix}\begin{bmatrix}Ind,\\ \text{না}\end{bmatrix}\right]]$$

This is the output of the system as a labeled based parse tree and probability of following sentence is *2.88E-05%*, the implementation of the operation shown in Fig. 2. Interrogative, imperative and exclamatory sentences are also implemented in this way.

B. Results

For evaluating the parser and grammar we used PARSEVAL measures [2]. In this paper, three basic measure are shown, Label precision, label recall, and F-measure. F-measure or F-score is a measurement technique of determining test accuracy of parser and grammar in statistical analysis. If the constituents in a hypothesis parse tree and a reference parse tree have the same starting point, ending point and non-terminal symbol, then sub-tree type labeled as correct, otherwise incorrect. As the diversity of assertive rule is more as a result in Fig .3, Lots of ups and down result can be shown in accuracy level and average 75% F-score found in correct sub-tree on the other hand the average accuracy level in interrogative and imperative sentences are about (85 to 90)%. But the evaluation of incorrect sub-tree is not pleasant to all types of sentences. Figure shown that the imperative sentences accuracy are better than the assertive and interrogative sentences. If we take more and more constitute grammar rules and lexicons into our existing model, we can improve the system performance.



(a)

(b)

(c)

Fig. 3. Graphical representations of constituency parsing with respect to correct sub-tree and incorrect subtree of Bangla (a) assertive (b) interrogative and (c) imperative sentences

V. Conclusion

The fundamental part of NLP and machine translation is sentence parsing. This model can take a vital part in Bengali machine translation system. The main part of the proposed system is using Bayesian method of assign probability effectively to parse the sentences including assertive, interrogative, imperative and exclamatory in a probabilistic way. For detecting the ambiguity of Bangla sentences more accurately the proposed model can be used, For measuring the parser accuracy we used standard PARSEVAL technique and the average F-score of the system is about 85%. We can improve the accuracy of the system by enlarging the corpus size. There is also a tremendous amount of research scope has in statistical NLP of Bangla, as a very small amount of work has been done. Statistical parsing with lexicalized PCFG can be a future extension of this work as well as stochastic semantic parsing can be extended.

References

[1] C. D. Manning and Schutze, *Foundations of Statistical natural language processing*, The MIT Press, 2001, pp. 381-423.

[2] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, USA: Prentice-Hall, 2007.

[3] M. Collins, "Three generative, lexicalised models for statistical parsing", In Proc. *8th Conference on European Chapter of the Association for Computational Linguistics*, 7 July 1997.

[4] M. M. Hoque and M. M. Ali, "A Parsing Methodology for Bangla Natural Language Sentences", In Proc. *International Conference on Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, pp. 277-282, 2003.

[5] R. Z. Rabbi, M. I. R. Shuvo, K. M. A. Hasan, "Bangla Grammar Pattern Recognition Using Shift Reduce Parser", *International Conference on Informatics Electronics & Vision (ICIEV)*, 2016.2006.

[6] S. Chakraborty, A. Sinha., S. Nath, (2018) "A Bengali-Sylheti Rule-Based Dialect Translation System: Proposal and Preliminary System." Proceedings of the *International Conference on Computing and Communication Systems. Lecture Notes in Networks and Systems*, 2018, vol 24. Springer, Singapore.

[7] S. Dasgupta, A. Wasif and S. Azam, "An optimal way of machine translation from English to Bengali", In Proc. *7th International Conference on Computer and Information (ICCIT)*, pp. 648-653, 2004.

[8] M. A. Karim, M. Kaykobad and M. Murshed, *Technical Challenges and Design Issues in Bangla Language Processing*, USA: IGI global, 2013.

[9] M. M. Hoque, M. O. Faruk, M. M. Hasan, M. K. Hassan and M. M. U. Karim, "An empirical framework for statistical parsing of Bangla sentences", *Computer Science & Engineering Research Journal*, vol. 04, pp. 29-38, 2006.

[10] A. Khatun, M. M. Hoque, "Statistical parsing of Bangla sentences by CYK algorithm", In Proc. *International Conference on Electrical, Computer & Communication Engineering*, February 16-18, 2017, Cox's Bazar, Bangladesh.

[11] P. P. Purohit, M. M. Hoque and M. K. Hassan, "An empirical framework for semantic analysis of Bangla sentences", *The 9th International Forum on Strategic Technology (IFOST)* , Oct. 2014.

[12] M. N. Hoque and M. H. Seddiqui, "Bangla Parts-of-Speech tagging using Bangla stemmer and rule based analyzer", In Proc. *18th International Conference on Computer and Information Technology (ICCIT)*, Dec. 2015.

[13] M. Johnson, T. Gritfiths and S. Goldwater, "Bayesian Inference for PCFG's via Markov Chain Monte Carlo", In Proc. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 139–146, 2007, Rochester, New York.

[14] M. S. Arefin, M. M. Hoque, M. O. Rahman and M. S. Arefin, "A Machine Translation Framework for Translating Bangia Assertive, Interrogative and Imperative Sentences into English", *2nd Int'l Conf. on Electrical Engineering and Information & Communication Technology (ICEEICT) 2015*, Jahangirnagar University, Dhaka-1342, Bangladesh, May 22-23, 2015.

[15] M. S. Arefin, L. Alam, S. Sharmin and M. M. Hoque, "An Empirical Framework for Parsing Bangia Assertive, Interrogative and Imperative Sentences", *1st International Conference on Computer & Information Engineering*, November 26-27, 2015, Rajshahi, Bangladesh.

[16] X. Song, S. Ding and C. Y. Lin, "Better binarization for the CKY parsing", In Proc. Empirical methods in NLP, October 25-27, 2008, Honolulu, Hawai.

[17] L. Hwng, H. Zhang, D. Gildea and K. Knight, "Binarization of Bynchronous context-free grammars", 2009 Association for computational linguistics, vol. 35, pp. 559-595, Dec 2009.