Statistical parsing of Bangla sentences by CYK algorithm

Conference Paper · February 2017

DOI: 10.1109/ECACE.2017.7912986

CITATION
1

READS
351

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

Statistical Parsing of Bangla Sentences by CYK Algorithm

Ayesha Khatun
Dept. of Computer Science & Engineering
Chittagong University of Engineering & Technology
(CUET), Chittagong, Bangladesh
ayeshankhatun@gmail.com

Mohammed Moshiul Hoque
Dept. of Computer Science & Engineering
Chittagong University of Engineering & Technology
(CUET), Chittagong, Bangladesh
moshiulh@yahoo.cm

***Abstract*—Statistical parsing is the task of enabling the parser to find the most probable parse of a sentence according to probabilistic context-free grammar. Crucial use of statistical parser is to solve the disambiguation problem. This paper proposes a statistical parser using probabilistic version of Cocke-Younger-Kasami (CYK) algorithm to parse different kinds of Bangla sentences. For improving parsing efficiency, this model also uses left binarization technique to grammar. Rule probability and word probability is used to generate different probabilities for the same structure of a sentence. Experiment results with different kinds of sentence shows the effectiveness of the propose parser with reasonable accuracy.**

Keywords— Statististical parsing, probabilistic context-free grammar, rule generator, Chomsoky normal form, binarization.

I. INTRODUCTION

Natural language sentences are ambiguous by nature and sentences have multiple parses. A statistical model is a systematic platform which assigning the score to the parse trees and chooses the one which has height score. The score is defined in term of probabilistic value. The syntactical ambiguity is a crucial problem for parsing, it is very difficult to manually define a grammar whose rules find out only one parse from an exponential number of possible parses and probabilistic model provide a well-established method for selection between the alternatives [1]. The concept of the statistical parser is related to probabilistic rules learning from a corpus text. The probability of the parse tree is calculated by multiplying the probabilities of all words and grammatical rules, those grammatical rules related to creating a parse tree. The statistical parser is a dynamic programming technique which makes able to parse most probable parse tree of a sentence. This proposed model used the probabilistic version of CYK [2] algorithm for statistical parsing of sentences dynamically. Bangla language processing is a very difficult task because of variation of sentences and many ambiguities occurred during parsing a sentence. In order to produce accurate parse structure of Bangla sentences, statistical parsing would be the best choice. The most crucial use of the statistical parser is in Bangla machine translation system. The statistical parser is help to translate Bangla to other language more accurately. It will also play an important role in the modeling of language for Bangla.

Parsing can be mainly classified into three categories, rule-based, statistical based and generalized parsers. The production rules are recursively applied in the rule-based parsing, as a result many ambiguities may arise. To detect or resolve ambiguities, it is very difficult task to write a complex grammar rule. The statistical parsing resolves ambiguity with the help of experience or by training corpus. The traditional parsing methods are used to find correct parse tree where statistical parser help to find the best parse tree using statistical information.

Bangla has a rich historical and cultural background. To keep Bangla history, culture, Literature existent and to introduce it globally we have to digitalize Bangla language. A statistical model can play a vital role for this purpose. Bangla is the fourth largest language of the world and in spite of having over 245 million native speakers, still now Bangla language have a negligible amount of work on statistical natural language processing.

In this paper, a statistical approach is proposed to parse different kinds of Bangla sentences statistically. To achieve the goal, stochastic or probabilistic context-free grammar is introduced. As we used dynamic CYK algorithm, ambiguity can easily detect. For increasing parsing efficiency, this model used left binarization technique. To parse more accurate structure, this model considered the rule probability as well as word probability. The proposed system is evaluated by the wide range of sentences with changeable word lengths and show that this probabilistic parser can parse Bangla sentence successfully.

II. RELATED WORK

Many works have done on the statistical parsing of English sentence. Sampson proposed the first Stochastic or probabilistic approach to parse sentences in [3]. A series of five statistical model of translation is described in [4]. Michael Collins developed a statistical parser [5] which plays a tremendously powerful role in NLP. Statistical parsing system based the Penn Wall Street Journal Treebank proposed in [6]. However, parsing is done based on the statistical decision for complex grammar is shown in [7]. Some work has been performed on the parsing methodology of Bangla. A parsing methodology of top down and bottom up parsing approach by using phrase structure rules to parse Bangla simple sentence in [8]. While a rule based

978-1-5090-5627-9/17/$31.00 ©2017 IEEE

intelligent Bangla parser has proposed in [9] that controlled semantic issues in machine translation. An automatic naming word identifier for Bangla sentence is introduced in [10], however, a stochastic language model is used to automatic Bangla word prediction in [11]. Dependency grammar for Bangla is considered in [12]. Bangla predictive parser based on top-down parsing method to avoid the left recursive factoring is presented in [13]. CYK algorithm has been presented in [14]. Many researches are done in Bangla Language Processing [15]. However, very few activities have been done on statistical parsing and PCFG of Bangla grammar. Statistical parsing of Bangla sentences using left corner (LC) parsing algorithm has proposed in [16]. It also modified the LC parsing algorithm to add the probabilistic features. This method used rule probability, not the word probability; as a result, same structure of sentences provided the same probability. This mechanism was also restricted to parse statistically the Bangla simple sentences. Complex, compound sentences are represented by semantic manner in [17] but for probabilistic semantic parsing need a strong statistical model. Existing works are more concentrated on deterministic parsing, not probabilistic or statistical parsing. However, many improvements occurred in statistical natural language processing in other languages. In this paper, we proposed a statistical approach of parsing different kinds of Bangla sentences using probabilistic version of CYK algorithm.

III. PROPOSED FRAMEWORK

Proposed system mainly contains five modules with input and output representation. This statistical parser is take a Bangla sentence as an input, with the help of other modules it finds out the most probable parse tree based on assigning the probability to grammar rules. Fig. 1 represents the schematic diagram of statistical model of our proposed system.

A. Input Sentences:

A parser needs an input sentence for processing, as the source language is Bangla, so a Bangla sentence is taken as input to the proposed model. For example, firstly we choose a simple sentence, "goru sobuj ghas khay (গরু সবুজ ঘাস খায়)" as an input.

B. Lexical Analyzer:

Lexical Analyzer is simply a function which receives a sentence as an input and broke into distinct words usually called Token [2]. It is a fundamental part of any kind of parser. For further use, tokens are collected in the list. For existing input sentence, the output of scanner will be represented as

TOKEN = goru (গরু), sobuj (সবুজ), ghas (ঘাস), khay (খায়).

C. Lexicon:

A lexicon is a database just like a dictionary of words that contains appropriate parts of speech (POS) tags [17] and probabilities. The words are taken from Bangla training corpus. We calculate the probability of lexicon using maximum likelihood estimation. How many time a word come within a training corpus, is helping to determine the word probabilities of words. The module can generate an error message if any word not existent in the current database. For following input sentence, lexicons will be shown as

goru (গরু) → Noun (0.6); sobuj (সবুজ) → Adjective (0.8);

gash (ঘাস) →Noun (0.4); khay (খায়) → Verb (1.0)

An example, goru (গরু) is a token or lexicon which parts of speech is noun (N) and probability of being goru (গরু) is a noun is 0.6. Lexicon with probability example for is illustrated in Table I.

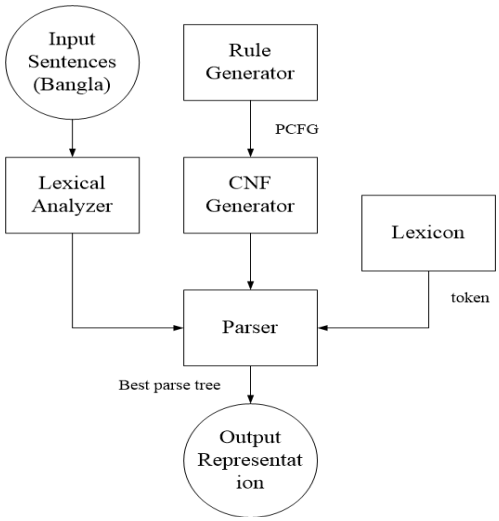

Fig. 1. The Schematic diagram of statistical model of proposed system.

D. Rule Generator:

Rule generator generates grammar rules, represented by the set of a context-free grammar (CFG). CFG rules says, how the parts of speech can put together to make a well-constructed grammatical sentence [18]. As we choose statistical approach, we have to construct a sophisticated probabilistic context-free grammar rules.

1) A probabilistic context-free grammar (PCFG): A probabilistic context-free grammar is the most natural probabilistic model for generating tree structure and it is the simplest model too [19]. A PCFG consists of CFG and parameters [2] which is defined by four tuples G=(N,∑,S,R). Where G is representation of grammar, N is a set of non-terminal symbol, ∑ is a set of terminal, R is a set of rules or productions, where each of the rule from $A \rightarrow \beta\ [p]$, here A is a nonterminal and p represente the probability of the production and β is a strings of symbol , and S express as the start symbol. Total probability of all possible expansion of non-terminal must be 1 [20].

$$\sum P(A \rightarrow \beta) = 1$$

2) Deriving a PCFG from a Bangla corpus: Treebank [6] is help to get the PCFG rule probabilities. The Tree-bank of the proposed system is simply collection of Bangla parse trees. Parse trees are generated from 2025 diffenet type of sentences and those sentences are collected from Bangla blogs, news papers, literatures etc. We constructed it using different types of Bangla sentence including simple, complex, compound. The

computation of probability of each expansion of non-terminal according to a treebank is define over [20] :

$$P(\alpha \rightarrow \beta|\alpha) = \frac{Count(\alpha \rightarrow \beta)}{\sum Count(\alpha \rightarrow \gamma)} = \frac{Count(\alpha \rightarrow \beta)}{Count(\alpha)}$$

This is a standard technique of estimate the probability of sentences. Here $Count(\alpha \longrightarrow \beta)$ means the number of times $\alpha \longrightarrow \beta$ rule is comes in Bangla training corpus and $Count(\alpha)$ means the number of non-terminal α is comes in the training corpus. This equation helps to determine the probability of each grammar rules and lexicon. A simple part of PCFG of the Bangla grammar and lexicon is represented in Table I.

TABLE I. A SMALE PART OF PCFG OF THE BANGLA GRAMMAR AND LEXICON

Rank	Probabilistic Context-Free Grammars	
	Rules	***Probability***
1	S → SS | CS | COMS	0.65 | 0.2 | 0.25
2	SS → NP VP	1.0
3	NP → N	0.5
4	NP → NP NP	0.06
5	NP → AP NP	0.18
6	NP → SPR AP N	0.06
7	NP → PRO	0.2
8	VP→ V	0.32
9	VP → NP VP	0.55
10	VP → AP VP	0.13
11	COMS → SS CONJ SS	0.85
12	COMS → SS CONJ CS	0.05
13	COMS → CS CONJ CS	0.10
14	CS → DC IC	0.93
15	CS → IC DC	0.07
16	DC → (SUBORD) SS	0.92
17	DC → NP (SUBORD) VP	0.08
18	IC → (SUBCOM) SS	0.84
19	IC → NP (SUBCOM) VP	0.16
20	PRO → ami (আমি) | tumi (তুমি) | she (সে)	0.44 | 0.19 |0.37
21	N → goru (গরু) | ghas (ঘাস)	0.6 | 0.4
22	V → khay (খায়)	1.0
23	AP → sobuj (সবুজ) | valo (ভালো)	0.8 | 0.2
24	SPR → ekti (একটি)	1.0
25	CONJ → ebong (এবং) | kintu (কিন্তু)	0.6 | 0.4
26	SUBORD → jodi (যদি) | je (যে)	0.7 | 0.3
27	SUBCOM → tahole (তাহলে) | she (সে)	0.7 | 0.3

E. CNF Generator:

For statistical parsing using CYK, a PCFG should be written in Chomsky normal form (CNF) [19]. CNF generator simply converts PCFG into an equivalent grammar in CNF. An example of CNF conversion is given in the following. Consider a simple grammatical rule

S → SS (0.65);

SS → NP VP (1.0)

After CNF conversion grammar will be, S → NP VP (0.65). A conversion of probabilistic context-free grammar into equivalent CNF is represented in Table II.

TABLE II. A CONVERSION OF PROBABILISTIC CONTEXT-FREE GRAMMAR INTO CNF.

Rank	Chomsky Normal Form (CNF)	
	Rules	***Probability***
1	S→ NP VP	0.65
2	NP→ NP NP	0.06
3	NP→ AP NP	0.18
4	VP→ NP VP	0.55
5	VP → AP VP	0.13
6	NP→ গরু (goru)	0.3
7	NP→ ঘাস (ghas)	0.2
8	VP→ খায় (khay)	0.32
9	ADJ→ সবুজ (sobuj)	0.8

F. Statistical Parser:

The definition of statistical parser is that, it is a parser which takes a sentence as an input and finds out most probable parse tree. Like any other parser, we need a well-defined parsing algorithm. This model used probabilistic version of CYK algorithm for parsing Bangla sentences statistically. This proposed parser can parse binary grammar for better parsing efficiency.

1) Binarization: The binarization is a method of transforming an n-ary grammar into an equivalent binary grammar and it also affect the efficiency of the CYK parser [21, 22]. All types of the tabular parsing algorithm like CYK need to construct their grammar rule into binary branching form. It is necessary to convert grammar into equvalent binary grammar as a result the parser will generate a binary tree. The Binarization increase the efficiency of chart parser it also helps to speed up the parsing operation. Fig. 2 represente the binary tree after applying left binarization technique. An example of compound sentence is "she kal asbe ebong ami jabo(সে কাল আসবে এবং আমি যাব)", then PCFG is COMS → SS1 CONJ SS2 (0.85). Left binarization is a technique of selecting the pair of left two, it will not affect the probability of grammar. After applying left binarization, the grammar will be

COMS → @SS1_CONJ SS2 (0.85)

@SS1_CONJ → SS1 CONJ (1.0)

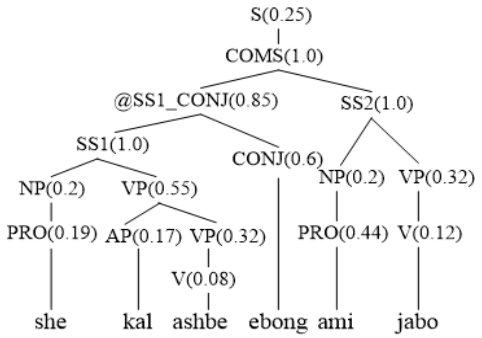

Fig. 2. Left binarization technique is applied in compound sentence.

2) Structural Ambiguity: The statistical parser is a dynamic way to parse sentence automatically, and its help to detect syntactic ambiguity like coordination ambiguity and attachment ambiguity [2]. In the parse for the sentence "goru sobuj ghas khay (গরু সবুজ ঘাস খায়)", we find more than one parse tree that means ambiguity occurred. The probability for the parse structure of Fig. 3 (a) 1.0*0.06*0.32*0.5*0.18*0.5*0.6*0.8*0.4 = 0.0002 and the probability for parsing structure of Fig. 3 (b) 1.0*0.5*0.55*0.18*0.32*0.6*0.8*0.4*0.5 = 0.002. Grammar rule NP→ NP NP has low probability then VP→NP VP. As a result, the parse structure that contains VP→NP VP rule has the higher probability. We can say that parse tree of Fig. 3 (b) is more accurate for current Bangla input sentence.

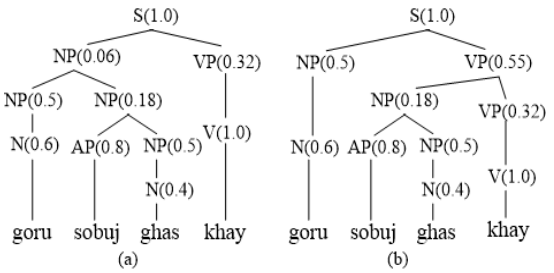

Fig. 3. Parse structure for following input where the rule for (a) is NP→NP NP and rule for (b) is VP→NP VP.

3) CYK algorithm: In this paper, probabilistic version of CKY algorithm is used for statistical parsing. It is a very popular algorithm for the probabilistic model. This algorithm is a bottom-up dynamic programming technique which saves the result into a chart and reuse these result in the finding the larger constituents, as a result, it also called chart parser. For the current input, probabilistic CKY algorithm dynamically parses the height probable parse tree which probability is 0.001. Fig. 4 show the operation of CKY parsing for the current input.

goru	sobuj	ghas	khai
N→goru (0.6) NP→N(0.5*0.6 =0.3)	--	NP→NP NP (0.06*0.3*0.28 =0.0005)	S→NP VP (1*0.3*0.005 =0.002)
	ADJ→sobuj (0.8)	NP→AP NP (0.18*0.8*0.2 =0.028)	VP→NP VP (0.55*0.028*0.32 =0.005) S→NP VP (0.005)
		N→gash (0.4) NP→N (0.5*0.4=0.2)	VP→NP VP (0.55*0.2*0.32 =0.0352) S→NP VP (0.0352)
			V→khay (1.0) VP→V (0.32*1.0=0.32)

Fig. 4. Probabilistic version of CKY algorithm for a simple Bangla sentence.

G. Output:

Finally, we find the output of the system. The output result is basically the most probable or the best parse tree with labeled based representation and the probability. The result shows the parse tree according to the PCFG. If any grammatical rule needs to convert into binary form then the output will show that the "Left Binarization" is used. The height probability of following input is 0.002, and the labeled parse tree is

S [NP [N, গরু] VP [NP [[AP, সবুজ] NP [N, ঘাস]] VP [V, খায়]]]

IV. EXPERIMENTAL RESULTS

A. System Requirements

To implement this system, personal computer with windows 8 operating system is used. Microsoft visual studio 2013 is helped to create GUI and programming language C sharp is used. For implementing lexicon or database, proposed system is used Microsoft SQL server 2012 to store large amount of Bangla lexicons. Siyam Rupali font is used as a Bangla font.

B. Implementations

Our proposed system assigns the probability for three types of sentence using a large Bangla training corpus. Firstly we experiment with a simple sentence "Hillary itihash gorlen (হিলারি ইতিহাস গড়লেন)", this sentence can parse easily with the help of simple PCFG of Bangla grammar. An Example of implementation of simple sentence is shown in Fig.5. The probability of this sentence is 0.00021. We can represent the parse tree as following list form.

S [SS [NP [N, হিলারি]] VP [NP [N, ইতিহাস] VP [V, গড়লেন]]]

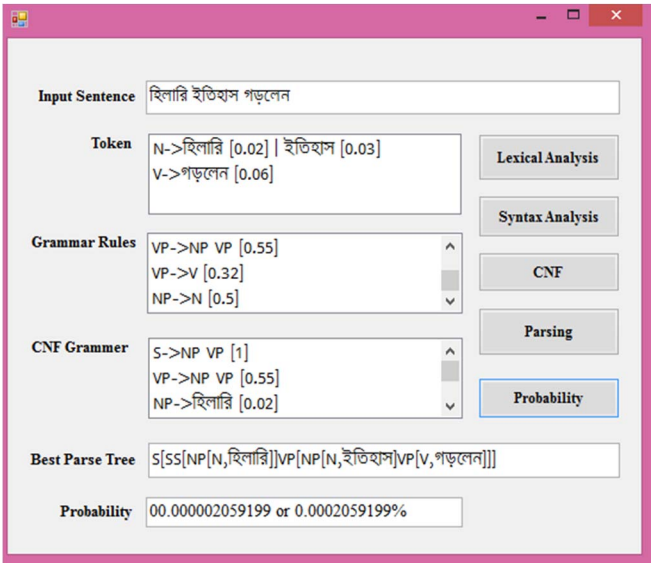

Fig. 5. Statistical parsing of Bangla simple sentence.

A compound sentence is simply, two simple sentence connected by conjunctions like ebong (এবং), kintu (কিন্তু), othoba (অথবা), notuba (নতুবা) etc. The grammar rule for compound sentence is needed to be in binary form for increasing parsing efficiency. An example of compound sentence is given in Fig. 6. The probability of this sentence structure is 0.00000018, the labeled base representation of parse tree for compound sentence is given below.

S[COMS[@SS1_CONJ[SS1[NP[PRO, সে]VP[[AP, কাল]VP[V, আসবে]]][CONJ, এবং]]SS2[NP[PRO, আমি]VP[V, যাব]]]]

The complex sentence is created by an independent clause (IC) and at least one dependent clause (DC). The probability of complex sentense is CS → DC IC (0.93), DC → SUBORD SS (0.92) , IC → SUBCOM SS (0.84), where jodi (যদি), jini (যিনি), jehetu (যেহেতু), tahole (তাহলে), sehetu (সেহেতু) are the example of subordinates clause.

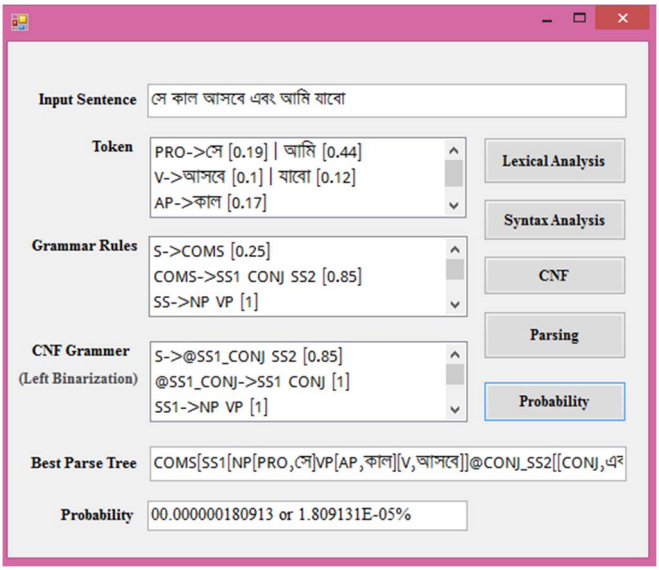

Fig. 6. Statistical parsing of Bangla compound sentence.

The grammar rule is already in binary form, so binarization technique is not needed for the grammar rules. A complex sentence "jodi korim taka day, tahole bari jabo (যদি করিম টাকা দেয় তাহলে আমি বাড়ি যাবো)" is taken as input. Fig. 7 show the implementation of statistical parsing operation of a complex sentence and the probability is the height probable parse tee with its probability 0.0000000515.

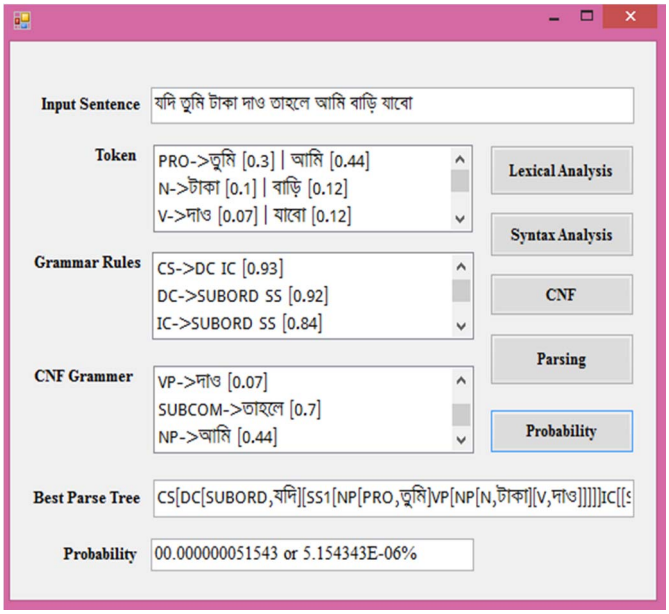

Fig. 7. Statistical parsing of Bangla complex sentence.

Parse tree for the complex sentence is represented in Fig. 8.

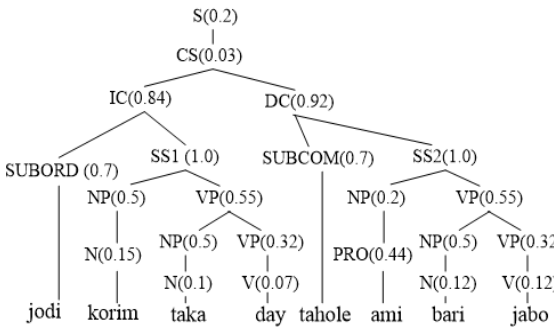

Fig. 8. Parse tree representation of complex sentence.

C. Result

Our proposed model is constructed to parse simple, complex and compound sentence of Bangla which sentences are collected from popular newspapers, textbooks, blogs, and literatures of Bangla. For this experiment, we tested the model with 2025 of different kinds of sentences. The accuracy of the proposed system is analyzed by using different kinds of sentences and word length. Table III indicates that the accuracy of the system is above 93%. The accuracy of given system is calculated by the

ratio of the total number of sentences is applied in the system and the number sentences which are successfully parsed.

TABLE III. PERFORMANCE EVALUATION OF PROPOSED SYSTEM

Sentence Type	Performance Evaluation			
	Word length	***No. of input sentences***	***No. of successful parsed sentences***	***Accuracy (%)***
Simple	3	280	280	100%
	4	200	200	100%
	5	200	190	95.0%
	6	120	105	87.50%
	7	80	65	81.25%
Complex	5	200	200	100%
	6	200	200	100%
	7	100	85	85.0%
	8	50	25	50.0%
Compound	5	220	220	100%
	6	180	180	100%
	7	120	100	83.33%
	8	50	32	64.0%
	9	25	9	36.0%
	Total	2025	1891	93.38%

Accuracy measurement is represented graphically in Fig. 9. The Accuracy versus word length graph shows that how accuracy level is decreased with the increase of the word length for different kinds of sentences. From this figure, we can say that simple sentence performed well then the complex and compound sentence. Here the accuracy of simple sentence is average 92.75% where the accuracy level for complex and compound sentence is average 83.75% and 76.66%. We can improve the performance of the system to include more constitute grammar and lexicon into existing probabilistic context-free grammar.

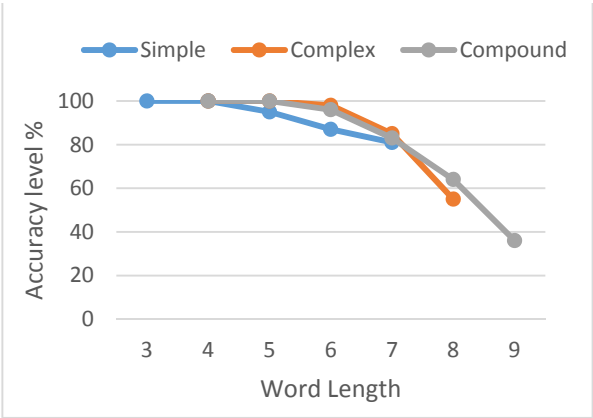

Fig. 9. Graphical representation of relation between accuracy and word length

V. CONCLUSION

Parsing is the most important part of natural language processing and machine translation system. A statistical parser can play a significant role in machine translation. The proposed system would be useful for Bangla machine translation system. The key part of the proposed system is to parse different kinds of Bangla sentences including simple complex and compound sentences in a statistical manner. We developed a stochastic context-free grammar and choose a well-known probabilistic CYK parser to parse Bangla sentences. We have shown how statistical parsing detects ambiguity of a sentence. Proposed module is capable of detecting structural ambiguity of the sentence. This model considers lexicon probability as well as rule probability as a result, we cannot get the same probability for the same structure. The accuracy of the system is tested with several sentences. The future extension can be possible to include statistical parsing with lexicalized PCFGs. Stochastic or probabilistic semantic parsing can be possible. This model can be extended by including Bangla idioms and phrases.

REFERENCES

[1] A. Clark, C. Fox and S. Lappin, *The Handbook of Computational Linguistic and Natural Language Processing*, Wiley-Blackwell, Aug. 2010, pp. 333-342.

[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] G. Sampson, "A stochastic approach of parsing", in *the 11th international conference on computational linguistics (COLING-86)*, 1986, pp. 151-155.

[4] P.F. Brown, VJD Pietra, SAD Pietra and RL Mercer, "The mathematics of statistical machine translation: Parameter estimation", Computational Linguistic, vol. 19, 1993, pp. 63-331.

[5] 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.

[6] E. Charniak, "Tree-bank grammars", In Proc. *13th National Conference on Artificial Intelligence*, AAAI Press/MIT Press, Memlo Park, 1996, pp. 1031-1036.

[7] D. M. Magerman, "Statistical decision-tree models for parsing", *33th Annual Meeting of the Association for Computational Linguastics*, pp. 276-283, Apr. 1995.

[8] 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.

[9] G. K. Saha, "Parsing Bengali Text: An Intelligent Approach", *ACM Ubiquity*, vol. 7, pp. 1-5, 2006.

[10] M. S. Islam and J. K. Das, "A new approach: automatically identify proper noun from Bengali sentence for universal networking language", *African Journal of Computing & ICTs*, vol. 8, pp. 97-106, june 2015.

[11] M. Haque, M. T. Habib, M. M. Rahman, "Automated word prediction in Bangla language using stochastic language model", *International Journal in fundation Computer Science & Technology (IJFCST)*, vol. 5, Nov. 2015.

[12] S. Chatterhi, T. M. Sarkar, P. Dhang, S. Deb, S. Sarker and A. Basu, "A dependency annotation scheme for Bangla treebank", *Language Resources and Evalution*, vol. 48, pp. 443-447, Sept. 2014.

[13] K. M. Azharul Hasan, A. Mondal, A. Saha, "A context free grammar and its predictive parser for Bangla grammar recognition" , In Proc. *13th International Conference on Computer and Information Technology (ICCIT)*, Dec. 2010.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] E. Rich, K. Knight,S. B. Nair, *Artificial Intelligence*, TATA McGraw-Hill, 2009.

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

[21] M. Collins, "Parameter estimation for statistical parsing methods : Theory and practice of distribution-free methods", in *Text, Speech and Language Technology*, vol. 23, pp. 19-55, 2005.

[22] X. Song, S. Ding, C. Lin, "Better binarization for the CKY parsing", In Proc. *Conference on Enmirical Methods in Natural Language Processing*, pp. 167-176, Oct. 2008.

[23] W. Wang, K. Knight and D. Marcu, "Binarizing Syntax Trees to Improve Syntax-Based Machine Translation Accuracy", In Proc. *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, June 2007, pp.746-754.