GUBJSE: ISSN: 2409-0476

Stochastic Approach of Parsing Bengali Sentences

Ayesha Khatun, Khadiza-Tul-Kobra Happy, Babe Sultana, Jahidul Islam and Sumaiya Kabir

Abstract**—** The parsing technique based on associate grammar rules as well as probability is called stochastic parsing. This paper suggested a probabilistic method to eliminate the uncertainty from the sentences of Bangla. The technique of Binarization is applied to increase the precision of the parsing. CYK algorithm is used in this paper. The work mainly focused on intonation-based sentences, for these reasons PCFGs (Probabilistic Context-Free Grammars) is based on proposed. About 30324 words are used to test the proposed system; average 93% accuracy is achieved.

Index Terms**—**Bangla Natural language processing, PCFGs, Binarization, Bayesian inference.

I. Introduction

HUMANS are independent of their language and they have regularity and mechanisms to describe the presentation of this language, but uncertainty, a lot of inconsistencies are a significant problem in the parsing of human words. The most probable parse is built by statistical parsing to eliminate uncertainty, which is really the primary objective of parsing [1]. The contemporary parser also aims people clarify several natural language functions, involving thematic function marking, knowledge discovery, text description and use in speech recognition [2]. Statistical parser connects grammatical rule with probabilistic data which is collected from a treebank. CYK is very common for parsing utility among all chart parser algorithms; as a result, The proposed system used the probabilistic CYK algorithm in this model. The Interface of Bayesian is applied in PCFGs to provide accurate probability to the grammar. CYK algorithm suggests us to find out the exact parse tree by using PCFGs as well as a dictionary in this proposed model. The Bangla language processing is more challenging part for a huge variety of sentence structures and ambiguities. The statistical model in the Bangla machine translation method is one of the most powerful parsers for language modeling as well. The parser can even be classified into three groups of natural language processing; (a) rule-based, (b) generalized, and another type is (c) statistical-based parser. In this parsing technique, ruled-based parser recursively applied grammatical rules for parsing sentences can occur as well as many ambiguities. This uncertainty, which is a very difficult problem, is overcome by large and complex grammar rules. On the contrast, by training a huge amount of corpus, the probabilistic parser which can accurately track the sentence ambiguity. The parser which is traditional can build parse trees, which help to find the height probable parse tree. This paper aims to parse all kinds of Bangla sentences automatically by using a statistical model. For improving the parsing efficiency, we used left binarization in our proposed system. The proposed stochastic approach is a golden standard method which can play a vital role in Bangla sentence parsing.

DOI: https://doi.org/10.3329/gubjse.v7i0.54021

This paper was received on 20 May 2020, revised on 10 March 2021 and accepted on 19 April 2021.

Ayesha Khatun in Dept. of CSE, GUB, Dhaka, Bangladesh, Email: ayesha@cse.green.edu.bd
Khadiza-Tul-Kobra Happy in Dept. of CSE, GUB, Dhaka, Bangladesh, Email: happy052409@gmail.com
Babe Sultana in Dept. of CSE, GUB, Dhaka, Bangladesh, Email: babe@cse.green.edu.bd
Md. Jahidul Islam in Dept. of CSE, GUB, Dhaka, Bangladesh, Email: jahid@cse.green.edu.bd
Sumaiya Kabir in Dept. of CSE, GUB, Dhaka, Bangladesh, Email: sumaiya@cse.green.edu.bd

II. Background Study

Statistical NLP is an affluent part of natural language processing and a huge number of works already published in this area. Foundation text is the first extensive start to statistical natural language processing. Algorithms and theory for building NLP tools provide in [1]. Speech recognition, computational linguistics and language processing discussed in [2]. A generative model of lexicalized context-free grammar with the probabilistic treatment of both subcategorization and wh-movement by using the statistical parsing model shown in [3] act an immensely strong role in the area of Statistical NLP. We know that Bangla sentence have tree major types. In this paper they shown the parsing these three types of sentences like simple, compound as well as complex sentences [4]. Bangla grammar recognition parsing technique described in [5]. The basic architecture of machine translation between bangle to Sylheti was proposed in the research work [6]. A transfer architecture with optimal time complexity presented in [7]. The design wise procedure in Bangla language processing has been discussed in this research paper [8]. Many developments occurred in the area of statistical NLP. Statistical parsing of Bangla sentences by using a probabilistic approach and CYK algorithm proposed in [9, 10]. The system can detect the ambiguity of the sentences. Several advances have taken place in the field of NLP statistical analysis. In this report, we developed an analytical approach to parsing according to the structure and intonation which is tasted with various types of Bangla phrases. This article used Bayesian Inference to obtain a more precise probability of generating PCFG validated for the parse tree with the F-score calculation strategy. This allowed providing a reasonably reasonable outcome.

III. PROPOSED FRAMEWORK

In our suggested framework with input and output configuration, there have been five parts and the first module is Syntax analyzer, second one is Rule generation, third one is Statistical parser, the fourth one is Error Controller, and the fifth module is Lexicon. The diagram statistical parsing of Bangla sentences is shown in Fig. 1.

A. Syntax Analyzer

When a sentence is take as input and divided this sentence into small part which called token [2]. As our proposed model design is on Bangla language for that reason, Bangla sentences are chosen as an input sentence. For example, firstly we chose simple sentence "একটি ছোট ছেলে মাছ ধরছে (ekti coto cele mash dhorche)" and the outcome of this module is presenting as

TOKEN= ekti (একটি), coto (ছোট), cele (ছেলে), mash (মাছ), dhorche (ধরছে).

B. Dictionary Module

This module defines as a collection of a word database with its relevant POS tag sections as well as associated probabilistic data [11]. Until placing it in the parse tree, it will examine each term and if the input word somehow doesn't fit in the database or if the grammatical structure is not understood, then an error message will be generated. The lexicon will be shown as the input sentence, according to

ekti (একটি) → APR [0.4];coto (ছোট) → AP [0.14];

cele (ছেলে) → N [0.4]; dhorche (ধরছে) → V [0.06]

here *ekti (একটি)* a specifier (SPR) and 0.4 is the probability of being *ekti (একটি)* as a specifier. The Bangla word dictionary or lexicon is briefly shown in Table I.

C. Rule Generation:

Grammar laws are generated by a rule generator represented by a context-free grammar set (CFG). A well-constructed grammatical statement is generated by CFG [12]. We want to develop advanced probabilistic context-free grammar rules when we choose the statistical approach.

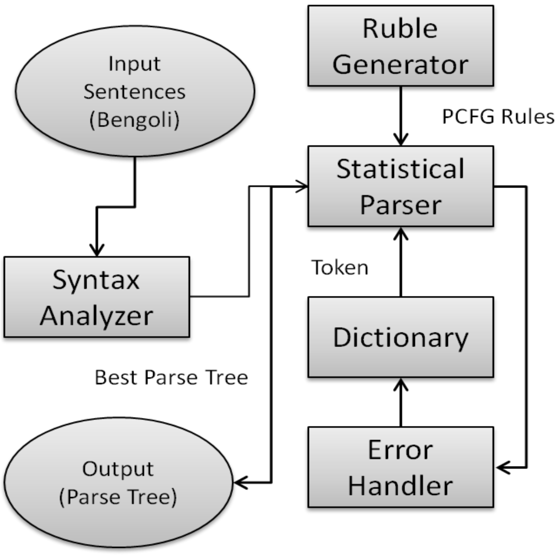


Fig. 1. The diagram of our proposed system of statistical model for Bangla sentences.

D. PCFG (Probabilistic Context-Free Grammar)

The certain probabilistic data are assigning to each rule in the probabilistic context-free grammar. Let G = (N, T, S, P) be the CFG, where finite set of terminals symbol are T, N is a finite set of symbols, S is the beginning symbol, and P is a finite set of output symbols of A→ BC or A →ω, where A, B, C∈N and ω∈T. The probabilistic context-free grammar (G, θ) is a pair of Gandθ. G is show as context-free grammar on the other hand θ represented as real-valued vector of length |R| and θ> 0, other all nonterminal are

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

The β represented as a variable ranging on $(N \times N) \cup T$. The string w which represented the probability of all tree which also shown in this research work[13].

Bayesian PCFG inference: Given a set of text terminals $w = (w_1, w_2, w_3 \ldots\ldots\ldots w_n)$ generated by a known CFG is G, the output probability θ is inferred and Bay's rule is offered. The chance would be,
$P(\theta|w) \propto P_G(w|\theta)P(\theta)$, were

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

It is possible to quantify and marginalize,

$$P(t,\theta|w) \propto P(w|t)P(t|\theta)P(\theta)$$

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

Where t is referred to as a sequence of syntax trees of w, and here's a common portion of the Bangla grammar PCFG and the intonation-based lexicon such as interrogative, imperative, assertive etc. [15, 14] is described in Table I.

TABLE I. A SMALE OF PROBABILISTIC CFG OF THE BANGLA GRAMMAR AS WELL AS LEXICON

Rank	**Probabilistic Context-Free Grammars**	
	Rules	***Probability***
1	S → AS | IRS | IS | ES	0.65 | 0.2 | 0.25
2	AS → NPh VPh	1.0
3	IRS → NPh VPh	0.5
4	IS → NPh VPh	0.06
5	ES → NPh VPh	0.18
6	ES → INJ NPh	0.06
7	INTJ → INTJ PN	0.2
8	NPh → N	0.32
9	NPh→ PRO	0.55
10	NPh → NPh NPh	0.13
11	NPh → SPR AP N	0.85
12	NPh → WH AP	0.05
13	NPh → NPh WH	0.10
14	NPh → AP NPh	0.93
15	NPh → NPh PN	0.07
16	VPh → V	0.92
17	VPh → NPh VPh	0.08
18	VPh → AP V	0.84
19	VPh → V Ind	0.16
20	VPh → VPh PN	0.44 | 0.19 |0.37
21	VPh → 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
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

E. *Statistical Parser Module:*

The Statistical parser assigned probabilities to the possible parse of the sentence and generate the best parse tree as an output. Using the rule generator module for PCFG is the statistical parsing method of the proposed model. With the assistance of a lexicon and lexical analyzer, it generates the most probable parse tree.

1) Binarization Technique: Binarization technique can convert binary grammar into n-arry and can do vice-versa and it also can parse the effective parse tree which have an o(n3) time complexity [16]. For all types of chart parsing algorithms such as CYK algorithm have to be convert into the form of binary. The Binarization approach optimizes the tabular parser's parser computationally and it also enables the parser time effectively. There we directly binarize the CFG into the CNF technique which is required by the CYK algorithm. Here we have given an example of a simple sentence [17] is "একটি ছোট ছেলে মাছ ধরছে (ekti coto cele mash dhorche)" and also the Probabilistic CFG of a sample grammar attached below.

NPh → SPR AP N [0.85]

There the specifier, the noun, and the adjective are expressed as AP, N and SPR. The grammatical rule will be after applying left binarization,

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

The proposed system used left binarization technique here that selects the leaving two pair and this strategy does not impact the possibility of CFG.

2) CYK algorithm (Probabilistic): The CYK algorithm [2] is known as bottom-up chart parser which is in polynomial form and it used in a table(n × n) to record the evaluation of a phrase substring called $s = (w_1, w_2, w_3 \ldots\ldots\ldots w_n)$. CYK algorithm's complexity is $O(n^3)$ where n is representing the total length of the text which is to be parsed. In this article, the probabilistic CYK algorithm is used for statistical parsing. The probabilistic CYK algorithm dynamically parses the high-level probabilistic parse of a sentence.

F. *Output:*

Ultimately, the production of the proposed method is created. There the highest probable parse tree is generated by the system which is presented as a structure based on labeling. If any grammar rule uses the left binarization technique, the output also presented accordingly. 1.65E-8% percent is the maximum likelihood and the parse tree's labeled-based structure is

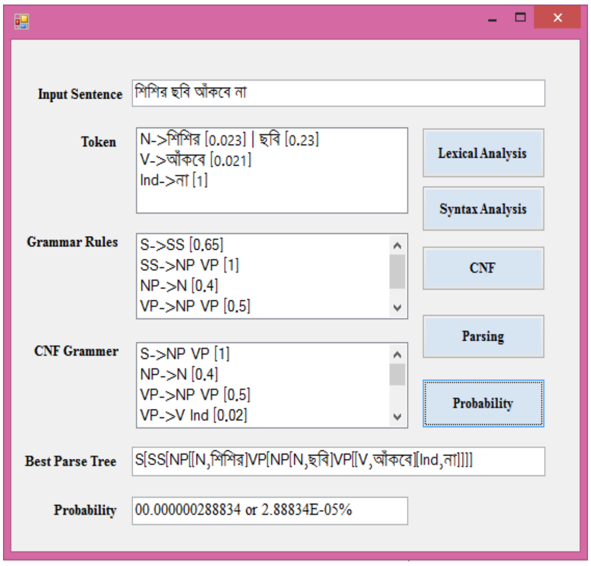
$$S[NP[[SPR,$$

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

III. IMPLEMENTATION AND EXPERIMENTAL RESULTS

A. *Implementations*

We used the MS SQL server to execute the scheme and store the tree bank. The Microsoft visual studio was kept for Interface development. Here, C Sharp is the programming language. We used Shonar Bangla font and also Siyam Rupali font for analyzing the data. To parse interrogative, assertive, imperative and exclamatory, the suggested model is built. The tree bank is gathered from Bangladesh's textbooks, popular newspapers, blogs, literature and novels. We validated the models for different kinds of phrases.



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 assertive sentences from Bangla.

We have shown assertive sentences in the following figure and we use the special symbol in PCFG for doing this. It have produced the best output with the help of the algorithm. The input sentence "shishir chobi akbe na (শিশির ছবি আঁকবে না)". First of all, it displays the tokens of the following statements with the corresponding probability and then the PCFG. Afterwards, the CNF type of syntax will be shown and finally, the optimal parsing tree will be shown as a result with the likelihood of input text.

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

This is now the performance of the method as a named parse tree and the likelihood of 2.88E-05 percent of the following statement. The system implementation is shown in Fig. 2. In this way, exclamatory, interrogative and imperative sentences are also used.

B. *Results*

We used PARSEVAL measurements [2] for testing the parser and grammar. Three simple measures, Label accuracy, label recall, and F-measure, are shown in this paper. In statistical analysis, F-measure or F-score is a calculation tool to assess the accuracy rate of the parser and grammar. If the components in the parse tree of the hypothesis and the parse tree of the relation have the same reference point, finishing point, and non-terminal symbol, the sub-tree form is called right, otherwise incorrect. As the spectrum of assertive law is more a consequence in Fig .3,

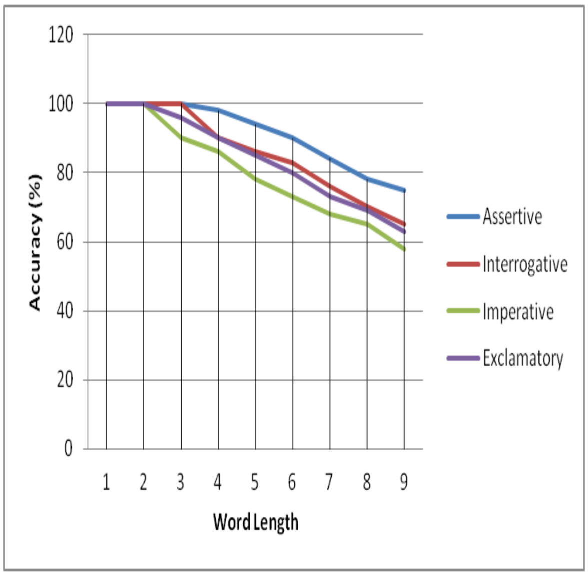


Fig. 3. Number of word length vs accuray (%) graph.

Many ups and downs can be seen to demonstrate the degree of accuracy and average. From this figure we show that parsing accuracy of assertive is high, where exclamatory sentence accuracy is not satisfactory. In order to provide more grammar and lexicon in existing PCFG, we can boost the system's efficiency.

IV. CONCLUSION

The syntax parser is the fundamental part of NLP processing as well as the machine translation system. This design may play a challenging role in the Bangla device translation method. In the key part of the development scheme, the Bayesian approach is used which ultimately allocates probability to evaluate various forms of phrases, comprising assertive, interrogative, imperative and exclamatory phrases, in a statistical manner. The proposed model detects the ambiguity of Bangla sentence more efficiently. The Standard PARSEVAL technique we used for measuring the accuracy rate of a parser. The system's average F-score is approximately 93%. By elaborating on the corpus size, we can boost the system's accuracy. By using statistical natural language, there is also a vast amount of research potential in Bangla. In this area, a very limited amount of work has been completed. A potential extension of this work can be statistical parsing with lexicalized PCFG and semantically parsing the sentences can be extended.

ACKNOWLEDGEMENT

This work was supported in part by the Center for Research, Innovation, and Transformation (CRIT) of Green University of Bangladesh (GUB).

REFERENCES

[1] Schutze H, Manning C, Foundations of statistical NLP. MIT press; 1999 May 28.

[2] Martin, J.H and Jurasky, D., 2000. Speech and Language Processing: *Computational Linguistics and Speech Recognition*, An introduction to natural language Processing. *Prentice Hall, New Jersey*.

[3] M. Collins, "There are Three generative, lexicalized models for statistical parsing", In Proceedings of International Conference of 8th Conference on European Chapter of the Association for Computational Linguistics, 7th July 1997.

[4] M. M. Hoque, M. M. Ali, 2003, December. A parsing methodology for Bangla natural language sentences. In *Proc. on Computer and Information Technology (ICCIT), Dhaka, Bangladesh* (pp. 277-282).

[5] Rabbi RZ, Shuvo MI, Hasan KA. Bangla grammar pattern recognition using shift reduce parser. In Proc. On 2016 5th International Conference on Informatics, Electronics and Vision Conference 2016 May 13 (pp. 229-234).

[6] Chakraborty S, Sinha A, Nath S. a Bengali Sylheti rule-based dialect translation system: Proposal and preliminary system. In Proceedings of the International Conference on Computing and Communication Systems 2018 (pp. 451-460). Springer, Singapore.

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

[8] Karim MA, editor. Technical challenges and design issues in bangla language processing. IGI Global; 2013 Apr 30.

[9] Hoque, M.M., Faruk, M.O., Hasan, M.M., Hassan, M.K. and Karim, M.M.U., 2006. An empirical framework for statistical parsing of Bangla sentences. *Computer Science & Engineering Research Journal*, *4*, pp.29-38.

[10] Khatun, A. and Hoque, M.M., 2017, February. Statistical parsing of Bangla sentences by CYK algorithm. In *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 655-661). IEEE.

[11] Purohit, P.P., Hoque, M.M. and Hassan, M.K., 2014, October. An empirical framework for semantic analysis of Bangla sentences. In *2014 9th International Forum on Strategic Technology (IFOST)* (pp. 34-39). IEEE.

[12] M. N. Hoque, M. H. Siddiqui , 2015, December. rule based analyzer and Bangla Parts-of-Speech tagging using Bangla stemmer. In *2015 18th International Conference on Computer and Information Technology (ICCIT)* (pp. 440-444). IEEE.

[13] Johnson M, Griffiths TL, Goldwater S. Bayesian inference for pcfgs via markov chain monte carlo. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference 2007 Apr (pp. 139-146).

[14] M.S. Arefin, M. M. Hoque, M. O. Rahman, and Arefin, M.S., 2015, May. interrogative and imperative sentences into English and the machine translation framework for translating Bangla assertive. In Proc. On 2015 *International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)* (pp. 1-6).

[15] Alam, L., Arefin, M.S., Hoque, M.M and Sharmin, S., 2015, November. For parsing Bangla assertive, interrogative and imperative sentences an empirical framework is designed. In *2015 International Conference on Computer and Information Engineering (ICCIE)* (pp. 122-125). IEEE.

[16] Song, X., Ding, S. and Lin, C.Y., 2008, October. Better binarization for the CKY parsing. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 167-176).

[17] Huang, L., Zhang, H., Gildea, D. and Knight, K., 2009. Binarization of synchronous context-free grammars. *Computational Linguistics*, *35*(4), pp.559-595.

Ayesha Khatun was born in Dhaka, Bangladesh in 1994. She received the B. Sc. Degree in Computer Science & Engineering, Chittagong University of Engineering & Technology (CUET). At present she is working as a Lecturer and Program Coordinator (Day), Dept. of CSE, Green University of Bangladesh. She achieved scholarship for Higher Study, Wide space Bangladesh Limited, Merit scholarship every year, Department of CSE, CUET. She was also the 2nd Runners up in ICT National Android Application Development Training 2015, Ministry of Information and Communication Technology Division. Her research interests include application of Natural Language Processing, Bangla Language Processing, Data Mining, Artificial Intelligence, and Internet of Things.

Khadiza-Tul-Kobra Happy was born in Chittagong, Bangladesh in 1996. She received her B.Sc. Degree in Computer Science& Engineering from Green University of Bangladesh (GUB). She achieved seven Vice Chancellor's Awards and one Deans Award for University result. Her research interests include application of NLP, Bangla NLP, Data Mining, Internet of Things as well as Artificial Intelligence.

Babe Sultana was born in, Cox’s Bazar, Bangladesh, in 1994. She received her B.Sc. Degree in Computer Science and Engineering (CSE) from Green University of Bangladesh in the year of 2018. At present she is working as a Lecturer, Dept. of CSE, Green University of Bangladesh. Also, her publication was about” Multi-mode Project Scheduling with Limited Resource and Budget Constraints which is published in *International Conference on Innovation in Engineering and Technology (ICIET)* 27-28 December, 2018 and she also got Best Paper Award as well as IEEE Best Paper Award of the following conference. Her research interests include Theory of Optimization, Natural language Processing, Speech Recognitions, Image Processing.

Md. Jahidul Islam was born in Sirajganj, Bangladesh, in 1991. He received the B.Sc. and M.Sc. degrees in Computer Science and Engineering from the Jagannath University, Dhaka, in 2015 and 2017 respectively. Currently, he is working as a Lecturer at Computer Science and Engineering (CSE), Green University of Bangladesh (GUB), Dhaka, Bangladesh science May 2017 to present. He is a member of Computing and Communication and Human-Computer Interaction (HCI) research groups, CSE, GUB. His research interests include Internet of Things (IoT), Blockchain, Software Defined Networking (SDN), Natural Language Processing (NLP), Digital Forensic Investigation (DFI), Vehicular Ad-Hoc Networking (VANET), Wireless Mesh Networking (WMN) and Data Mining.

Ms. Sumaiya Kabir was born in Barisal, Bangladesh, in 1989. She received the B.Sc. in Computer Science and Engineering from Patuakhali Science and Technology (PSTU) in 2012 and M.Sc. in CSE from (EWU) in 2017. At present she is working as an Assistant Professor & program coordinator (Day) of department of Computer Science and Engineering in Green University of Bangladesh (GUB) from 2013 to present. She is a member of Systems & Security research cell in CSE, GUB. Her research interest includes semantic web, web 3.0 architecture, ontology designing and semantic knowledge engineering.