UNVEILING THE POWER OF TAG USING STATISTICAL PARSING FOR NATURAL LANGUAGES

Pavan Kurariya, Prashant Chaudhary, Jahnavi Bodhankar, Lenali Singh and Ajai Kumar

Centre for Development of Advanced Computing, Pune, India

ABSTRACT

The Revolution of the Artificial Intelligence (AI) has started when machines could decipher enigmatic symbols concealed within messages. Subsequently, with the progress of Natural Language Processing (NLP), machines attained the capacity to understand and comprehend human language. Tree Adjoining Grammar (TAG) has become powerful grammatical formalism for processing Large-scale Grammar. However, TAG mostly rely on Grammar which is created by Languages expert and due to structural ambiguity in Natural Languages computation complexity of TAG is very high o(n^6). We observed that rules-based approach has many serious flaws, firstly, language evolves with time and it is impossible to create grammar which is extensive enough to represent every structure of language in real world. Secondly, it takes too much time and language resources to develop a practical solution. These difficulties motivated us to explore an alternative approach instead of completely rely on the rule-based method. In this paper, we proposed a Statistical Parsing algorithm for Natural Languages (NL) using TAG formalism where Parser makes crucial use of data driven model for identifying Syntactic dependencies of complex structure. We observed that using probabilistic model along with limited training data can significantly improve both the quality and performance of TAG Parser. We also demonstrate that the newer parser outperforms previous rule-based parser on given sample corpus. Our experiment for many Indian Languages, also provides further support for the claim that above mentioned approach might be an awaiting solution for problem that require rich structural analysis of corpus and constructing syntactic dependencies of any Natural Language without much depending on manual process of creating grammar for same. Finally, we present result of our on-going research where probability model will be applying to appropriate selection of adjunction of any given node of elementary trees and state chart representations are shared across derivation.

KEYWORDS

Artificial Intelligent (AI), Natural Language Processing (NLP), Tree Adjoining Grammar (TAG), Natural Languages (NL)

1. INTRODUCTION

Natural Language Parsing is the process of analysing a string of symbols as a human would a sentence of natural language. TAG Parser is a core component of NLP research having important role in multiple thematic areas like Machine Translation, Information extraction and retrieval, semantic analysis and human computer interaction. At initial phase of Natural Language Processing, Numerous algorithms were discovered for analysing the Natural Languages. Earlier, NLP Programmer need to rely on grammar rules where they analyze structure of source language and convert it into Parse derivation. As part of our research on Tree Adjoining Grammar, we

David C. Wyld et al. (Eds): AI&FL, SCM, NLPTT, DSCC -2023
pp. 57-68, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131407

have developed Multithreaded TAG Parser which is implementation of the 'Early-Type Parsing Algorithm' originally proposed by Arvind Joshi [1]. Basically, TAG Parser is a software program that analyze the source in order to determine its grammatical structure with respect to a given formal grammar. There are two types of basic trees in TAG - Initial trees and Auxiliary trees. Any sentence of the language can by represented using a derived tree, constructed from initial and auxiliary tree, by Adjunctions and/or Substitutions. We found that the computational complexity of TAG is extremely high because of structural ambiguity in natural languages and differences between the elementary trees of various languages. Our earlier research also described complexity and challenges beneath parsing-generation process [2]. To address the challenges in the existing TAG Parser, we explored an alternative approach employing a probabilistic model using advanced algorithms, and we observed substantial enhancements in both the performance and efficiency of numerous natural language processing (NLP) applications

Therefore, we explored Statistical Parsing algorithm to facilitate the translation of sentences from one language to another. Statistical Parser operates through two separate models: the H1 Model and the H2 Model. The H1 Model serves as a tagging probability model, responsible for determining the most appropriate tree to be chosen during the parsing process. Its primary objective is to identify the tree that best suits the sentence being translated. Conversely, the H2 Model acts as a parsing probability model. Its main purpose is to determine the probabilistic tree that should be employed for Adjunction or Substitution at a specific node within the parsing process. By utilizing probability calculations, the H2 Model assists in making informed decisions regarding tree selection and manipulation during the translation process. Overall, the Statistical Parser, with its H1 and H2 Models, offers an advance approach to Machine translation, enabling the conversion of English sentences into various Indian languages.

2. LITERATURE SURVEY

In this paper, initially, we analyze the standard definition of TAG Introduced by Arvind Joshi [3] and Lexicalized Tree Adjoining Grammar L-TAG [4], Then Dr. Joshi and Srinivas defined syntactic annotation based on L-TAG [5] where they introduced super-tagging approach for enhance the TAG derivation. We have also proposed Virtual research Lab [6] for TAG related research where we have described extension of TAG Derivation. Earlier, we had also made effort to improve Performance of TAG based Machine Translation [7]. We found one close work related to our research by Vijay-Shanker [8] where they have described LTAGs and their application using statistical parsing. We clearly observe that there has been a fair amount of work done to explore parse tree derivation for predict the future parse structure. Some earlier work we notice in the area has been discussed by Michael Collins [9][10] where he discussed about statistical parser based on bigram lexical dependencies. Philip Resnik [11] also proposed framework for Statistical Natural Processing while Yves Schabes [12] introduced Parse derivation where multiple auxiliary trees can be adjoined at a single node and extended notion of derivation and its formal definition. We have also observed that in more recent research, TAG has also been experimented with Neural network models by Kasai[13] and Kuncoro[14] has demonstrated results that suggest that introducing structure information into LSTM is beneficial. Hence Neural Network based model has shown its potential to substantially improve performance over conventional Parsing Algorithm and also open new thread to TAG based research.

3. ADVANCEMENTS AND INNOVATION IN STATISTICAL TAG PARSER DEVELOPMENT

The foundation of Statistical TAG consists of essential building blocks, encompassing the following key elements:

3.1. H1 Module

3.1.1. Language Model

The H1 Model incorporates a sophisticated language processing approach by utilizing a five-gram model as its language model. This five-gram model analyses text by considering sequences of five consecutive words, allowing it to capture a more comprehensive understanding of the language. Within this model, a notable feature is the selection of a specific tree structure based on the context of the word and the sequence of Part-of-Speech (POS) tags. This selection process ensures that the appropriate tree, which represents the syntactic structure of the sentence, is chosen to accurately capture the intended meaning.

By considering both the context of the word and the POS sequence, the H1 Model enhances its ability to generate coherent and contextually appropriate responses. This approach enables more accurate language understanding and generation, leading to improved overall performance and naturalness in communication.

3.2. H2 Module

3.2.1. Initial Tree Prefix_SubSentence_And_Postfix_SubSentence

The module conducts an analysis specifically on the left subsentence of the initial tree. It further organizes this subsentence in a linear order based on priority. Similarly, the module performs the same analysis on the postfixed subsentence. All of this valuable information is efficiently stored within a single data structure.

3.2.2. Statistical Parser

The parsing process employed in this system relies on a probability-based approach. At the outset, the parser starts with the root node, distinguished by the TOP label. At each node, a probability model is employed to determine the appropriate tree attachment, or no attachment at all if the model yields no result. To maintain the current state at every node, a state is created and stored in the state chart. The system utilizes logical attachment for parsing while simultaneously capturing and storing the state information in the state chart. This stored information proves valuable in constructing a derivation parser.

3.2.3. Probability Model

A probability model is employed to determine the appropriate tree for attachment at each node, considering both the Initial operation and Auxiliary operation. This model utilizes a Probability table that assigns probabilities based on the current category, current tree, adjuncted category, and adjuncted tree, facilitating the decision-making process for adjunction at each node. It is worth noting that this Probability Model relies on the Initial Tree Prefix_SubSentence_And And_Postfix_SubSentence model, as the priority of tree selection is a crucial factor in the model's functioning.

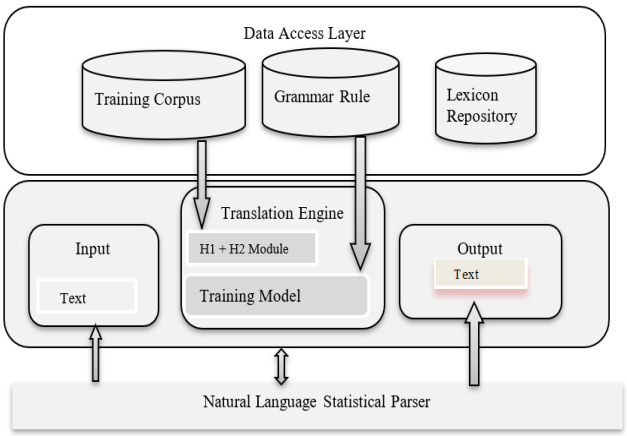

Figure 1. Statistical Parser Architecture

4. TRAINING MODEL

4.1. Training of the H1 Model

In the H1Model, the training process begins by training the category and Tree sequence using the parse Table. Once this initial training is complete, we proceed to apply various n-gram models such as bi-gram, trigram, four gram, and five grams. These models help capture the linguistic patterns and dependencies within the data. After training the Language model table, the next step involves applying probability calculation on the probability table. This calculation leverages the trained model to determine the likelihood of certain sequences or combinations occurring based on the observed patterns and frequencies in the data. Overall, the H1Model follows a sequential process of training the category and Tree sequence, applying n-gram models, and utilizing probability calculations to enhance the understanding and prediction capabilities of the language model.

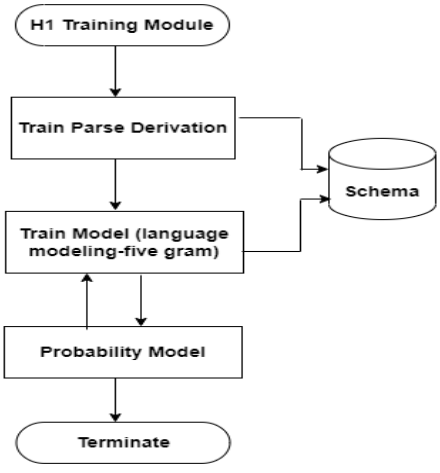

Figure 2. H1 Training module flow

4.2. Training of the H2-Model

4.2.1. Word Based probability model

This conventional approach relies on a vast corpus for training and is a time-intensive process. It depends on the lexical items present in the trained corpus. Word-Based Probability Model entails the utilization of two tables to facilitate the calculation of probabilities.

4.2.1.1. TreeWordTable

This table records the frequency of occurrences for combinations of tree and word categories.

4.2.1.2. Probability table

This table stores the probability associated with each word, incorporating an overall probability and a back-off model. It encompasses the probability calculation for word-category-tree combinations.

4.2.2. Category Based Probability model

We have adopted an alternative approach where probability is determined independently of the lexical item. Instead, it operates based on the categories present in the trained corpus. Category-Based Probability Model utilizes two tables to facilitate the calculation of probabilities.

4.2.2.1. Tree-Category Table

The purpose of the Tree-Category Table is to meticulously track the frequency with which specific combinations of tree and category occur. It acts as a repository that keeps a comprehensive record of the number of times a particular tree is associated with a specific category within the parsed data or corpus. By observing and analysing these frequencies, valuable insights can be gained regarding the relationship between tree structures and their corresponding categories. This information forms a crucial part of statistical parsing, aiding in the accurate prediction and generation of syntactically valid sentences.

4.2.2.2. Probability table

The Probability Table is a crucial component of the statistical parsing system. It serves as a repository for probabilities associated with each category present in the parsed data or corpus. This table is constructed based on an overall probability calculation and a back-off model, incorporating the concept of tree-category probability.

The probabilities stored in this table provide insights into the likelihood of a particular category being assigned to a given syntactic structure. These probabilities are derived from extensive training and analysis of linguistic data, enabling the parser to make informed decisions during the parsing process.

By incorporating tree-category probability, the Probability Table enhances the accuracy and reliability of the statistical parsing system.

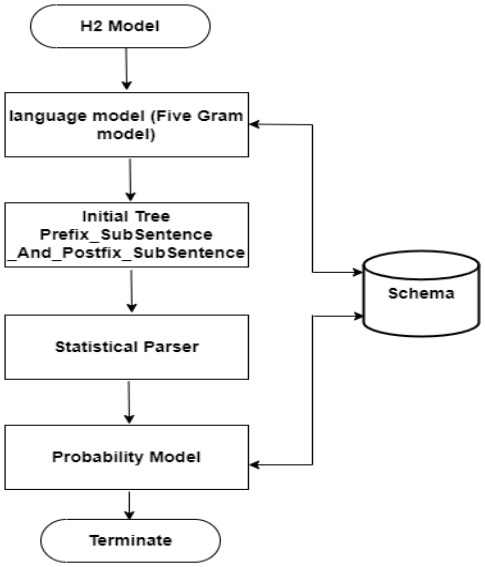

Figure 3. H2 Training module flow

5. ER DIAGRAM OF TRAINING MODEL

5.1. ER-Diagram for H1- model

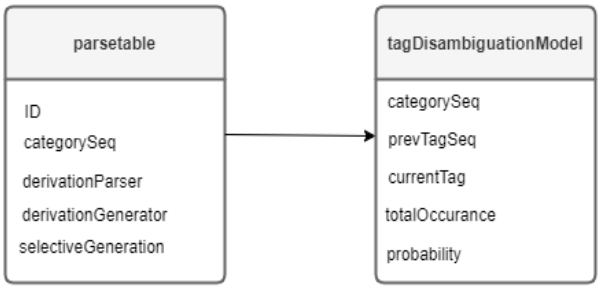

Fig 4: Hi-model ER diagram

5.2. ER-Diagram for Word Based probability H2- model

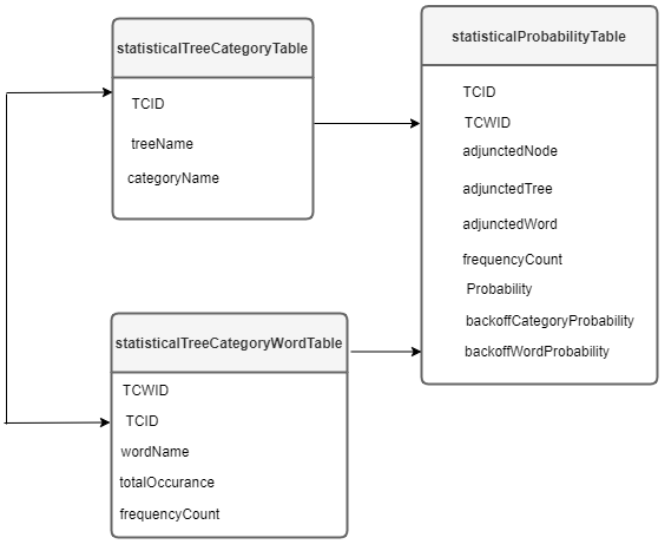

Fig 5: H2-model ER diagram

6. PROBABILITY

6.1. H2 Probability Equation

Direct Probability:
Pr(£',W',P'|Node,£,W,P) = Pr(£'| Node,£,W,P)* Pr(P'|£'Node,£,W,P)* Pr(P',£',Node,£,W,P)
Pr (£'| Node, £, W, P) = (Count (Node, £, W, P, £') +alpha)/
(Count (Node, £, W, P) +k*alpha)
Where: k is the diversity of Adjunction
Alpha = 1/ (10,000)

Back-Off Probability:
E1: Lexicalized Level Model:
Pr(£',W',P'|Node,£,W,P) = Pr(£'| Node,£,W,P)* Pr(P'|£'Node,£,W,P)* Pr(P',£',Node,£,W,P)
E2: Back-off Level Model:
Pr(£',W',P'|Node,£,P) = Pr(£'| Node,£,P)* Pr(P'|£'Node,£,P)* Pr(P',£',Node,£,P)
Count(e1)= C.
delta(C)e1+(1-delta(C))
Where: delta(C) =C/(C+D)
D: diversity of e1.

6.2. H1 Probability Equation

Final Probability = wordEmitProbability+contextual probability

Where:
WordEmtProbability= ΣWiTi/ΣWi;
Contextual Probability = Σ(Ti)|(Ti-2*Ti-1)/ Σ (Ti-2|Ti-1)

7. EXPERIMENT OF STATISTICAL PARSER WITH TREE BANK

Statistical approach is based on the tree adjoining grammar working on probability calculation of Training Model. STP works in the two models: H1 Model and H2 Model. H1 model is a tagging probability model responsible to identify most appropriate tree to pick up for parsing process and H2 model is a parsing probability model, finds probabilistic tree to adjunct/substitute at given node.

Multilingual Tree Bank which was we used for this experiment was created by language expert for different languages using TAG Grammar based research Lab [11] were shown in figure 5.

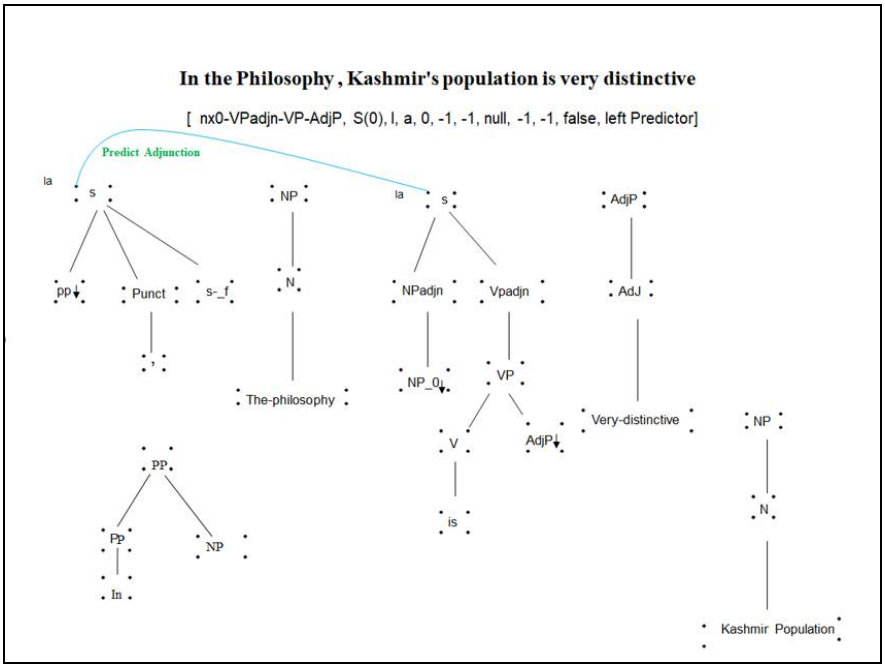

Fig 6: Statistical TAG parsing

The LISP Notation converted into Object and stored into database into encrypted form. Some of the multi lingual trees from the grammar are shown in below graph:

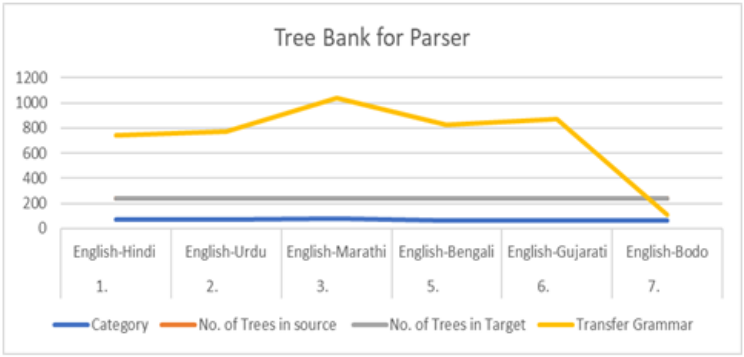

Fig 7: Multilingual grammar for statistical parser

- 7500 out of 8000 sentences are trained with H1 Model.
- 5430 out of 7500 Sentences (trained on H1 model) are Parse and generated well on H2 Model.
- Speed of Parser has been examined and it is observed that Parser is taking approx. 3:45 minute to Parse (along with H1 and H2 model) 11000 sentences all together.

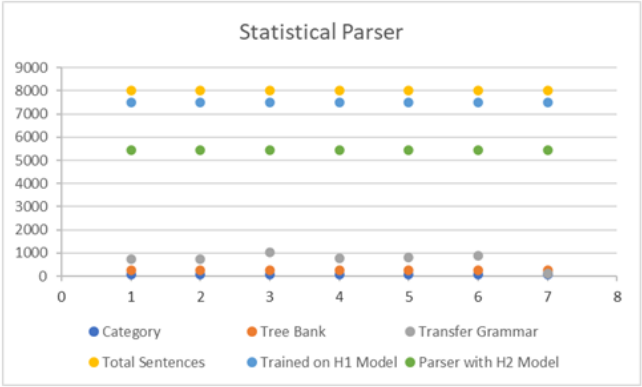

Fig 8: Multilingual grammar for statistical parser

8. EXTENSION OF TAG TOWARDS NEURAL NETWORK

Traditionally, Tree adjoins Grammar (TAG) was based on rule-based approach but we have explored its potential with statical model where data driven model utilized to create automatic derivation Parser. Recently, Neural Networks (NN) based approach are also getting popular to process the Natural Languages. At initial level, we have also investigated TAG with Neural Network which is created by using TAG Tree Bank and that information can be used during the operations of TAG, which enable the parser to construct derivation. Some prior studies also suggest that structure information can be transform as Long Short-Term Memory (LSTM) which is advance version of recurrent neural network (RNN) [Kuncoro et al. (2018) demonstrated LSTM can utilize during prediction parse derivation. Cohen et al., 2011 has experimented that by utilizing Neural networks with TAG has the potential to capture complex dependencies and long-range interactions between multiple clauses, improving the performance of TAG parsing compared to conventional rule-based approaches. The neural network-based derivation parser leverages the power of machine learning to auto learn the grammar and structure of sentences directly from data. By capturing the relationships between words and utilizing training model, it can generate parse trees that represent the syntactic structure of sentences in a more accurate and efficient manner compared to traditional rule-based parsers. Neural networks are also capable of capturing non-linear relationships in the data. This is particularly useful in syntactic parsing, as the relationships between words and their syntactic structure can be highly complex. The neural network can learn these hidden non-linear dependencies and capture them in its parameters, leading to improved parsing performance.

9. CONCLUSIONS

In this paper, firstly, we describe the original definition of TAG and its implementation of Early Type TAG parsing as proposed by Joshi then we also study in detail efforts by various researched to advancement in Parsing techniques. We also examined their advantages and Structural ambiguity, the dissimilarity between the elementary trees of different grammars also is another obstacle in identifying the link information between these grammars which required to further improvement in evolving parsing techniques. We have also propped new algorithm of TAG

parsing using statistical model. We also described an implementation of a statistical parser for TAGs in detail along with its advantages over conventional early type TAG parsing. We ran some empirical tests by running the parser on 7749 gold English sentences from the General domain. We used Treebank Grammar along with statistical model parse these sentences. We showed in the experiment that the time complexity of the parser on these sentences improves compare to conventional rule-based algorithm TAGs. During the Investigation of Parse derivations produced by the parser, we observed that variation in the number of derivations for the same sentence length also reduced in compare to conventional TAG Parser. We presented results that indicate that the number of trees selected by the given node during the adjunction operation (a measure of the syntactic lexical ambiguity of a sentence) is a better predictor of complexity in statistical model-based parsing.

We have demonstrated that our end-to-end statistical parsing algorithm outperforms our conventional early type implementation of TAG Parser. These results illustrate that TAG is a viable formalism for comprehensive syntax analysis of rich structural corpus.

REFERENCES

[1] K. Joshi, L. S. Levy, and M. Takahashi, "Tree adjunct grammars," Journal of Computer and System Sciences, vol. 10, no. 1, pp. 136-163, 1975.

[2] P. Kurariya, P. Chaudhary, J. Bodhankar, L. Singh, A. Kumar, and H. Darbari, "TREE ADJOINING GRAMMAR BASED 'LANGUAGE INDEPENDENT GENERATOR'," in Proceedings of the 17th International Conference on Natural Language Processing (ICON), Dec. 2020, pp. 138-143.

[3] K. Joshi, "An introduction to tree adjoining grammars," Mathematics of Language, vol. 1, pp. 87-115, 1987.

[4] Y. Schabes and A. Joshi, "An Earley-type parsing algorithm for tree adjoining grammars," in 26th Annual Meeting of the Association for Computational Linguistics, Jun. 1988, pp. 258-269.

[5] Sarkar, "Practical experiments in parsing using tree adjoining grammars," in Proceedings of the Fifth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+ 5), May 2000, pp. 193-198.

[6] P. Kurariya, P. Chaudhary, J. Bodhankar, L. Singh, A. Kumar, and H. Darbari, "VTAG: Virtual Lab for Tree-Adjoining Grammar-Based Research," in International Conference on Information and Communication Technology for Competitive Strategies, pp. 765-777, Singapore, Springer Nature Singapore, 2022.

[7] P. Kurariya, P. Chaudhary, P. Jain, A. Lele, A. Kumar, and H. Darbari, "File model approach to optimize the performance of Tree Adjoining Grammar based Machine Translation," in 2015 International Conference on Computer, Communication and Control (IC4), Sept. 2015, pp. 1-6.

[8] U Sarkar and A. Joshi, "Tree-adjoining grammars and its application to statistical parsing," Data-oriented parsing. CSLI, 2003.

[9] M. Collins, "A new statistical parser based on bigram lexical dependencies," arXiv preprint cmp-lg/9605012, 1996.

[10] M. Collins, "Three generative, lexicalised models for statistical parsing," arXiv preprint cmp-lg/9706022, 1997.

[11] P. Resnik, "Probabilistic tree-adjoining grammar as a framework for statistical natural language processing," in COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics, 1992.

[12] Y. Schabes and S. M. Shieber, "An alternative conception of tree-adjoining derivation," arXiv preprint cmp-lg/9404001, 1994.

[13] J. Kasai, R. Frank, R. T. McCoy, O. Rambow, and A. Nasr, "TAG parsing with neural networks and vector representations of supertags," in Conference on Empirical Methods in Natural Language Processing, Sept. 2017, pp. 1712-1722.

[14] Kuncoro, C. Dyer, J. Hale, D. Yogatama, S. Clark, and P. Blunsom, "LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jul. 2018, pp. 1426-1436.

AUTHORS

Mr. Pavan Kurariya is Joint Director of the Applied Artificial Intelligence Group, C-DAC (Centre for Development of Advanced Computing is the premier R&D organization of the Ministry of Electronics and Information Technology (MeitY), Government of India) and have more than 15 years of experience working in Natural Language Processing and Cyber Security. He is a distinguished researcher, and his expertise lies in various domains such as Natural Language Processing, Cyber security, Cryptography and Quantum Computing. He has contributed significantly to the advancements of Machine Translation and Cyber Security, Quantum Computing. His primary area of interest centres around Machine Translation and Cryptography where he investigates novel techniques and cutting-edge methodologies to enhance the accuracy and efficiency of the various applications.

Mr. Prashant Chaudhary is Joint Director of the Applied Artificial Intelligence Group, C-DAC (Centre for Development of Advanced Computing is the premier R&D organization of the Ministry of Electronics and Information Technology (MeitY), Government of India) and have more than 15 years of experience working in Natural Language Processing and Cyber Security. He is a distinguished researcher, and his expertise lies in various domains such as Natural Language Processing, Machine Translation, and Speech Technology. Though his numerous research papers, he has made significant contributions to the field of Tree Adjoining Grammar (TAG), by investigating the theoretical aspects and practical applications of TAG. His primary area of interest centres around Machine Translation where he investigates cutting-edge techniques and methodologies to enhance the accuracy and efficiency of the various NLP applications.

Ms. Jahnavi Bodhankar is Associate Director of the Applied Artificial Intelligence Group, C-DAC (Centre for Development of Advanced Computing is the premier R&D organization of the Ministry of Electronics and Information Technology (MeitY), Government of India) and have more than 18 years of experience working in Natural Language Processing and Cyber Security. She is a distinguished researcher, and her expertise lies in various domains such as Machine Translation, Electronic based signature, Machine Learning and Blockchain Technology. She has contributed significantly to the advancements and understanding of NLP and E-Signature, blockchain through her numerous research papers and intricate work.

Ms. Lenali Singh is Associate Director of the Applied Artificial Intelligence Group, C-DAC (Centre for Development of Advanced Computing is the premier R&D organization of the Ministry of Electronics and Information Technology (MeitY), Government of India) and have more than 20 years of experience working in Natural Language Processing. Her key role is in initiating and execution of various projects in the area of Natural Language Processing. She is a distinguished researcher, and her expertise lies in various domains such as Machine Translation, Speech Technology. She has contributed significantly to the advancements and understanding of NLP filed through her numerous research papers and intricate work.

Dr. Ajai Kumar is Senior Director and Head of the Applied Artificial Intelligence & GIST Group, C-DAC (Centre for Development of Advanced Computing is the premier R&D organization of the Ministry of Electronics and Information Technology (MeitY), Government of India) and have more than 20 years of experience working in Natural Language Processing including Machine Translation, Speech Technology and Information Extraction & Retrieval and E-learning systems. His key role is in initiating mission mode consortium projects in the area of Natural Language Processing, Speech Technology, Video Surveillance etc. Through his meticulous research, he aims to bridge the gap between different languages and enable seamless communication across linguistic boundaries.