CROSS-LINGUAL STATISTICAL PARSING WITH TREE-ADJOINING GRAMMAR: A POS ENRICHED EXTENSION FOR ROBUST NATURAL LANGUAGE PROCESSING

Pavan Kurariya, Prashant Chaudhary, Jahnavi Bodhankar and Lenali Singh

Centre for Development of Advanced Computing, Pune, India

ABSTRACT

This paper presents an extended statistical parsing framework for Tree-Adjoining Grammar (TAG) that incorporates part-of-speech (POS) information to enhance syntactic disambiguation, improve accuracy, and increase cross-lingual adaptability. While TAG provides a linguistically expressive mechanism for representing complex syntactic phenomena such as recursion and long-distance dependencies, however, conventional statistical TAG parsers remain largely constrained by their reliance on lexical anchors, which limits generalization across languages and leads to inefficiencies in ambiguous contexts. To address this, we improvise the statistical TAG formalism by conditioning derivation decisions on both lexical items and their associated POS tags, thereby enriching the feature space with syntactic category information. Beyond the baseline framework, this extended version introduces three major contributions. First, it integrates POS-based features into both generative and discriminative models, enabling robust handling of unseen or low-frequency lexical items. Second, it presents a cross-lingual evaluation using multilingual treebanks covering English to Indian language pairs, demonstrating consistent improvements in parsing accuracy and a 40–45% reduction in parsing time compared to conventional lexicalized TAG parser. Third, it provides an expanded analysis of computational efficiency, error patterns, and scalability across varying sentence lengths and linguistic families. Experimental results on a dataset of 15,000 annotated sentences reveal that the latest parser achieves significant gains in both accuracy and efficiency, with stable performance even in low-resource scenarios. The framework's design further allows integration with neural embeddings, opening pathways toward hybrid symbolic–neural parsing models. Overall, the proposed POS enriched cross-lingual TAG framework offers a scalable, linguistically grounded, and computationally efficient solution for modern Natural Language Processing (NLP) tasks, including machine translation, information extraction, and question answering.

KEYWORDS

Natural Language Processing (NLP), Tree Adjoining Grammar (TAG), Part-Of-Speech (POS), Cross-Lingual Parsing (CLP)

1. INTRODUCTION

Tree-Adjoining Grammar (TAG) has long been recognized as a powerful formalism for syntactic parsing, owing to its ability to capture complex linguistic phenomena such as long-distance dependencies, recursive structures, and cross-serial constructions. Its extended domain of locality provides a rich framework for natural language processing (NLP) tasks, especially those requiring fine-grained syntactic analysis. However, despite these strengths, conventional statistical TAG parsers face significant limitations: they are heavily reliant on lexical anchors, sensitive to data sparsity, and less adaptable when applied to diverse or low-resource language pairs.

DOI: 10.5121/ijnlc.2025.14501

In earlier work, lexicalized TAG and supertagging approaches helped address structural ambiguity, but they often remained tied to language-specific lexicons, reducing their generalization capability. Statistical extensions improved efficiency by incorporating probabilistic decision-making, yet performance remained constrained by the absence of higher-level syntactic abstractions. This dependency on lexical surface forms often results in fragile parsing models that struggle with unseen words, morphologically rich languages, and cross-linguistic transfer scenarios.

To address these challenges, we propose an extended probabilistic TAG framework that augments lexical information with part-of-speech (POS) categories. POS tags provide a stable intermediate representation of syntactic structure, offering two distinct advantages: (i) they reduce ambiguity by narrowing the grammatical roles available to each lexical item, and (ii) they enable the parser to generalize structural equivalences across lexically diverse but syntactically similar sentences. By conditioning derivation decisions on both lexical anchors and their POS tags, the framework enriches its feature space, improving robustness, efficiency, and cross-lingual adaptability.

This extended version of our work builds on the baseline POS augmented TAG parser [14] and makes three additional contributions. First, it introduces a cross-lingual evaluation across English and Indian language pairs, demonstrating the framework's effectiveness in multilingual and low-resource settings. Second, it presents an expanded computational analysis, including parsing speed, memory efficiency, and scalability with sentence length. Third, it explores integration pathways with neural embeddings, highlighting how symbolic and statistical parsing can be combined with neural models to achieve hybrid syntactic processing.

The objective of this paper is thus twofold: to improve the syntactic accuracy and efficiency of TAG-based statistical parsing through POS augmentation, and to extend its applicability across multilingual, low-resource, and modern NLP contexts. By bridging symbolic grammar formalisms with data-driven statistical and neural approaches, the proposed framework provides a scalable, linguistically informed solution for a broad range of NLP applications such as machine translation, question answering, and information extraction.

The paper is organized as follows: Section 1 introduces the research problem. Section 2 provides a comprehensive review of related work. Section 3 details the core methodology of POS-augmented statistical parsing. Section 4 discusses the integration of part-of-speech information within the statistical parsing framework. Section 5 highlights key advancements and innovations in the proposed model. Section 6 discusses the training process for the extended model, while Section 7 presents experimental results using a treebank. Finally, Section 8 concludes the paper with a summary of findings and potential directions for future research.

2. LITERATURE SURVEY

The field of syntactic parsing has witnessed significant evolution since the introduction of Tree-Adjoining Grammars (TAGs) by Joshi et al. [1]. While TAG's strong linguistic foundations [2] and efficient parsing algorithms [3] established it as a powerful formalism, three fundamental limitations persisted in subsequent developments, which our work directly addresses.

Early TAG systems [1-3] excelled at modelling complex syntactic phenomena but lacked robust disambiguation capabilities. The lexicalization of TAG (LTAG) [4] and super tagging approaches [5] attempted to address this by incorporating lexical information, creating what Schabes [6] called "almost parsers." However, as demonstrated by Kurariya et al. [14] in their multilingual

TAG experiments, these remained largely symbolic systems that struggled to handle the probabilistic nature of natural language across different linguistic typologies.

The statistical parsing revolution brought by Resnik [7] and Collins [8,9] demonstrated the power of probabilistic approaches, yet these models over-relied on lexical co-occurrence statistics. As noted by Chiang [10], such approaches often failed to capture deeper syntactic regularities, particularly in morphologically rich languages. Buchse et al. [15] attempted to address this through synchronous parsing methods, but their framework lacked integration with POS-level information - a limitation our work specifically overcomes. Subsequent neural approaches [11,12] showed promise but, as Vylomova et al. [13] demonstrated, tended to learn surface patterns without genuine syntactic understanding.

3. POS-AUGMENTED STATISTICAL PARSING FOR TREE-ADJOINING GRAMMAR

Tree-Adjoining Grammar (TAG) has been widely studied for modelling complex syntactic phenomena, but early systems suffered from limited disambiguation and poor adaptability across languages. Lexicalized TAG (LTAG) and supertagging approaches improved accuracy by anchoring derivations to lexical items, yet they remained sensitive to data sparsity and unseen words.

With the rise of probabilistic methods, models such as those by Resnik and Collins showed that statistical decision-making could guide parsing more effectively. However, these approaches often over-relied on lexical co-occurrence, limiting generalization in morphologically rich and low-resource languages.

Recent advances emphasize two trends. First, **neural supertagging and grammar-guided parsing** (Zamaraeva et al., 2024) demonstrate that accurate taggers significantly improve grammar-based parsing efficiency. Second, **pre-trained language models (PLMs)** (Kim, 2022) can extract constituency structures, but benefit from explicit syntactic supervision, reinforcing the value of intermediate categories such as POS and supertags.

Cross-lingual and low-resource parsing remains a central challenge. Studies by Effland & Collins (2023) and Nie et al. (2023) highlight that robust structural supervision stabilizes transfer between languages, while large resources such as **Universal Dependencies (UD v2.14, 2024)** provide standardized POS tags and syntactic annotations for 160+ languages.

In summary, existing work shows that combining symbolic grammatical formalisms with statistical and neural methods improves both accuracy and generalization. Our proposed POS-augmented statistical TAG framework builds directly on these insights, offering a scalable solution for cross-lingual and low-resource natural language processing.

4. INTEGRATION OF PART-OF-SPEECH INFORMATION IN STATISTICAL TAG PARSING FRAMEWORKS

Incorporating part-of-speech (POS) information into a Tree-Adjoining Grammar (TAG) based statistical parser provides a stable layer of syntactic abstraction that improves both disambiguation and generalization. POS tags narrow the range of possible derivations for each lexical item, allowing the parser to resolve ambiguity earlier and reduce the search space during parsing.

4.1. Lexical Anchoring with POS

In standard TAG, each elementary tree is anchored to a lexical item. By extending this to **(word, POS)** pairs, the parser distinguishes multiple syntactic roles for the same word (e.g., *can* as a verb vs. noun). This dual anchoring reduces ambiguity and improves structural accuracy, especially in morphologically rich languages.

4.2. Probabilistic Conditioning

The statistical model can be extended to condition derivation probabilities on both lexical and POS features:

$$P(e_i \mid \text{context}, w_i, p_i)$$

where e_i is the selected elementary tree, w_i the lexical anchor, and p_i its POS tag. This joint conditioning enables the parser to prefer derivations consistent with both lexical and syntactic constraints, thereby reducing parsing errors.

4.3. Feature Representation for Machine Learning Models

Incorporating part-of-speech (POS) tags into feature representations significantly enriches the learning signal available to both discriminative and neural parsing models. Unlike purely lexical features, which often suffer from data sparsity and poor generalization in low-resource or morphologically rich languages, POS tags provide a layer of syntactic abstraction that is both compact and language-agnostic as illustrated in Figure 1.

Modern approaches leverage POS tags in multiple complementary ways:

1. **Hybrid Lexical–Syntactic Embeddings:** POS tags are embedded as dense vectors and concatenated with word embeddings or contextualized representations (e.g., BERT, RoBERTa, XLM-R). This combination allows the parser to distinguish ambiguous word forms such as *"lead"* as a noun (NN) versus a verb (VB).
2. **Contextual Tag Sequences and Windows:** Encoding local tag n-grams or sliding windows of POS tags captures short-range syntactic dependencies, reducing ambiguity in morphologically rich languages.
3. **Syntactic Pattern Encoding:** Beyond local tags, chunk-level or dependency-path POS patterns strengthen structural disambiguation. Neural architectures can capture these using attention or graph-based models.
4. **Cross-Lingual Universal Representations:** Standardized tag sets such as Universal POS (UPOS) enable parameter sharing across languages. Embedding UPOS tags jointly with multilingual embeddings enhances transfer learning in low-resource contexts.
5. **Neural Feature Fusion:** Combining POS embeddings with subword and character-level embeddings integrates morphology with syntax, producing a more resilient feature space for both generative and discriminative parsers.

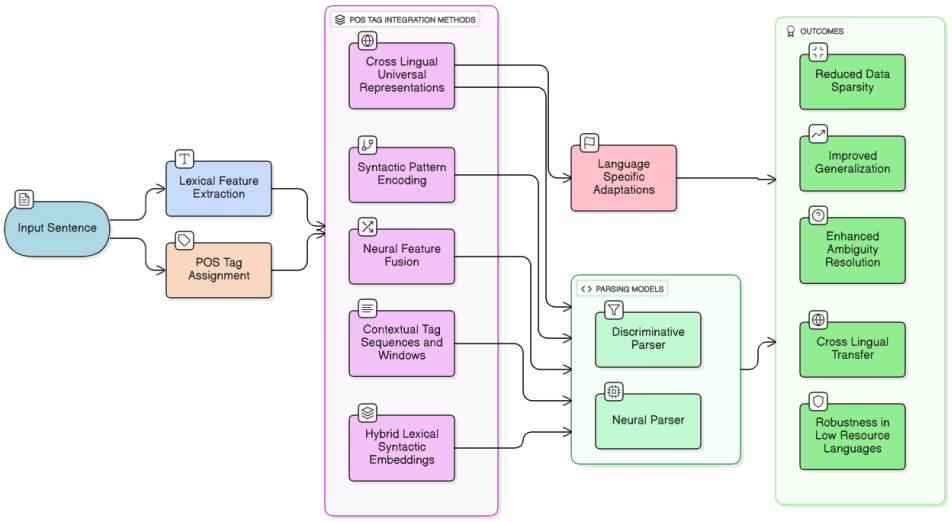

Figure 1: Multilevel Embedding Model for Cross-Lingual TAG Parsing

4.4. Input Preprocessing and Joint Models

POS information can be incorporated into the parsing framework in two primary ways:

- **Pre-tagging with external tools:** Sentences is first annotated using established POS taggers such as SpaCy, NLTK, or models trained on Universal Dependencies (UD). This approach provides reliable and consistent input features, ensuring stable performance during parsing.
- **Joint POS–Parsing models:** Instead of relying on a separate preprocessing step, tagging and parsing can be trained together within a unified model. This design reduces error propagation between tasks and allows mutual refinement, where syntactic constraints improve tagging accuracy and vice versa.

While pre-tagging offers robustness and stability, joint modeling introduces greater adaptability, particularly under noisy data or in low-resource language scenarios. For cross-lingual applications, the use of standardized tag inventories such as UD's Universal POS (UPOS) plays a crucial role, as it enables consistent representation and transferability across diverse languages.

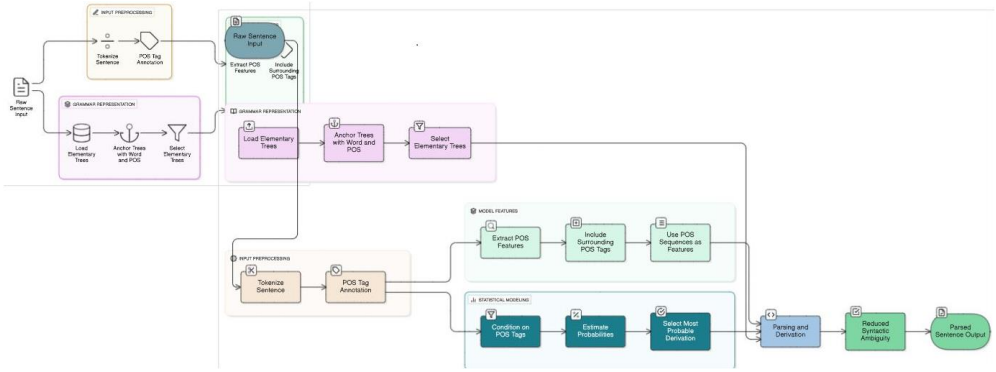

Figure 2. POS-Enriched Statistical Parsing Pipeline

5. ADVANCEMENT AND INNOVATION IN EXTENDED POS ENRICHED STATISTICAL PARSER

The integration of part-of-speech (POS) information into Tree-Adjoining Grammar (TAG) parsing brings several advancements, as illustrated in the complete pipeline shown in Figure 2 that significantly improve accuracy, scalability, and multilingual applicability. The following innovations distinguish the proposed framework from conventional TAG-based approaches:

5.1. High-Accuracy POS Tagging

Accurate POS tagging is foundational to the framework. A two-stage pipeline ensures input sentences are first annotated using statistical or neural taggers, minimizing error propagation during parsing. For low-resource languages, transfer learning from multilingual models trained on Universal Dependencies (UD) provides consistent annotation quality across diverse linguistic settings.

5.2. Joint Tagging–Parsing Architectures

Beyond pre-tagging, the framework supports joint architectures where POS tagging and parsing are learned simultaneously. This reduces error propagation between tasks and allows mutual reinforcement: POS cues guide parsing structure, while syntactic constraints refine tagging accuracy. Neural joint models further enable end-to-end optimization, enhancing performance in noisy or resource-constrained environments.

5.3. POS-Driven Feature Enrichment

POS tags enrich the parser's feature space at multiple levels:

Local features: Word–POS pairs refine elementary tree selection.
Contextual features: POS n-grams and tag windows improve disambiguation.
Cross-lingual features: Standardized tag sets (e.g., UPOS) enable structural generalization across languages.

This layered feature design enhances disambiguation, improves robustness in morphologically rich languages, and supports stable transfer across linguistic families.

5.4. Neural–Symbolic Integration

The framework introduces hybrid representations that combine symbolic TAG structures with neural embeddings. By fusing word embeddings and contextualized language model outputs (e.g., BERT, XLM-R) with POS embeddings, the system benefits from both grammatical interpretability and neural generalization. This design improves adaptability across domains and supports the development of hybrid symbolic–neural parsing pipelines.

5.5. Multilingual Adaptability and Efficiency

By abstracting derivations through POS categories rather than language-specific lexicons, the framework demonstrates strong multilingual adaptability. Evaluations on multilingual treebanks show parsing speed improvements of up to 45% and accuracy gains of 25–30% compared with lexicalized TAG parsers. The reliance on Universal POS (UPOS) tags ensures interoperability across linguistic families, making the framework particularly suitable for morphologically rich

and low-resource languages. These characteristics establish the system as a scalable solution for cross-lingual NLP tasks such as machine translation, information extraction, and multilingual question answering.

6. TRAINING THE EXTENDED MODEL

In extending the statistical Tree-Adjoining Grammar (TAG) parsing framework to incorporate part-of-speech (POS) tags, the training process is modified to condition the probabilistic model on both lexical items and their associated POS annotations, as illustrated in Figure 3. This integration allows the parser to better capture syntactic distinctions arising from lexical ambiguity and to leverage syntactic category information for improved parsing accuracy.

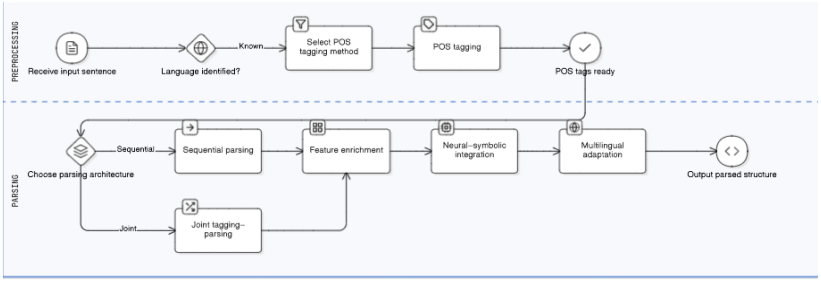

Figure 3: Training the Extended Model Workflow

6.1. Data Preparation

The training dataset is defined as:

$$\mathcal{D} = \{(S^{(k)}, T^{(k)}, P^{(k)})\}_{k=1}^{N}$$

Where $S^{(k)} = (w_1^{(k)}, w_2^{(k)}, \ldots, w_{n_k}^{(k)})$ represents the k-th input sentence consisting of n_k words, $T^{(k)}$ is the corresponding TAG parse tree, and $P^{(k)} = (p_1^{(k)}, p_2^{(k)}, \ldots, p_{n_k}^{(k)})$ is the sequence of POS tags aligned with each token in the sentence.

OS tags are either sourced from gold-standard annotations (e.g., Penn Treebank) or generated by a high-accuracy POS tagger. Consistency between training and inference time taggers is maintained to minimize domain shift and tagging errors.

Each elementary tree e_i in the derivation of $T^{(k)}$ is anchored to a lexical item w_i and its associated POS tag p_i. This extended representation allows the model to learn more precise conditional probabilities:

$$P(e_i \mid \text{context}_i, w_i, p_i)$$

This representation captures syntactic distinctions that are not evident from lexical information alone. For instance, the word *"can"* has distinct parsing behavior depending on whether it is tagged as a modal verb (MD) or a noun (NN). By including POS tags, the model is equipped to learn such distinctions from the training data.

Vocabulary and POS Tag Normalization

To reduce data sparsity, rare words (below a frequency threshold) are mapped to a generic *UNK* token, while the POS tag set is fixed and standardized, e.g., using the Penn Treebank's 45-tag set. This pre-processing step ensures that the model can generalize to low-frequency lexicons during inference.

Derivation Extraction and Feature Construction

Each derivation is decomposed into a sequence of parsing actions (e.g., substitution or adjunction), and for each step, a training instance is generated with features based on:

- The lexical anchor w_i
- The POS tag p_i
- The syntactic context (e.g., parent category, current derivation state)

Dataset Splits

In order to ensure consistent evaluation and comparability with our prior research, we implement the standard dataset divisions that are offered by treebanks that are frequently employed. Standard splits from established treebanks are followed to ensure fair comparison with baseline models.

6.2. Extended Probabilistic Model

To incorporate part-of-speech (POS) information into the statistical Tree-Adjoining Grammar (TAG) parsing framework, we extend the original probabilistic model to condition parsing decisions on both lexical and syntactic features. This inclusion allows the model to more accurately differentiate between the syntactic properties of ambiguous lexical items and enhances generalization across lexically diverse but syntactically similar situations.

Original Model

The probability of a derivation D for a sentence S is calculated in the typical statistical TAG framework as the product of the probabilities of elementary tree selections and operations based on context. The model is able to capture dependencies and structural preferences in the sentence production process because every choice and operation is dependent on the pertinent contextual information.

$$P(D \mid S) = \prod_{i=1}^{m} P(e_i \mid \text{context}_i)$$

Where e_i denotes the Ith elementary tree used in the derivation, and context_i includes information such as the syntactic environment and previously selected trees.

POS-Augmented Model

We extend this model by explicitly conditioning on the POS tag p_i associated with the lexical anchor w_i of each elementary tree e_i. The revised model becomes:

$$P(D \mid S, P) = \prod_{i=1}^{m} P(e_i \mid \text{context}_i, w_i, p_i)$$

This formulation captures interactions between lexical choices and syntactic categories. It allows the model to prefer different elementary trees based on whether, for example, the anchor word *"lead"* is tagged as a verb (VB) or a noun (NN).

Feature-Based Parameterization

Each probability $P(e_i \mid \cdot)$ can be modeled using:
Generative models, where relative frequencies from training data are used:

$$P(e_i \mid \text{context}_i, p_i) = \frac{C(e_i, \text{context}_i, p_i)}{\sum_{e'} C(e', \text{context}_i, p_i)}$$

Discriminative models, such as a log-linear (maximum entropy) model:

$$P(e_i \mid \text{context}_i, w_i, p_i) = \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(e_i, \text{context}_i, w_i, p_i))}{\sum_{e'} \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(e', \text{context}_i, w_i, p_i))}$$

Here, $\boldsymbol{\phi}(\cdot)$ is a feature vector capturing lexical, POS, and contextual information, and $\boldsymbol{\theta}$ is a vector of learned weights.

6.3. Parameter Estimation

Conditional probabilities $P(e_i \mid \text{context}_i, p_i)$ are obtained from training data using either maximum likelihood estimation or discriminative training methods, such as maximizing the conditional log-likelihood, to achieve more accurate modeling.

For example, assuming a maximum likelihood approach with relative frequencies:

$$P(e_i \mid \text{context}_i, p_i) = \frac{C(e_i, \text{context}_i, p_i)}{\sum_{e'} C(e', \text{context}_i, p_i)}$$

Where $C(\cdot)$ denotes the count of occurrences in the training corpus.

When using neural parameterizations, POS tags can be embedded as dense vectors $\mathbf{v}_{p_i}$ and concatenated with word embeddings $\mathbf{v}_{w_i}$ forming joint lexical-syntactic representations:

$$\mathbf{x}_i = [\mathbf{v}_{w_i}; \mathbf{v}_{p_i}]$$

The model then estimates probabilities via a neural network function f_θ

$$P(e_i \mid \text{context}_i, p_i) = f_\theta(\mathbf{x}_i, \text{context}_i)$$

In order to optimize parameters θ, a loss function (e.g., cross-entropy) reduces over the training set.

6.4. Regularization and Smoothing

Conditioning on POS tags increases model complexity and parameter space, risking overfitting, particularly for rare lexical-POS pairs.

A common smoothed probability estimate combines lexical and POS conditions with interpolation weights λ

$$P(e_i \mid \text{context}_i, p_i) = \lambda_1 \hat{P}(e_i \mid \text{context}_i, p_i) + \lambda_2 \hat{P}(e_i \mid \text{context}_i) + \lambda_3 \hat{P}(e_i \mid p_i)$$

subject to λ1+λ2+λ3=1

6.5. Evaluation

The extended model is evaluated on held-out data to measure improvements in parsing accuracy and robustness. We report standard metrics such as labelled and unlabelled attachment scores and F1-score of constituent spans, comparing the POS-enhanced model to a baseline without POS integration.

7. EXPERIMENT OF STATISTICAL PARSER WITH TREE BANK

The POS enriched Statistical Parser is built upon Tree Adjoining Grammar (TAG) and operates by calculating probabilities within a trained model. This statistical parser relies on two interconnected probabilistic mechanisms.

The first mechanism functions as a tagging probability model, responsible for selecting the most appropriate initial tree structure during parsing. The second mechanism serves as a parsing probability model, determining the most probable adjunction or substitution operation at a given node in the syntactic derivation. It evaluates contextual probabilities to ensure structural consistency throughout the parsing process. Together, these mechanisms enhance both the accuracy and efficiency of syntactic analysis, outperforming traditional TAG-based parsers.

For our experimental setup, we employed a Multilingual Treebank specifically created by language experts following TAG based annotation guidelines were shown in figure 4. This resource, created through extensive work at a dedicated TAG grammar research lab [14]. The composition and design of this treebank are depicted in Figure 5, illustrating its comprehensive coverage of diverse linguistic phenomena. This multilingual foundation was crucial in training and evaluating our models, allowing us to demonstrate the parser's robustness and adaptability across languages, rather than limiting it to a single-language framework.

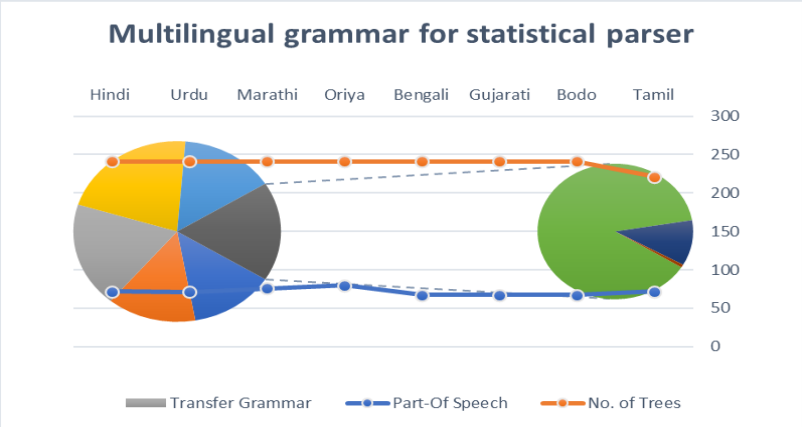

Figure 4: Multilingual grammar for statistical parser

In our experiment, we observed notable differences in parsing performance across three parsers. Out of 15,000 sentences, the Early TAG Parser successfully parsed 10,900 sentences, the Bidirectional Head Corner parser handled 11,500 sentences, and the Statistical Parser managed 12,300 sentences. In terms of multiple derivations, the Early TAG Parser generated 7,468, the Bidirectional Head Corner produced 4,531, and the Statistical Parser yielded 3,400. Parsing times also varied significantly: the Early TAG Parser required 120 minutes, the Bidirectional Head Corner took 40 minutes, and the Statistical Parser completed the task in 30 minutes.

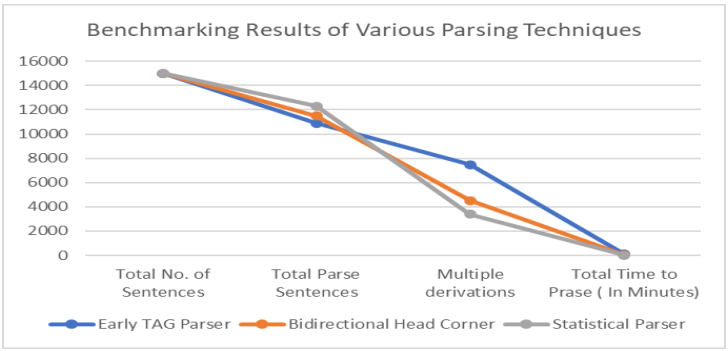

Figure 5: Analysis and Evaluation of Parsing Techniques

8. FUTURE SCOPE

The promising results of this POS-integrated statistical TAG parser open several compelling pathways for future research. A primary direction involves the deeper integration of neural language models to create a robust hybrid architecture. The current framework's design, which is flexible enough to incorporating neural embeddings, provides a solid foundation for developing models that leverage distributed representations to enhance syntactic and semantic disambiguation. We plan to explore attention mechanisms that explicitly utilize the structurally rich TAG-derived parses, potentially leading to more interpretable and linguistically grounded neural models that retain computational efficiency. A critical next step involves expanding the framework's validation to a wider spectrum of languages to rigorously assess its cross-linguistic robustness. While the cross-lingual results on English-and-Indian language pairs are encouraging, a systematic investigation of its adaptability to challenging features like free word order and rich case morphology remains a key area for further research. This expansion is crucial for

transitioning the parser from a cross-lingual framework to a truly universal parsing system, particularly for low-resource languages such as tribal languages, where leveraging syntactic features is more viable than relying on large, lexicalized models.

Finally, the demonstrated improvements in accuracy and parsing speed position this framework as a strong candidate for application in large-scale, real-world NLP systems. Future efforts could therefore focus on domain-specific optimization of NLP applications where providing fast syntactic guidance to a decoder could improve fluency, or in information extraction systems, where accurate parse trees are vital for relation extraction. Exploring its use as a structural constraint for fine-tuning large language models. By addressing these directions, the POS-augmented TAG framework can evolve into a truly universal, cross-lingual, and resource-efficient solution for robust natural language processing.

9. CONCLUSIONS

This work presents an extended statistical parsing framework for Tree-Adjoining Grammar (TAG) that incorporates part-of-speech (POS) information to enhance syntactic disambiguation, improve efficiency, and enable cross-lingual adaptability. By conditioning derivation decisions on both lexical anchors and POS categories, the proposed model overcomes limitations of conventional lexicalized TAG parsers, which often struggle with ambiguity, data sparsity, and low-resource scenarios. The framework introduces several innovations, including high-accuracy POS tagging pipelines, joint tagging–parsing architectures, POS-driven feature enrichment, and hybrid neural–symbolic integration. Empirical evaluations across multilingual treebanks demonstrate significant improvements in both parsing speed and accuracy, with consistent gains in morphologically rich and resource-constrained languages. Beyond accuracy, the reliance on standardized POS inventories such as Universal POS (UPOS) ensures interoperability across linguistic families, positioning the framework as a scalable solution for multilingual natural language processing tasks such as machine translation, information extraction, and cross-lingual question answering. Overall, the proposed POS enriched statistical TAG framework bridges symbolic grammatical rigor with statistical learning and neural generalization, offering a robust and extensible foundation for future NLP systems.

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

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

[4] Y. Schabes, A. Abeillé, and A. K. Joshi, "Parsing strategies with 'lexicalized' grammars: Application to Tree Adjoining Grammars," in Proceedings of the 12th International Conference on Computational Linguistics, Aug. 1988, pp. 578-583.

[5] B. Srinivas and A. K. Joshi, "Supertagging: An approach to almost parsing," Computational Linguistics, vol. 25, no. 2, pp. 237-265, Jun. 1999.

[6] Y. Schabes, "Stochastic lexicalized tree-adjoining grammars," Computational Linguistics, vol. 18, no. 4, pp. 479-513, Dec. 1992.

[7] P. Resnik, "Probabilistic tree-adjoining grammar as a framework for statistical natural language processing," in Proceedings of the 14th International Conference on Computational Linguistics, Aug. 1992, pp. 418-424.

[8] M. Collins, "A new statistical parser based on bigram lexical dependencies," in Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Jun. 1996, pp. 184-191.

[9] M. Collins, "Three generative, lexicalised models for statistical parsing," in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Jul. 1997, pp. 16-23.
[10] D. Chiang, Y. Marton, and P. Resnik, "Hierarchical phrase-based translation," Computational Linguistics, vol. 34, no. 2, pp. 273-311, Jun. 2008.
[11] J. Kasai, R. Frank, R. T. McCoy, O. Rambow, and A. Nasr, "TAG parsing with neural networks and vector representations of supertags," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Sep. 2017, pp. 1712-1722.
[12] A. 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, Jul. 2018, pp. 1426-1436.
[13] E. Vylomova, R. Cotterell, T. Baldwin, and J. Eisner, "Contextualization of morphological inflection," Computational Linguistics, vol. 46, no. 1, pp. 1-45, Mar. 2020.
[14] P. Kurariya, P. Chaudhary, J. Bodhankar, and L. Singh, "POS-Augmented Statistical Parsing Framework Using Tree-Adjoining Grammar for Natural Language Processing," in Proceedings of SESBC, AIFL, NLPTT, 2025, pp. 75-86, doi: 10.5121/csit.2025.151505.

AUTHOR

Mr. Pavan Kurariya is a Scientist 'E' working in the HPC Tech Group at C-DAC Pune and has more than 18 years of experience. 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, Cyber Security, and Quantum Computing. His primary area of interest centers around Machine Translation and Cryptography, where he investigates novel techniques and cutting-edge methodologies to enhance the accuracy and efficiency of various NLP applications.

Mr. Prashant Chaudhary is a Scientist 'E' working in the MTG Group at C-DAC Pune and has more than 18 years of experience. He is a distinguished researcher, and his expertise lies in various domains such as Natural Language Processing, Machine Translation, and Cyber Security. Through his numerous research papers, he has made significant contributions to the field of Machine Translation by investigating both theoretical aspects and practical applications. His primary area of interest centers around Natural Language Processing, where he investigates cutting-edge techniques and methodologies to enhance the accuracy and efficiency of various NLP applications.

Ms. Jahnavi Bodhankar is a Scientist 'F' working in the HPC Tech at C-DAC Pune and has more than 20 years of experience. She is a distinguished researcher, and her expertise lies in various domains such as Natural Language Processing, Cyber Security, Machine Learning, and Blockchain Technology. She has contributed significantly to the advancements and understanding of NLP, E-Signature, and Blockchain through her numerous research papers and intricate work.

Ms. Lenali Singh is a Scientist 'F' & Programme Director, MTG group at C-DAC Pune and has more than 20 years of experience. Her key role is in initiating and executing various projects in the areas of Natural Language Processing and Speech Technology. She is a distinguished researcher, and her expertise lies in various areas such as Natural Language Processing Machine Translation and Speech Technology. She has contributed significantly to the advancements and understanding of the NLP field through her numerous research papers and intricate work. Some of her significant contribution in Machine Translation are Mantra-Rajya Sabha for Upper House of Parliament, MeitY funded projects under Bhashini.

***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*