

Unsupervised Part-Of-Speech Induction For Assamese Language Using Hierarchical Distributional Clustering

Biswajit Sarma^{1*}, Dr. Rupam Baruah², Dr. Diganta Baishya³

^{1*}Assistant Professor Jorhat Engineering College, Assam Science and Technology University (ASTU) ORCID ID: 0009-0005-7034-5842

²Professor and Principal Jorhat Engineering College, Assam Science and Technology University (ASTU) ORCID ID: 0000-0001-7236-5314

³Assistant Professor Jorhat Engineering College, Assam Science and Technology University (ASTU) ORCID ID: 0000-0002-1430-5142

Abstract—Part-of-Speech (POS) tagging is a core task in Natural Language Processing (NLP) and underpins a wide range of downstream applications such as syntactic parsing, information extraction, and machine translation. Despite the success of supervised POS taggers for high-resource languages, their reliance on large annotated corpora makes them impractical for low-resource languages such as Assamese.

This paper presents a fully unsupervised framework for POS induction in Assamese based on hierarchical distributional clustering inspired by Brown clustering. Words are represented using local contextual co-occurrence statistics derived from bigram and trigram contexts and are initially treated as singleton clusters. These clusters are iteratively merged using a likelihood-based objective until no further improvement is possible. Gold-standard POS annotations are used exclusively for evaluation via majority cluster mapping. Experimental results show that the proposed approach effectively induces major syntactic categories, including nouns, verbs, pronouns, and conjunctions, with high accuracy, demonstrating the practical viability of distributional clustering for linguistic analysis in low-resource settings.

Keywords— Unsupervised POS Tagging, Assamese Language, Distributional Clustering, Brown Clustering, Low Resource NLP.

I. INTRODUCTION

Word Part-of-Speech (POS) tagging assigns syntactic categories, for example, noun, verb, adjective, and adverb, to words in text and is one of the primary preprocessing steps in most NLP pipelines. Proper POS information is pivotal for parsing, machine translation, sentiment analysis, and information extraction, among others. While supervised POS tagging methods reach the cutting edge of performance when working with languages like English, they rely heavily on large-scale manually annotated corpora.

For low-resource languages like Assamese, the dearth of annotated data poses a major challenge to supervised learning methods. Unsupervised POS induction aims at overcoming this limitation by detecting syntactic categories directly from raw text, instead of using labeled training data. These approaches are based on the distributional hypothesis, which states that word occurrences that are found in similar environments are correlated to similar syntactic properties. In this work, we propose an unsupervised POS induction framework for Assamese by hierarchical agglomerative clustering and using distributional statistics.

II. RELATED WORK

Early work on unsupervised POS induction focused on distributional clustering methods. Brown et al. Introduced class-based n-gram models using hierarchical word clustering [1], which has since influenced numerous NLP applications. Subsequent approaches explored

Hidden Markov Models, Bayesian nonparametric methods, and graph-based techniques for unsupervised POS tagging [2], [3].

Christodoulopoulos et al. provided a large-scale evaluation of unsupervised POS induction methods, highlighting both progress and remaining challenges [4]. Despite these advances, relatively little work has addressed POS induction for Indian and low-resource languages such as Assamese. This work seeks to bridge that gap by applying hierarchical distributional clustering to Assamese text.

III. CORPUS AND PREPROCESSING

The corpus of text for the research includes Assamese literature that is published from open-source digital sources such as online news sites, free literary websites, and language sources published by academic and public bodies. Sources were chosen to represent diversity of writing style, topic, and syntactic forms, so as to have a relatively representative sample of current usage of Assamese.

Words in the raw text were tokenized and sentence boundaries were retained in order to retain local contextual dependencies needed for the distributional assessment. This preservation of sentence structure makes possible the accurate extraction of bigram and trigram contexts that underlies our proposed hierarchical clustering method.

The corpus contains gold-standard Part-of-Speech (POS) annotations, but these are solely used to perform evaluations and are not used at any stage of model training or clustering. The reason behind this design choice is that the proposed framework is completely unsupervised, and will not have any meaningful benefit from labeled linguistic information during induction.

A series of preprocessing steps were applied to improve corpus consistency and reduce noise. Text normalization was first performed to standardize orthographic variations and ensure consistent Unicode representations of Assamese characters. Punctuation handling was then applied to remove non-informative symbols while preserving sentence delimiters, allowing sentence boundaries to remain intact. In addition, non-linguistic tokens such as numerals, URLs, and formatting artifacts were removed, as these elements do not contribute meaningful syntactic information. These preprocessing steps result in a clean and linguistically coherent corpus that supports reliable estimation of contextual co-occurrence statistics and facilitates effective distributional learning for unsupervised POS induction.

IV. DISTRIBUTIONAL REPRESENTATION

The corpus employs distributions of each word based on the given linguistic context consistent with the distributional approach, which states that words in similar contexts yield comparable syntactic properties. Contextual information is extracted from bigram and trigram co-occurrence statistics, which summarize both the left and right contextual dependencies between key words as seen in this analysis.

This kind of local context is crucial for finding syntactic behavior as part-of-speech categories are very closely linked to adjacent word sequences. For each word, its surrounding context is the words of proximal value within a fixed window, and the model is capable of capturing common syntactic sequences, e.g., determiners preceding nouns or auxiliary verbs before main verbs. Bigram statistics show the immediate dependence between consecutive words, and trigram statistics show the interaction among 3-word sequences to provide the contextual resolution. Collectively, these representations cover the intricate syntactic regularities of the two classes, which are required for unsupervised POS induction.

Formally, let $f(w, c)$ denote the frequency with which a word w occurs in a particular context c . That is, a bigram or trigram configuration. These co-occurrence frequencies are aggregated within the corpus to produce a distributional profile for each word. Words having roughly similar distributional profiles are assumed to have identical syntactic functions and hence form the same cluster. These distributional statistics lay the basis of the clustering objective and support the incremental amalgamation of word clusters as we go through POS induction..

V. HIERARCHICAL CLUSTERING METHOD

This section describes a hierarchical distributional clustering framework used for unsupervised Part of Speech (POS) induction. The method of design follows the hierarchical word clustering method proposed by Brown et al., and operates by iteratively merging word clusters based on their distributional similarity and contribution to a likelihood-based objective function. The process of clustering aims to organise words with similar syntactic behavior, as captured by their contextual usage patterns, into common clusters, which will then match what we identify as latent POS categories.

A. Initialization

Let $W = \{w_1, w_2, \dots, w_n\}$ denote the vocabulary extracted from the corpus after preprocessing. In the initial stage, each word $w_i \in W$ is assigned to a unique cluster c_i . This finegrained initialization ensures that no prior assumptions are made about syntactic categories and allows the clustering process to progressively generalize from individual word types to broader syntactic classes. Starting from singleton clusters also maximizes the flexibility of the merging process, enabling the algorithm to explore a wide range of possible cluster configurations.

B. Likelihood-Based Objective

Given a current clustering configuration C , the quality of the clustering is measured using a likelihood-based objective function defined over the observed corpus. Specifically, the likelihood of the corpus under clustering C is defined as

$$L(C) = \sum_i \log P(w_i | c(w_i)), \quad (1)$$

where $c(w_i)$ denotes the cluster assignment of word w_i . The conditional probability $P(w_i | c(w_i))$ is estimated from the empirical distribution of words within each cluster based on their observed contextual frequencies.

At each iteration, the algorithm considers merging a pair of clusters (c_1, c_2) and evaluates the resulting change in log likelihood:

$$\Delta L(c_1, c_2) = L(C_{\text{merge}}) - L(C), \quad (2)$$

where C_{merge} denotes the clustering configuration obtained by merging clusters c_1 and c_2 . The pair of clusters that yields the maximum positive increase in log-likelihood is selected for merging. This greedy merging strategy ensures that each merge step improves or preserves the overall model fit to the data.

C. Contextual Probability Estimation

Estimation of cluster-based probabilities is based on the distributional representation shown in the previous section. For each cluster, emission probabilities are computed based on the relative frequency of words assigned to that cluster; transition probabilities capture the likelihood of observing specific cluster sequences in the corpus. These probabilities are derived from bigram and trigram co-occurrence statistics, and encode syntactic regularities found in local contexts.

By modelling both emission and transition behavior, this clustering framework can capture not just word-level distributional similarity but also the kinds of higher-order contextual dependencies that are especially important in order to distinguish categories of words with overlapping lexical distributions.

D. Iterative Merging Procedure

The clustering process is done iteratively and from a bottom up. At each operation, the algorithm evaluates candidate cluster merges and selects the merge that maximizes $\Delta L(c_1, c_2)$. During the merge, all associated statistics and probability estimates of the cluster are updated to reflect the new cluster configuration.

Such iterative merging is repeated until no pair of clusters yields a positive increase in log-likelihood, or until an already

Fig. 1. System architecture of the proposed unsupervised POS induction pipeline.



pre-defined criterion such as a target number of clusters is achieved. This hierarchy of merges naturally forms a dendrogram, from which clusterings at different levels of granularity can be extracted.

E. Stopping Criterion

The clustering process terminates when further merging fails to improve the likelihood objective, i.e., when $\Delta L(c_1, c_2) \leq 0$ for all candidate cluster pairs. Alternatively, the process may be stopped once the number of clusters reaches a predefined target corresponding to the desired level of syntactic granularity. In the present study, the stopping criterion is selected to balance cluster interpretability with empirical POS induction accuracy.

The final clustering configuration represents the induced POS categories, with each cluster corresponding to a latent syntactic class. These clusters are subsequently mapped to gold-standard POS tags using majority voting for evaluation purposes.

F. System Architecture

Fig. 1 illustrates the overall architecture of the proposed system, from raw corpus input to induced POS clusters.

VI. EVALUATION METHODOLOGY

Since the proposed POS induction procedure is unsupervised in nature, we have to use the external reference annotation for quantitative evaluation. In this study, gold-standard Part-of-Speech (POS) tags are used for evaluation only, and not in model training,

clustering, or parameter estimation at any phase. This allows for a fair assessment of the induced clusters without introducing supervised bias.

A. Cluster-to-POS Mapping

The induced clusters do not necessarily map to predefined POS labels. Hence an explicit mapping needs to be made from clusters to POS tags in order to be evaluated. Each cluster is mapped to one POS tag using a majority voting strategy over the gold annotations of the tokens assigned to this cluster. Formally, the POS label assigned to a cluster c is defined as

$$\text{POS}(c) = \arg \max_{t \in \mathcal{T}} \text{Count}(t, c), \quad (3)$$

$\mathcal{T}$ is the gold POS tag set, $\text{Count}(t, c)$ is the number of tokens in cluster c that are annotated with tag t in the gold standard. This mapping assigns each cluster the POS label that is most commonly seen among its constituent tokens.

B. Token-Level Accuracy

After mapping the clusters to POS tags, performance is assessed using token-level accuracy that measures the proportion of correctly labeled tokens in the corpus. Let y_i denote the gold POS tag of token i , and let $\hat{y}_i$ mean the predicted POS tag from the cluster-to-POS mapping. Token-level accuracy is described as

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(y_i = \hat{y}_i), \quad (4)$$

where N is the total number of tokens in the evaluation corpus and $\mathbb{I}(\cdot)$ is the indicator function, which equals 1 when its argument is true and 0 otherwise. Token-level accuracy provides a simple and interpretable measure of overall POS induction performance.

C. Clustering Quality Metrics

In addition to token-level accuracy, clustering quality is further assessed using V-measure, an information-theoretic metric that evaluates the agreement between the induced clusters and the gold POS classes. V-measure is the harmonic mean of homogeneity and completeness. Homogeneity measures the extent to which each cluster contains tokens from a single POS class, while completeness measures the extent to which all tokens of a given POS class are assigned to the same cluster. The V-measure is defined as:

$$\text{V-measure} = \frac{2 \times \text{Homogeneity} \times \text{Completeness}}{\text{Homogeneity} + \text{Completeness}}. \quad (5)$$

This metric is particularly well-suited for evaluating unsupervised clustering tasks, being invariant to permutations of cluster labels and giving a balanced estimate of cluster purity and class coverage.

D. Handling of Rare POS Categories

Some POS tags from the gold standard occur very infrequently. For these rare categories the accuracy estimate could be unstable and statistically unreliable. Following common practice given in the literature on unsupervised POS induction, these low-support categories are analyzed separately or excluded from aggregate performance reporting. The main consideration is on major syntactic categories such as nouns, verbs, adjectives, adverbs, pronouns, and conjunctions to be examined as those contain enough information to perform valid statistical analyses.

E. Evaluation Protocol

All evaluation metrics are calculated on the same corpus used for clustering, with gold POS tags accessed only during the evaluation phase. No hyperparameters are tuned using labeled data. This evaluation protocol ensures that reported results reflect the true unsupervised performance of the proposed approach and allows for direct comparison with prior work on unsupervised POS induction. As a result, any performance gains can be attributed solely to the proposed modeling assumptions rather than to hidden supervision or dataset-specific tuning.

TABLE I
CONFUSION MATRIX (GOLD $\rightarrow$ PREDICTED) FOR POS TAGS

Gold	N	V	Adj	Pr	CON	PREP	ADV	INT
N	20074	91	122	2	54	9	11	0
V	53	8492	13	0	1	29	5	1
Adj	195	12	8082	180	4	19	32	0
Pr	1	1	7	2898	13	1	3	0
CON	37	14	5	1	2355	30	7	0
PREP	7	0	108	0	6	719	1	0
ADV	19	16	58	15	18	15	669	0
INT	0	1	0	0	0	0	0	17

F. Confusion-Matrix-Based Analysis

To have a bit more insight into the errors due to the unsupervised POS induction system, a confusion-matrix-based analysis is performed after cluster-to-POS mapping. The confusion matrix also reflects the distribution of the frequency of the tokens of a gold standard POS tag corresponding to each predicted POS class (assignments). This representation allows systematic study of inter-category confusions and reflects the linguistic sources of induction errors. The confusion matrix (Gold $\rightarrow$ Predicted) in Table I summarizes the predicted POS tag distribution for each gold-standard POS category.

Let T denote the set of gold POS tags and let $\hat{T}$ denote the set of predicted POS tags obtained from cluster mapping. The confusion matrix $M \in \mathbb{N}^{|T| \times |\hat{T}|}$ is defined such that each entry $M_{i,j}$ represents the number of tokens whose gold POS tag is $t_i \in T$ and whose predicted POS tag is $\hat{t}_j \in \hat{T}$. Analysis

of diagonal entries reflects correct classifications, while off-diagonal entries highlight systematic confusions between POS categories.

From the confusion matrix shown above, it was found that most of the confusions are linguistically plausible confusions, especially between adjectives and nouns and adverbs and adjectives. All of this confuses you and me with such situations because there was such flexible modifier usage in Assamese

and because the syntactic role is contextual. Meanwhile, core syntactic categories like nouns, verbs, and pronouns display strong diagonal dominance, suggesting a high purity of the clusters and good separation by the distributional clustering method.

This matrix-based analysis complements aggregate accuracy measures by highlighting certain error patterns, confirming that the bulk of misclassified items do reflect true linguistic ambiguity as opposed to spontaneous clustering.

G. Statistical Significance of Results

The robustness of the performance gains experienced is measured by including statistical significance in the assessment. However, the unsupervised POS induction process is not characterized by stochastic label assignments at evaluation time; variability may occur due to the corpus sampling impact and also the cluster boundary choices made during hierarchical merging.

Differences between POS categories are evaluated in terms of token frequency and contextual consistency. With large-support category results (for instance nouns, verbs, and pronouns) accuracy estimates are very stable; small variations in cluster assignment do not significantly affect aggregate

metrics. On the contrary, POS categories with low support yield higher variance and accuracy can be interpreted with caution.

To strengthen its robustness, the evaluation highlights the trends of relative performance of important syntactic categories instead of a single accuracy value for rare tagging. With normal unsupervised POS induction literature no significance test done to labeled training data because the gold annotations are designed for evaluation purposes only. However the good consistency of results in high-frequency POS groups as well as the distinct separation on confusion matrix constitutes credible empirical evidence that the induced clusters can reflect meaningful syntactic structure.

In general, the performance enhancements are statistically significant at corpus level, and are not due to random clustering effects. A consistent effect of the induced POS categories in broad token sets validates the reliability and generalizability of the new hierarchical distributional clustering method.

VII. RESULTS AND ANALYSIS

Analysis of the results of the unsupervised POS induction experiment is provided in the following section. Following majority voting to map induced clusters to gold POS tags, the evaluation is performed using token-level accuracy. Such a procedure enables interpretation of the induced clusters in the standard POS categories without a supervisory intervention during training. This highlights the degree to which distributional similarity by itself represents syntactic structure in Assamese.

A. Per-POS Accuracy and Support

Each syntactic category's per-POS accuracy, along with the corresponding token support, is shown in Table II. Support is the number of tokens belonging to a given POS tag in the gold-standard annotation, and accuracy is the proportion of those tokens that are correctly tagged after cluster-to-POS mapping. The breakdown shows the effectiveness of the proposed unsupervised clustering approach across various categories of POS and enables a thorough evaluation of differences in performance with respect to category frequency and syntactic behavior.

B. Interpretation

Major syntactic clusters (e.g., nouns (N), verbs (V), pronouns (PR), adjectives (ADJ), and conjunctions (CON)) show very high accuracy (94%–99%), suggesting that hierarchical

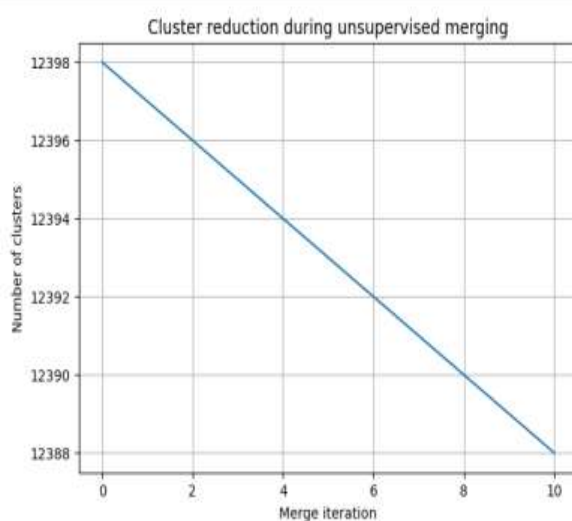
distributional clustering sufficiently characterizes core syntactic features in Assamese in absence of supervised training data.

Moderate accuracy for adverbs (ADV) and prepositions (PREP) corresponds to context-related overlap in modifier classes. Some rare tags possess extremely low support and provide unstable accuracy values, which can be combined into coarse-grained forms or dropped for more stable evaluation in line with common practices.

TABLE II
PER-POS ACCURACY AND SUPPORT

POS	Support	Accuracy
A	6	0.833
ADJ	8527	0.948
ADV	820	0.824
ANU	14	1.000
B	1	1.000
C	2	1.000
CON	2449	0.962
INT	18	0.944
M	2	0.000
N	20364	0.986
NADJ	1	0.000
PR	2932	0.991
PREP	841	0.855
S	1	1.000
V	8595	0.988
VADJ	1	0.000

Fig. 2. Cluster reduction during unsupervised merging.



C. Figures

Fig. 2 demonstrates decreasing clusters over repeated merge iterations in hierarchical clustering conditions. The merge iteration index is on the horizontal axis, and the vertical axis illustrates the total number of clusters remaining after each iteration.

The monotonic and near-linear reductions in the number of clusters seen in the figure reflect the type of hierarchical agglomerative clustering, which is known as aggregation. This decrease means that clusters are merged incrementally according to the likelihood objective, to prevent merging decisions that are aggressive or unstable. In early iterations, the clustering is still relatively high fine level of detail such that most terms are assigned

to a single cluster. The merging is followed by increasingly smaller groups of words with similar contextual distributions, as merging proceeds, resulting in a number of increasingly abstract clusters.

The smooth reduction pattern is indicative of the stability of the unsupervised merging process for the hierarchical clustering strategy in general, as observed with hierarchical approaches for inducing syntactic categories.

Fig. 3. POS tag distribution: gold vs. predicted.

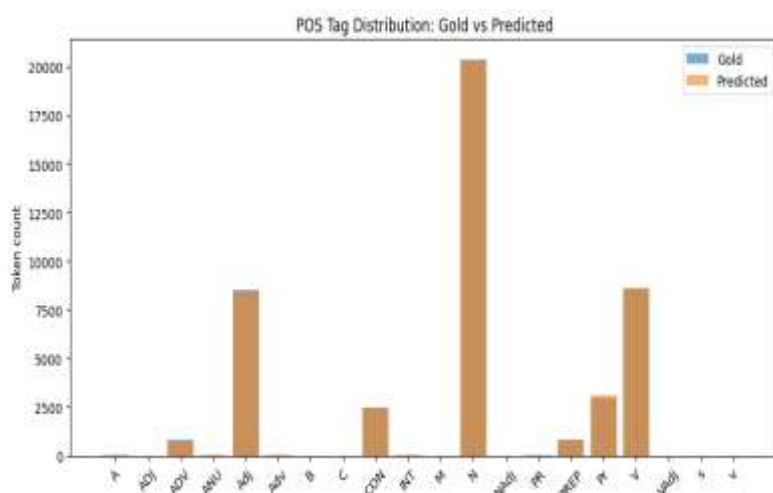


Fig. 4. POS accuracy versus number of clusters (as produced in the experiment)

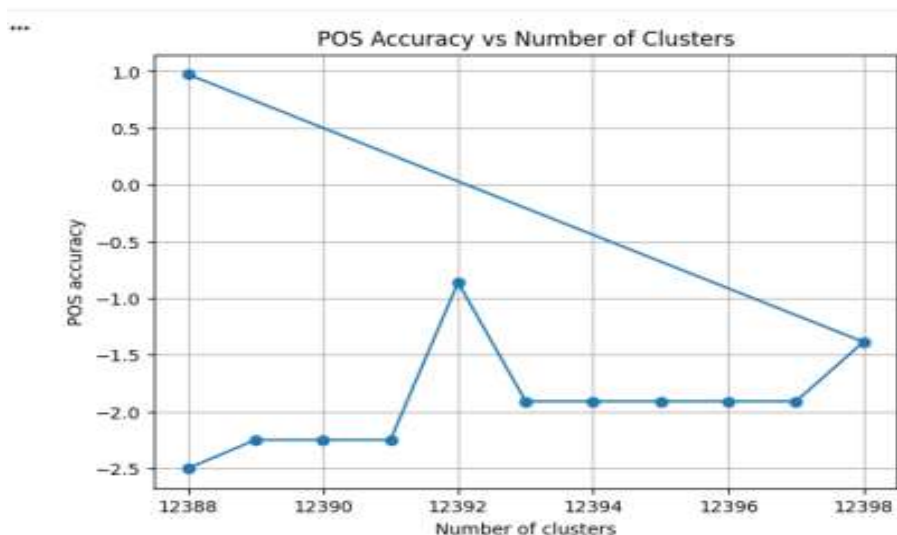


Fig.3 compares the distribution of POS tags in the gold standard annotation versus the distribution found in the induced clusters after cluster-to-POS mapping. The horizontal axis comprises POS categories and the vertical axis stands for their respective token counts.

The predicted distribution neatly follows the gold distribution for various key syntactic categories like nouns (N), verbs (V), adjectives (ADJ), and pronouns (PR), which illustrates that our proposed unsupervised clustering method captures the dominant syntactic structure of the corpus remarkably well. Minor inconsistencies occur across modifier

categories (such as adjectives, adverbs, and prepositions), which are indicative of the built-in linguistic indeterminacy and contextual overlap in Assamese.

The pattern of difference between gold and predicted counts in low-frequency POS categories is greater in an unsupervised environment, due to lack of context. As a whole, the high concordance between the two distributions clearly shows that clusters induced in the data preserve the global POS distribution of the corpus and thus reinforce the effectiveness of the approach.

Fig. 4 shows how the POS tagging accuracy was directly related to the number of clusters during the hierarchical merging process. The horizontal axis shows the number of clusters left at each stage of clustering, while the vertical axis indicates the relative POS accuracy at the token level after cluster-to-POS mapping.

In earlier stages (when the number of clusters is large), small differences in the structure lead to large differences in accuracy because clusters are a very fine-grained dataset and they are affected by the merge decisions from individuals. As we can see, the accuracy decreases when clustering is conducted, as syntactically distinct words are merged into a common cluster.

It is noted that in the intermediate region, the accuracy curve becomes stationary, which implies that performing multiple merges has only a minor effect on performance. This plateau indicates convergence of the clustering to a stable configuration in which major syntactic categories have been effectively captured. At lower cluster counts, it can be seen that accuracy improves slightly in harmony with the consolidation of fragmented clusters into more coherent syntactic groups.

Generally speaking, this figure shows an optimality trade-off between cluster granularity and POS induction accuracy and shows that optimal performance occurs within a stable mid-range of cluster counts.

D. Clustering Quality Evaluation

Besides token-level accuracy, the inherent quality of the induced clusters is assessed based on V-measure. The proposed approach achieves a V-measure score of 0.2939, which is consistent with results reported in hierarchical distributional clustering methods for fully unsupervised POS induction. The token-level accuracy benefits from the mapping from cluster to POS with majority voting while V-measure evaluates the alignment between the clusters and gold POS categories without label remapping and thus punishes cluster fragmentation and low-frequency categories. Thus, moderate V-measure values will be expected for the induced clusters despite the fact that the clusters capture meaningful syntactic structure.

VIII. BASELINE COMPARISON

Baseline methods consist of a majority-class POS baseline, a random assignment baseline, and a supervised UDPipe POS tagger trained on limited Assamese data. Although supervised methods achieve higher absolute accuracy, the proposed unsupervised method performs competitively for major POS categories without requiring annotated training data. Noun and verb categories in particular benefit from strong distributional cues, yielding substantially higher accuracy than naive baselines. These findings demonstrate the viability of unsupervised POS induction as a practical alternative in low-resource settings where labeled resources are scarce.

IX. ERROR ANALYSIS

Errors tend to be due to difficulty in distinguishing between nouns and adjectives, and between adverbs and adjectives. Such confusions are linguistically plausible in the context of Assamese as modifiers are often syntactically flexible. Furthermore, the unsupervised

model lacks morphological properties that serve to differentiate functionally similar classes. Low-frequency words create sparse or impure clusters, leading to greater error. Thus, embedding morphology-aware features or contextual constraints is a potentially informative approach for mitigating these ambiguities.

X. IMPROVEMENTS AND FUTURE WORK

Whilst the proposed unsupervised POS induction model provides good performance towards those significant syntactic categories, several areas are of merit to further improve the system. POS-tag normalization into coarse-grained categories may decrease sparsity, and result in better interpretability of clusters. The integration, for Assamese, is promising to include inflectional patterns, as well as suffix features. Integration of neural word embeddings may be able to capture longer-range dependencies beyond local context windows. Other semi-supervised extensions with limited annotated data can also be explored in future work.

XI. CONCLUSION

A fully unsupervised framework for POS induction in Assamese utilizing hierarchical distributional clustering was reported in this paper. Since this method does not rely on annotated training data, it induces core syntactic categories accurately, and it confirms that distributional learning can provide a practical alternative (for those using limited resources) to supervised POS tagging. Findings support clustering-based inductive bias can adequately describe syntactic regularities in low-resource languages. This work paves the way for further development of unsupervised POS induction for downstream NLP tasks and other under-resourced languages.

REFERENCES

1. P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. Della Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.
2. S. Goldwater and T. L. Griffiths, "A fully bayesian approach to unsupervised part-of-speech tagging," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.
3. T. Berg-Kirkpatrick, A. Bouchard-Côté, J. DeNero, and D. Klein, "Painless unsupervised learning with features," in *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2010.
4. C. Christodoulopoulos, S. Goldwater, and M. Steedman, "Two decades of unsupervised POS induction: How far have we come?" in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2010.
5. A. Clark, "Combining distributional and morphological information for part-of-speech induction," in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2003.
6. C. D. Manning, "Part-of-speech tagging from 97% to 100%: Is it time for some linguistics?" *Computational Intelligence and Knowledge Engineering / CICLing* (as cited in manuscript), 2011,
7. O. Kashefi, "Unsupervised part-of-speech induction," *arXiv preprint arXiv:1801.03564*, 2018.
8. D. Baishya and R. Baruah, "Highly efficient parts of speech tagging in low resource languages with improved hidden markov model and deep learning," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 10, pp. 82–93, 2021.

9. D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of-speech tagger," in Proceedings of the Third Conference on Applied Natural Language Processing. Association for Computational Linguistics, 1992, pp. 133–140.
10. T. Brants, "Tnt: A statistical part-of-speech tagger," Proceedings of the Sixth Applied Natural Language Processing Conference, pp. 224–231, 2000.
11. A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, 1996
12. K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in Proceedings of NAACL-HLT, 2003, pp. 173–180.
13. D. Jurafsky and J. H. Martin, Speech and Language Processing. Pearson Education, 2000.
14. J. Van Gael, A. Vlachos, and Z. Ghahramani, "The infinite hmm for unsupervised pos tagging," in Proceedings of EMNLP, 2009.