



Subword Representations Close the Out-of- Vocabulary Gap in Minimal-Supervision Part-of-Speech Tagging for Assamese

Biswajit Sarma¹, Rupam Baruah², Diganta Baishya³

¹Assistant Professor, Department of Computer Science and Engineering, Jorhat Engineering College, Assam Science and Technology University (ASTU), Assam, India ORCID: 0009-0005-7034-5842 Email: eduneristbiswaj@gmail.com

²Professor and Principal, Jorhat Engineering College, Assam Science and Technology University (ASTU), Assam, India ORCID: 0000-0001-7236-5314 Email: rupam.baruah.jec@gmail.com

³Associate Professor, Assam Engineering College, Assam Science and Technology University (ASTU), Assam, India ORCID: 0000-0002-1430-5142 Email: dbaishya@gmail.com

ABSTRACT

Minimal-supervision part-of-speech (POS) tagging pipelines for low-resource languages are typically improved by refining the learning algorithm, the supervision signal, or the labelled seed. We report a study in which five such refinements were applied to an established evidence-gated tagging pipeline for Assamese and *all five failed*, four of them reducing accuracy, before a diagnostic revealed why: on a 44,572-token Assamese corpus with a labelled seed of only 142–228 sentences, 82–86% of all remaining errors fall on word types absent from the pipeline's accumulated evidence pool. The classifier is already accurate on words it has seen (94.4%) and collapses on words it has not (74.4%); the widely-reported weakness of the adjective category is shown to be largely an artifact of this split, since adjective accuracy is 0.894 on seen types but 0.481 on unseen ones. Motivated by this localization, we add sub word representations—character n -gram indicator features and FastText embedding trained on unlabelled in-domain text—to the pipeline's backfill classifier. The combination improves tagging accuracy by +1.78 points at a seed of three sentences per distinct sentence length and

+1.46 points at five, improving in all ten paired replicate draws, and consumes no additional annotation. The gain is concentrated exactly where the diagnostic predicted: unseen-type accuracy rises from 0.744 to 0.774 and adjective accuracy on unseen types from 0.492 to 0.533, the first intervention in this line of work to move adjective recall at all. We report the five negative results alongside the positive one, since each delineates a class of technique that does not transfer to this regime, and we show that character n -grams alone—requiring no embedding and no training step—recover roughly 88% of the benefit.

Keywords: Assamese, part-of-speech tagging, low-resource NLP, minimal supervision, subword representations, character n -grams, FastText, out-of-vocabulary, negative results

1. INTRODUCTION

ART-OF-SPEECH tagging for a low-resource language is constrained less by algorithmic sophistication than by the cost of producing labelled training data. A trained annotator's time is scarce, and a tagger that reaches most of a fully supervised system's accuracy from a small fraction of its labelled data is often more valuable to a language community than a marginally more accurate system demanding an order of magnitude more annotation effort. This reframes the central question from accuracy maximisation to annotation efficiency: given a fixed, small labelling budget, how much tagging accuracy can be recovered?

Our prior work established a three-stage answer for Assamese. An unsupervised hierarchical distributional clustering method induces major syntactic categories from raw text with no labels [1]. A cluster-anchored weak-supervision frame-work converts those clusters into labelling functions for a discriminative tagger, reaching 90.59% accuracy without manual annotation of a training corpus [2]. An evidence-gated cascade with classifier backfill then compresses the labelled requirement drastically, recovering 84.7–86.0% accuracy at full coverage from only 142–228 labelled sentences—roughly 5–8% of the corpus—against a

90.8% baseline trained on 2,840 sentences [3]. A subsequent study examined whether the composition of that small seed matters, finding a modest benefit from preferring sentences with mutually distinct gold tag sequences [4].

That line of work also produced an unusual density of negative results. Distributional word clustering, Viterbi sequence decoding, and self-training were each tested inside the minimal-supervision regime and each failed to deliver the benefit they are known for at larger scale [3]. In every case the diagnosis was data-volume-driven rather than algorithmic. That pattern motivated the present study, which asks a narrower question: where, precisely, does the remaining error live, and what class of intervention can reach it?

The answer turns out to be neither a better decision rule nor a better supervision signal. Section IV shows that 82–86% of the remaining error falls on word types the classifier has never observed, and that the well-documented weakness of the adjective category is largely a consequence of that split rather than an independent phenomenon. Section V introduces subword representations as the natural response, and Section VII shows they deliver +1.78 and +1.46 accuracy points at the two seed sizes, improving in all ten paired draws, at no annotation cost.

Our contributions are:

- 1) A quantitative localisation of the residual error in a minimal-supervision tagging pipeline, establishing out-of-vocabulary word types as the binding constraint and showing that a per-category error analysis alone is misleading without conditioning on vocabulary coverage (Section IV).
- 2) A demonstration that subword representations recover +1.78/+1.46 accuracy points with no additional la-belled data, with the gain verified to fall on out-of-vocabulary tokens specifically (Sections V and VII).
- 3) Five documented negative results—prior calibration, mined morphological rules, stacked neighbour-tag features, a dedicated adjective/noun discriminator, and hand-specified linguistic rules—each with a measured failure mechanism, which together delineate the classes of technique that do not transfer to this regime (Section VIII).
- 4) Evidence that character n-gram indicator features alone, requiring no embeddings and no training, capture approximately 88% of the benefit (Section VII-D).

II. RELATED WORK

Distributional clustering and weak supervision. Brown et al. [5] introduced class-based n-gram models in which words are merged to maximise the mutual information of a class-based language model, a technique that underlies the unsupervised induction stage of [1]. Ratner et al. [6] formalised weak supervision as data programming, combining multiple noisy labelling sources into a single probabilistic training signal; that formalism underlies the cluster-anchored labelling functions of [2]. Christodoulopoulos et al. [7] provide the standard large-scale evaluation of unsupervised POS induction, establishing that distributional methods recover coarse syntactic groupings well while underperforming on fine-grained evaluation.

Subword and character-level representations. Bojanowski et al. [8] extended the skip-gram model so that each word is represented as a bag of character n-grams, allowing a vector to be composed for a word never seen during training. This property—out-of-vocabulary handling by construction—is precisely what the present work requires, and is especially suited to morphologically rich languages, as their evaluation across nine languages demonstrates. Earlier, Ling et al. [9] showed that composing word representations from characters benefits POS tagging in morphologically rich languages, and Plank et al. [10] demonstrated that combining word and character representations improves multilingual POS tagging across 22 languages, with the largest gains on languages having high out-of-vocabulary rates. Our findings are consistent with these results but arrive at them from a different direction: rather than adopting subword modelling as an architectural default, we identify out-of-vocabulary error empirically as the dominant residual and then show that subword features address it.

Low-resource and Assamese POS tagging. The foundational computational treatment of Assamese POS tagging is that of Saharia et al. [11], who developed a 172-tag tagset and a hidden Markov model tagger over roughly 10,000 manually annotated words, reporting approximately 87% accuracy—a labelled-data scale much closer to the present minimal-supervision regime than to a high-resource setting. Tagset standardisation across Indian languages traces to Sankaran et al. [12]. More recent Assamese work has moved toward larger corpora and deep learning: Pathak et al. [13] present a neural tagger using pre-trained word embeddings with a two-phase training procedure, and later report an ensembling approach combining BiLSTM-CRF with multiple tagger types on a 404k-token corpus [14], explicitly motivated by the observation that resource-poor languages often lack enough annotated data to train deep taggers outright. Baishya and Baruah [15] report a related Assamese-specific line combining improved hidden Markov models with deep learning. Our corpus, at 44,572 tokens, is roughly an order of magnitude smaller than that used in the most recent of these works, which is precisely the gap the minimal-supervision framing addresses. Error analysis and annotation efficiency. Settles' survey [16] catalogues query strategies for selecting which in-stances to label, a concern directly adjacent to seed selection [4]. Our diagnostic contribution is complementary: rather than asking which instances to label next, we ask which property of

the already-labelled set limits performance, and find it to be vocabulary coverage rather than syntactic-pattern coverage.

III. CORPUS, TAGSET AND TASK SETUP

All experiments use the manually annotated Assamese corpus introduced in [1] and used unchanged in [2], [3], [4]: 3,549 sentences, 44,572 tokens and 12,401 distinct word types in word/TAG format. Sixteen raw tag variants in the source annotation are normalised to nine canonical categories: N (noun), V (verb), ADJ (adjective), PR (pronoun), CON (con-junction), PREP (preposition/postposition), ADV (adverb), IN (interjection) and OTHER (residual).

We adopt without modification the fixed test-set protocol of [3]: a held-out set of 709 sentences (9,018 tokens), with the remaining 2,840 sentences forming the trainable pool from which all seed and validation sentences are drawn. No test sentence is available to any stage of any system reported here.

A. The Evidence-Gated Pipeline

The base system is the evidence-gated cascade of [3], summarised here for self-containment. Its governing principle is precision-first: each stage assigns a tag only where it has direct

TABLE I. BACKFILL ACCURACY BY WHETHER THE TOKEN'S WORD TYPE OCCURS IN THE EVIDENCE POOL (SEED=5, DRAW 1; 3,139 EVIDENCE WORD TYPES).

Tag	Sup.	P	R	F ₁	P	R	F ₁
N	4238	0.864	0.922	0.892	0.889	0.922	0.905
V	1717	0.926	0.893	0.909	0.922	0.927	0.924
ADJ	1661	0.811	0.714	0.759	0.817	0.733	0.773
PR	575	0.892	0.872	0.882	0.907	0.914	0.911
CON	505	0.948	0.878	0.912	0.938	0.896	0.916
PREP	173	0.750	0.799	0.773	0.741	0.821	0.778
ADV	147	0.510	0.574	0.540	0.579	0.585	0.582

evidence and otherwise abstains. A labelled seed is selected by taking k sentences of each distinct sentence length ($k = 3$ or 5 , giving 142 or 220–228 sentences). Four further stages extend coverage where evidence supports it: distributional word clustering; a seed-ambiguity audit with context-based disambiguation; confidence-gated cluster labelling with a Laplace-smoothed posterior and threshold $\tau = 0.80$; and empirically validated suffix, prefix and grammar rule cascades. These stages together tag roughly 20% of tokens at approximately 95% accuracy, deliberately leaving the remainder untagged.

Full coverage is reached by a classifier backfill: a logistic-regression model trained on the pipeline's own accumulated evidence pool and applied only to positions left bare. Its feature set comprises word identity, word shape, one- to three-character prefixes and suffixes, a cluster identifier, and the immediately preceding and following words. It never overwrites an evidence-based tag. Because it is responsible for roughly 80% of test tokens at substantially lower accuracy than the evidence stages, the backfill classifier is where the residual error concentrates, and it is the sole component modified in this paper.

IV. DIAGNOSTIC: LOCATING THE RESIDUAL ERROR

A. Per-Category Analysis Is Misleading in Isolation

A per-tag error decomposition of the published seed=5 results suggests an obvious target. Writing the error mass attributable to a gold category as $\text{support} \times (1 - \text{recall})$, the adjective category accounts for approximately 38% of all errors, more than any other. The confusion matrix, measured here over five paired replicate draws, sharpens this: ADJ→N accounts for 368 tokens and N→ADJ for 216, so roughly 48% of all ~1,227 errors lie on a single, bidirectional class boundary.

Bidirectionality is decisive. A prior shift, class weight or decision-threshold adjustment can only move a boundary; it trades errors in one direction for errors in the other. Section VIII confirms this empirically: across five separate interventions, adjective recall never left the band 0.696–0.715.

B. Conditioning on Vocabulary Coverage

The decisive split is not by category but by whether a token's word type appears anywhere in the pipeline's accumulated evidence pool. Table I reports backfill accuracy under that split. Unseen types constitute 50.8–56.7% of bare tokens across draws and carry 82.1–86.2% of all bare-position errors. On words for which it has evidence the classifier is already strong; essentially all remaining error is out-of-vocabulary error.

This reframes the adjective problem. Adjective accuracy is 0.894 on seen types and 0.481 on unseen ones, and roughly half of adjective tokens are unseen. The aggregate adjective weakness is therefore substantially

a composition effect. Any intervention that does not improve unseen-type prediction cannot materially improve adjective recall, which explains the uniform failure documented in Section VIII.

C. Ruling Out Annotation Inconsistency

A competing explanation is that the gold annotation is itself inconsistent on the noun/adjective boundary. We tested this directly. Only 191 word types are gold-tagged as both N and ADJ anywhere in the corpus, covering 4.8% of test N/ADJ tokens (8.3% of ADJ tokens, 3.5% of N tokens). This is an upper bound of approximately 285 tokens against roughly 584 ADJ↔N errors, so annotation inconsistency accounts for at most half and probably considerably less. It is not the dominant cause.

V. METHOD: SUBWORD REPRESENTATIONS

Subword representations are the one family of techniques that addresses out-of-vocabulary tokens by construction: a word's representation is composed from its character sub-strings, so a word never seen in training still receives an informative representation assembled from pieces that were seen. For a morphologically rich language such as Assamese this is a natural fit. We evaluate two forms, separately and combined.

A. Character n-gram Indicator Features

For a word w we form the padded string $\langle w \rangle$ and extract every character substring of length $n \in [2, 5]$, adding each as a binary indicator feature. Formally, the feature set contributed by w is

$$\mathcal{G}(w) = \bigcup_{n=2}^5 \{ s_{i:i+n} : 0 \leq i \leq |s| - n \}, \quad s = \langle w \rangle. \quad (1)$$

Boundary padding follows [8] and distinguishes word-initial and word-final material from internal material. This requires no training and no external data.

The baseline feature set already contains one- to three-character prefixes and suffixes, but these are anchored at the two word edges only. $\mathcal{G}(w)$ additionally captures internal substrings and extends the window to five characters, which for Assamese covers derivational and inflectional material the edge-anchored features miss.

B. FastText Embeddings

We train a FastText skip-gram model [8] using gensim on the raw, unlabelled text of the trainable pool only: 2,840 sentences, vocabulary 10,698 types, dimension 100, subword range $n \in [2, 5]$, window 5, minimum count 1, 30 epochs.

Training takes approximately six seconds. Each token's vector is L2-normalised and concatenated to the sparse feature matrix. Because a FastText vector is the sum of its subword vectors, a word absent from the training vocabulary still receives a composed vector rather than a zero or an UNK placeholder—the property this application requires.

Two points of protocol deserve emphasis. First, training consumes no labelled data, so the minimal-supervision annotation budget is unchanged. Second, training on the trainable pool only is stricter than the baseline it is compared against: the published pipeline's own clustering stage consumes raw text from the entire corpus, including test sentences.

C. Variants Evaluated

- baseline — the published backfill feature set.
- charngram — baseline + $\mathcal{G}(w)$.
- fasttext — baseline + the dense 100-dimensional vector.
- ft charngram — baseline + both.
- fasttext only — the dense vector plus shape and neighbour features, but with word identity and cluster identifier removed, testing whether the embedding can substitute for lexical identity rather than supplement it.

VI. EXPERIMENTAL PROTOCOL

Every configuration reuses the evidence-gated pipeline, the fixed 709-sentence test set, and the per-draw random seeding scheme of [3], [4] unchanged. Five replicate draws are run at each of $k = 3$ and $k = 5$ sentences per length. Because the seeding scheme matches, each variant is compared against a baseline that used byte-identical seed sentences and an identical evidence pool: all comparisons are paired.

Significance is assessed with an exact paired permutation test over all $25 = 32$ sign flips of the per-draw differences, together with an exact sign test. With five draws the smallest attainable two-sided p-value is 0.0625; that value therefore indicates a perfectly consistent direction rather than conventional significance, and we interpret it accordingly. Reporting these tests addresses a limitation explicitly acknowledged in [3], [4]. Because the hypothesis concerns out-of-vocabulary tokens specifically, every variant is additionally reported with accuracy split by evidence-pool membership.

VII. RESULTS

A. Overall Accuracy

Table II reports mean accuracy over five paired draws at each seed size, and Table III the paired differences. The combined variant improves in all ten paired draws across both seed sizes. Absolute accuracy rises from 84.80% to 86.58% at $k = 3$ and from 86.40% to 87.86% at $k = 5$. For scale, this single change is worth approximately as much as in-creasing the annotation budget from three to five sentences per length, which buys +1.25 points [3], but costs no additional labelled data.

TABLE II. MEAN ACCURACY AND WEIGHTED F1 OVER FIVE PAIRED REPLICATE DRAWS.

Seed	Variant	Accuracy	Std	Weighted F_1
$k = 3$	baseline	0.8480	0.0006	0.8468
	charngram	0.8637	0.0027	0.8624
	fasttext	0.8552	0.0010	0.8539
	ft_charngram	0.8658	0.0026	0.8644
	fasttext_only	0.8476	0.0012	0.8473
$k = 5$	baseline	0.8640	0.0011	0.8630
	charngram	0.8773	0.0015	0.8762
	fasttext	0.8694	0.0013	0.8684
	ft_charngram	0.8786	0.0014	0.8775
	fasttext_only	0.8620	0.0010	0.8618

TABLE III. PAIRED DIFFERENCES AGAINST THE BASELINE. p IS THE EXACT PAIRED PERMUTATION TEST; 0.0625 IS THE MINIMUM ATTAINABLE WITH FIVE DRAWS.

Seed	Variant	Δ (pts)	Improved	p
$k = 3$	charngram	+1.57	5/5	0.0625
	fasttext	+0.72	5/5	0.0625
	ft_charngram	+1.78	5/5	0.0625
	fasttext_only	-0.04	1/5	0.4375
$k = 5$	charngram	+1.34	5/5	0.0625
	fasttext	+0.54	5/5	0.0625
	ft_charngram	+1.46	5/5	0.0625
	fasttext_only	-0.20	0/5	0.0625

TABLE IV. ACCURACY SPLIT BY EVIDENCE-POOL MEMBERSHIP, AND ADJECTIVE ACCURACY ON UNSEEN TYPES (MEANS OVER FIVE DRAWS).

Seed	Variant	Seen	Unseen	ADJ unseen
$k = 3$	baseline	0.9422	0.7300	0.4855
	charngram	0.9485	0.7598	0.5290
	fasttext	0.9425	0.7458	0.4829
	ft charngram	0.9486	0.7645	0.5234
$k = 5$	baseline	0.9440	0.7438	0.4918
	charngram	0.9504	0.7704	0.5332
	fasttext	0.9446	0.7565	0.4934
	ft charngram	0.9501	0.7740	0.5322

B. The Gain Falls on Out-of-Vocabulary Tokens

Table IV verifies the mechanism.

Seen-type accuracy moves only marginally (0.9440 $\rightarrow$ 0.9501 at $k = 5$), consistent with its already being

close to ceiling. Unseen-type accuracy rises by approximately three points (0.7438 $\rightarrow$ 0.7740), and adjective accuracy on unseen types—the weakest cell in the entire system and the origin of the apparent adjective problem—improves from 0.4918 to 0.5322. The improvement is therefore attributable to the mechanism hypothesised, not to an unrelated regularisation effect.

C. Per-Category Results

Table V gives per-tag means at $k = 5$. Every category with non-trivial support improves in F1. Adjective F1 rises from 0.759 to 0.773, driven by recall (0.714 $\rightarrow$ 0.733): this is the first intervention in this line of work to move adjective recall at all, five prior attempts having left it within the band 0.696–0.715. Adverb F1 improves most in relative terms (0.540 $\rightarrow$ 0.582), consistent with adverbs having both low support and a high unseen-type rate.

TABLE V. PER-TAG PRECISION, RECALL AND F1 AT $k = 5$, MEANS OVER FIVE DRAWS. B = BASELINE, S = FT CHARNGRAM

Tag	Sup.	Baseline			ft_charngram		
		P	R	F_1	P	R	F_1
N	4238	0.864	0.922	0.892	0.889	0.922	0.905
V	1717	0.926	0.893	0.909	0.922	0.927	0.924
ADJ	1661	0.811	0.714	0.759	0.817	0.733	0.773
PR	575	0.892	0.872	0.882	0.907	0.914	0.911
CON	505	0.948	0.878	0.912	0.938	0.896	0.916
PREP	173	0.750	0.799	0.773	0.741	0.821	0.778
ADV	147	0.510	0.574	0.540	0.579	0.585	0.582

D. Component Ablation

Character n-grams alone account for +1.57 of the +1.78 total at $k = 3$ and +1.34 of +1.46 at $k = 5$ —approximately 88% of the benefit—while requiring no embeddings, no training step and no external dependency. FastText contributes a smaller but consistent independent gain (+0.72/+0.54), and the two are not redundant, since their combination exceeds either alone at both seed sizes.

The fasttext only variant is the informative negative case. Removing word identity and the cluster identifier in favour of the embedding improves unseen-type accuracy relative to baseline (0.7403/0.7518 vs 0.7300/0.7438) but loses more on seen types (0.9278/0.9309 vs 0.9422/0.9440), for a small net loss. Subword representations therefore supplement lexical identity rather than replace it: exact identity remains the better signal for known words, subword composition for unknown ones.

VIII. WHAT DID NOT WORK

Five interventions preceded the successful one. All are reported here because each localises a class of technique that does not transfer to this regime, and because together they explain why the diagnostic of Section IV was necessary. All use the identical paired protocol.

Validation-tuned prior calibration. Per-class additive logit offsets, fitted by coordinate ascent on the disjoint validation set, yield +0.18 and +0.09 points, improving in 4/5 and 2/5 draws respectively—indistinguishable from zero. The published backfill classifier already uses balanced class weighting, so this is a second-order correction on an already-corrected model; more fundamentally, a bidirectional boundary cannot be improved by moving it.

Automatically mined morphological rules. Mining all character suffixes of length ≤ 5 occurring at least four times in the seed with dominant-tag purity ≥ 0.85 yields 149–218 validated rules per draw. Injecting them into the pipeline’s rule stage costs –1.36 and –1.12 points, 0/5 draws improving. The mechanism is that rule outputs enter the evidence pool that trains the backfill classifier, so a rule 85% pure on four seed occurrences injects systematically wrong labels into the classifier’s training data. Restricting to adjective rules reduces the damage roughly sevenfold but does not reverse it. This indicates the conservative rule-acceptance threshold of [3] is near-optimal rather than over-cautious.

Stacked neighbour-tag features. Re-training the classifier with the first pass’s neighbouring predicted tags as additional features costs –0.56 and –0.47 points, 0/5 improving at both sizes. First-pass predictions are only ~86% accurate on exactly the bare positions where they are needed, so the model learns to rely on a noisy feature and propagates its own errors. Together with the failures of Viterbi decoding and self-training reported in [3], this is a third independent mechanism by which sequence structure fails to be

exploitable at this annotation budget.

A dedicated adjective/noun discriminator. A binary classifier with extended affixes, neighbour tags carrying provenance flags, an attributive next-is-noun feature, lexical priors and positional features, invoked only on low-margin N/ADJ decisions, yields +0.02 and -0.08 points. Notably, in seven of ten draws the validation-tuned margin threshold was selected as zero—the model-selection procedure disabled the discriminator entirely. Rerouting all N/ADJ decisions through it costs -16 to -17 points. Its strongest features are word identity and lexical priors, which are by definition unavailable for the unseen types that carry the errors.

Hand-specified linguistic rules. Four rules were specified on linguistic grounds: a word followed by a noun is an adjective; and words carrying case markers, plural markers or classifier morphemes are nouns. All four reduce accuracy. Measured against gold on the tokens each fires on, and with base rates $P(N) = 0.470$ and $P(ADJ) = 0.184$: the case-marker rule attains 0.654 precision (genuinely informative, but far below the 0.88–0.98 the classifier already achieves); the plural rule 0.469 (indistinguishable from base rate); the classifier-morpheme rule 0.404, meaning such words are gold-tagged adjective more often than noun—the rule is inverted, independently reconfirming at corpus scale a discrepancy first noted for a single numeral-classifier rule in [3]; and the followed-by-noun rule 0.238, since a token preceding a noun is twice as likely to be another noun, Assamese noun-noun compounding being pervasive. Structurally, all these markers are one to three characters and therefore already present as features, so hard-coding them replaces a learned probabilistic weighting with a cruder deterministic one.

IX. DISCUSSION

Diagnosis before intervention. The five failures share a cause: each addressed a plausible symptom rather than the binding constraint. Per-category error analysis identified adjectives; the confusion matrix identified a bidirectional boundary; only conditioning on vocabulary coverage revealed that both are largely consequences of out-of-vocabulary composition. We would emphasise this methodologically: in minimal-supervision settings, per-category error tables can be actively misleading unless conditioned on whether the model has lexical evidence for the tokens involved.

Why sub word features are the right instrument here. With a labelled seed of 142–228 sentences, the evidence pool covers only 2,555–3,214 word types of a 12,401-type vocabulary. Roughly half of test tokens are therefore lexically unknown. Techniques that reweight, gate or post-process decisions operate on information the model already has; sub-word composition supplies information it does not. This also explains the one prior positive result in this line—a genuine distributional clustering feature, worth +0.1 to +0.2 points—as a weaker instance of the same principle: it too generalises across word types, but only for types frequent enough to have a usable context vector, leaving the long tail untouched.

Position on the annotation-efficiency curve. Combining the subword features with the previously validated distributional cluster feature brings the pipeline to approximately 86.6% at $k = 3$ and 87.9% at $k = 5$ on the fixed test set. The gap to the 2,840-sentence weak-supervision baseline of 90.8% narrows from 4.83 to approximately 2.9 points at $k = 5$, while still using roughly 8% of its labelled data. Since part of what additional seed sentences previously bought was vocabulary coverage, and that is now partly supplied without annotation, the seed-size return reported in [3] should be re-measured under the new feature set.

X. LIMITATIONS

Results derive from a single corpus and a single language; we make no claim that the magnitudes reported transfer to other low-resource languages, though the mechanism—out-of-vocabulary dominance under a small labelled budget—plausibly generalises. Five replicate draws per condition suffice to establish a perfectly consistent direction ($p = 0.0625$, the minimum attainable) but constitute a small sample for stronger claims. The subword hyper parameters ($n \in [2, 5]$, dimension 100, 30 epochs) were fixed a priori rather than swept, so the reported gain is a lower bound on what tuning might achieve. FastText trained on the 2,840-sentence trainable pool only; training on larger unlabelled Assamese text, which is readily available, would likely improve the embedding component further. Finally, the evidence-gated pipeline's own hyper parameters were inherited unchanged, and their interaction with subword features has not been explored.

XI. CONCLUSION

We have shown that the residual error in a minimal-supervision POS tagging pipeline for Assamese is dominated not by any syntactic category but by out-of-vocabulary word types, which constitute roughly half of bare tokens and carry 82–86% of all errors. Conditioning error analysis on vocabulary coverage reveals that the apparent weakness of the adjective category is largely a composition effect. Subword representations—character n -gram indicators combined with FastText embeddings trained on unlabelled in-domain text—improve accuracy by +1.78 points at three seed sentences per length and +1.46 at five, improving in all ten paired replicate draws, with the gain verified to fall on out-of-vocabulary tokens and adjective recall moving for the first time across six studies. Character n -grams alone, requiring no training

and no external resource, recover approximately 88% of that benefit. Five documented negative results delineate the classes of technique—decision reweighting, rule expansion, sequence-structure exploitation and specialised discriminators—that do not transfer to this regime, and clarify why: they operate on information the model already possesses, whereas the binding constraint is information it lacks.

REFERENCES

- [1] B. Sarma, R. Baruah, and D. Baishya, “Unsupervised part-of-speech induction for Assamese language using hierarchical distributional clustering,” *Int. J. Adv. Sig. Img. Sci.*, vol. 12, no. 2s, pp. 2420–2431, 2026.
- [2] B. Sarma, R. Baruah, and D. Baishya, “Sequence-smoothed semi-supervised POS tagging for Assamese language using cluster-anchored weak supervision and HMM Viterbi decoding,” *Int. J. Sci. Res. Eng. Manag. (IJSREM)*, vol. 10, no. 6, 2026.
- [3] B. Sarma, R. Baruah, and D. Baishya, “Minimal-supervision part-of-speech tagging for Assamese language: An evidence-gated cascade with classifier backfill,” *Int. J. Comput. Inf. Syst. Ind. Manag. Appl. (IJCISIM)*, vol. 18, no. 12s, pp. 1196–1206, 2026. doi:10.70917/ijcisim-2026-3991
- [4] B. Sarma, R. Baruah, and D. Baishya, “Tag-pattern-diverse seed selection for minimal-supervision Assamese part-of-speech tagging,” manuscript, 2026.
- [5] 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,” *Comput. Linguist.*, vol. 18, no. 4, pp. 467–479, 1992.
- [6] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Re’, “Snorkel: Rapid training data creation with weak supervision,” *Proc. VLDB Endowment*, vol. 11, no. 3, pp. 269–282, 2017.
- [7] C. Christodoulopoulos, S. Goldwater, and M. Steedman, “Two decades of unsupervised POS induction: How far have we come?” in *Proc. EMNLP*, 2010, pp. 575–584.
- [8] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, 2017.
- [9] W. Ling, C. Dyer, A. W. Black, I. Trancoso, R. Fernandez, S. Amir, L. Marujo, and T. Lu’is, “Finding function in form: Compositional character models for open vocabulary word representation,” in *Proc. EMNLP*, 2015, pp. 1520–1530.
- [10] B. Plank, A. Søgaard, and Y. Goldberg, “Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss,” in *Proc. ACL*, 2016, pp. 412–418.
- [11] N. Saharia, D. Das, U. Sharma, and J. Kalita, “Part of speech tagger for Assamese text,” in *Proc. ACL-IJCNLP Conf. Short Papers*, 2009, pp. 33–36.
- [12] B. Sankaran, K. Bali, M. Choudhury, T. Bhattacharya, P. Bhattacharyya, G. N. Jha, S. Rajendran, K. Saravanan, L. Sobha, and K. V. Subbarao, “A common parts-of-speech tagset framework for Indian languages,” in *Proc. LREC*, 2008.
- [13] D. Pathak, S. Nandi, and P. Sarmah, “AsPOS: Assamese part of speech tagger using deep learning approach,” in *Proc. IEEE/ACS 19th Int. Conf. Comput. Syst. Appl. (AICCSA)*, 2022.
- [14] D. Pathak, S. Nandi, and P. Sarmah, “Part-of-speech tagger for As-samese using ensembling approach,” *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 10, art. 235, pp. 1–22, 2023.
- [15] D. Baishya and R. Baruah, “Highly efficient parts of speech tagging in low resource languages with improved hidden Markov model and deep learning,” *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*, vol. 12, no. 10, pp. 82–93, 2021.
- [16] B. Settles, “Active learning literature survey,” Univ. Wisconsin–Madison, Dept. Comput. Sci., Tech. Rep. 1648, 2009.