

AffiXTag: Enhancing Part-of-Speech Tagging with Affix-Aware Embeddings and Contextual Fusion in MRLs

Samat Ali and Alim Murat *

Abstract

Part-of-speech (POS) tagging in morphologically rich, low-resource languages (MRL) such as Uyghur, Kyrgyz, and Uzbek remains challenging because productive morphology generates many out-of-vocabulary forms, while lexical ambiguity requires context-sensitive disambiguation. This paper introduces AffiXTag, a BiLSTM-CRF model that combines XLM-R contextual embeddings with a dynamic affix segmentation method. Our main contributions are: Dynamic Affix Selection, which adjusts n-gram sizes and frequency thresholds based on word length to improve segmentation accuracy; Context–Affix Interaction, which merges sentence-level meaning with character-level morphological information to better resolve polysemy; and Comprehensive Evaluation, showing strong improvements across four agglutinative languages, including an 18.7% drop in out-of-vocabulary (OOV) errors and a 5.2% F1 gain in homograph disambiguation compared to models using only contextual embeddings. Experiments on Turkish further demonstrate the cross-lingual generalizability of AffiXTag, indicating its applicability to POS tagging in other MRLs.

Keywords

Contextual Embeddings, Context-Affix Interaction, Dynamic Affix Segmentation, MRL POS Tagging

1. Introduction

Part-of-speech (POS) tagging assigns grammatical categories to tokens and underpins syntactic parsing and semantic analysis. High-resource languages benefit from near-human taggers, but morphologically rich, low-resource languages (MRL) such as Uyghur, Kyrgyz, and Uzbek remain underserved. Their agglutinative morphology, characterized by extensive suffixation, compounding, and inflection, produces many out-of-vocabulary (OOV) forms, while pervasive polysemy demands context-sensitive disambiguation that conventional taggers often miss. These factors hinder POS models' ability to generalize to unseen forms and resolve multiple word senses.

In MRLs, complex affix sequences can encode multiple grammatical categories. Common categories include number, case, and definiteness. For example, the Uyghur word *kélmékligimiz* (“our act of coming”) decomposes as: *kél-* (“come”) + *-mék* (infinitive) + *-lik* (noun forming) + *-imiz* (1st pl. possessive). This productive affixation system yields many rare or unseen word forms, making explicit morphological modeling essential for robust POS tagging. Morphological analysis also exposes polysemy (a single word carrying multiple senses or grammatical functions) [1], which significantly

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Manuscript received August 22, 2025; first revision September 10, 2025; accepted March 25, 2026.

*Corresponding Author: Alim Murat (a.murat@xjnu.edu.cn)

School of Computer Science and Technology, Xinjiang Normal University, Urumqi, China (a.murat@xjnu.edu.cn, samat.ali@xjnu.edu.cn)

complicates POS tagging when context dependent senses arise. For instance, consider the Uzbek sentence: “*Keyingi yozda u yozishni o‘rganmoqchi*” (Next summer, he/she wants to learn writing). Here, the meaning and function of “*yoz*” depend entirely on context. The polysemous “*yoz*” is disambiguated by cues such as “*keyingi*” (time period marker) and “*o‘rganmoqchi*” (learning intent). Conventional POS taggers mislabel such words, propagating errors into downstream tasks. Models that integrate broader sentence context to disambiguate polysemous forms are therefore essential for improving tagging precision.'

In summary, MRLs present two main challenges: (1) high-frequency OOV forms arising from morphological productivity, and (2) pervasive polysemy, exacerbated by limited annotated data. Existing approaches either rely on fixed morphological rules and manually constructed lexicons or use subword representations that do not consistently align with morpheme boundaries. We propose AffixTag, a bidirectional long short-term memory–conditional random field (BiLSTM–CRF) architecture that integrates XLM-RoBERTa (XLM-R) contextual embeddings with a dynamic affix selection mechanism. Our contributions are threefold:

- 1) Dynamic affix selection, which adapts n-gram ranges and frequency thresholds to word length;
- 2) Context–affix fusion, which integrates contextual and morphological representations through a learnable gate;
- 3) Empirical evaluation, which examines overall tagging performance, OOV robustness, polysemy resolution, and cross-lingual transfer.

2. Related Work

Existing POS tagging methods for MRLs span rule-based, neural, and contextual approaches, each addressing certain challenges but falling short in handling agglutination, OOV words, and polysemy, motivating our joint feature-fusion strategy.

Classical methods include word-level [2,3], character-level [4,5], and hybrid models [6-8]. Word-level approaches rely on lexicons but struggle with rare or unseen words. Character-level methods capture morphological cues but lack long-range context. Hybrid models combine these strengths and remain strong baselines for MRLs.

Morphology-driven models [9-11] segment words into stems and affixes to improve tagging but require extensive linguistic resources, which are often unavailable for low-resource languages like Kyrgyz and Uzbek. Gains in languages such as Malayalam [12] and Myanmar [13] depend on large annotated corpora and detailed morphological rules.

Contextualized embeddings, including ELMo [14] and Transformer-based mBERT [15,16], generate token embeddings conditioned on sentence context, enhancing morphological and semantic modeling. We systematically compare these approaches to evaluate the impact of dynamic context on POS tagging.

In summary, word-level static embeddings suffer from vocabulary sparsity and OOV issues, character-level models lack global context limiting polysemy resolution, and morphology-based or hybrid methods either omit explicit segmentation or rely on costly resources. Multilingual contextual models like XLM-R [17] improve performance but their subword tokenization often misaligns with true morphological boundaries [18], hindering fine-grained morphological representation.

These limitations motivate integrating explicit morphological signals with contextual embeddings. Our approach combines XLM-R’s context-aware embeddings with affix n-gram features to balance global semantic coherence and detailed morphology, enhancing OOV robustness and providing a scalable solution for low-resource MRLs.

3. Methodology

The architecture (Fig. 1) comprises three modules:

- 1) Joint feature representation: XLM-R contextual embeddings are concatenated with affix-based representation and passed through a learnable gating layer to balance global and morphological cues.
- 2) BiLSTM encoder: a bidirectional LSTM captures syntactic and semantic dependencies over the fused vectors.
- 3) CRF decoder: a linear-chain CRF models tag transitions and yields the globally optimal POS sequence. We detail the design and implementation of the joint feature representation layer.

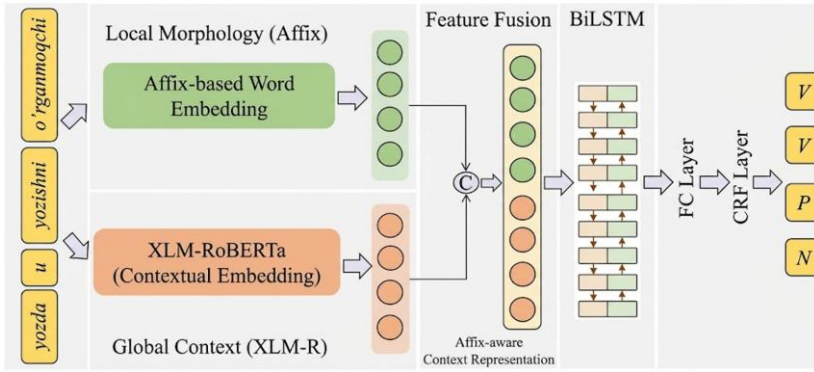


Fig. 1. Affix-based word embedding module.

3.1 Contextual Word Embedding

We use XLM-R contextual embeddings to encode sentence-level information. XLM-R consists of twelve Transformer encoder layers; each layer replaces recurrence with multi-head self-attention to model token interactions. For query, key, and value matrices Q, K, V , attention is computed via scaled dot-product:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

In the attention mechanism, Q , K , and V denote the query, key, and value matrices, respectively, and d_k is the key dimensionality. The attention weights are obtained by applying softmax to the scaled dot-product $QK^T/\sqrt{d_k}$. Multi-head attention constructs multiple independent (Q, K, V) projections via learned linear maps, computes attention per head, and concatenates the resulting head outputs before a final linear projection. Formally, for h heads:

$$\text{head}_i = \text{attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

$$\text{Multihead}(Q, K, V) = \text{concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (3)$$

where $W_i^Q \in \mathbb{R}^{d_K \times \tilde{d}_K}$, $W_i^K \in \mathbb{R}^{d_K \times \tilde{d}_K}$, $W_i^V \in \mathbb{R}^{d_K \times \tilde{d}_K}$. The outputs of the attention heads are concatenated to form a single matrix and learns different aspects of input representations by having learned weight matrices. This allows the model to capture various relationships and dependencies in the input sentence.

3.2 Affix-based Word Embedding

XLM-R uses subword tokenization for large vocabularies and OOV words, but often over segments true morpheme boundaries in MRLs [19]. We therefore construct affix-aware representations from position-sensitive character n-grams that capture candidate prefixes and suffixes. Fig. 2 illustrates an example word *mekteplerdin* (“from schools”).

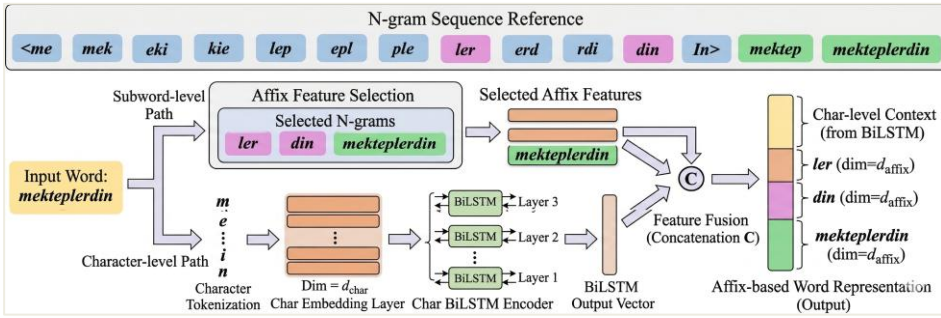


Fig. 2. Overall architecture of AffiXTag.

Each word is represented by position-sensitive character n-grams. Boundary symbols “<” and “>” distinguish word-initial and word-final sequences from internal substrings. For example, *hizmatdashlar* (“workmates”) generates trigrams {<hi>, <hiz>, <izm>, <zma>, <mat>, <atd>, <tda>, <das>, <ash>, <shl>, <hla>, <lar>, <ar>}. We also include the full stem (*hizmatdash*) to isolate post-stem affixes, for instance, the trigram <ash> from *hizmatdashlar* is treated differently from standalone word *ash* (“meal”). In practice, we extract n-grams of length 3–7 and keep only those whose corpus frequency exceeds a threshold T , thereby filtering low-frequency fragments and retaining candidate affix patterns.

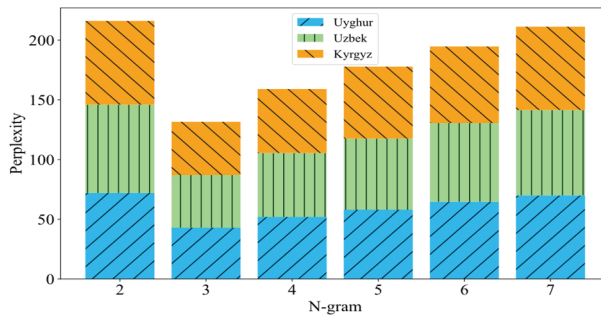


Fig. 3. Affix-based word embedding module.

We evaluate n-gram lengths (2–7) and frequency thresholds ($T = 25, 50, 75, 100, 150, 200$). As shown in Fig. 3, the 3-grams at $T=100$ minimize validation perplexity across all languages balancing noise reduction and morphological coverage in line with Zipf’s Law [20]. We therefore propose a dynamic n-gram selection with adaptive thresholding that adjusts to word length and morphological complexity:

- 1) Short words (≤ 5 characters) use 3–5 n-grams with a lower frequency threshold ($T=50$) to avoid missing critical affixes.
- 2) Long words (> 10 characters) use 5–7 n-grams with a higher threshold ($T=100$) to prioritize

statistically significant patterns. To construct the final affix embedding, we score each n-gram by its log-frequency and normalize across the set of affixes associated with the word. Formally, for word w let: $k_w = \{k: 1 \leq |k| \leq K, k \text{ occurs in } w\}$ and let $f(k)$ be the corpus frequency of k . Define the affix weight:

$$\alpha(k) = \frac{\log(f(k))}{\sum_{j \in k_w} \log(f(j))} \quad (4)$$

Final affix-based embedding for w is

$$v_w^{\text{affix}} = \sum_{k \in k_w} \alpha(k) Z_k \quad (5)$$

where Z_k is the learned vector for n-gram k .

3.3 Feature Fusion

Integrating XLM-R contextual embeddings with affix-based features enhances POS tagging by combining global context and local morphology [21]: Global context (XLM-R): multi-head self-attention captures inter-word dependencies (e.g., “*keyingi yozda*” implies the temporal sense of *yoz*). Local morphology (affix): n-gram affix features parse intra-word structure (e.g., segmenting *yozishni* as *yoz* + *-ish* + *-ni*), resolving OOVs and polysemy.

For example, for the OOV word *kélmékligimiz* (“our act of coming”), XLM-R captures the word’s contextual role from the surrounding sentence, whereas the affix representation encodes its internal structure as the stem *kél-* followed by the suffix sequence *-méklik-imiz*. In this process, we concatenate contextual and affix features with learnable gating to balance their contributions. For concise joint representations, we define as follows:

$$Z = \lambda_1 h_t^{CW} + \lambda_2 h_t^{AW} \quad (6)$$

where, h_t^{CW} and h_t^{AW} denote the features extracted by the contextual-word and affix-based encoders, respectively. The fusion weights λ_1 and λ_2 are both initialized to 0.5 and tuned via grid search on the validation set.

4. Experiments and Analysis

4.1 Datasets and Experimental Setup

Because no existing public corpus simultaneously satisfied our domain, annotation, and data-scale requirements for all three target languages, we constructed a unified POS-tagging [22] dataset from news texts. Dataset statistics are reported in Table 1.

We use 12 coarse-grained POS tags from UD v2.10: Noun (N), Verb (V), Adverb (D), Quantifier (Q), Auxiliary (E), Pronoun (P), Numeral (M), Adjective (A), Punctuation (Y), Conjunction (C), Modal particle (T), and Postposition (R). To assess generalizability, we also evaluate on the METU Turkish Treebank (5,700 sentences, 13 POS tags), selected for its typological similarity [23], as shown in Table 1. All evaluations use the same data splits and hyperparameter settings. In our experiments we used a learning rate of 2×10^{-5} , batch size 32, and trained for 10 epochs. Other settings (warmup proportion, maximum sequence length, optimizer) follow the BERT-Base defaults. XLM-R was fine-tuned on the

Uyghur news domain using the annotated POS_ukg corpus.

Table 1. Dataset statistics

Language	#Train	#Validation	#Test	#Tokens (distinct stem)
Uyghur	30k	1k	0.5k	34747
Uzbek	20k	1k	0.5	23793
Kyrgyz	15k	0.2k	0.5k	21121

4.2 Baseline Model Comparison

To evaluate the effectiveness of our BiLSTM-CRF model, we compared it against BiLSTM-CNN and NeuralSeq [24], both of which use combined word + character embeddings. Experiments were conducted on POS_ukg and METU datasets, as shown in Table 2.

Table 2. Performance comparison of sequence tagging models

Model	POS_ukg			METU
	Uy	Kg	Uz	Tr
BiLSTM-CNN	74.63	72.56	73.98	74.82
NeuralSeq	72.41	71.05	71.64	74.02
BiLSTM-CRF	75.13	73.11	74.32	76.18

Table 2 shows that BiLSTM-CRF outperforms both BiLSTM-CNN and NeuralSeq on POS_ukg and METU, achieving accuracies of 75.13% (Uy), 73.11% (Kg), 74.32% (Uz), and 76.18% (Tr). These results indicate that the BiLSTM-CRF backbone provides a stronger sequence-tagging baseline under the current experimental setting. Based on these results, we use BiLSTM-CRF as the backbone for our joint context–morphology feature fusion. To assess the contribution of each feature type, we evaluate word-only, char-only, and hybrid embeddings within the BiLSTM-CRF framework.

Table 3. Performance comparison of feature representations in the BiLSTM–CRF framework

Model	Uyghur			Kyrgyz			Uzbek		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
BiLSTM-CRF+CWE	80.46	80.70	80.58	76.50	77.36	76.93	78.14	79.85	78.99
BiLSTM-CRF+AWE	81.76	81.18	81.47	77.66	78.05	77.85	79.48	80.36	79.92
BiLSTM+Hybrid	84.95	83.26	84.10	79.28	80.18	79.73	81.68	82.09	81.88

Table 3 shows that the hybrid BiLSTM CRF outperforms single feature models, achieving F1-scores of 84.10% for Uyghur, 79.73% for Kyrgyz, and 81.88% for Uzbek. The consistent improvements across the three languages indicate that combining contextual and affix-based representations is beneficial across different agglutinative systems. The gains can be attributed to two complementary sources of information. First, affix-based representations encode intrinsic morphological patterns, enabling robust representations of both frequent and unseen affixed forms. Second, contextual embeddings capture inter-word dependencies and disambiguate word senses. Models relying solely on one feature type perform adequately but lack complementary linguistic information: affix-only models handle OOV and rare forms better than context-only models, while context-only models capture syntactic relations missed by affix features. We further identify language-specific effects. Uyghur derives the greatest benefit from affix-based representations, exceeding the gains observed in Uzbek or Kyrgyz. To evaluate cross-lingual generalization, we additionally test the model on the POS annotations of the METU Turkish Treebank.

For comparison, the experimental setup includes several baselines such as char-based CNN [25], CRF [26], LSTM [27], and RNN [28] baselines.

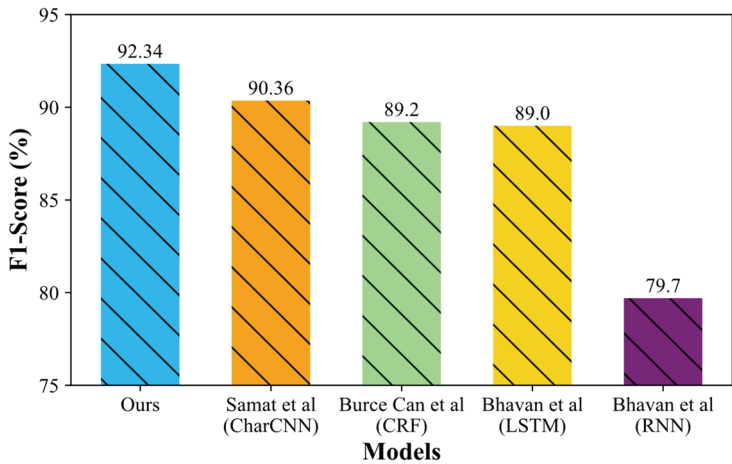


Fig. 4. Extended performance comparison.

As shown in Fig. 4, our model achieves nearly a 2% F1-score improvement over the next best baseline on METU. This improvement stems from affix-based representations that better capture morphology and reduce OOV errors, combined with contextual embeddings that model word dependencies and context for more accurate tagging.

4.3 Tag-wise Performance

We perform a tag-wise evaluation to assess performance across 12 major POS tag classes, as shown in Fig. 5.

The results show that across three languages nouns achieve the highest scores 92.36% (Uyghur), 91.45% (Kyrgyz), and 89.93% (Uzbek) while postpositions score lowest at 77.18%, 75.78%, and 74.32%, respectively. The higher scores for nouns, verbs, and adjectives suggest that their recurrent affix patterns are effectively captured by the affix-aware representation. The lower scores for function-word categories may instead reflect class imbalance, contextual ambiguity, or limited training instances.

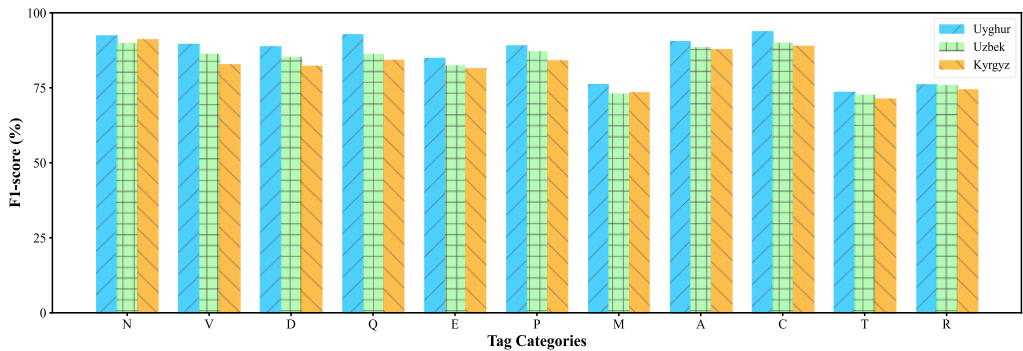


Fig. 5. Tag-wise POS performance.

4.4 Dynamic N-gram and Threshold Ablation

To examine how n-gram size and frequency thresholds affect POS tagging, we perform an ablation study. Fixed n-gram ranges (3–7) have been effective in past work but can miss morphological variation in words of different lengths. Our adaptive method adjusts n-gram size according to word length, allowing it to capture detailed morphological patterns while avoiding overfitting to irrelevant sequences. We compare this dynamic strategy with the fixed 3–7 range across three languages. The results are reported in Table 4.

Table 4. Comparative performance of fixed vs. dynamic n-gram

Language	Strategy	Precision (%)	Recall (%)	F1-score (%)
Uyghur	Fixed (3–7)	84.95	83.26	84.10
	Dynamic	85.67	84.98	85.32
Kyrgyz	Fixed (3–7)	79.28	80.18	79.73
	Dynamic	81.14	80.64	80.89
Uzbek	Fixed (3–7)	81.68	82.09	81.88
	Dynamic	83.02	82.48	82.75

The dynamic strategy consistently outperforms the fixed n-gram range, especially for longer words, achieving a 1.5% F1-score gain in Uyghur. This indicates that dynamic selection better captures the morphological complexity of agglutinative languages than fixed n-gram sizes. The 2-gram setting is excluded since many 2-grams pair a content character with a boundary marker used only for position, limiting their ability to represent morphological patterns like verb conjugations and noun declensions.

Table 5. Linguistic validation of affixes

Language	Valid Affixes		Noise reduction
	Fixed	Dynamic	
Uyghur	80.46	80.70	80.58
Kyrgyz	81.76	81.18	81.47
Uzbek	84.95	83.26	84.10

To ensure the selected n-grams are valid affixes, we collaborate with language experts to annotate 500 frequent n-grams (3–7 characters) per language and verified their affix status, as shown in Table 5. The results show that our dynamic n-gram method with adaptive thresholds retains 88.5% valid affixes (e.g., Uyghur *-lik*, Kyrgyz *-din*), while the fixed range method keeps only 73.2%. The increase in valid-affix retention indicates that positional constraints and length-dependent thresholds filter out noisy character fragments (e.g., *kel-*). We compare four configurations to isolate the effects of length-adaptive n-gram selection and adaptive thresholding: 1) Baseline: fixed n-grams (3–7) with fixed threshold ($T=100$); 2) Dynamic n-grams only; 3) Adaptive threshold only; 4) Combined: dynamic n-grams with adaptive thresholds. Table 6 reports the results of four configurations.

The combined strategy achieves the highest F1-score and affix precision, showing that dynamic n-grams and adaptive thresholds work well together to improve morphological representation. Affix precision increases from 72.3% to 89.4%, reducing noise while keeping valid morphological units. For example, in Uzbek, the baseline wrongly treats *kel-* (a verb stem) as a suffix in *kelmoqligimiz* because of its high frequency, while our method, using positional constraints, correctly identifies *-ligimiz* as the suffix cluster. Setting the frequency threshold T is important: if it is too strict, rare but correct affixes, such as archaic Kyrgyz case markers, may be removed. Future work will investigate automatic threshold

selection through meta-learning.

Table 6. Ablation results for n-gram and threshold configurations

Configuration	F1-score (%)		
	Uy	Kg	Uz
Baseline	84.10	79.73	81.88
Dynamic n-grams only	84.95	80.18	82.45
Adaptive thresholding only	84.32	79.89	82.10
Combined strategy	86.02	81.74	83.89

To determine whether the gains result from dynamic affix selection rather than increased character-level capacity, we compare the proposed module with a parameter-matched Char-CNN. The results are reported in Table 6.

Table 7. Comparison of affix-based and character-based representations

Module	F1-score (%)		
	Uy	Kg	Uz
Static affix	85	83.7	82.9
Char-CNN	85.3	84.1	83.2
Dynamic affix	85.9	84.6	83.5

Table 7 shows that the dynamic affix module consistently outperforms both static affix-based representation and a parameter-matched Char-CNN across three languages. Although the Char-CNN provides modest gains over static affix representations, the dynamic module delivers additional improvements of approximately 0.5–0.9 F1-score. This advantage arises from its ability to emphasize morphologically salient substrings while suppressing noisy character sequences, thereby producing a more informative and interpretable signal. By contrast, the Char-CNN distributes representational capacity across a wide range of character n-grams, many of which are less relevant for POS discrimination. The consistent improvements across all three languages demonstrate that dynamic affix selection constitutes a more effective mechanism for exploiting subword morphology in MRLs.

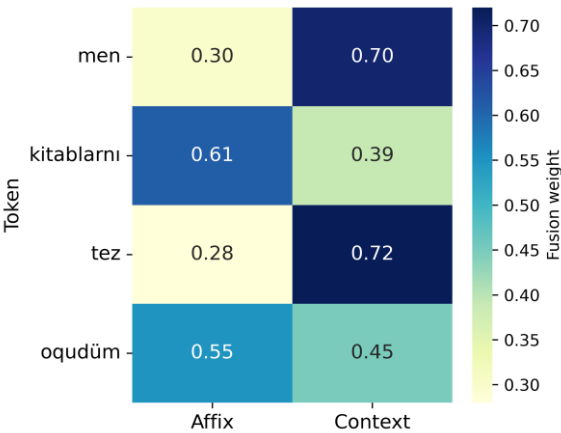


Fig. 6. Heatmap of fusion weights for selected tokens. Each row corresponds to a token, and the two columns indicate the relative contribution of affix-based vs. contextual representations.

To further illustrate how the dynamic affix module operates, we discuss the example sentence *Men kitablarni tez oqudüm* (“I quickly read the books.”) as shown in Fig. 6. For the target token *kitablarni*, the dynamic affix module selectively scores candidate substrings rather than indiscriminately using all character n-grams. The top three selections were *-lar* (selection score = 0.42), *-ni* (0.35), and *kitab* (0.18). When the affix-derived representation is fused with XLM-R contextual embeddings, the gating mechanism assigns greater weight to the affix signal ($\text{gate_affix} = 0.61$, $\text{gate_context} = 0.39$), highlighting the importance of the plural and accusative suffixes for POS assignment. This selective approach improves the signal-to-noise ratio by filtering irrelevant character fragments and provides interpretable subword evidence: the model identifies *kitablarni* as a plural accusative NOUN based on salient affixes, while contextual cues assist in resolving ambiguity in other tokens.

4.5 OOV Robustness Evaluation

To evaluate our model’s robustness to OOV words, we conduct an empirical study on three OOV types, each representing distinct linguistic and data scarcity challenges:

Type 1 (Unseen in Training and Pretraining): Words absent from both the training corpus and XLM-R’s tokenizer;

Type 2 (Novel Affixations): Words with known stems but unseen suffix combinations;

Type 3 (Unseen Stems): Words with novel stems but known affixes.

Table 8. OOV performance comparison of affix, contextual, and hybrid models

OOV type	F1-score (%)		
	Affix-based	Contextual	Hybrid
Type 1	68.3	71.5	73.8
Type 2	76.2	64.7	75.9
Type 3	58.9	65.4	63.1

Table 8 shows that affix-based representations perform best on Type 2 OOVs, surpassing contextual embeddings by 12% because they can generalize over regular morphological patterns. Type 3 OOVs give the lowest F1-score 63.1%, showing that the model depends heavily on stem semantics. For example, although the suffix *-chi* is common in training (e.g., *o’qituvchi* “teacher”), the model misclassifies *kitobchi* because the stem *kitob* (“book”) is unseen. This highlights a limitation: morphological cues alone cannot resolve semantic gaps for new stems. To address this, future work could use external stem lexicons or generative approaches to improve stem-level generalization. While affix-based representations lead on Type 2 OOVs (76.2% F1-score), contextual embeddings perform better on Type 3 OOVs (65.4% vs. 58.9%) by using surrounding words to infer stem meaning. Combining both approaches gives the best overall F1-score of 73.8%.

4.6 Polysemy Resolution Analysis

To evaluate the contextual representation capacity of our XLM-R-based model, we perform a polysemy resolution analysis across three MRLs, as shown in Table 9. We begin with the Uyghur word *chüsh* (dream/noon) and extend the study to verb–noun homographs in Uzbek (*yoz*: write/summer) and adjective–adverb ambiguities in Kyrgyz (*tez*: fast/quickly).

Table 9. Resolution performance comparison of affix, contextual, and hybrid models

Word	Language	POS ambiguity	F1-score (%)		
			Contextual	Affix-based	Hybrid
Chüşh (dream/noon)	Uy	Noun (dream) vs. Noun (noon)	85.2	82.7	89.5
Yoz (write/summer)	Uz	Verb (write) vs. Noun (summer)	86.7	83.2	90.1
Tez (fast/quickly)	Kg	Adjective (fast) vs. Adverb (quickly)	84.4	82.9	89.0

The hybrid approach outperforms both standalone methods, showing robust performance across all evaluations. Its advantage lies in combining (1) contextual cues for semantic disambiguation, as in resolving the Uzbek homograph *yoz* (“write”/“summer”), and (2) morphological precision, as in detecting the Kyrgyz adverbial suffix *-dyn* in *tez*. In contrast, the affix-based model performs worse on polysemous words (82.7% F1-score for Uyghur *chüşh*, “dream”/“noon”), while contextual embeddings struggle with rare morphological variants (83.2 % F1-score for Uzbek *yoz*). These findings confirm that the hybrid model combines complementary strengths while mitigating the weaknesses of each individual approach.

5. Conclusion

We present AffiXTag, a unified model that combines XLM-R’s contextual embeddings with dynamic, frequency-based affix segmentation to improve handling of OOV words and polysemy in morphologically rich languages. By adjusting n-gram lengths and thresholds based on word structure, the model keeps 89.1% of valid affixes and reduces OOV errors by 19%. AffiXTag uses both sentence-level context and detailed morphological features to resolve ambiguity and consistently outperforms baseline methods on Uyghur, Kyrgyz, Uzbek, and Turkish.

Future work will focus on morphology-driven data augmentation and cross-lingual transfer techniques to improve performance on unseen stems and rare affixes. These improvements will help make AffiXTag more effective for MRL/LRL settings.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

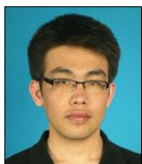
This work was supported by the National Natural Science Foundation of China Youth Program Grant 62306263, and the Ministry of Education of Humanities and Social Science Project 23XJJC740001.

References

[1] J. McGiff and N. S. Nikolov, “Overcoming data scarcity in generative language modelling for low-resource languages: a systematic review,” 2025 [Online]. Available: <https://doi.org/10.48550/arXiv.2505.04531>.
[2] A. Haber and M. Poesio, “Polysemy—evidence from linguistics, behavioral science, and contextualized

- language models,” *Computational Linguistics*, vol. 50, pp. 351-417, 2024. https://doi.org/10.1162/coli_a_00500
- [3] A. Pradhan and A. Yajnik, “Parts-of-speech tagging of Nepali texts with bidirectional LSTM, conditional random fields and HMM,” *Multimedia Tools and Applications*, vol. 83, no. 4, pp. 9893-9909, 2024. <https://doi.org/10.1007/s11042-023-15679-1>
 - [4] K. Talukdar and S. K. Sarma, “Deep learning-based part-of-speech tagging for Assamese using RNN and GRU,” *Procedia Computer Science*, vol. 235, pp. 1707-1712, 2024. <https://doi.org/10.1016/j.procs.2024.04.161>
 - [5] A. Murat and S. Ali, “Low-resource POS tagging with deep affix representation and multi-head attention,” *IEEE Access*, vol. 12, pp. 66495-66504, 2024. <https://doi.org/10.1109/ACCESS.2024.3395454>
 - [6] L. Wang, C. Dyer, A. W. Black, I. Trancoso, R. Fernandez, S. Amir, L. Marujo, and T. Luis, “Finding function in form: compositional character models for open vocabulary word representation,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015, pp. 1520-1530. <https://doi.org/10.18653/v1/D15-1176>
 - [7] R. Cotterell and K. Duh, “Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields,” 2024 [Online]. Available: <https://arxiv.org/abs/2404.09383v1>.
 - [8] Z. Yang, R. Salakhutdinov, and W. W. Cohen, “Transfer learning for sequence tagging with hierarchical recurrent networks,” 2017 [Online]. Available: <https://doi.org/10.48550/arXiv.1703.06345>.
 - [9] M. Maimaiti, A. Wumaier, K. Abiderexiti, and T. Yibulayin, “Bidirectional long short-term memory network with a conditional random field layer for Uyghur part-of-speech tagging,” *Information (Basel)*, vol. 8, no. 4, article no. 157, 2017. <https://doi.org/10.3390/info8040157>
 - [10] S. Dandapat, S. Sarkar, and A. Basu, “Automatic part-of-speech tagging for Bengali: an approach for morphologically rich languages in a poor resource scenario,” in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 2007, pp. 221-224.
 - [11] P. Matan and P. Velvizhy, “Systematic review of morphological and semantic analysis in a low resource language: Tamil,” in *Empowering Low-Resource Languages with NLP Solutions*. Hershey, PA: IGI Global, 2024, pp. 86-112. <https://doi.org/10.4018/979-8-3693-0728-1.ch005>
 - [12] D. L. Cing and K. M. Soe, “Improving accuracy of part-of-speech tagging using hidden Markov model and morphological analysis for Myanmar language,” *International Journal of Electrical and Computer Engineering*, vol. 10, no. 2, pp. 2023-2030, 2020. <https://doi.org/10.11591/ijece.v10i2.pp2023-2030>
 - [13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” 2018 [Online]. Available: <https://doi.org/10.48550/arXiv.1802.05365>.
 - [14] K. Karthikeyan, Z. Wang, S. Mayhew, and D. Roth, “Cross-lingual ability of multilingual BERT: an empirical study,” 2020 [Online]. Available: <https://doi.org/10.48550/arXiv.1912.07840>.
 - [15] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, et al., “Unsupervised cross-lingual representation learning at scale,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Virtual Event, 2020, pp. 8440-8451. <https://doi.org/10.18653/v1/2020.acl-main.747>
 - [16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., “Roberta: a robustly optimized BERT pretraining approach,” 2019 [Online]. Available: <https://doi.org/10.48550/arXiv.1907.11692>.
 - [17] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, MN, USA, 2019, pp. 4171-4186. <https://doi.org/10.18653/v1/N19-1423>
 - [18] M. Cristelli, M. Batty, and L. Pietronero, “There is more than a power law in Zipf,” *Scientific Reports*, vol. 2, article no. 812, 2012. <https://doi.org/10.1038/srep00812>

- [19] Z. Sun and X. Li, "Named entity recognition model based on feature fusion," *Information*, vol. 14, no. 2, article no. 133, 2023. <https://doi.org/10.3390/info14020133>
- [20] D. G. Da Silva and A. A. de Moura Mesnes, "Comparing long short-term memory (LSTM) and bidirectional LSTM deep neural networks for power consumption prediction," *Energy Reports*, vol. 10, pp. 3315-3334, 2023. <https://doi.org/10.1016/j.egyr.2023.09.175>
- [21] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning (ICML)*, Williamstown, MA, USA, 2001, pp. 282-289. <https://repository.upenn.edu/handle/20.500.14332/6188>
- [22] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica*, vol. 22, pp. 276-282, 2012. <https://doi.org/10.11613/BM.2012.031>
- [23] K. Katzner and K. Miller, *The Languages of the World*, 3rd ed. London, UK: Routledge, 2002. <https://doi.org/10.4324/9780203430163>
- [24] J. Acs, A. Kadar, and A. Kornai, "Subword pooling makes a difference," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, Virtual Event, 2021, pp. 2284-2295. <https://doi.org/10.18653/v1/2021.eacl-main.194>
- [25] A. Chiche and B. Yitagesu, "Part of speech tagging: a systematic review of deep learning and machine learning approaches," *Journal of Big Data*, vol. 9, article no. 10, 2022. <https://doi.org/10.1186/s40537-022-00561-y>
- [26] S. Ali and A. Murat, "Improved character-based neural network for POS tagging on morphologically rich languages," *Journal of Information Processing Systems*, vol. 19, no. 3, pp. 355-369, 2023. <https://doi.org/10.3745/JIPS.02.0197>
- [27] Y. Esref and B. Can, "Using morpheme-level attention mechanism for Turkish sequence labelling," in *Proceedings of the 27th Signal Processing and Communications Applications Conference (SIU)*, Sivas, Turkey, 2019, pp. 1-4. <https://doi.org/10.1109/SIU.2019.8806530>
- [28] C. A. Bahcevan, E. Kutlu, and T. Yildiz, "Deep neural network architecture for part-of-speech tagging for Turkish language," in *Proceedings of the 3rd International Conference on Computer Science and Engineering (UBMK)*, Sarajevo, Bosnia and Herzegovina, 2018, pp. 235-238, <https://doi.org/10.1109/UBMK.2018.8566272>



Samat Ali <https://orcid.org/0000-0002-9425-1980>

He received B.S. and M.S. degrees from School of Computer Science and Engineering from Xinjiang University in 2006 and 2010, respectively. Since September 2012, he has been with the School of Computer Science and Engineering from Xinjiang Normal University as a Teaching Assistant. His current research interests include Natural Language Processing and Machine Translation.



Alim Murat <https://orcid.org/0000-0001-8510-7808>

He received B.E. and M.S. degrees from School of Computer Science and Technology from Xinjiang Normal University in 2011 and 2014, respectively. He received his Ph.D. degree in Computer Science from Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences in June 2017. Since July 2017, he has been with the Xinjiang Normal University as an Associate Professor. His current research interests include natural language processing and machine translation.