Deep BiLSTM POS Tagging for Morphologically Rich Languages

Abstract

This project presents the development of an advanced Part-of-Speech (POS) tagging system using a Deep Bidirectional Long Short-Term Memory (BiLSTM) architecture, specifically designed for morphologically rich languages (MRLs). Such languages often contain complex word formations, multiple affixes, flexible grammar, and high out-of-vocabulary (OOV) rates, making accurate POS tagging significantly more challenging compared to languages with simpler morphology.

To address these challenges, our system integrates multiple enhancements into a unified architecture. The core model uses a deep multi-layer BiLSTM that learns contextual dependencies from both left-to-right and right-to-left directions, enabling stronger understanding of sentence structure. To better capture morphological richness, the system incorporates character-level CNN embeddings, which effectively learn subword patterns such as roots, stems, prefixes, and suffixes. Additionally, pretrained FastText / Word2Vec embeddings strengthen word-level semantics, while an optional attention mechanism and layer normalization improve contextual representation. A CRF - based structured prediction layer ensures globally consistent tag sequences, especially for complex languages.

The model is trained and evaluated on Universal Dependencies (UD) datasets for Hindi, Gujarati, and English, enabling cross - lingual robustness. English serves as a simpler morphological baseline, while Hindi and Gujarati represent highly inflectional, morphologically rich Indo - Aryan languages. The enhanced system achieves up to 96% accuracy on Hindi - English code - mixed text, and showing similarly high performance on Gujarati and multilingual combinations, outperforming several low-resource MRL baselines such as Limbu (94%), Assamese, and Bodo (~92%). Regularization strategies (dropout, weight decay), learning-rate scheduling, and data augmentation further improve generalization and stability across the three languages.

Overall, this project demonstrates that combining deep sequence modeling with subword - level morphological features significantly boosts POS tagging performance for MRLs. The architecture provides a strong, extensible foundation for future expansions, including integration with transformer - based encoders, lemmatization-based multi-task learning, and improved handling of multilingual or code-switched data.