Unsupervised Subword Segmentation for POS Tagging in Low-Resource Agglutinative Languages

Jonas Oliveira Pereira[1], **Lilian Teixeira de Sousa**[1], **Marlo Souza**[1]

[1]Federal University of Bahia (UFBA)
Campus Ondina, 40.170-110 Salvador, Brazil

{jonas.pereira, lilian.sousa, msouza1}@ufba.br

Abstract. *While Part-of-speech tagging is considered a well-understood task in the literature, most work has focused on indo-european languages with fusional morphology. For low-resource agglutinative languages, such as brazillian indigenous languages, the task is commonly constrained by small annotated corpora, high lexical sparsity, and morphological patterns that are poorly represented by word-level models. This paper investigates the impact of unsupervised subword segmentation techniques on POS tagging for brazillian indigenous languages. We compare a word-level baseline with Byte Pair Encoding, Morfessor, and FlatCat on Bororo, Nheengatu, and Tupinamba corpora, including a controlled experiment on the size of the training corpus. Our findings suggest that unsupervised segmentation can reduce sparsity in low-resource POS tagging, although its benefit depends on the language, corpus size, and segmentation method.*

Keywords: POS tagging; low-resource; agglutinative; tokenization; indigenous.

1. Introduction

Agglutinative languages pose a difficult challenge for natural language processing due to its morphological richness [Tsarfaty et al. 2013, Özçift et al. 2021]. More yet, for those in low-resource settings, such as for Brazilian Indigenous languages, annotated corpora are often too small to support reliable supervised training [Mager et al. 2018]. This is especially problematic for part-of-speech (POS) tagging since a single word may contain several morphologically relevant affixes, carrying grammatical information that helps determine the word's tag. As a result, the same stem can appear in many different surface forms, increasing lexical sparsity and making rare or unseen forms more common during training and evaluation [Mager et al. 2018, Mager et al. 2022].

For Latin American Indigenous and other under-documented languages, this problem is intensified by practical constraints: digital corpora are small, linguistic expertise is scarce, and annotation resources are expensive to produce [Tonja et al. 2024, Mager et al. 2018]. A word-level POS tagger trained under these conditions may fail to generalize across related surface forms, even when those forms share recurring stems or affixes [Eskander et al. 2022, Libovický and Helcl 2024]. Unsupervised subword segmentation offers a possible way to expose these recurring units without requiring manually segmented data [Mager et al. 2022, Virpioja et al. 2013].

This paper investigates whether unsupervised subword segmentation improves POS tagging for low-resource agglutinative languages. The remainder of the paper is organized as follows. Section 2 presents related work. Section 3 describes the segmentation methods, corpora, and experimental setup. Section 4 reports and interprets the results, and Section 5 concludes the paper.

2. Related Work

Prior work has studied subword and morphological segmentation as a way to reduce sparsity in low-resource morphologically complex languages. [Mager et al. 2022] compare BPE with supervised and unsupervised morphological segmentation for machine translation in four polysynthetic languages, finding that unsupervised morphological segmentation often outperforms BPE for the task. [Saleva and Lignos 2021] evaluate morphology-aware segmentation for low-resource neural machine translation, but find no consistent advantage over BPE across language pairs. More recently, [Libovický and Helcl 2024] propose lexically grounded segmentation methods and report clearer gains for POS tagging than for machine translation. These results suggest that segmentation effects are task-dependent and especially relevant for morphologically sensitive tasks.

For POS tagging, [Eskander et al. 2022] study cross-lingual POS tagging for low-resource morphologically rich languages by projecting tags through stems and morphemes instead of full word forms. Their results show that replacing sparse surface words with stem- or morpheme- based representations improves projection-based tagging. [Keren et al. 2022] compare character-level and subword-level pretrained models on morphologically rich languages, including POS tagging and morphological disambiguation, and find that subword models are not uniformly better across tasks.

Subword segmentation methods differ in the assumptions they make about word structure. BPE learns frequent symbol merges and is widely used to handle rare words in neural NLP [Sennrich et al. 2016]. Morfessor models unsupervised morphological segmentation from raw word forms [Virpioja et al. 2013], while FlatCat extends the Morfessor family with an HMM-based model that assigns morph categories, targeting unsupervised and semi-supervised morphology learning [Grönroos et al. 2014]. Although these methods are often evaluated in translation or segmentation tasks, their effect on POS tagging in limited resource languages, particularly Latin American Indigenous-languages corpora remains less explored.

3. Methodology

We compare four input representations for POS tagging: a word-level baseline and three subword segmented variants produced by BPE, Morfessor, and FlatCat. This design tests whether subword information helps the tagger handle sparse word forms while still predicting POS tags at the word level.

3.1. Word-level baseline

We employ a BiLSTM-CRF [Huang et al. 2015] tagger, a competitive approach for sequence tagging [Souza et al. 2019] without requiring large pretrained models. As a baseline we train and evaluate the model on the original word-level tokens, without subword segmentation. Thus, all gains reported in the results are computed against this baseline under the same dataset split, evaluation unit, and model configuration.

3.2. Corpora and Datasets

The experiments use three agglutinative languages that have publicly available datasets based on Universal Dependencies: Bororo [Gerardi et al. 2023], Nheengatu [de Alencar 2024], and Tupinamba [Gerardi 2023]. Since among the languages studied, only the Bororo corpus is sufficiently large to allow further segmentation, we conduct a study on the impact of the training corpus size in the performance, by selecting segments from the corpus of different sizes for training (see Table 1). This will be used to provide a clearer view of how segmentation behaves as training data becomes scarcer.

Table 1. Corpus statistics and language classification. Token counts refer to the original word-level baseline representation.

Dataset	Family/branch	Train sent.	Train tok	Dev sent.	Dev tok
Bororo 1k	Bororoan / Macro-Jê	1,000	7,856	2,138	15,678
Bororo 5k	Bororoan / Macro-Jê	5,000	37,923	2,138	15,678
Bororo full	Bororoan / Macro-Jê	17,107	128,418	2,138	15,678
Nheengatu	Tupian / Tupi-Guarani	1,077	11,811	1,043	10,002
Tupinamba	Tupian / Tupi-Guarani	464	3,593	117	915

3.3. Experimental Setup

In all experiments, we employ a BiLSTM-CRF POS tagger. The model uses a learned embedding dimension of 100, hidden dimension of 128, dropout of 0.2, AdamW optimization with learning rate 10^{-3}, batch size 16, gradient clipping at 5.0, and 10 training epochs. Each condition is trained with three random seeds: 13, 42, and 2026. For each corpus, segmentation models are trained only on the corresponding training tokens and then applied to the development set. BPE is learned with 1,000 merge operations and minimum frequency 2. Morfessor and FlatCat are trained for 10 epochs using their default settings.

The main experiments use an aggregation strategy to convert the segmented input back into a word-level sequence. Each sentence is initially represented as subtokens. For each original word, the model averages its constituent subtoken embeddings to produce a single word-level vector. The BiLSTM then processes this sequence of averaged vectors, and the CRF predicts one POS tag per word. Consequently, the loss is computed entirely at the word level. This setup maintains a standard word-level tagging task while allowing the input representations to incorporate internal subword structure.

In the following, we report and discuss the difference in word-level accuracy and macro-F1 (after removing punctuation) from the investigated methods to the baseline on the evaluation sets. Accuracy gives the overall POS tagging performance, while macro-F1 indicates whether improvements extend beyond the most frequent classes.

4. Results and Discussion

Figure 1 presents the results of our experiments, depicting the mean difference in Accuracy and Macro-F1 from the methods evaluated against the baseline. Segmentation improves accuracy in nearly every condition, with the only decrease occurring for BPE on Nheengatu. The largest gains appear in the smallest corpora: Tupinamba improves

from 0.600 to 0.684 ($+8.4pp$) with Morfessor and to 0.681 ($+8.1pp$) with BPE, while Bororo 1k improves from 0.676 to 0.732 ($+5.5pp$) with BPE. Nheengatu also shows a large gain with Morfessor, rising from 0.827 to 0.878 ($+5.1pp$). More yet, the standard deviation between the samples decreases with the size of the corpus, providing evidence that the morphological information captured by these methods improves accuracy.

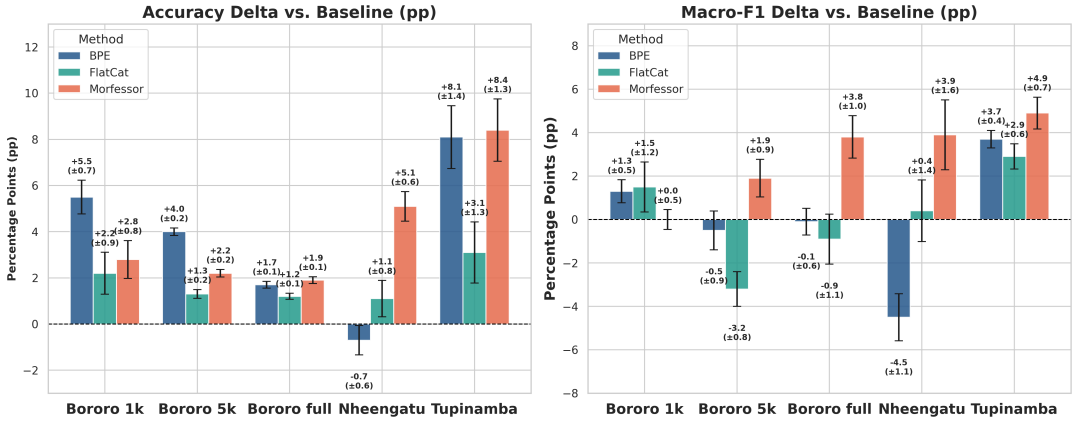

Figure 1. Mean difference (percentage points) in Accuracy and F1 from word-level baseline over three seeds, with standard deviation in error bars.

The best method varies by dataset. BPE gives the highest accuracy for the reduced Bororo settings, with gains of $+5.5pp$ in Bororo 1k and $+4.0pp$ points in Bororo 5k. Morfessor gives the highest accuracy for Bororo full, Nheengatu, and Tupinamba, and also produces the strongest macro-F1 results in most settings. FlatCat improves over the baseline in several cases, but it does not produce the best accuracy in any dataset.

These results suggest that segmentation is most useful when training data is scarce and word-level sparsity is high. The gains for Tupinamba and the Bororo data reduced settings show that subword units can help the tagger generalize across related forms. The effect, however, is not consistent across methods studied, with Morfessor being more consistent across both metrics. Notice that macro-F1 is susceptible to low performance on infrequent classes on the training data - which is often the case for low-data scenarios and a large number of classes, such as in our experiments, an aspect which we are intentionally evaluating.

These findings should be interpreted with caution. The training data scenarios explored for the Bororo language provide controlled evidence, but cross-language results also reflect differences in corpus size, annotation, domain, and morphology. Since the segmentations are evaluated only through POS tagging, the results show usefulness for tagging, not necessarily linguistically correct morphological analyses.

5. Conclusion

The results show that segmentation can improve word-level POS tagging, with Morfessor and BPE producing the strongest gains depending on the dataset and its size. Future work should examine other strategies for combining subword representations into word-level predictions, evaluate per-tag behavior in more detail, compare unsupervised segmentations against linguistically informed morphological analyses where such resources are available.

References

de Alencar, L. F. (2024). A Universal Dependencies treebank for Nheengatu. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, volume 2, pages 37–54.

Eskander, R., Lowry, C., Khandagale, S., Klavans, J., Polinsky, M., and Muresan, S. (2022). Unsupervised stem-based cross-lingual part-of-speech tagging for morphologically rich low-resource languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4061–4072, Seattle, United States. Association for Computational Linguistics.

Gerardi, F. F. (2023). UD_Tupinamba-TuDeT. Universal Dependencies treebank documentation.

Gerardi, F. F., Toribio, L., and Sollberger, D. (2023). UD_Bororo-BDT. Universal Dependencies treebank documentation.

Grönroos, S.-A., Virpioja, S., Smit, P., and Kurimo, M. (2014). Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014*, pages 1177–1185.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Keren, O., Avinari, T., Tsarfaty, R., and Levy, O. (2022). Breaking character: Are subwords good enough for MRLs after all? *CoRR*, abs/2204.04748.

Libovický, J. and Helcl, J. (2024). Lexically grounded subword segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7420, Miami, Florida, USA. Association for Computational Linguistics.

Mager, M., Gutierrez-Vasques, X., Sierra, G., and Meza-Ruiz, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mager, M., Oncevay, A., Mager, E., Kann, K., and Vu, T. (2022). BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961–971, Dublin, Ireland. Association for Computational Linguistics.

Özçift, A., Akarsu, K., Yumuk, F., and Söylemez, C. (2021). Advancing natural language processing (nlp) applications of morphologically rich languages with bidirectional encoder representations from transformers (bert): an empirical case study for turkish. *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije*, 62(2):226–238.

Saleva, J. and Lignos, C. (2021). The effectiveness of morphology-aware segmentation in low-resource neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 164–174, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Souza, F., Nogueira, R., and Lotufo, R. (2019). Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Tonja, A., Balouchzahi, F., Butt, S., Kolesnikova, O., Ceballos, H., Gelbukh, A., and Solorio, T. (2024). Nlp progress in indigenous latin american languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6972–6987.

Tsarfaty, R., Seddah, D., Kübler, S., and Nivre, J. (2013). Parsing morphologically rich languages: Introduction to the special issue. *Computational linguistics*, 39(1):15–22.

Virpioja, S., Smit, P., Grönroos, S.-A., and Kurimo, M. (2013). Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Technical report, Aalto University.