Byte Pair Encoding for Roman Urdu: Challenges and Implementation

1 Introduction

Byte Pair Encoding (BPE) is a widely used subword tokenization technique, particularly effective for morphologically rich and low-resource languages. This report presents an implementation of BPE for Roman Urdu, addressing key challenges such as spelling variations, lack of standardization, and out-of-vocabulary (OOV) words. The approach is informed by existing research on subword tokenization [1], with adaptations for the unique characteristics of Roman Urdu.

2 Challenges and Methodology

Roman Urdu presents distinct tokenization challenges due to its informal nature and lack of a standardized orthography. The methodology is designed to address these issues through a combination of lexicon mapping, character-based tokenization, and iterative merging of subword units.

2.1 Handling Spelling Variations

Unlike standardized languages, Roman Urdu words exhibit multiple spellings (e.g., "zindagi" vs. "zindagee"). Without normalization, the BPE model would treat such variations as distinct tokens, reducing its effectiveness. To address this:

- A lexicon-based approach was adopted, clustering spelling variations and mapping them to a canonical form before tokenization. This ensures consistency, allowing BPE to learn meaningful subword patterns rather than redundant variations.
- This approach is supported by studies focusing on lexical normalization of Roman Urdu, which highlight the importance of addressing spelling variations to improve text processing tasks [2].

2.2 Detecting Frequent Subword Patterns

Identifying meaningful subwords is challenging due to Roman Urdu's flexible word structure. Inspired by [1], a character-level BPE approach was selected:

- The corpus is first tokenized at the character level, ensuring a minimal initial vocabulary.
- The most frequent adjacent character pairs are iteratively merged, forming larger subword units.
- This method effectively captures common morphemes while keeping the vocabulary size manageable.

2.3 Reducing Out-of-Vocabulary (OOV) Words

A key goal of subword tokenization is to handle OOV words effectively. Without BPE, unseen words would result in `<UNK>` tokens, reducing model robustness. To mitigate this:

- The vocabulary is expanded dynamically by merging frequent subwords, enabling the model to generate subword representations even for unseen words.
- The encoding function maps input text to learned subword tokens, while decoding reconstructs the original sequence using stored merge operations.

3 Evaluation

The effectiveness of the implementation is assessed through:

- **OOV Rate:** Measuring the proportion of unseen tokens in test data to determine vocabulary coverage.
- **Decoding Accuracy:** Comparing reconstructed text with the original input to assess information preservation.
- **Corpus Compression Ratio:** Evaluating the reduction in token count, as a lower count indicates better compression and generalization.

These evaluation metrics are commonly used in assessing subword tokenization models, as discussed in [3].

4 Conclusion

This work explores the challenges of applying BPE to Roman Urdu and presents a customized approach integrating lexicon mapping for spelling normalization. The methodology effectively reduces OOV rates while preserving meaningful

subword structures. Future work includes refining lexicon mapping strategies and incorporating phonetic similarity techniques to further enhance consistency in tokenization.

References

[1] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.

[2] M. Hanif and S. A. Khokhar, "Lexical Normalization of Roman Urdu Text," *International Journal of Computer Science and Network Security*, vol. 17, no. 12, pp. 199-205, 2017.

[3] M. Gungor, "A Comprehensive Analysis of Subword Tokenizers for Morphologically Rich Languages," M.S. thesis, Dept. of Computer Engineering, Bogazici University, 2020.