Canonical Morphological Segmentation for Nguni Languages using Hard Attention

**Simbarashe Mawere & Jan Buys**

Department of Computer Science

University of Cape Town, South Africa

MWRSIM003 at myuct dot ac dot za

# Abstract

The task of Morphological Segmentation is one explored in Linguistics concerning the decomposition of words into their composite morphemes. The project focused on the task of Canonical Morphological Segmentation of Nguni Languages to extract underlying morphemes from words utilising both soft-attention and hard-attention machine learning, sequence-to-sequence models. The models included a hard-attention Neural Transducer (HNT) and a soft-attention Transformer (ST), in one implementation, in comparison to another soft-attention Transformer baseline (STB) from another implementation. Both the HNT and the ST outperformed the baseline with the HNT outperforming both transformers with an optimised average F1 score of 75.11% on all four of the languages. The expectation is that the explored models will be better optimised in the future through an iterative experimental process for use in the natural language processing (NLP) of Nguni languages and other low-resource or agglutinative languages.

# Introduction, Problem Statement and Motivation / Formulation of aims

In the computational linguistics field of NLP, the task of segmentation is an important one that involves splitting structures such as words, phrases, or sentences into composite substructures. Morphological segmentation refers to the decomposition of words into morphemes which are the smallest meaning-bearing units of language, while canonical morphological segmentation is segmentation that involves the extraction of underlying morphemes which are not visible on the surface. The purpose of the extraction of morphemes is in the preservation of languages, mainly the Nguni languages and other African languages, which are low-resource and are not optimally explored in NLP.

This morphological segmentation can then be used to help linguists in the preservation of these languages by automating tasks which would otherwise have been performed manually by human beings, costing time and human resources. These languages also contain rich information in the form of morphemes and analysis of these can be used to explore further NLP tasks such as web-crawling using search engines and more acute translation of these languages. The morphemes would be used to extract the meaning of the word and find translations which would have been difficult to decipher without the segmentation.

The is illustrated in the table below:

|  |  |  |
| --- | --- | --- |
| Word | Surface segmentation | Canonical segmentation |
| zobomi | zo-bo-mi | za-u-bu-omi |

In this project, the aim was to explore models for undertaking the abovementioned task for Nguni languages using character-level sequence-to-sequence models then compare the performance of hard-attention to soft attention as implemented by Moeng et al.

# Literature Review

# Methodology

# Results

# Conclusions and Future extensions of work

# Bibliography