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. The HNT and the ST outperformed the baseline with the HNT outperforming both transformers with an optimised average F1 score of 75.11% in all four 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, segmentation is an important task 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 that 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 that would have been difficult to decipher without the segmentation.

This is illustrated in the Table 1 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 with the expectation to improve on performance. Furthermore, the project also had the aim of improving the performance of the Transformer with regular soft attention as compared to the baseline by substantial percentage points.

The models explored in the experiments were used for the analogous grapheme-to-phoneme (g2p) sequencing, which takes input as a sequence of space-separated characters and outputs a similar sequence with the phoneme version of the grapheme. This task is almost identical to the task of canonical segmentation except for the inputs and outputs represent. Since canonical segmentation extracts underlying morphemes which may not be present on the surface therefore the input and output sequences probably differ in the length.

The experiments’ results showed that the HNT and the ST models were outperforming the baseline Transformer achieving average F1 scores of \_\_\_\_% and \_\_\_\_\_% respectively, across all four languages, when evaluated using the metrics and testing data.

# Literature Review

Talk about the models: The HNT and the ST.

* howow they were used for the task of g2p tasks
* How they work
* How they implemented hard attention

Morph Segment

* What is achieved
* What baseline did it present
* Canonical Segmentation

Data

Figure 1

# Methodology

Data Processing

The data as aforementioned was acquired from the NCHLT Speech Corpus in the form given in Figure 1. Of the three character-level sequence tasks explored by Wu et al, the g2p conversion task was most similar due to similarity in both sequence length and form. There was no annotation as with the morphological inflection task nor transliteration with the named-entity transliteration task hence the only focus was the input and output token sequences. The models took the input in the format as shown in Table 2 below meaning that the input from the speech corpus had to be moulded into an identical format to match the expectation.

|  |  |  |
| --- | --- | --- |
| Study | Input sequence | Output sequence |
| Wu et al | a c t i o n | AE K SH AH N |
| Mawere | n g o m t h o m b o | n g a - u - m u - t h o m b o |

The input files were formatted to have the input and expected output tab separated while the individual tokens were space-separated. The data loader for the model loaded all the individual unique tokens for use in the model.

The data for each language were adapted as used by the MORPH\_SEGMENT code base and the splits that were used were identical. This was to minimise the variations between the baseline and the experiments. The training, development, and testing dataset sizes for the four languages are shown in the table below

|  |  |  |  |
| --- | --- | --- | --- |
| Language | Training | Development | Testing |
| isiZulu | 17778 | 1777 | 3298 |
| isiXhosa | 16879 | 1688 | 3004 |
| isiNdebele | 12929 | 1119 | 2553 |
| siSwati | 13278 | 1080 | 1347 |

The data are split into an approximately 78:8:14 split for training, dev and testing respectively with the siSwati data set being the outlier with an approximate split of 85:7:8.

Preliminary Experimentation

Due to the utilised models being complex and involving multiple hyperparameters, preliminary experiments were conducted using the scripts provided for each model. It was due to this preliminary experimentation and model training that the g2p conversion task models were chosen over the other two, and the format for the data was acquired.

Experimental Setup

Both the hard attention and the soft attention data were formatted in the Table 2 format since the data loader was the same for the g2p conversion task, adapted for canonical segmentation. The models were implemented by adapting the code base of Wu et al which was free for use on GitHub. As shown in the sample output sequence, the generated output contains the hyphen (“-“) character as it is recognised as the morpheme demarcation boundary for the individual morphemes in the output. The empirical part of the project was to determine the better model from the present options based on recall, precision and F1 metrics. Firstly, the model architectures will be discussed, and then the hyperparameters used in the model and finally the results will be presented and analysed.

Architectures

There are three architectures considered in the experiments: the soft attention tran

Architectures

Tinkering the model to deduce inner workings

* Acquiring the data
* Trying to get the model to work with data set
* Reprocessing the data to get it into working condition for the models to use.
* Running the experiments
* Running several iterations with multiple changing parameters/

# Results

# Conclusions and Future extensions of work

# Bibliography