MORPHOLOGICAL SEGMENTATION

Morphological Segmentation is a Linguistic Operation wherein words are separated into their composite morphemes. Morphemes are the smallest possible building blocks of language that also have meaning when alone. This is a useful operation because it facilitates the study of words at a granular level. We investigated two ways in which words could be segmented, **Surface Segmentation** and **Canonical Segmentation**. Using surface segmentation a given word will be segmented into a sequence of sub-strings, which when concatenated will form the original word. Using canonical segmentation, the word will be analyzed and segmented into a sequence of canonical morphemes, where each canonical morpheme corresponds to a surface morpheme as its orthographic representation. After the segments have been predicted, the models can also be used to label the generated segments according to their function to the word as a whole.

Word: *attainability*
Surface Segmentation*: attain-abil-ity*
Canonical Segmentation: *attain-able-ity*

We implemented three machine learning models to perform these tasks on four Nguni Languages: isiNdebele, isiXhosa, isiZula and siSwati.

CONDITIONAL RANDOM FIELDS (CRF)

CRFs are a class of discriminative probabilistic supervised machine learning models. CRFs accept data sequences as input and output the corresponding label sequences. For each item in the input sequence, one label will be predicted. As such CRFs cannot be used to predict canonical segments therefore we used them to predict surface segments, and to label those generated segments

SEQ2SEQ MODEL

Sequence models (Seq2Seq) are a type of supervised learning model which are able to take in a variable length input and output a varbiale length output. This means they are able to segment words into their canonical form. In this project, three different Seq2Seq models were trained to segment words into the canonical form. A mechanism of *Attention* is introduced in order to evaluate its future potential of accurately segmenting words.

ENTROPY-BASED MODEL

The entropy-based model uses a Long-Short Term Memory character-level language model to determine the probability distribution over the successive character in a token. This probability distribution is then used to compute the information entropy at a given position in a word. The model is trained and evaluated the dataset, producing the left-entropy, and the dataset's reverse, producing the right-entropy. Both left and right entropy is used by an objective function to determine if there is a morpheme boundary. This model is unsupervised.

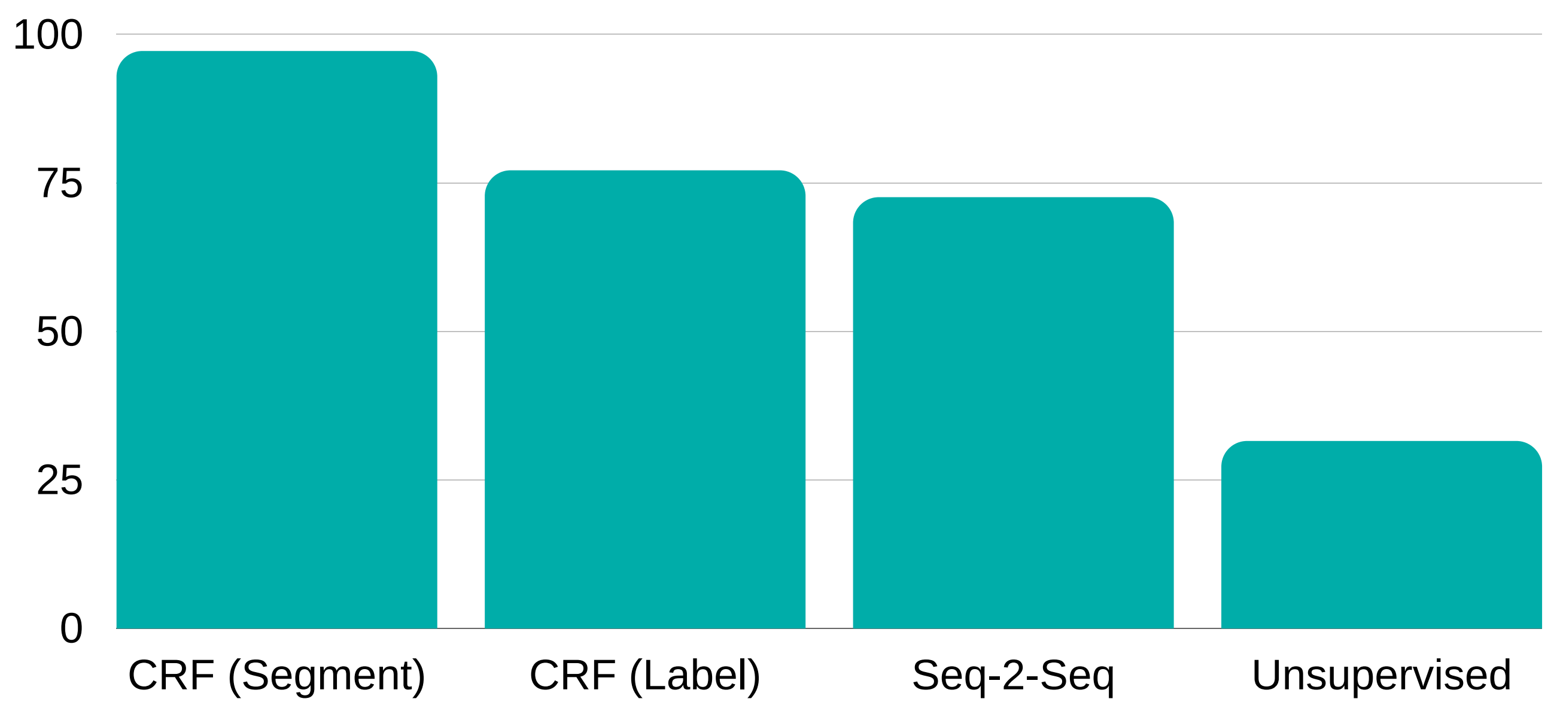

Figure: Comparison of results between the three different approaches using the F1 Score metric.
The CRF yielded the best results while the Seq2Seq came in second in the supervised category.

RESULTS

CRF:
In the task of surface segmentation, the best CRF performed with an F1 score of ***97.14%*** *, and the labeling of correct segments performed with an F1 score of* ***77.06%****. This performance shows that the CRF is an appropriate tool in the task of morphological segmentation, and that it is useful in the task of labeling.*

Entropy-based model:
The unsupervised entropy-based model was marginally surpassed by Morfessor-Baseline in early evaluations. After some adjustment the entropy-based model outperformed Morfessor-Baseline with an average F1 Score of ***31.51%*** *compared to Morfessor-Baseline's average,* ***27.69%****.*

Seq2Seq Results:
The best performing model was the Transformer model which was able to achieve an F1 score of ***72.54%****, which is an improvement of 11.95% over the baseline. The Bi-LSTM+Attention model was able to achieve an F1 score of* ***65.41%****.*

Computer Science Department
University of Cape Town
Private Box X3
Rondebosch
7701

Supervisor:
Dr. Jan Buys
Co-reader:
Mr. Zola Mahlaza

Team Members:
Aaron Daniels
Tumi Moeng
Sheldon Reay