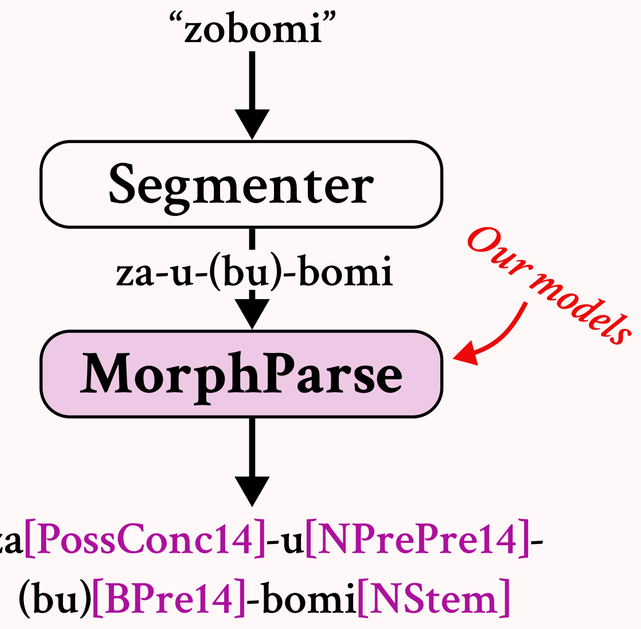
Morph Parse

Deep Learning for the Morphological Parsing of Nguni Languages

Background

- The **Nguni languages** - **IsiNdebele** (NR), **SiSwati** (SS), **IsiXhosa** (XH) and **IsiZulu** (ZU) - are a family of four South African languages.
- These languages lack **quality linguistic tools** for **Natural Language Processing** despite their large **.**
- **Morphemes** are the **smallest unit of linguistic meaning** in language (e.g. “wind” as part of “windings”).
- **Morphological parsing** is the process of splitting words into morphemes then **tagging the morphemes with grammatical tags.**
- We apply **deep learning techniques** to the **tagging step** of this problem.

Example



Approaches

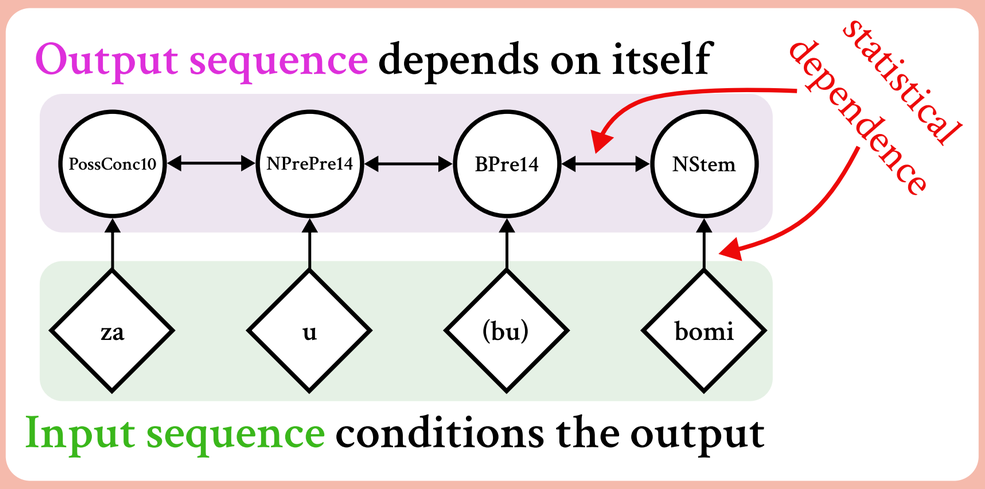
Models Trained from Scratch

Bi-LSTM: Bidirectional Long Short-Term Memory

- A type of recurrent neural network which considers the past and future of sequences for context.

Bi-LSTM CRF: Bi-LSTM Conditional Random Field

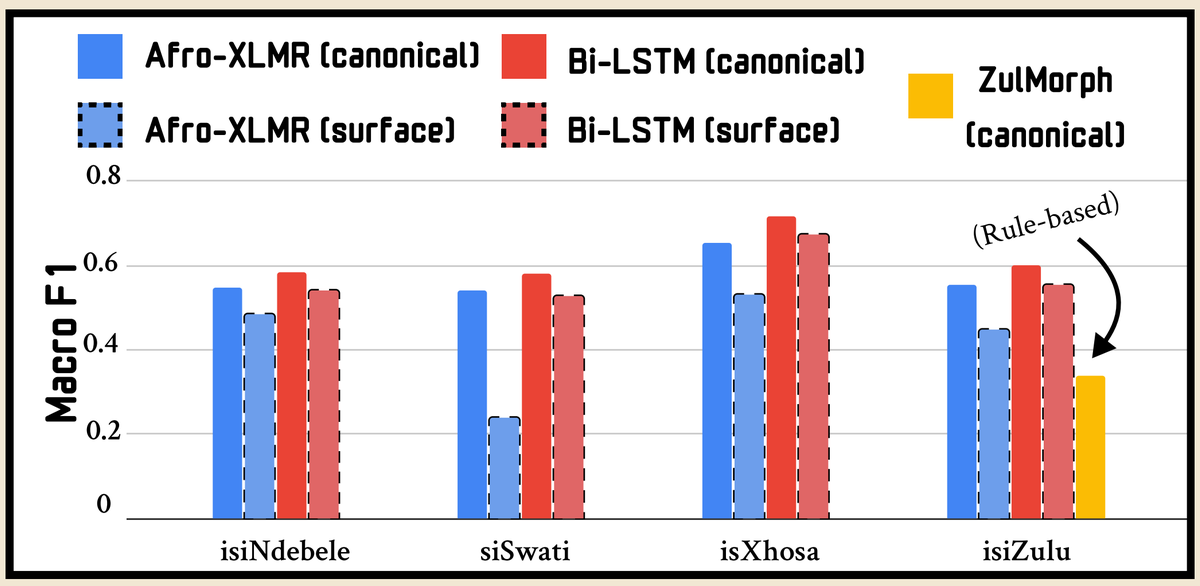
- Uses a bi-LSTM to generate features.
- Considers how output sequence tokens depend on input sequence *as well as themselves*:



Pre-Trained Language Models

PLMs leverage transfer learning to extend knowledge gained in one pre-training task (and language) onto a different task in possibly different languages. The PLMs are based on the Transformer architecture and sourced from the Hugging Face library. Our experiments used **PLMs pre-trained on increasing levels of inclusion** of the Nguni languages as shown in the table below:

model	nr	ss	xh	zu
XLM-R			✓	
Afro-XLMR			✓	✓
Nguni-XLMR	✓	✓	✓	✓



Results

- Our **deep-learning approaches** (MorphParse) **outperformed rule-based baseline** (ZulMorph).
- **Models trained from scratch outperformed pre-trained language models.**
- **Sentence-level models outperformed their word-level counterparts.**
- The models performed significantly better on **canonical segmentations** than on **surface segmentations.**
- IsiXhosa’s **linguistic compactness** showed better performance than the other Nguni languages.
- **Future work:** investigate better tokenisers for pre-trained language mdoels.

Conclusions & Contributions

- We demonstrated **feasibility of deep-learning for morphological parsing.**
- Developed **state-of-the-art morphological parsers** for use in **linguistic analysis** or **downstream NLP tasks.**

Project team

Cael Marquard
mrqcae001@myuct.ac.za

Simbarashe Mawere
mwrsim003@myuct.ac.za

Supervisor

Francois Meyer
francois.meyer@uct.ac.za