Subword Segmental Language Modelling for Nguni Languages

Francois Meyer Jan Buys

University of Cape Town

Motivation

- Subword segmenters like BPE and ULM are widely used, but are sub-optimal for morphologically rich languages.
- The Nguni languages of South Africa have agglutinative morphology: Words are formed by stringing together morphemes.
- They are also under-resourced: Available datasets are small, so held-out datasets contain rare or previously unseen words.

Instead of viewing subword segmentation as a preprocessing step, we let our model learn subword segmentation during training. This enables the model to learn subwords that optimise its training task.

Dynamic Programming

Conditioning the segment probability on all possible segmentation histories is computationally intractable, so we introduce a conditional semi-Markov assumption:

$$p(s_{ij}|\mathbf{s}_{\leq \mathbf{i},<\mathbf{j}}) \approx p(s_{ij}|\mathbf{x}_{<\mathbf{k}}),$$

where $\mathbf{x}_{<\mathbf{k}}$ is the unsegmented text before segment s_{ij}. The dynamic program computes the marginal likelihood of eq. 1 incrementally by computing the forward scores:

$$\alpha_t = \sum_{k}^{t} \alpha_k p(s = \mathbf{x}_{\mathbf{k}:\mathbf{t}}|\mathbf{x}_{<\mathbf{k}}),$$

where k is the start index of the current word. Finally, $p(\mathbf{w}) = \alpha_{|\mathbf{x}|}$.

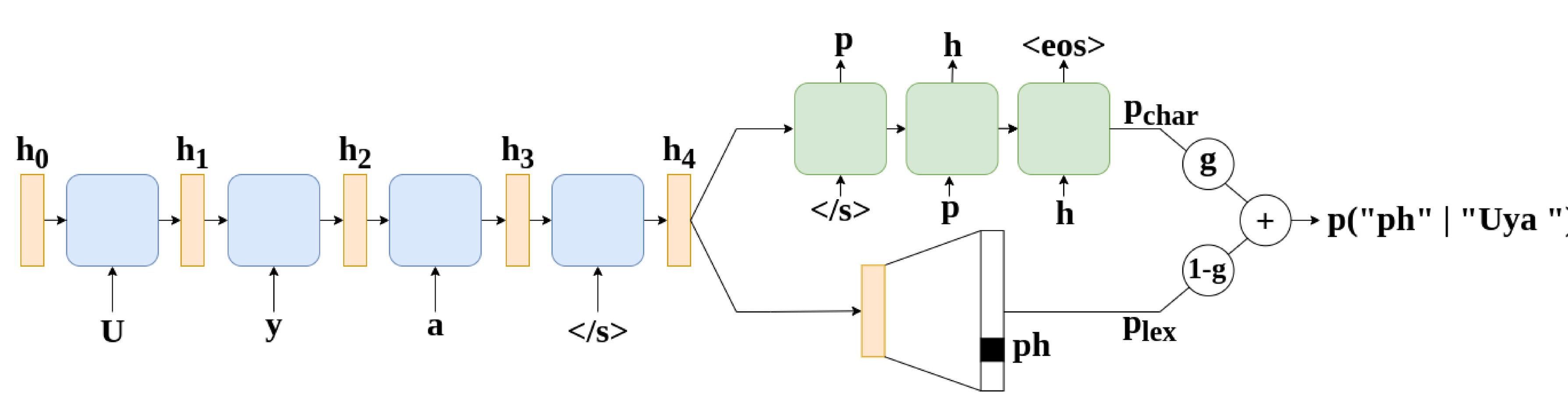
Main Findings

- **Entropy-based** segmentation with LSTM or Transformer LMs already outperform Morfessor as well as subword models such as BPE.
- The **word-level SSLM** outperforms the long-range SSLM (modelling the full context) as a morphological segmenter.
- The long-range SSLM has high recall and low precision, indicating that SSLM **tends to over-segment**.
- A **medium-sized subword lexicon** works best for both language modelling and segmentation.

Contributions

1. A subword segmental language model (SSLM) that **jointly learns subword segmentation** and **autoregressive language modelling**.
2. SSLM is designed specifically for **low-resource** languages that are **morphologically complex**, which includes many African languages.
3. We compile and release **curated LM datasets** for the 4 Nguni languages of South Africa: isiXhosa, isiZulu, isiNdebele, and Siswati.
4. SSLM **outperforms standard segmenters like BPE** across the 4 languages, when evaluated on intrinsic LM performance.
5. On **unsupervised morphological segmentation** SSLM outperforms all baselines across all 4 languages.
6. Analysis shows that the **subword lexicon** is critical to the model's success.

Subword Segmental Language Model (SSLM)



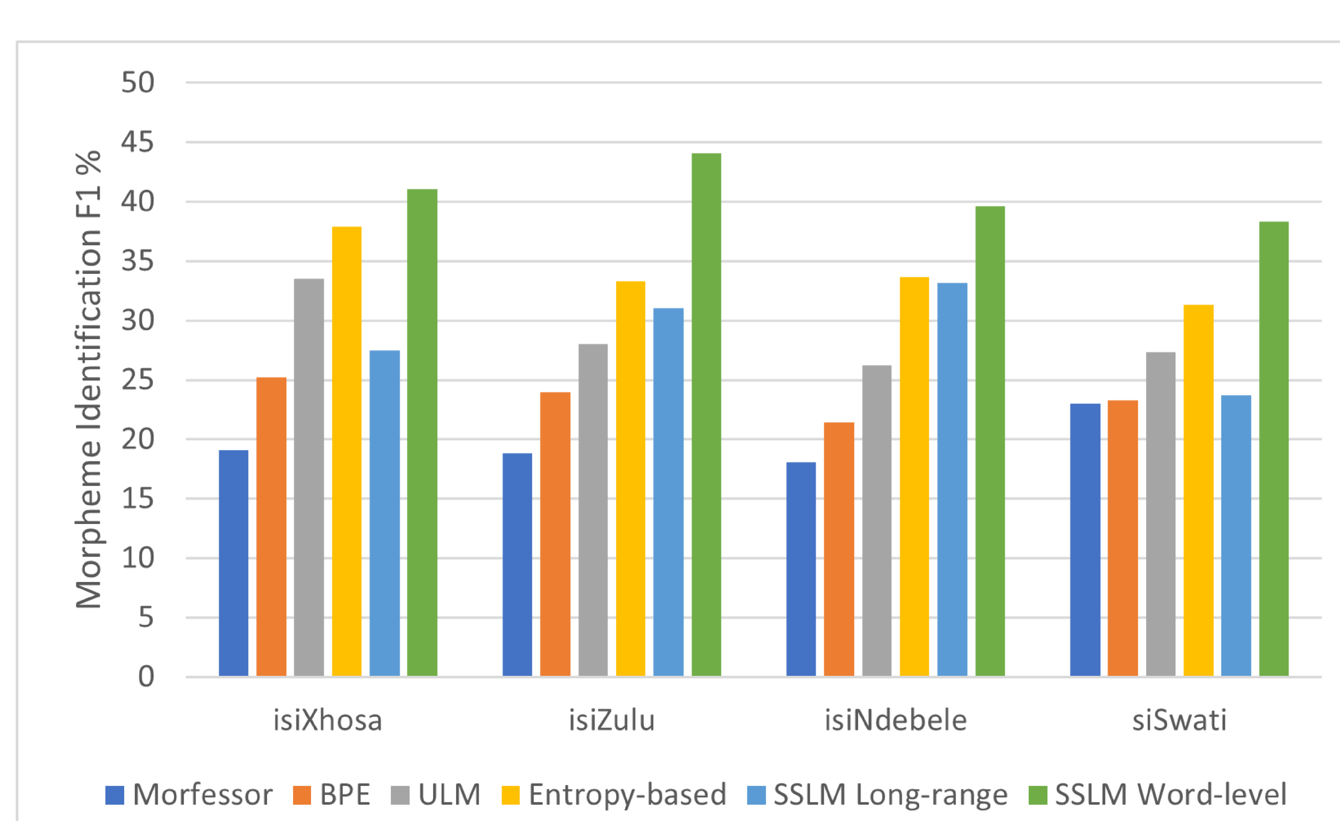
The SSLM generates a sequence of words $\mathbf{w} = \mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_n}$. Each word $\mathbf{w_i}$ is a sequence of subwords $\mathbf{s_i} = s_{i1}, s_{i2}, \ldots, s_{i|\mathbf{s_i}|}$. We marginalise over all possible word segmentations:

$$p(\mathbf{w}) = \sum_{\mathbf{s}:\pi(\mathbf{s})=\mathbf{w}} \prod_{i=1}^{|\mathbf{w}|} \prod_{j=1}^{|\mathbf{s_i}|} p(s_{ij}|\mathbf{s}_{\leq \mathbf{i},<\mathbf{j}}) \quad (1)$$

Each segment probability is mixture of the subword lexicon p_{lex} and a character LSTM p_{char}:

$$p(s_{ij}|\mathbf{s}_{\leq \mathbf{i},<\mathbf{j}}) = g_k p_{\text{char}}(s_{ij}|\mathbf{h_k}) + (1 - g_k) p_{\text{lex}}(s_{ij}|\mathbf{h_k}) \quad (2)$$

Results: Unsupervised Morphological Segmentation



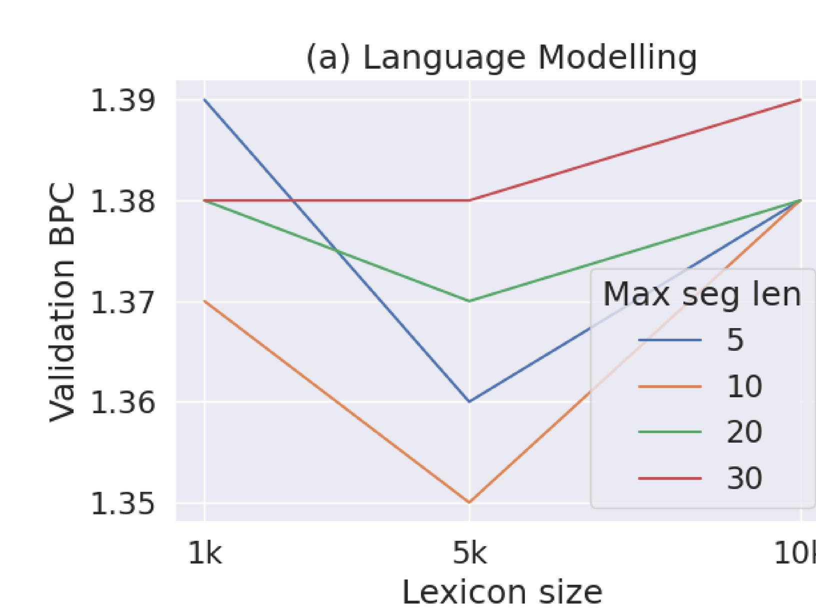
Subword Segmentation Example

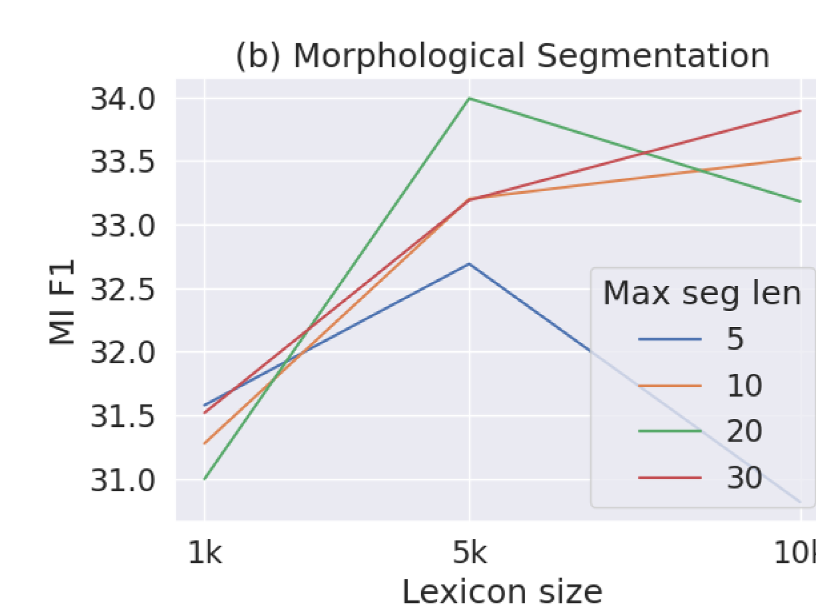
Sentence	Sibuye sithokoze khulu kwamanikelela emphakathini weentjhabatjhaba ngesekelo labo elinganakuzaza emzabalazweni wethu.
Morphemes	Si-buy-e si-thokoz-e khulu k-w-amanik-elel-a e-m-phakath-ini weentjhabatjhaba nge-sekelo eli-nga-nakuzaza e-mzabalazw-eni w-ethu.
SSLM	Si-buy-e s-i-t-h-oko-z-e k-hulu kwam-a-nikele-l-a e-m-phakath-i-n-i ween-tjhaba-tjhab-a n-g-e-sekelo e-l-i-ngana-kuz-az-a e-mz-abal-az-w-e-n-i w-e-thu.
BPE	Si-bu-ye si-tho-ko-ze khulu kwa-m-ani-k-elela em-phaka-thini ween-tjhaba-tjhaba nge-se-k-elo eli-ng-ana-ku-za-za em-za-bala-z-weni we-thu.
ULM	Si-bu-ye si-tho-ko-ze khulu kwama-nikele-la emphakathin-i w-eentjhabatjhab-a nge-se-ke-lo e-lingana-ku-za-za em-za-ba-la-zwe-ni we-thu.

Results: Language Modelling

Bits-per-character (BPC)				
Model	**xh**	**zu**	**nr**	**ss**
Char-LSTM	1.32	1.26	1.39	1.30
BPE-LSTM	1.30	1.22	1.39	**1.28**
ULM-LSTM	**1.25**	1.27	1.39	1.31
Char-Transformer	1.53	1.48	1.47	1.43
BPE-Transformer	1.33	1.27	1.36	1.30
ULM-Transformer	1.34	1.27	1.36	1.29
SSLM	1.27	**1.21**	**1.35**	**1.28**

Analysis: Subword Lexicon





Acknowledgements

This work is based on research supported in part by the National Research Foundation of South Africa (Grant Number: 129850), the South African Centre for High Performance Computing, and the Hasso Plattner Institute for Digital Engineering through the HPI Research School at UCT. Some computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing.

University of Cape Town Natural Language Processing Group

http://www.janmbuys.com/uctnlp

- Authors: MYRFRA008@myuct.ac.za, jbuys@cs.uct.ac.za