Low-resource NLP for African Languages: Initial Explorations

Jan Buys

Department of Computer Science

University of Cape Town

UNIVERSITY OF CAPE TOWN

IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

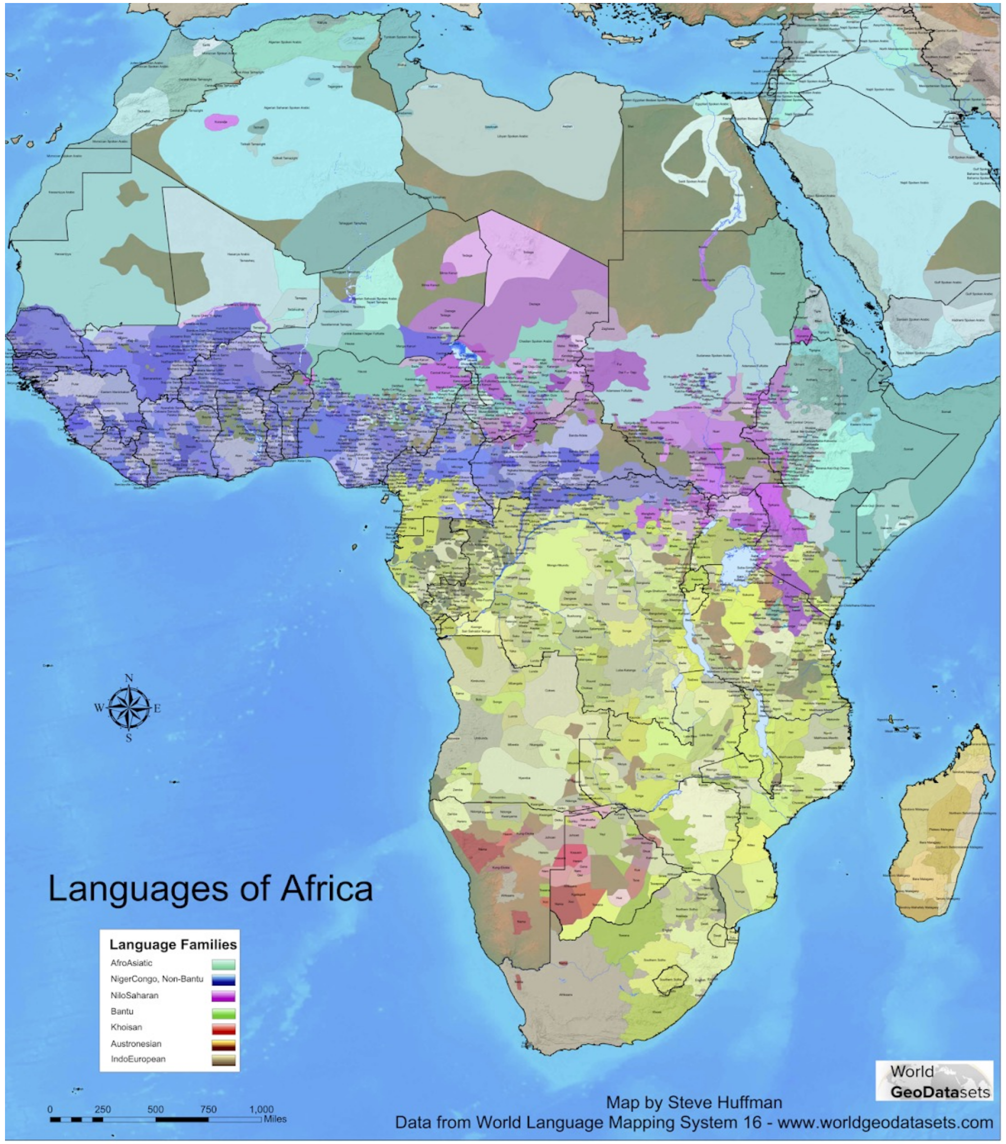
African Languages

Linguistic diversity in Africa:

- >2000 languages (Ethnologue)

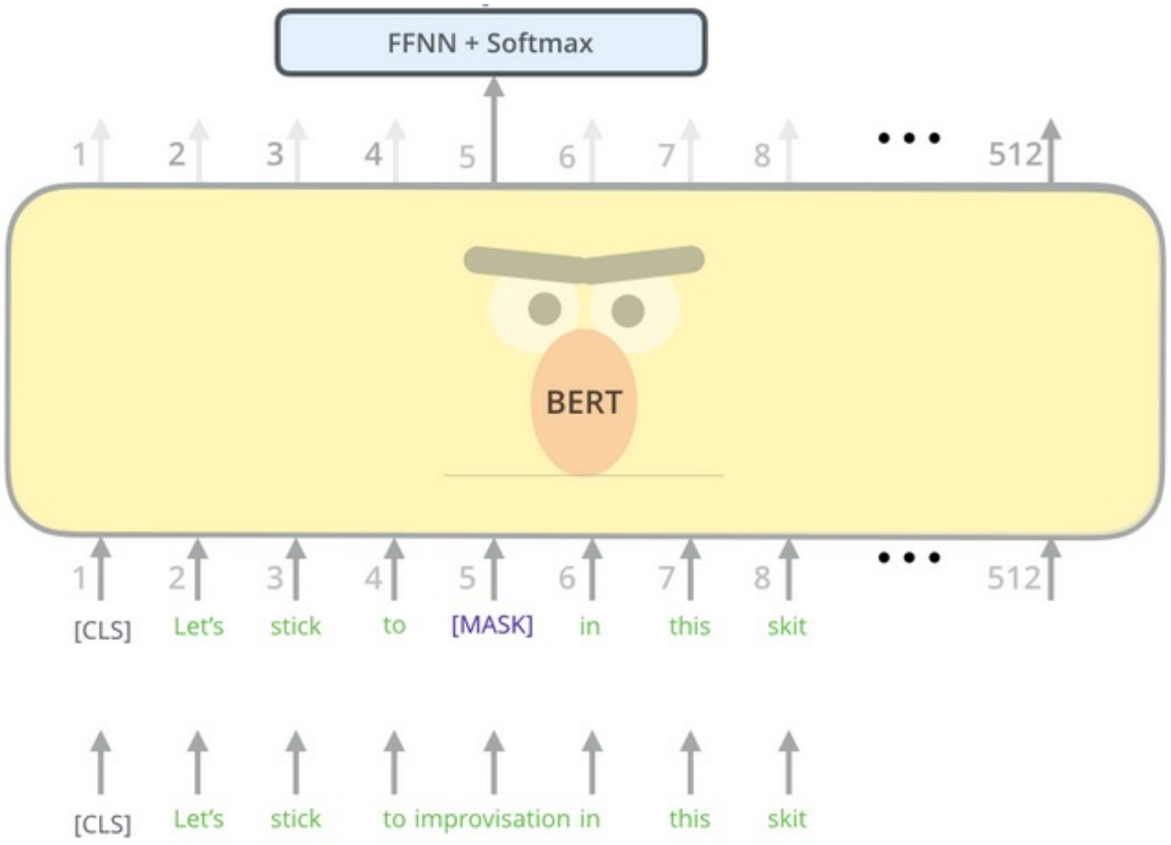
Major language families:

- Atlantic-Congo
- Afroasiatic
- Nilo-Saharan
- Khoisan



Language Modelling

- Large pretrained language models (BERT, GPT-3, etc.) have been very successful for both language understanding and language generation
- Can we transfer some of this success to NLP for African languages?



http://jalammar.github.io/illustrated-bert/

Language Modelling

Challenge 1: Low resource languages

- Size of easily available datasets:

English (C4):	**10.4 TB**	HIGH RESOURCE
isiZulu (C4):	**839 MB**	LOW RESOURCE
isiZulu (NCHLT):	**12 MB**	LOW RESOURCE
Sepedi (NCHLT):	**9.9 MB**	LOW RESOURCE

Language modelling

Challenge 2: Rich morphology

- Words may consist of multiple small meaningful units (morphemes)

Examples (isiZulu):

- wukutholakala

-> wu u ku thol akal a

- negzinkonzo

-> nga i zin konzo

This talk

- Low-resource language modelling for South African languages
- Morphological segmentation for Nguni languages

Low-resource language modelling

- Predict the next word in a word sequence

https://jalammar.github.io/illustrated-word2vec/

Low-resource language modelling

- Language modelling: Assign a probability to a sequence of words, one word at a time

$$P(W_1^n) = \prod_{k=2}^{n} P(w_k | W_1^{k-1})$$

- Example:

$\underbrace{\text{Ubusuku obuhle namaphupho}}_{\text{context}}$ amamnandi

Prediction target

Low-resource language modelling

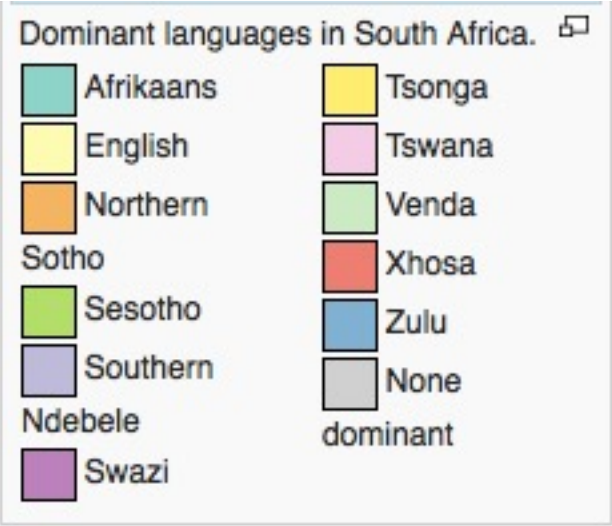
- Example applications of language models
  - Machine Translation
    - P("**high** winds tonight") > P("**large** winds tonight")
  - Spell Correction
    - P("about fifteen **minutes** from") > P("about fifteen **minuets** from")
  - Speech Recognition
    - P("recognize speech") > P("wreck a nice beach")
    - P("I saw a van") >> P("eyes awe of an")
    - P("I ate a cherry") >> P("eye eight uh Jerry")
- In deep learning, vector representations learned by language models are used as the "foundation" of downstream models

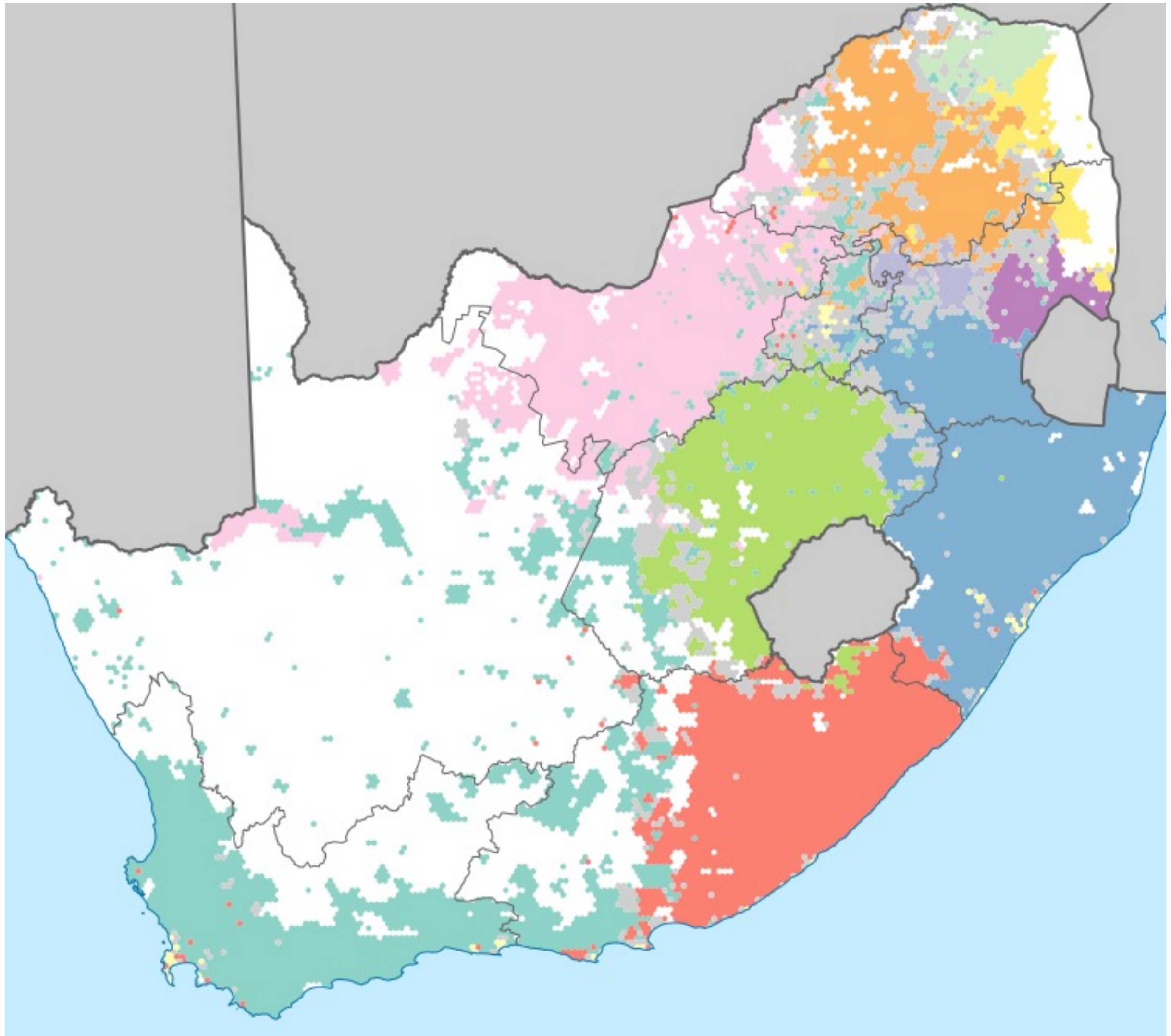
South African Languages

- 11 Official languages

Two largest language groups:

- Nguni languages
- Sotho/Tswana languages





Wikipedia

Low-resource language modelling

- Language modelling for South African Atlantic-Congo languages

Corpus	Training Tokens (000's)	Valid/test Tokens (000's)
NCHLT (isiZulu)	978.6	122.3
Isolezwe (isiZulu)	940.2	117.5
NCHLT (Sepedi)	1357.3	169.7

- Focus on isiZulu and Sepedi, but some experiments using all 9 languages

Low-resource language modelling

Language models:

- n-gram model (modified Knesser-Ney smoothing)
- Feed-forward neural networks
- Recurrent neural networks – LSTMs and QRNNs
- Transformers

Goals:

- Tune and evaluate models systematically to determine which kind of model is most suited for this setup
- Determine if multilingual modelling has advantages

Low-resource language modelling

Open vocabulary language modelling

- The languages are agglutinative, which creates some problems for using the word as fundamental unit
- Split words into subwords using byte pair encoding (BPE)
- Unseen words can then be split in the same way, eliminating the unknown word problem
- The subword vocabulary size is a hyperparameter

Byte-pair encoding

Example on a toy corpus

- The initial vocabulary is the set of characters

corpus		**vocabulary**
5	l o w _	_, d, e, i, l, n, o, r, s, t, w
2	l o w e s t _	
6	n e w e r _	
3	w i d e r _	
2	n e w _	

Byte-pair encoding

- Iteratively merge the most frequent pair of adjacent characters/subwords

Merge	Current Vocabulary
(ne, w)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new
(l, o)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo
(lo, w)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low
(new, er_)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low, newer_
(low, _)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low, newer_, low_

Jurafsky and Martin (2020) Ch. 2

Low-resource language modelling

- Byte-pair encoding for isiZulu and Sepedi:

Ubusuku obuhle namaphupho amamnandi!

➡ **Ubu _suku obu _hle nama _phupho ama _mnandi !**

Robalang gabotse

➡ **R _o _ba _la _ng gabotse**

Low-resource language modelling

n-gram language models

- These models make a Markov assumption: The probability of a word is conditioned only on a fixed number of previous words:

$$P(w_1 w_2 \dots w_n) \approx \prod_i P(w_i \mid w_{i-k} \dots w_{i-1})$$

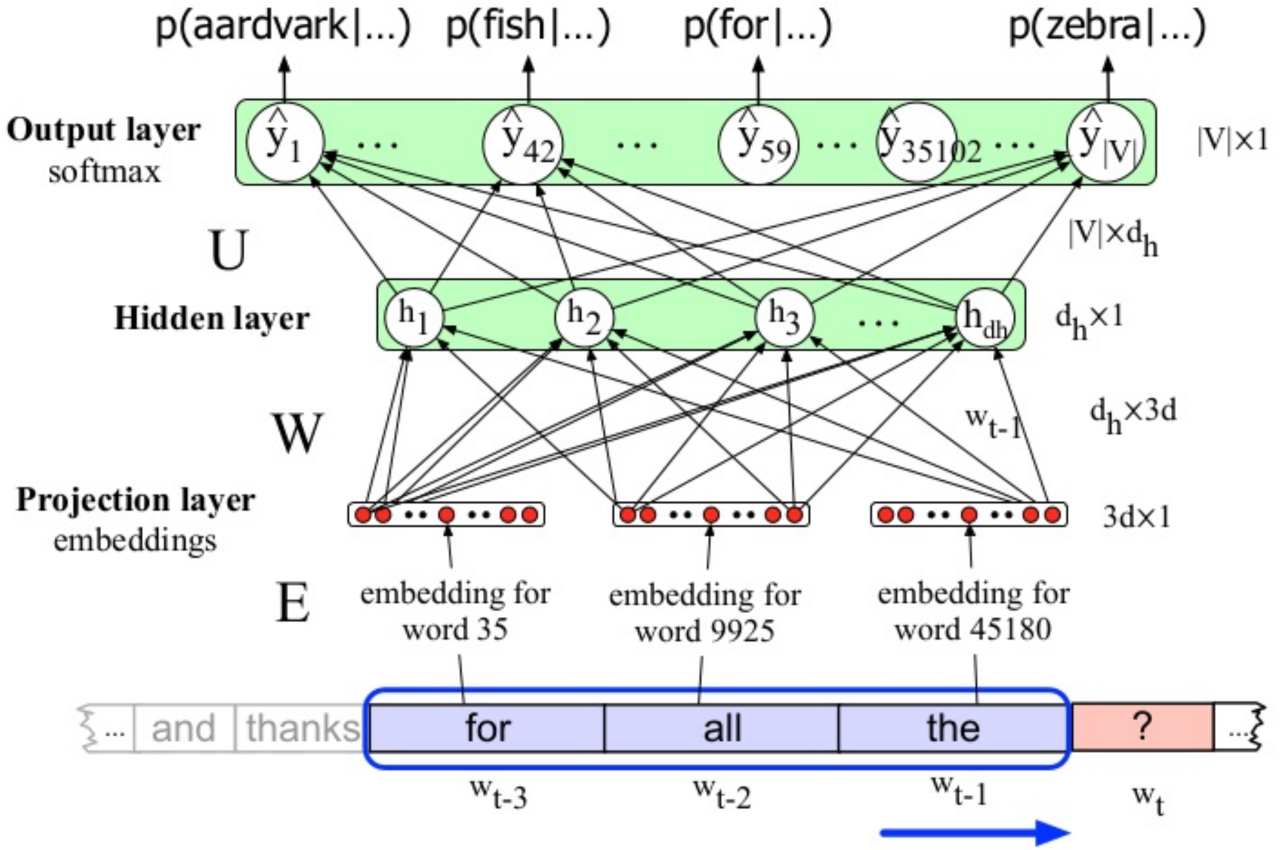
- In other words, each next word probability is approximated as

$$P(w_i \mid w_1 w_2 \dots w_{i-1}) \approx P(w_i \mid w_{i-k} \dots w_{i-1})$$

- Probabilities are estimated by counting *n*-grams and "smoothing" the counts to improve the estimates

Low-resource language modelling

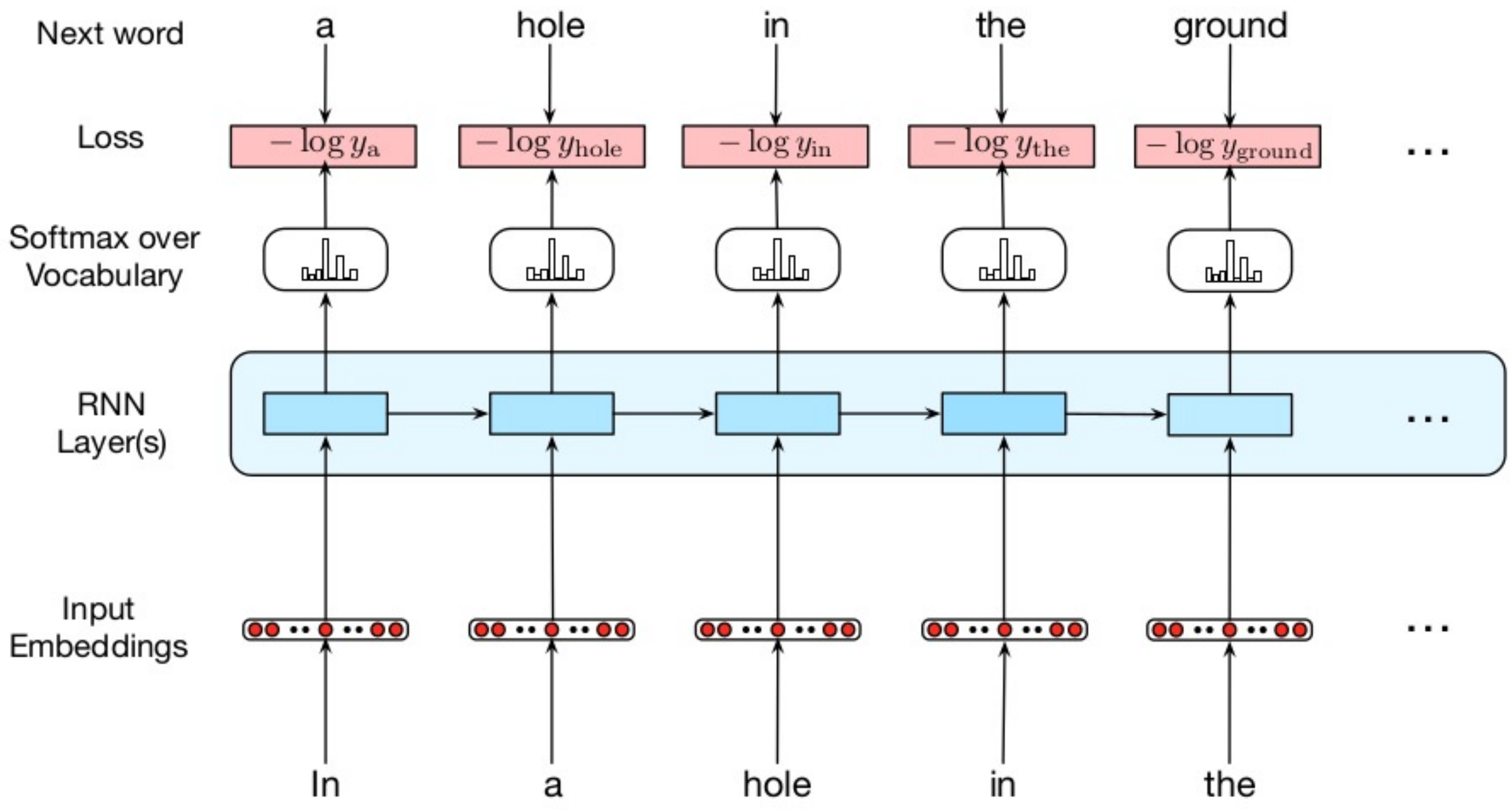
- Feedforward neural network language model:



Jurafsky and Martin (2020) Ch. 9

Low-resource language modelling

- Recurrent neural network (RNN) language model:



Jurafsky and Martin (2020) Ch. 9

Low-resource language modelling

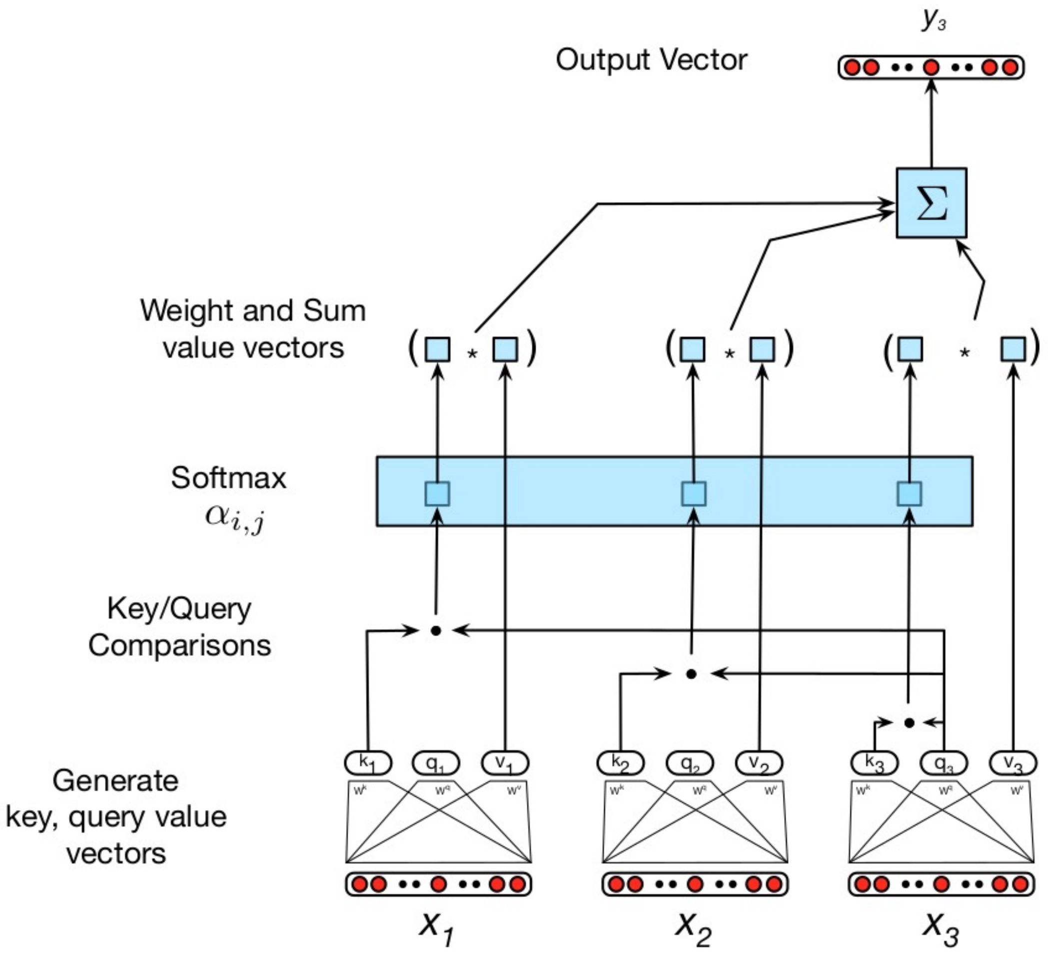
RNN Language models:

- Basic LSTM: Standard model with input/output dropout
- AWD LSTM (Merity et al., 2018):
  - DropConnect for hidden-to-hidden connections
  - Variational dropout over inputs and outputs
  - Word dropout
  - Variable length backpropagation
  - L1 and L2 regularization
- Quasi-RNN (Bradbury et al., 2017):
  - More efficient model
  - Similar regularization and optimization to AWD LSTM

Low-resource language modelling

Transformer language model

- Based on GPT-2 architecture
- Dropout over all parameters
- Model was tuned extensively, but does not use the more sophisticated techniques of AWD LSTM



Jurafsky and Martin (2020) Ch. 9

Low-resource language modelling

Evaluation:

- Intrinsic evaluation of LMs is based on test set entropy

$$H(W_1^n) = -\frac{1}{n}\log_2 P(W_1^n)$$

- Word-based models uses perplexity, but for open-vocabulary models we use bits per character (BPC) – normalize by number of characters *c*

$$\mathrm{BPC}(W_1^n) = \frac{n}{c}H(W_1^n)$$

Low-resource language modelling

- Results: NCHLT (isiZulu)

Model	# Params	Vocab	BPC
n-gram	7.5M	500	1.588
FFNN	4.7M	8000	1.572
Basic LSTM	3.3M	5000	1.548
AWD LSTM	29.8M	5000	1.325
QRNN	29.5M	10000	**1.323**
Transformer	8.6M	8000	1.391

Low-resource language modelling

- Results: Isolezwe (isiZulu)

Model	# Params	Vocab	BPC
n-gram	6.9M	500	1.544
FFNN	5.7M	10000	1.532
Basic LSTM	3.3M	5000	1.677
AWD LSTM	29.8M	5000	**1.259**
QRNN	29.5M	10000	1.264
Transformer	8.6M	8000	1.320

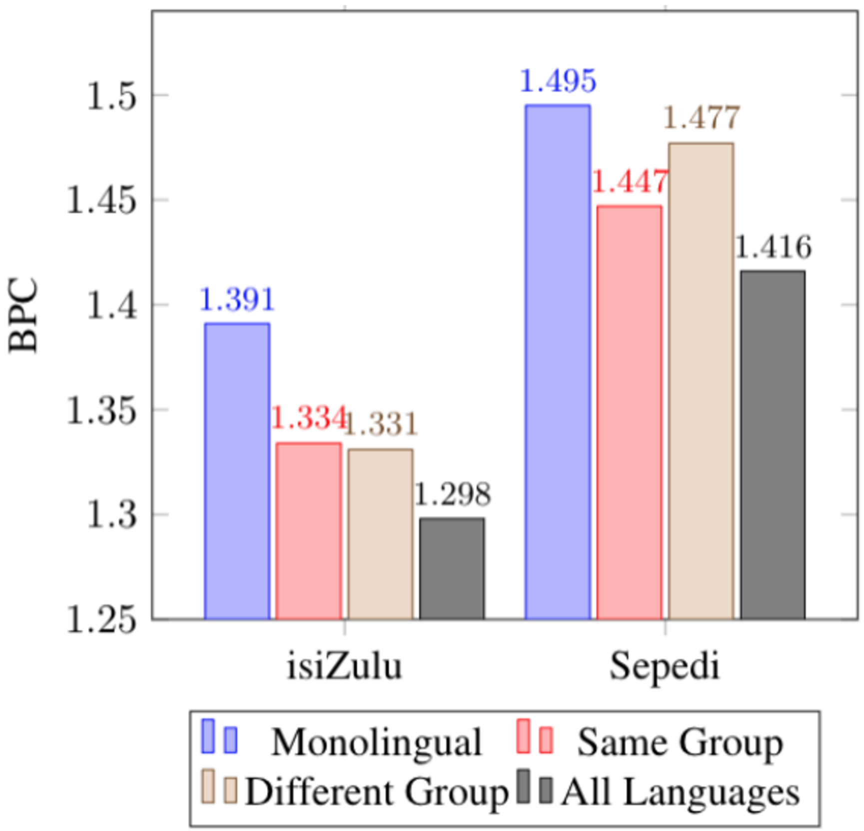
Low-resource language modelling

- Results: NCHLT (Sepedi)

Model	# Params	Vocab	BPC
n-gram	5.7M	2000	1.656
FFNN	5.1M	8000	1.723
Basic LSTM	3.3M	5000	1.625
AWD LSTM	29.8M	5000	1.421
QRNN	29.5M	5000	**1.421**
Transformer	7.1M	2000	1.495

Low-resource language modelling

- Multilingual models: Train on all 9 languages, or on all languages from the same language group (Nguni or Sotho-Tswana)



Low-resource language modelling

Conclusions:

- AWD-LSTM and QRNN outperformed other models with minimal adjustment from the hyperparameter ranges of English models
  - May be due in particular to sophisticated regularization techniques
- Smaller Transformer models come close in BPC
- Relatively similar performance across languages
- Multilingual training improves performance without any architectural changes

Morphological Segmentation

- Task of splitting words into *morphemes*
- Goal: Develop data-driven models for segmentation (previous work on SA languages was rule-based)

- Here we focus on the South African Nguni languages: isiZulu, isiXhosa, isiNdebele, and siSwati
- The Nguni languages are *agglutinative* and written *disjunctively*

Morphological Segmentation

Two types of segmentation:

- Surface segmentation: a word *w* is segmented into a sequence of substrings. The concatenation of the substrings reproduces the original word *w*.
- Canonical segmentation: a word is analysed and segmented into a sequence of canonical morphemes, Each canonical morpheme corresponds to a surface morpheme as its orthographic representation.

Example:

ngezinkonzo

nge-zin-konzo

nga-i-zin-konzo

Morphological Segmentation

Data gives canonical segmentation: process to induce surface form

- Dataset sizes (number of words):

Language	Train	Dev	Test
isiZulu	17 778	1 777	3 298
isiXhosa	16 879	1 688	3 004
isiNdebele	12 929	1 119	2 553
siSwati	13 278	1 080	1 347

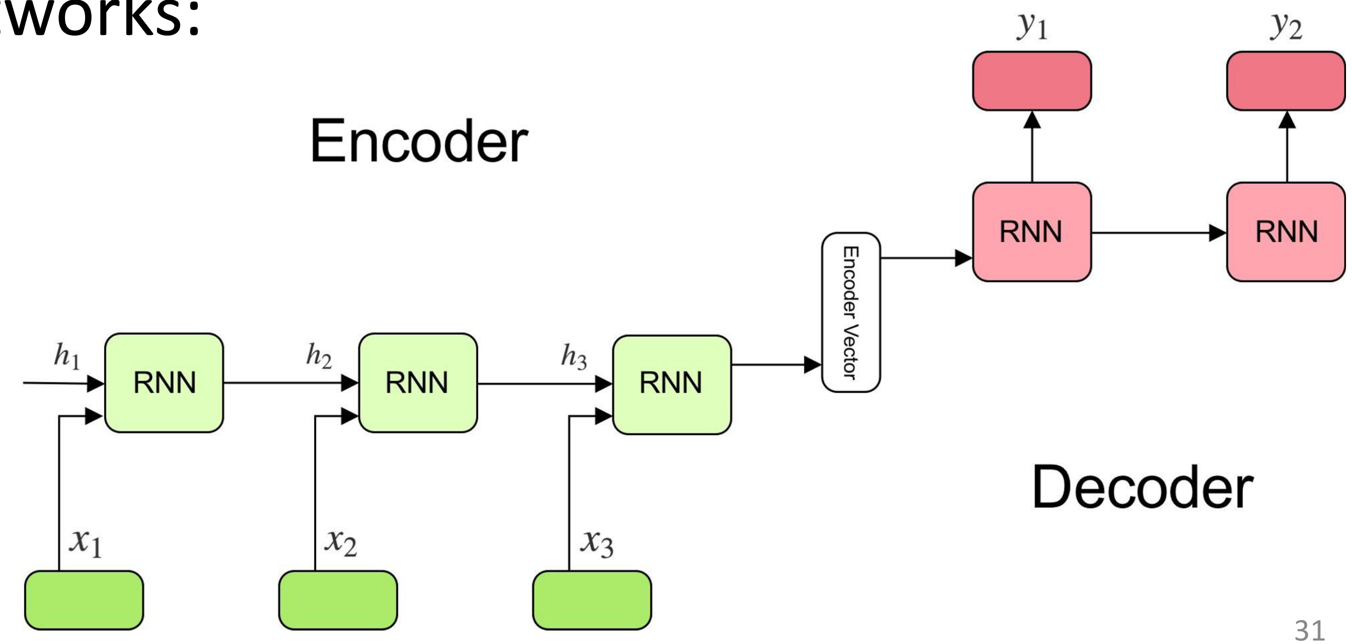
Morphological Segmentation

- Canonical segmentation: Frame as a sequence-to-sequence problem

```
Input:      selayisense
Prediction:   sa-i-li-layisense
```

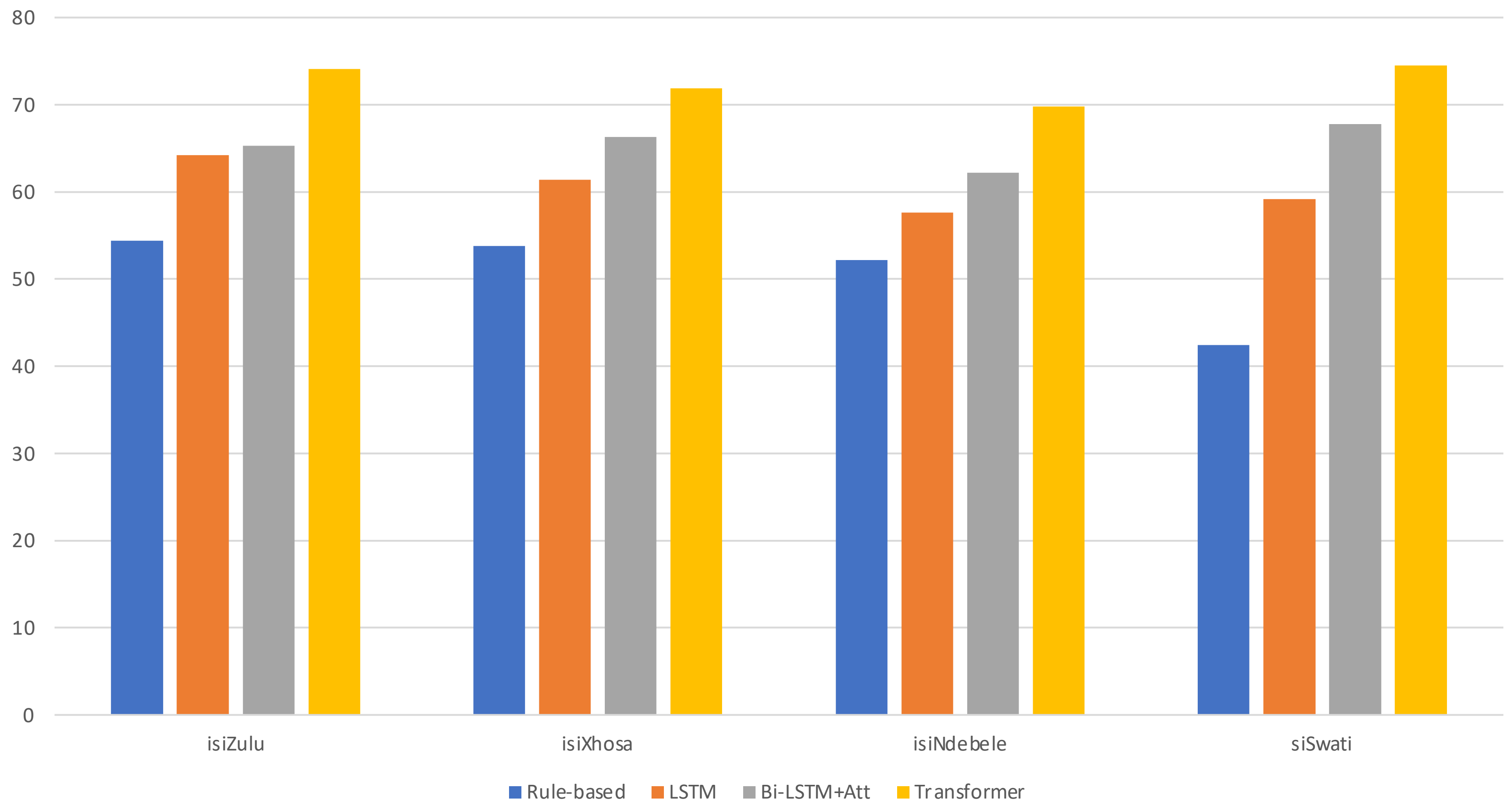
Encoder-decoder neural networks:

- LSTM
- BiLSTM with attention
- Transformers



Morphological Segmentation

- Results: Canonical segmentation (F1 score)



Morphological Segmentation

Analysis

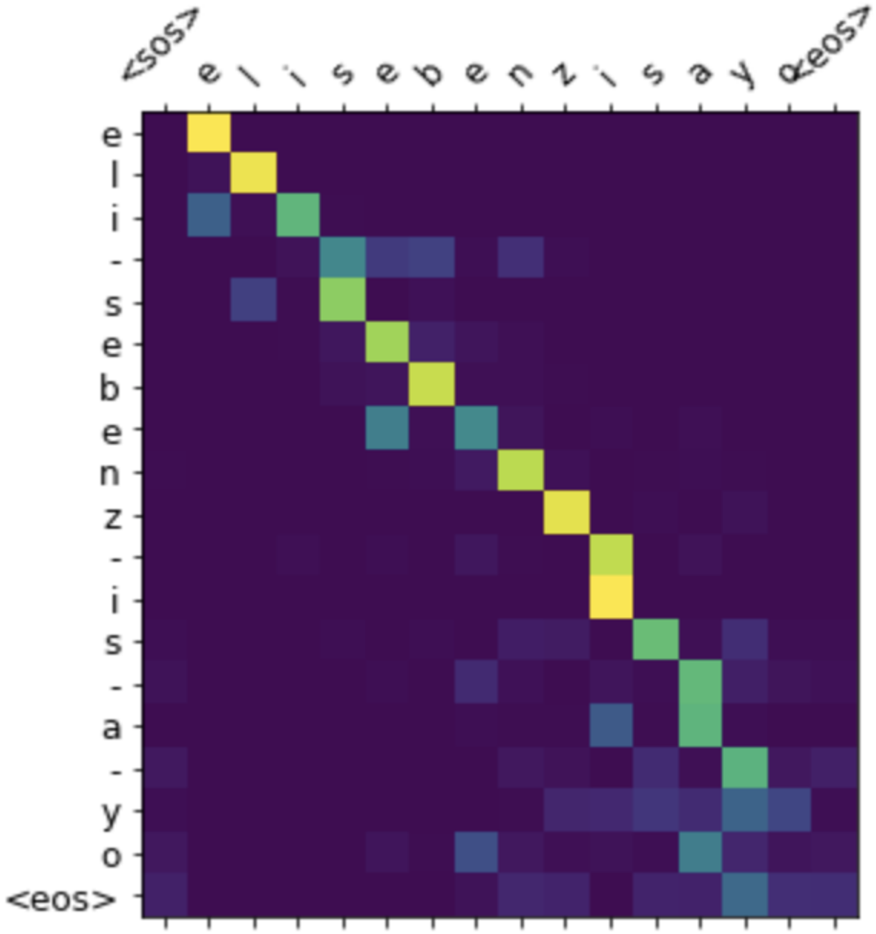
- Sample outputs:

```
Input:   ngaphansi
Prediction:   nga-phansi
```

```
Input:   nemisebenzi
Prediction:   na-i-mi-sebenzi
```

```
Input:   elisebenzisayo
Prediction:   eli-sebenz-is-a-yo
```

Transformer Attention:



Morphological Segmentation

Surface segmentation:

- Frame segmentation as a sequence labelling problem with BIO tagging

B I I B I I B I I I I

n g e z i n k o n z o

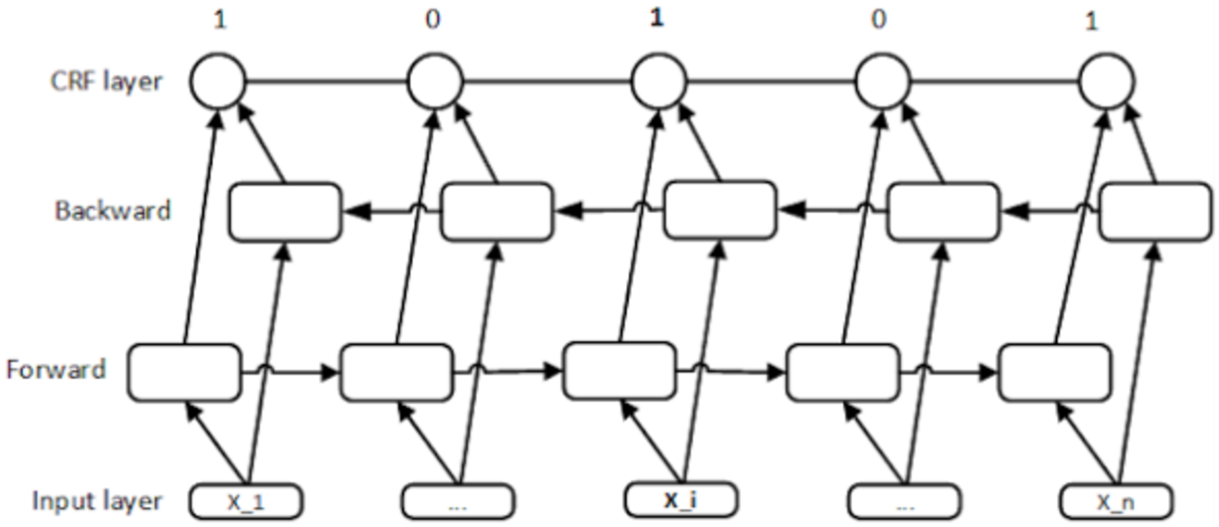
- Use Conditional Random Fields (CRFs) as sequence labelling model

$$S(X, Y) = \sum_{i=0}^{n-1} s(X, y_i, y_{i+1}) \qquad p(Y|X) = \frac{e^{S(X,Y)}}{\sum_{\widetilde{Y} \in Y^{|X|}} e^{S(X,\widetilde{Y})}}$$

Morphological Segmentation

Surface segmentation with Conditional Random Fields (CRFs)

- **Traditional CRFs:** The features are hand-crafted
- **Bi-LSTM-CRFs:** The Bidirectional LSTM Recurrent Neural Network component generates the features



Morphological Segmentation

- Results: Surface segmentation

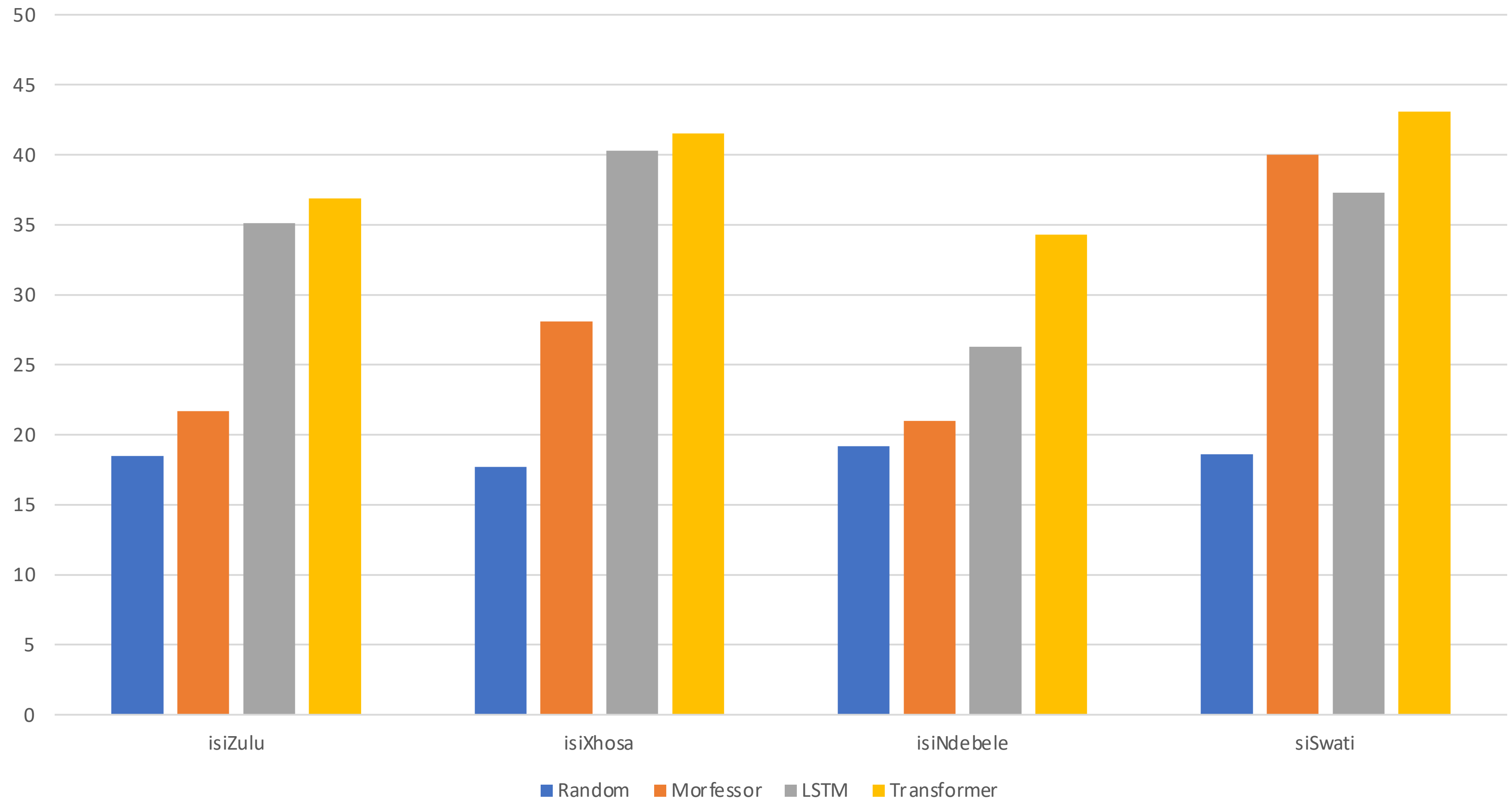
Model	Language	Precision	Recall	F1 Score
Baseline CRF	isiNdebele	97.94	96.62	97.27
	isiXhosa	97.16	97.13	97.14
	isiZulu	97.88	96.82	97.35
	siSwati	97.17	96.40	96.78
Bi-LSTM-CRF	isiNdebele	96.59	96.21	96.40
	isiXhosa	94.88	95.61	95.24
	isiZulu	96.64	96.64	96.64
	siSwati	90.59	91.48	91.03

Morphological Segmentation

- Unsupervised segmentation: How well can segment without any annotated morphological segmentations?
- Entropy-based model:
  - Train a character-based language model
  - Intuition: At the start of a new morpheme the entropy will increase (less predictable), while inside a morpheme the entropy will decrease (more predictable)
  - Different entropy-based segmentation criteria can be formulated
  - Extend Mzamo et al. (2019) to use neural language models instead of n-gram language models
- Train on larger text corpora

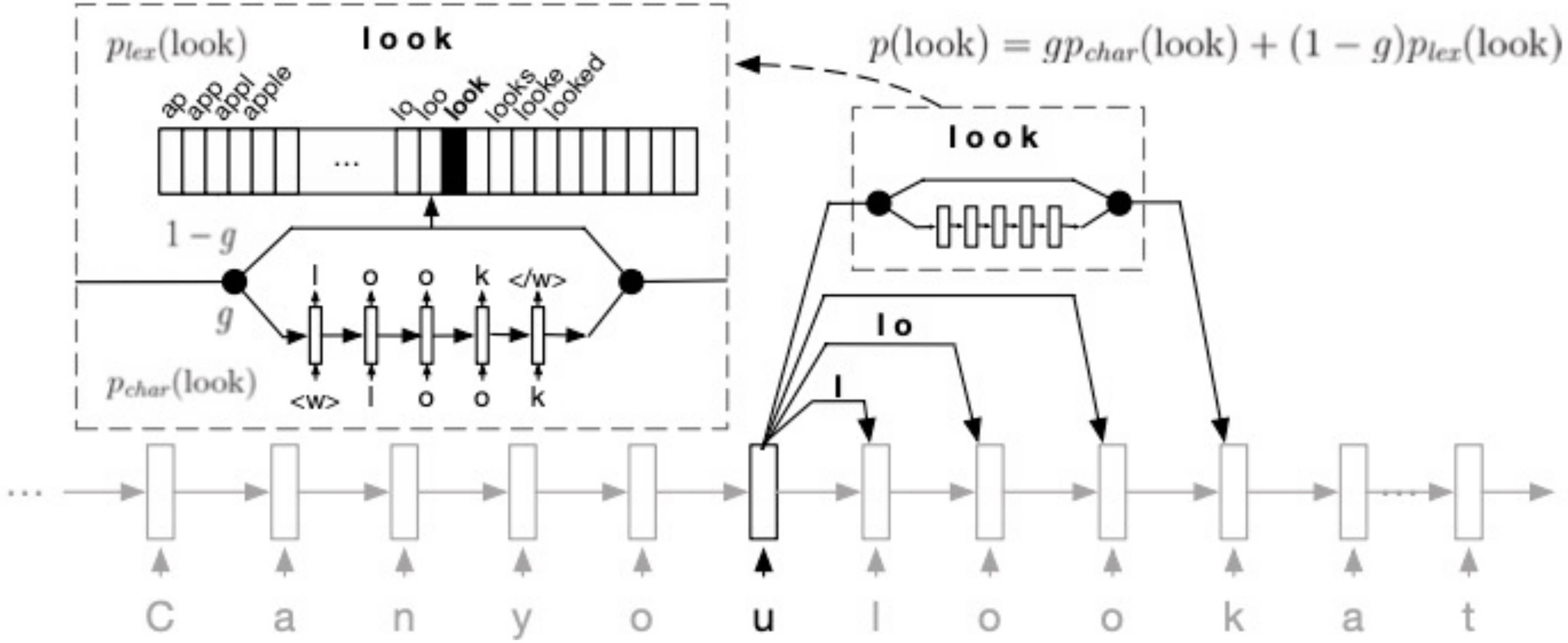
Morphological Segmentation

- Results: Unsupervised Segmentation



Morphological Segmentation

- Work in progress: Joint language model and (unsupervised) morphological segmentation model
- Extend previous work on unsupervised word segmentation (Kawakami et al., 2019)



Ongoing Research

Neural Machine Translation for South African languages

- Projects on both Nguni and Sotho-Tswana languages

Investigating different data augmentation techniques:

- Backtranslation
- Multilingual training (related languages)
- Related language word replacement

As a baseline, also apply the same techniques to phrase-based Statistical Machine Translation

References

- Low-Resource Language Modelling of South African Languages
  Stuart Mesham, Luc Hayward, Jared Shapiro, Jan Buys.
  AfricaNLP workshop at EACL 2021.

- Canonical and Surface Morphological Segmentation for Nguni Languages
  Tumi Moeng, Sheldon Reay, Aaron Daniels, Jan Buys.
  AfricaNLP workshop at EACL 2021.

- Some additional experiments by Francois Meyer

Enkosi kakhulu

- jbuys@cs.uct.ac.za