Language identification *for* ***South African Bantu Languages*** *using* ***Rank Order Statistics***

Meluleki Dube and Hussein Suleman
Digital Libraries Lab, Department of Computer Science, School of IT
University of Cape Town
6 November 2019

Language Identification

- What is the language of a segment of text?
- Significant for:
  - Machine translation
  - Natural language processing
  - Information retrieval
- Essentially a classification problem.
  - given text T => predict language L

South African Languages

Key Issues

- Mixed text.
  - e.g., Text is not written in *slegs een taal* (a single language).
- Low-resource languages have few NLP algorithms and corpora.
  - e.g., Bantu languages
- Short texts.
  - e.g., Tweets, social media posts

Related Work

- Naïve Bayesian Classifier [10]
- Language models [4]
- Support Vector Machines [1]
  - NB and SVM yield 99.4% can accuracy with 100 characters of test data.

Rank Order Statistics

- Proposed by Cavnar and Trenkle [2] as a counting technique instead of a network model.
- Algorithm:
  - Separately count n-grams in training and test data.
  - Sort both lists in order, and discard n-grams after rank M.
  - Similarity = ∑(differences in rank) over n-grams.
    - Where n-gram is only in one list, difference=M

Rank Order Statistics Example

Trigrams for the testing data that is in isiNdebele arranged in the order of their frequencies (highest to lowest)	Trigrams for the training data (model for isiNdebele language) arranged in the order of their frequencies (highest to lowest)	Out of order number for the model and the testing data given by the absolute value of the difference between rank in mode- rank in testing data
nga	nga	|0-0|=0
la	oku	|1-2|=1
oku	la	|2-1|=1
a n	ela	Max
ana	ana	|4-4|=0
enz	nam	|5-6|=1
ela	enz	|6-3|=3
		$\therefore distance = 0 + 1 + 1 + Max + 0 + 1 + 3 = 6 + Max$

Our Goal

- Ignore English and Afrikaans.
  - Over-studied, and potentially biases results.
- Differentiate among other African languages.
  - So we can build an African language digital library with automatic language detection for submissions.
- Test how well this works with small texts and noisy training/test data.
  - Because social media is the new “sliced bread”.

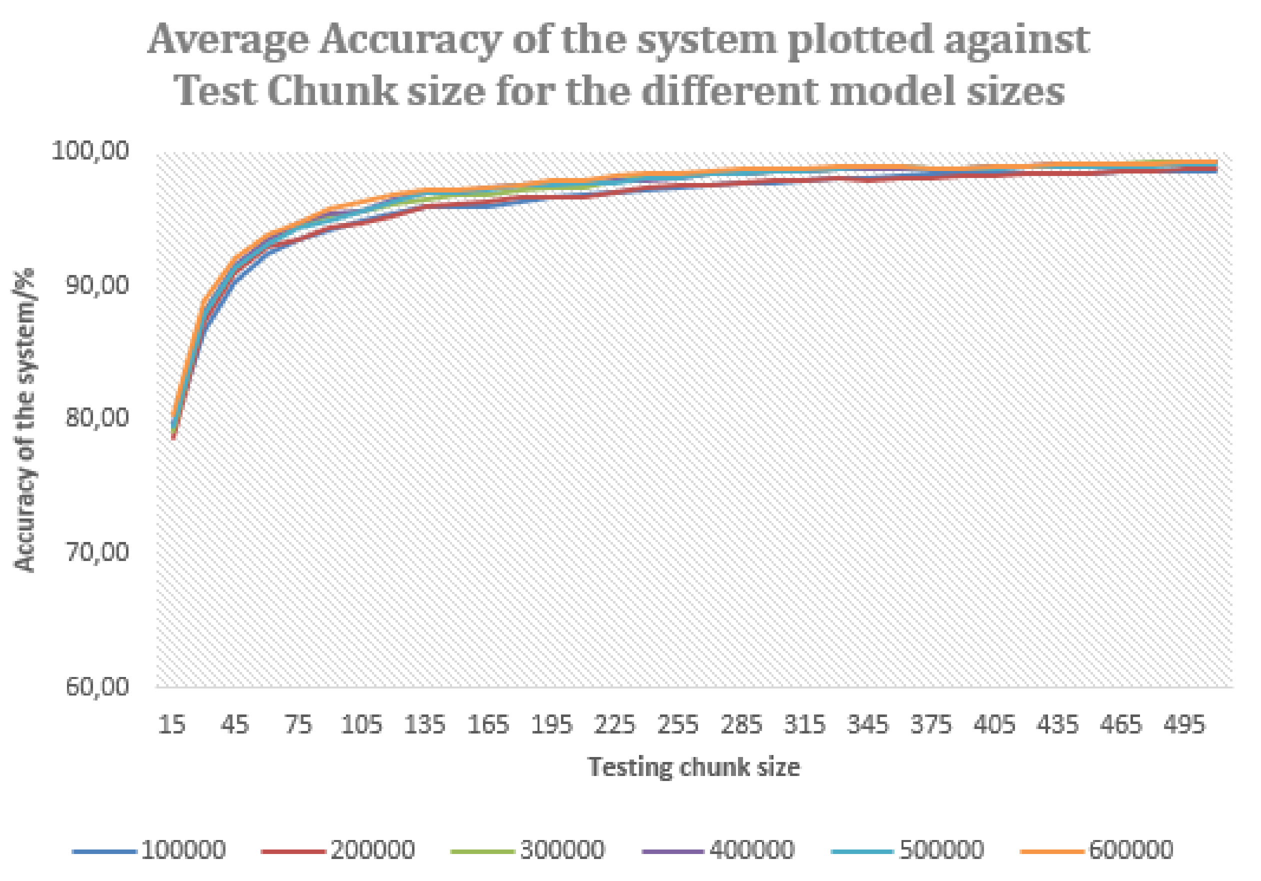
Approach

- Use Rank Order Statistics.
  - Easy to re-train/update/explain.
- Use M=300.
  - Only use the top 300 n-grams.
  - Initial tests showed little benefit in increasing this.
- Obtain test/training data from Sadilar project, which is building an archive of text corpora.

Experiments

- 10-fold cross-validation.
- Training data sizes from 100000-600000 characters, in 100000 increments.
- Test data sizes from 15-495 characters, in 30-character increments.

Results 1/3


Average Accuracy of the system plotted against Test Chunk size for the different model sizes
100,00
90,00
80,00
70,00
60,00
Accuracy of the system/%
15
45
75
105
135
165
195
225
255
285
315
345
375
405
435
465
495
Testing chunk size
100000
200000
300000
400000
500000
600000

Results 2/3

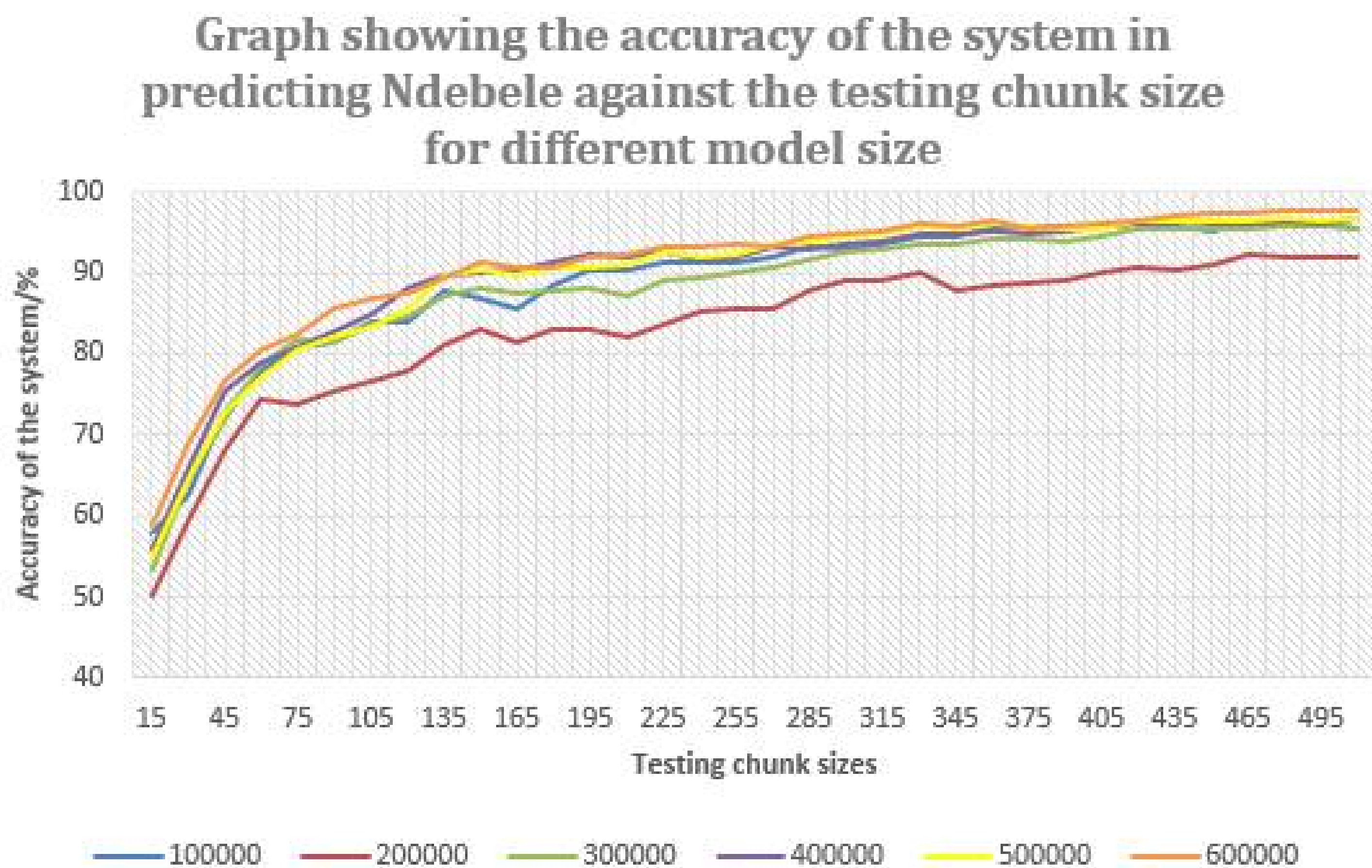
Predicted

Actual

	Ndebele	Pedi	Sotho	Tswana	Swati	Tsonga	Venda	Xhosa	Zulu
Ndebele	519	1	1	3	16	10	5	98	247
Pedi	3	786	22	80	4	2	1	2	0
Sotho	9	7	783	90	5	0	2	2	2
Tswana	0	51	100	737	1	5	2	0	4
Swati	40	3	4	2	788	6	3	27	27
Tsonga	11	2	2	9	8	854	11	0	3
Venda	4	1	2	1	2	10	873	3	4
Xhosa	84	1	5	1	41	6	5	519	238
Zulu	105	1	2	1	63	2	3	156	567

test=100000, training=15

Results 3/3


Graph showing the accuracy of the system in predicting Ndebele against the testing chunk size for different model size
Accuracy of the system/%
100
90
80
70
60
50
40
15
45
75
105
135
165
195
225
255
285
315
345
375
405
435
465
495
Testing chunk sizes
100000
200000
300000
400000
500000
600000

Conclusions

- 99.3% accuracy with 495 characters of test data and 600000 characters of training data.
- 78.72% accuracy with 15 characters of test data and 100000 characters of training data.

- This algorithm works sufficiently well to differentiate among African languages.
  - Even with noise and short texts, with substantial language similarity, and little training data.

that's all folks!