Language identification for South African Bantu languages Using Rank Order Statistics

Meluleki Dube and Hussein Suleman[0000−0002−4196−1444]

University of Cape Town, South Africa
{DBXMEL004@myuct.ac.za,hussein@cs.uct.ac.za}
http://dl.cs.uct.ac.za/

Abstract. Language identification is an important pre-process in many data management and information retrieval and transformation systems. However, Bantu languages are known to be difficult to identify because of lack of data and language similarity. This paper investigates the performance of n-gram counting using rank orders in order to discriminate among the different Bantu languages spoken in South Africa, using varying test and training data sizes. The highest average accuracy obtained was 99.3% with a testing size of 495 characters and training size of 600000 characters. The lowest average accuracy obtained was 78.72% when the testing size was 15 characters and learning size was 200000 characters.

Keywords: N-grams · Bantu languages · Rank Order statistics.

1 Introduction

Language identification is the process of automatically determining the natural language of any electronic text [8]. It is often the precursor to language-dependant processing such as machine translation, search algorithms, text-to-speech, etc. [6]. It is currently a hard task to identify text written in Bantu languages due to the lack of support for these languages on the Web. This, in turn, has led to the Bantu languages being excluded from services like translation services and voice synthesis services and also means that processing of these languages is made harder [3].

The South African Bantu languages used in this paper include: isiZulu, isiNdebele, isiXhosa, siSwati, Setswana, Sesotho, Pedi, Tsonga and Venda; these are all national languages of South Africa, excluding the 2 languages of European origin (English and Afrikaans). The first four languages (isiZulu, isiNdebele, isiXhosa and siSwati) belong to the same family, which is the Nguni languages [1] that share much vocabulary. The goal of this paper is to explore the degree to which texts in these Bantu languages can be differentiated using algorithmic language identification.

There are currently different techniques in use to identify languages. These include: Naïve Bayesian, normalized dot-product, centroid-based, and relative entropy and n-gram counting using rank order statistics [1]. In this paper, rank order statistics, as described by Cavnar and Trenkle [2], are applied to South

African Bantu languages and evaluated. Three key aspects are investigated: the effect of training data size; the effect of test data size; and the impact of language similarity.

2 Literature Review

2.1 N-gram Count Using Rank Orders

An n-gram is a sequence of n consecutive characters from some text, with the value of n often being 1, 2, or 3 [9]. N-grams are usually called unigrams when n=1, bigrams when n=2 and trigrams when n=3 e.g. trigrams for the Ndebele word, *'bhala'* (translation: *'write'*) are *bha, hal, ala*.

Cavnar and Trenkle [2] devised an approach for text categorisation that is tolerant of textual errors, achieving 99.8% accuracy (correct classification rate) in one of their tests. Their system also achieved 80% accuracy in classifying articles from a different computer-oriented newsgroup. This approach is adopted in this research to investigate whether it is possible to achieve high accuracy ratings for classifying Bantu languages. The system entails calculation and comparisons of profiles of n-gram frequencies, where a profile of n-gram frequencies refers to the n-gram frequencies ranked in descending order [2]. The reason this is possible is due to Zipf's principle of least effort, which states that the frequency of occurrence of a word is almost equal to the inverse of its rank [7]. This then suggests that 2 separate texts in a particular language, say Ndebele, will have almost the same n-gram distribution. Thereby if we can compare 2 profiles and determine that they are not far apart by some measure, there is then a high probability that the 2 profiles belong to the same classification group. The exact method of computing the distances is described within the methodology section in this study.

2.2 Related Work on Bantu Languages

Zulu, Botha, and Barnard [11] investigated if it is possible to quantify the extent to which the South African languages differ, given that the 9 different languages cluster into families. For example, isiNdebele, isiXhosa, siSwati and isiZulu are all in the Nguni sub-family. Measurements for similarity were made and, for example, isiNdebele and isiZulu had a distance of 232 between them, isiNdebele and isiXhosa had a distance of 279 between them and isiNdebele and siSwati had a distance of 257 between them. In contrast, English and isiNdebele had a distance of 437, while English and isiZulu had a distance of 444. It can be seen that there is a strong affinity within the group and a marked difference from English, which is not Nguni.

Combrinck and Botha [4] presented a statistical approach to text-based automatic language identification, which focused on discrimination of language models as opposed to the representation used. They used a subset of South African Bantu languages (isiZulu, isiXhosa, Tswana, Sepedi), as well as Swazi.

Botha and Barnard [1] looked at the factors that affected language identification with a focus on all South African languages, including English and Afrikaans. They included multiple algorithms, but concluded that SVM and Naïve Bayes performed the best, with an accuracy of up to 99.4% for a test data size of 100 characters. In contrast, this paper considers only the Bantu languages and investigates a rank-order algorithm that is arguably simpler and faster to update/train.

3 Methodology

The methodology employed in this paper is adapted from the work by Cavnar and Trenkle [2]. Given that their system achieved a maximum accuracy of 99.8% for classifying languages, the goal is to find out if a combination of parameters can result in a comparable accuracy.

3.1 Corpora

Text corpora for the 9 South African Bantu languages were acquired from the South African Centre for Digital Language Resources. The files were first manually cleaned to remove metadata.

A maximum training size of 600000 was decided on to test how well these algorithms work for languages with lesser amounts of data. The corpus for each language was sampled to create these subsets, by extracting 2000-character segments at periodic intervals from within the full corpora to ensure maximum variability in the data.

3.2 Language detection algorithm

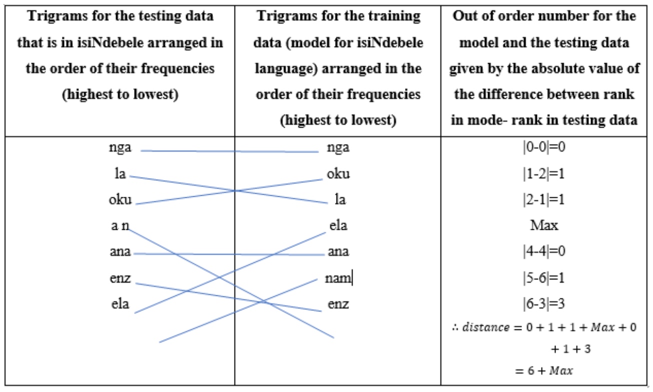
Given a set of corpora for different languages, a model is created for each language in the training dataset. Each such model is a vector of the most highly-occurring trigrams, sorted in order of frequency, for the language. In order to detect the language of some unknown test data, a similar list of frequent trigrams is calculated. The test data trigram vector is then compared against each of the language models and the most similar is chosen as the language of the test data.

This calculation for distance/similarity between 2 language models is presented in Figure 1 as an example. The ranks of each pair of corresponding n-grams is subtracted, and the sum of these differences is then the similarity metric. There is a maximum of 300 n-grams in each model and this maximum value is used where one list contains an n-gram but the other does not.

3.3 Experimental Design

In evaluating the system, 10-fold cross validation was employed.

One language dataset was divided into 10 sections/folds. 9 folds were used as training data and 1 fold was held back and used for testing. Testing was performed by dividing the test fold into small chunks and determining the language



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
		$\therefore distance = 0 + 1 + 1 + Max + 0 + 1 + 3 = 6 + Max$

Fig. 1. Example of similarity metric calculation

of each chunk. This was repeated 10 times, using different testing folds each time. During this process, the other language models were held constant based on the complete dataset for each. The final predicted language accuracy is the average over all the languages predicted in each of the 10 tests. This process was then repeated for each language.

The major parameters in this experiment were the different training data sizes and and test chunk sizes. In this experiment, the test chunk size was varied from 15 characters to 510 characters and the training data size was varied from 100000 characters to 600000 characters.

4 Results and Analysis

The maximum achieved accuracy was 99.3%. This was when the language model size was 600000 characters and the testing chunk size was 450 characters. The minimum accuracy achieved, on the other hand, was 78.72%, which was obtained when the training data size was 200000 and the testing chunk size was 15 characters.

The results, to a certain extent, adhere to the hypothesis. Figure 2 shows that both increasing the testing data size and the data size for creating the models will effectively increase the accuracy of the system. The increase in accuracy due to increase in training data size is seen in the shift to the right of the different graphs. The system never reaches a 100% accuracy and the 100% accuracy line acts as an asymptote of the graph. However, after a certain point in time, increasing the testing data size becomes less useful, as there is no increase in the overall accuracy of the system. Therefore, the combination that yields the optimum results will be taken as the combination that yields the maximum accuracy.

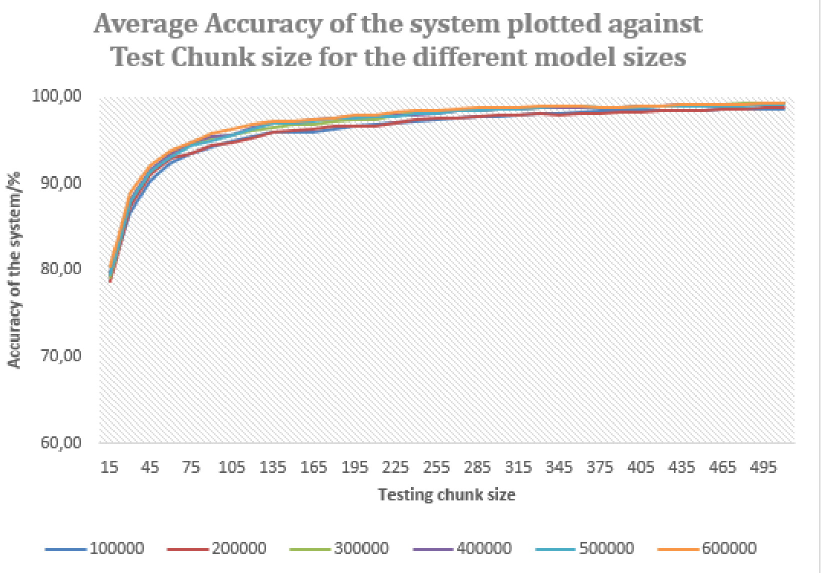


Fig. 2. Average accuracy of the system plotted against the testing chunk size for the different training data sizes

Consider the matrix in Figure 3, which shows the number of times the system mistakenly recalled one language as the other, for when the model size was 100000 and testing chunk size was 15.

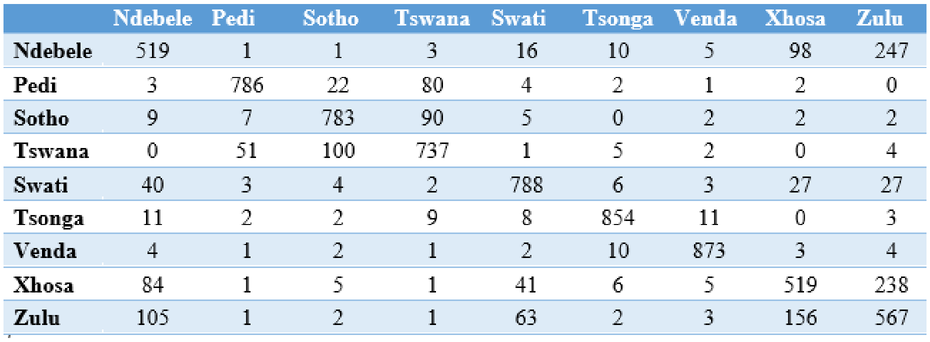
As seen from Figure 3, the reason for the low average accuracy for the system is the failure to properly identify some of the Nguni languages, in particular isiNdebele. The system managed to correctly identify isiNdebele 57.67% of the time within the 900 trials. Other languages achieved 81.89% and higher accuracies for the same testing chunk size and model size. This behaviour also persisted for larger training data sizes.

When using training data in isiNdebele, this effect is more clearly seen. As already stated, the low accuracies are mostly due to failure to correctly classify the Nguni languages, which share considerable vocabulary so are difficult to tell apart.

This effect was also reported by Zulu, Botha and Barnard [11] when using smaller model sizes and minimal test chunk sizes.

5 Conclusions and Future Works

In this paper, we set out to investigate if the method of counting n-grams using rank orders can be used to determine the language of text in a Bantu language. The similarity metric proposed by Cavnar and Trenkle [2] was adopted.



	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

Fig. 3. Predicted languages when different languages were used as testing model. The heading/first row indicates the predicted languages and the heading/first column indicates the languages used for testing. Testing data size used was 15 characters and training data size used was 100000 characters.

A key goal was to determine combinations of testing chunk size and training data size that would give the optimum accuracy values. A 2-factor experiment was conducted, with n-fold cross-validation to determine the average system accuracy and the accuracy within each parameter combination. The results indicate that it is indeed possible to use rank orders to perform language identification among the South African Bantu languages, provided that the model sizes are sufficiently large to ensure a relatively high accuracy across all languages.

As expected, the accuracy for languages within a family of related languages was lower (e.g., in the case of isiNdebele) when smaller models were used for training and smaller test chunks were used. This is not necessarily problematic, as similar languages will share morphological analysis, translation and other tools, so the exact language may not impact on the eventual use case.

These results confirm prior results on different subsets of languages and using different techniques. But, in particular, they confirm that the rank order technique used can be applied to a regionally-defined set of related and unrelated low resource languages, as can be found in many parts of the world. This work was initially motivated by the creation of low resource language archives and multilingual search across low-resource languages. The results are promising in that they show even very short social media posts in low resource languages can be identified using these techniques.

Acknowledgements

This research was partially funded by the National Research Foundation of South Africa (Grant numbers: 85470 and 105862) and University of Cape Town. The authors acknowledge that opinions, findings and conclusions or recommendations expressed in this publication are that of the authors, and that the NRF accepts no liability whatsoever in this regard.

References

1. Botha, G.R., Barnard, E.: Factors that affect the accuracy of text-based language identification. Computer Speech & Language **26**(5), 307–320 (2012)
2. Cavnar, W.B., Trenkle, J.M., et al.: N-gram-based text categorization. In: Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval. vol. 161175. Citeseer (1994)
3. Chavula, C., Suleman, H.: Assessing the impact of vocabulary similarity on multilingual information retrieval for bantu languages. In: Proceedings of the 8th annual meeting of the Forum on Information Retrieval Evaluation. pp. 16–23. ACM (2016)
4. Combrinck, H.P., Botha, E.: Text-based automatic language identification. In: Proceedings of the 6th Annual Symposium of the Pattern Recognition Association of South Africa (1995)
5. Dunning, T.: Statistical identification of language. Computing Research Laboratory, New Mexico State University Las Cruces, NM, USA (1994)
6. Duvenhage, B., Ntini, M., Ramonyai, P.: Improved text language identification for the south african languages. In: 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech). pp. 214–218. IEEE (2017)
7. Li, W.: Random texts exhibit zipf's-law-like word frequency distribution. IEEE Transactions on information theory **38**(6), 1842–1845 (1992)
8. McNamee, P.: Language identification: a solved problem suitable for undergraduate instruction. Journal of Computing Sciences in Colleges **20**(3), 94–101 (2005)
9. Ndaba, B., Suleman, H., Keet, C.M., Khumalo, L.: The effects of a corpus on isizulu spellcheckers based on n-grams. In: 2016 IST-Africa Week Conference. pp. 1–10. IEEE (2016)
10. Poole, D., Mackworth, A.: Artificial intelligence foundations of computational agents. 2010 (2017)
11. Zulu, P., Botha, G., Barnard, E.: Orthographic measures of language distances between the official south african languages. Literator: Journal of literary criticism, comparative linguistics and literary studies **29**(1), 185–204 (2008)