Quantifying the effect of corpus size on the quality of automatic diacritization of Yorùbá texts

Tunde Adegbola
African Languages Technology Initiative
Ibadan Nigeria
taintransit@hotmail.com

Lydia Uchechukwu Odilinye
African Languages Technology Initiative
Ibadan Nigeria
oluceee@yahoo.com

Abstract

Yorùbá being a tone language requires tone information for the correct pronunciation of words in Text-to-Speech synthesis. Based on standard Yorùbá orthography, such information is held in tone marks, which applied to vowels and syllabic nasals as diacritical markings. However, the tone marks are not always correctly applied in many Yorùbá documents because appropriate input devices for the accurate application of the diacritic marks are not always available. Hence, the absence of tone marks in most written Yorùbá texts presents a major challenge in speech synthesis as the information required for applying the right tone sequences to synthesized Yorùbá speech may not always be available. This study proposes the use of Machine Learning techniques as a basis for the automatic application of tone marks as part of the pre-processing in high level synthesis. Being a resource-scarce language however, there is a lack of sufficiently large Yorùbá corpora for the training of an automatic diacritizer. The study therefore investigated the relationship between corpus size and the quality of automatic diacritization towards estimating the size of corpus required for an ideal level of accuracy.

1 Introduction

Yorùbá is the language of about 30 to 50 million speakers in Nigeria, Benin Republic and the Republic of Togo (Everyculture.com 2009). It is a tone language that makes heavy use of lexical tones which are indicated in orthography by the use of diacritics as tone marks. There are three tones in Yorùbá; namely LOW, MID and HIGH tones. Each vowel and syllabic nasal in the Yorùbá language carries one of these three tones and each of the syllables centered on these vowels and syllabic nasals have the capacity to change the meaning of the resulting word by virtue of the tones they carry.

For example the Yorùbá words; *ọwọ́*, *ọ̀wọ̀* and *ọ̀wọ́* are minimal pairs as *ọwọ́* means hand, *ọ̀wọ̀* means honour and *ọ̀wọ́* means group. The difference in the meanings of these words is caused only by the difference in the tones carried by the vowels. Since these tones are indicated in writing by the application of diacritical marks a misapplication of such diacritical marks would alter the meaning of the affected words. In addition to use of diacritics to indicate tones, Yorùbá orthography also uses dots under *s*, *e* and *o*, thus obtaining *ṣ*, *ọ* and *ẹ* as character symbols used for certain speech sounds in the sound inventory of the language. This implies a rather heavy use of diacritical marks in the Yorùbá orthography.

The heavy use of tones marks and other diacritical marks in Yorùbá orthography as well as the lack of appropriate input devices for their easy application in written Yorùbá presents a problem in the text-to-speech (TTS) synthesis of documents written in Yorùbá.

Developments in artificial intelligence now make it possible to use machine learning techniques to infer the appropriate tones to be applied to a given word based on the context in which such words are used in a given document. Such automatic diacritization could be incorporated as a standard pre-processing stage within the high-level synthesis for Yorùbá.

2 Automatic Diacritization

Various studies on automatic diacritization have been carried out for various languages whose orthographies use diacritics. Some of these studies approach diacritization at the character/phone level while others approach it from a word-based point of view. For example, Simard (1998) worked on French texts; Mihalcea (2002) under-

took a comparison of Czech, Hungarian, Polish and Romanian texts while Mohamed & Kubler (2008) worked on Arabic texts. All the above investigators approached their work from the character/phone point of view. On the hand, Elshafei et al. (2006) worked on Arabic using the word-based approach. Cocks and Keegan (2011) however worked on Maori comparing the character/phone approach to the word-based approach and came out with the conclusion that the word-based approach is better that the character-based approach for the Maori language. Mahalcea (2002) in comparing Czech, Hungarian, Polish and Romanian texts concluded that the character-based approach worked better for these languages.

DePauw et al (2007) as well as Scannell (2010) reported works on the diacritization of Yorùbá and a number of other African languages. Both investigators reflected on the lack of adequately sized corpus as a weakness in the automatic diacritization of Yorùbá texts.

Unfortunately, the unavailability of sufficiently sized corpus reported by these investigators still subsists as there haven't been much coordinated efforts to build sufficiently large corpora for use in Human Language Technologies (HLT) for the Yorùbá language. Hence, Yorùbá remains a resource-scarce language and the training of a machine learning system to be used for automatic diacritization is still problematic.

The objective of this study is to quantify the effect of corpus size on the quality of automatic diacritization and thereby determine the size of corpus needed for a certain level of accuracy as well as the level of accuracy expected from a given corpus size.

To achieve this objective, a Naïve Bayes diacritizer based on word trigram probabilities and using linear interpolation for smoothing was tested.

3 Theory

The evaluation of the accuracy of a machine learning system is normally conducted by the use of data that was not part of the training set. This is usually referred to as outside testing. In contrast, debugging and diagnostics of machine learning systems are usually carried out by inside testing. An inside test is one in which the test data is taken from within the set of training data while an outside test is one in which the test data is independent of the training data. In addition to its normal use in debugging and diagnostics, an inside test can also give an idea of the best possible behavior of an automatic diacritizer even though its accuracy results would be highly exaggerated.

According to Chomsky (1957), human language consists of a potentially infinite set of sentences. Since it is impossible to gather a training corpus of infinite size, the standard scientific approach to training for machine learning is to use a corpus of a sufficiently large size, which can be assumed to adequately represent the potentially infinite number of sentences in the language under examination. Hence, to evaluate the accuracy of an automatic diacritizer for a particular language, such an evaluation can only be based on a sufficiently large but finite subset of the potentially infinite set of sentences in the language.

As the size of the training corpus for a machine learning system increases, an outside test and an inside test will tend towards giving the same results. This is because an outside test, based on an infinitely large training corpus can be viewed as an inside test because both the training and test files would have been obtained from the infinitely large training corpus. Based on this argument therefore, we can argue that the accuracy produced by an inside test will be limit of the accuracy produce by an outside test. It would be possible therefore to use an inside test to determine the best possible performance of an automatic diacritizer and use the model of an outside test to determine the size of corpus requited to achieve such level of performance.

The fundamental assumption of such an extrapolation is that the accuracy of a diacritizer will grow logarithmically (rather than linearly, polynomially or exponentially) with increase in corpus size. The accuracy values for the outside test can therefore be modeled by a logarithmic function. The parameters of the model obtained from the inside test can then be modified by those obtained from the outside test so as to better reflect the accuracy expected from an outside test based on a corpus of sufficiently large size which still remains unavailable for Yorùbá.

4 Test

An automatic diacritizer based on word trigrams, using linear interpolation was trained in 10 stages by accumulating 10,000 word portions of a manually diacritized corpus of about 100,000 words. The training corpus consisted of three Yorùbá published texts.

WHY? WHY?

NO! Split your corpus!

Three test files were used to test the accuracy of the automatic diacritizer at each of the 10 stages of the training. The first test file consisting of about 5,000 words was obtained from a single segment of the training text while the second test file also consisting of 5,000 words was accumulated from each of the ten segments of the training text. The third test file consisting of about 4,000 words was obtained from a different document which is independent of the training files.

In order to be able to relate the accuracy of the diacritizer to the size of training corpus, the three test files were stripped of their diacritical marks and presented to the diacritizer for automatic diacritization. The automatically diacritized text that resulted was then compared with the original manually diacritized text. The accuracy of diacritization was measured as a percentage of the number of correctly diacritized words divided by the total number of words in the respective test files.

5 Observations

As the training corpus grew steadily from a modest 10,000 words to about 100,000 words, the accuracy of the automatic diacritizer grew from 66.7% to 95.9% for the inside tests and from 58.7% to 70.5% for the outside tests.

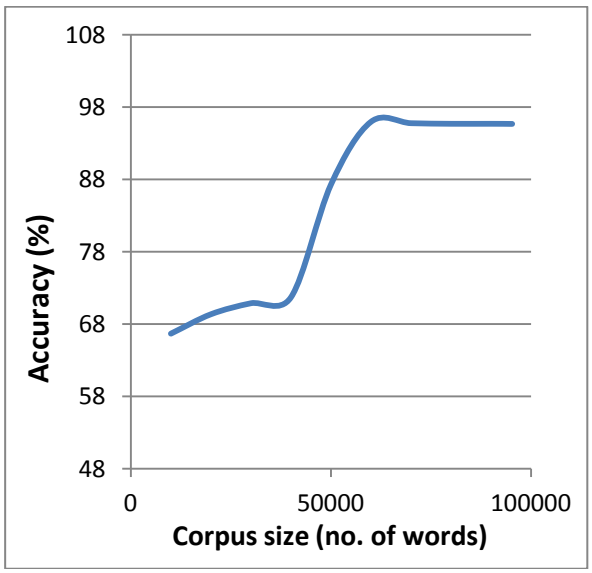

Figure 1: Corpus Size against Accuracy for inside test file from a single segment of training text

As shown in Figure 1, the plot of the results of the inside test based on the first test file consists of three distinct segments. The first segment lies between 10,000 words and 40,000 words with accuracy values between 66.7% and 71.7%, the second segment lies between 40,000 words and 60,000 words with accuracy values between 71.7% and 95.7% while the third segment lies between 60,000 words and 100,000 words with accuracy values between 95.7% and 95.9%. As earlier noted, the test file that was used for this inside test was obtained from the portion of the training file around the 40,000 to 60,000 words. This is reflected in the sharp rise in accuracy from 71.7% to 95.7% around this portion of the plot.

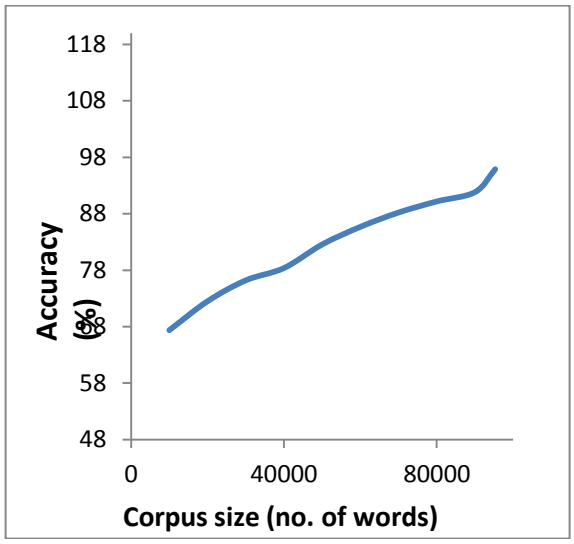

Figure 2: Corpus Size against Accuracy for inside test file from ten segments of training text

Figure 2 shows the plot of the results of the inside test file obtained from the ten segments of the training file. The plot consists of a single segment rising steadily in accuracy from 67.4% to 95.9%.

Figure 3 shows the result of the outside test which, in contrast to the first inside test consists of only one segment lying between 10,000 words and about 100,000 words of the training corpus and rising steadily in accuracy between 58.7% and 70.5%.

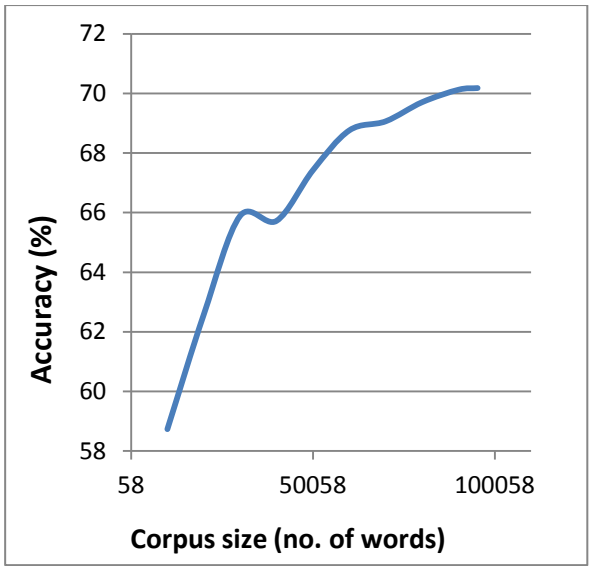

Figure 3: Corpus Size against Accuracy for the outside test file

It was also observed that the length of a word affects the probability of its accurate diacritization. As seen from Table 1 showing the distribution of wrongly diacritized words, 85.8% of wrongly diacritized words in the inside test were one-syllable words, 13.8% were two-syllable and three-syllable words while only 0.5% were 4-syllable or more words. As for the outside test, 43.5% of wrongly diacritized words were one-syllable words, 55.0% were two-syllable and three-syllable words while only 1.5% were four-syllable or more words. This shows that for the inside test, the source of error was primarily one-syllable words while for the outside test the source of error was more widely distributed; most of the wrongly diacritized words being two and three syllables words.

	Inside Test	Outside Test
One syllable	85.8%	43.5%
Two/Three syllables	13.8%	55.0%
Four/more syllables	0.5%	1.5%

Table 1: Distribution of wrongly diacritized words

Out of the total of about 7,500 distinct words in the 100,000 word training corpus, only 93 were diacritized wrongly in the inside test and only 262 were diacritized wrongly in the outside test.

Words of more than four syllables that manifested inconsistency between the manually diacritized and automatically diacritized words occurred mostly due to typographical errors originating from the test file. Specifically, there were instances in which a word in the test file was wrongly diacritized from source and the automatic diacritizer was able to correct the error. However, even though the automatic diacritizer undertook the correct diacritization, the mismatch between the manually and automatically diacritized words was seen as an inaccuracy in the automatic diacritization.

WHY? WHY?

NO! Split your corpus!

6 Discussion

By extracting the test file from the training corpus for the inside test, it is expected that the result would suggest the maximum possible levels of accuracy of automatic diacritization. We may be able to conclude therefore that for Yorùbá, the accuracy of an automatic diacritizer based on trigram probabilities may not be able to exceed figures beyond the region of 97% since the inside test guarantees that all words in the test file would have been encountered in the training session. Such a guarantee is not available for the outside test.

Even though the accuracy figure obtained from an inside test of a diacritizer trained on a 10,000 word corpus was about 66.7%, the equivalent outside test produced an accuracy figure of 58.7%. On the other hand, while the accuracy figures for an inside test for the diacritizer trained on a 100,000 word corpus went as high as 95.7% the accuracy of the outside test did not exceed 70.5%. This shows a level of exaggeration in the accuracy levels produced by the inside test and an increase in the differences between the accuracy levels of the inside and outside tests as the corpus size increases.

The plot in Figure 2 appears to be a stretched out version of the plot of Figure 1 because the test file of Figure 1 was obtained from one segment of the training corpus, while the test file for Figure 2 was distributed over the ten segments of the training corpus. The data that produced the plot in Figure 1 can be modelled by a logarithmic function given by:

$$y = 16.208\ln(x) - 89.226$$

The logarithmic function that best models the data that produced the plot in Figure 2 is given by:

$$y = 12.227\ln(x) - 48.141$$

Figure 4 compares these two models. It shows the similarity between the two functions within the range of 10,000 word and 100,000 words.

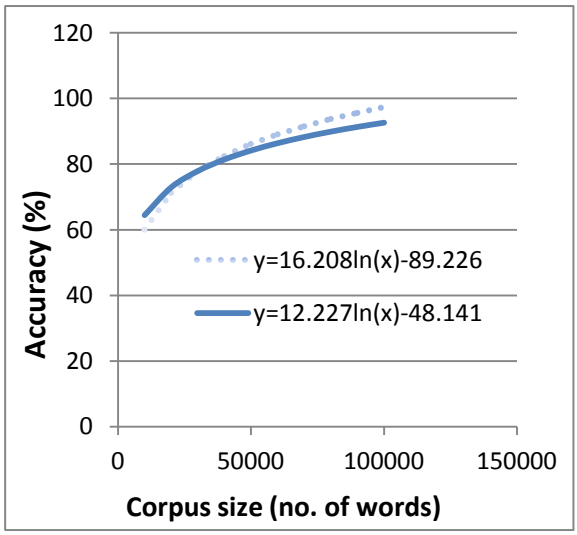

Figure 4: Comparison of models of single segment and ten segment test files

Having used the inside test to establish that the accuracy of a trigram based automatic diacritizer may not exceed 95.9%, it is possible to use the model of the outside test to calculate the corpus size that may produce this ideal accuracy level of 95.9%.

The accuracy values of the outside test are modeled by the logarithmic function:

$$y = 5.0818\ln(x) + 12.364$$

Hence, to get a *y* value of 95.9 from the above model we would require a 13.72×10^6 word corpus, which is a really large sized corpus. A more reasonably sized corpus of about 3,000,000 words would produce an accuracy value of 88.2%. Hence, a machine learning technique based solely on n-gram probabilities may not attain the ideal accuracy level of 95.9% based on a reasonably sized training corpus of about 3,000,000 words.

As shown in Table 1, the distribution of the wrongly diacritized words suggests that the probability of accurate diacritization of a word is proportional to its length. This suggests that the richer the contest the better the chances for accurate diacritization.

The preponderance of diacritization error among words of less than three syllables can be attributed to the lack of sufficient context for the identification of the right diacritical marks. Cocks and Keegan (2011) concluded that for Maori, the word-based approach is better that the character-based approach and our observation here suggests that word-based approach is also better for Yorùbá.

Two main factors affect the accuracy of diacritization. The first factor is the presence or absence of a given word in the training corpus while the second is the number of valid alternative arrangements of diacritics that can be applied to the vowels and syllabic nasals within a word. The observed shift in the source of error from one-syllable words to two and three syllable words between the inside and outside tests can be explained as follows. For an inside test, all words in the test file would have been encountered in the training process. Hence, the only prevailing error factor is the number of valid alternative arrangements of the diacritical marks that can be applied to the vowels and syllabic nasals within the words. Longer words have lower likelihood of many valid alternative diacritic arrangements because the length of the word promotes uniqueness in the pattern of arrangement of the diacritics. For one syllable words however, such contexts are not available. Since this is the only error factor in an inside test, it is not surprising that the bulk of the wrongly diacritized words in the inside test are one-syllable words. In the outside test however, both error factors contribute because there is no guarantee that every word in the test file would have been encountered in the training process.

OOV or In-Vocab

Since the diacritization errors are concentrated around words of less than three syllables, it would be instructive to construct n-grams of $n>3$ such as quadrigrams and quintigrams for this set of words. This will provide better context for their proper diacritization. The cardinality of the vocabulary obtained from the 100,000 word training test was about 7,500 while only 262 of these 7,500 words were wrongly diacritized. Given that much fewer words than the cardinality of the vocabulary would be used in the construction of these quadrigrams and quintigrams, the usual challenge of the manifestation of combinatorial explosion that discourages the use of n-grams beyond $n=3$ would have been addressed.

7 Conclusions

It has been demonstrated in this study that an automatic diacritizer based simply on trigrams may require corpora of more than 3,000,000 words for training in order to achieve significant accuracy levels for use in TTS. It would be nec-

essary therefore to seek complementary approaches to the simple n-gram approach in order to be able to use automatic diacritization for the effective application of tone marks on inadequately diacritized Yorùbá texts for use in TTS.

One possible way of addressing this challenge may be by the use of n-grams based on n *>3* for the few (e.g. 262) words that were most often diacritized wrongly. In addition, the use of morphological rules to increase the number of words covered by the automatic diacritizer may also be considered.

The literature suggests that for automatic diacritization, the character-based approach may be better than the word-based approach for certain languages while the revere may be true for some other languages. From our observation in this study, it would appear that the word-based approach is better for the Yorùbá language. A formal comparison of these two approaches to the automatic diacritization of Yorùbá texts would be a valuable future study.

References

Chomsky, N. 1957. Syntactic Structures. Mouton and Co. The Hague/Paris.

Cocks, J. and Keegan, T.T. 2011. A word-based approach for diacritic restoration in Maori. In Proceedings of Australasian Language Technology Association Workshop, pages 126-130

De Pauw, G., Wagacha, P.W. and De Schryver, G.-M. 2007. Automatic diacritic restoration for resource-scarce languages, In: V. Matousek and P. Mautner, ed., Proceedings of Text, Speech and Dialogue conference 2007, 170-179.

Elshafei M, Al-Muhtaseb H and Alghamdi M. 2006. Statistical methods for automatic diacritization of Arabic text. In: Proceedings of 18th national computer conference NCC'18, Riyadh, March 26–29, 2006.

www.everyculture.com. Countries and their Cultures. Consulted on the 27th January, 2012.

Mihalcea, R.F. 2002. Diacritics restoration: Learning from letters versus learning from words. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 339–348. Springer, Heidelberg

Mohamed, E. and Kubler, S. 2009. Diacritization for real-world Arabic texts. In Proceedings of Recent Advances in Natural Language Processing, Borovets, Bulgaria, 2009.

Scannell P. K. 2011. Statistical Unicodification Of African Languages, In: Language Resources and Evaluation, Issue 3, Vol 45,375-386.

Simard, M. 1998 Automatic Insertion of Accents in French Texts. In Ide & Vuotilainen (eds) Proceedings of the Third Conference on Empirical Methods in Natural Language Processing, Granada, Spain, 27-35.