Entropy-Based Generic Stopwords List for Yoruba Texts

Asubiaro, Toluwase Victor
E. Latunde Odeku Medical Library
University of Ibadan
Ibadan, Nigeria
Email: toluwaase {at} gmail.com

Abstract— **This research employed entropy based algorithm to identify stopwords candidate for Yoruba Language texts. Two sets of corpus of 756,039 Yoruba words were used; the diacritized and its undiacritized versions. All words whose entropy is greater than 0.6 but not a noun was considered as a stopword. A stopwords list of 256 words was drawn from the diacritized texts while a stopword list of 189 words was drawn from the undiacritized texts. For the diacritized texts, the removal of the stopwords reduced the full text by 65.91%. The full text of the undiacritized texts was also reduced by 67.46% when the stopwords were removed. Only 69.1 % of the stopwords have corresponding words in English stoplist. This suggests that existing English stoplist will not work optimally and could not be adopted for Yoruba language.**

Keywords- Stopwords, Yoruba Langauge, Information Entropy, Diacritics

I. INTRODUCTION

Removal of stopwords is one of the text preprocessing steps in Information Retrieval, text classification, document clustering and similar document analysis [1],[2]. Stopwords are the words that "appear frequently in documents and only serve syntactic function but they carry no usable information to aid learning tasks and are unlikely to assist in text classification, retrieval, clustering or analysis and hence are deleted during pre-processing" [3],[4],[5]. These words are considered as noise in information systems, hence there are research efforts to develop stoplists that are robust enough to contain these words and can help to efficiently manage noise in textual processing activities and information systems.

Therefore, stoplists that are either domain specific or language specific have emerged because of the idea of which words constitutes noise in a language or domain. The importance of these "customized " stop lists is well founded on the language differences in the languages or domains where there are specialized linguistic and morphological rules. Consequently, stoplist for a language may be inefficient for another depending on the similarity or differences between the languages. Lately, researchers compile stoplist that are time specific because of the time changing attribute of natural languages with human sophistication[3].

While some languages like English have a number of stoplist that have been developed both manually and automatically. However, the only reported stoplist for Yoruba was in Asubiaro[6] which was compiled by manual identification of redundant words from corpus that was collected from two religious websites. Yoruba language is spoken in some west African where the speakers occupy southwesten Nigeria, southern Benin republic and southern Togo. Like most African languages; it is a technologically resource-scarce language. Resource scarce languages lack necessary language technologies. This study therefore employs the automatic means of creating a language specific stoplist for Yoruba language by using word Entropy to identify likely stopwords in the language.

II. STOPLISTS

Over time, different metrics have been applied to determine noise constituents in a language or domain. Whether manually or automatically, researchers have employed the word frequency metric based on the principle that the frequency of words negatively correlate the information they bear [7], [8], [9]. Hence, high frequency correlates with high noise value. Although this method resulted in acceptable stoplists, but it does despise the effect of high frequent words in a number of documents, thereby making little of the evenness of the distribution of words across the set of documents. Another method employed is the syntactic analysis of words, where words considered as "fluff" or non information bearing words are considered as stopwords. This method was employed by Asubiaro [6] and Davarpanah, Sanji and Aramideh [10] for Yoruba and Farsi stop list generation respectively.

Most of the earliest studies on the creation of stop list were found to create generic stop list for a particular language.

Later studies have employed the automatic method of creating stoplist for languages. A very common metric for measuring information contained in words in most of the automatic methods is the Entropy. The rationale behind this methodology is that words with high Entropy were considered as stopword candidate. Zou Wang, Deng, Han and Wang [4] automatically created stoplist for Chinese language using the aggregated model which involves measuring the word frequency feature by statistical and informational models based on some statistics features and Entropy of words respectively. Stoplist developed using Entropy values of word include Alajmi et al [11]. Another study that used Entropy is Silva and Ribeiro [3] whose stoplist was specifically for the web.

Other researchers have employed some other methods. Choy [12] created English stoplist from twitter using combinatorial values. Another stoplist created was by Tsz-Wai and He [13] based on Kullback-Leibler divergence measure. Kiso Shimbo, and Matsumoto [14] also created a stoplist using the Klein-berg's HITS algorithm.

III. Corpus for the Study

A major challenge in language technology development for Yoruba language is the unavailability of sufficiently sized corpus which has been reported by several researchers in the past[16]. Hence, a corpus that is representative and qualitative enough was developed for this research. Bodies of Yoruba texts were gathered online and offline. A corpus of 756, 039 Yoruba words from 331 documents drawn from religious texts, published articles, online and offline newspapers and academic research projects.

Two sets of corpus were used; the diacritized and its undiacritized version. Diacritized texts were obtained from undiacritized versions by autodiacritizing them using a Yoruba autodiacritizer with minimal human corrections.

Yoruba orthography seriously use diacritics, but due to dearth of specialized input devices for the language, most writers type without the diacritics. It is noteworthy that diacritics provide morphological and lexical information in Yoruba. Therefore ignoring diacritics results in loss of information and specificity. For the diacritized texts, 23,185 unique words were present, while 18,453 unique words were present in the undiacritized texts.

IV. Information Entropy and Stopwords Generation

According to the information theory, the randomness of a word correlates its information bearing capacity. Shanon [16] posits that entropy is a measure of randomness. Subsequently, highly random words, which are also low entropy words are very informative. Since stopwords carry little information they are high entropy words. There are other functions such as word-document frequency, inverse-document frequency, that have been used to determine what constitute noise in Information systems. Harman [17] proved that of all the metrics employed in measuring information, the highest precision was achieved with the entropy of each individual word. Entropy measures the frequency variance of a given word for multiple documents, i.e. words with very high frequencies in some documents but low frequency in others will have high entropy. Entropy $W(w)$ of a given word w with respect to a given set of n documents is as follows:

$$W(w_j)=\Sigma\, P_{i,j}.log(1/P_{i,j}) \quad (1)$$

where,

$$p_i(w) = \frac{f_i(w)}{\sum_{j=1}^{n} f_j(w)} \quad (2)$$

and:

$f_i(w)$ = Frequency of word w in document i
n = number of documents.

The entropy of each word in the dataset was calculated, the resulting list was ordered by ascending entropy to reveal the words that have a greater probability of being noise words.

Proposed Stopword List

The general stopword list for Yoruba is expected to have the following features;

Firstly, the stopwords must be bad indexing terms, they cannot be used as keywords. Secondly, the stopwords when removed from the full text must reduce the full text by at least 50%. Thirdly, the stopwords must only perform syntactic function and they bear no meaning in themselves. Fourth, they must be general words that are not restricted to a particular field.

A certain level of arbitrariness was employed by removing all nouns from the stopword candidate list, this method was also used by Fox [8] and Savoy [9]. Therefore, all nouns were removed from the list of words whose Entropy is greater than 0.6. Nouns are found as information bearing words or keyword candidates

V. Results

Tables 1 and 2 contain stopwords from the two corpus used for this study. A list of 256 stopwords was drawn from the diacritized texts. The full text was reduced by 65.91% when the stopwords were removed. Comparing the stoplist with English stoplist compiled by [8], only 69.1% of the stopwords are present in the English stoplist. Also, some of the (Yoruba) stopwords have no corresponding words in English. This suggests therefore, adapting English stopwords list for Yoruba will not work optimally for the language.

For the undiacritized version of the texts, 189 stopwords were drawn. The full text was reduced by 67.46% when the stopwords were removed from the full text.

Table 1: Stop list of Yoruba words with diacritics
tí, ni, wọn, àwọn, tó, pé, ń, ṣe, náà, ó, kò, sí, bá, wá, fi, kí, lọ, yìí, kan, a, jẹ́, sì, fún, ti, bí, yóò, sọ, àti, rẹ, láti, wọ́n, ẹ, ní, í, ló, máa, òun, gbà, nínú, rí, gbogbo, nígbà, lè, ọmọ, gbé, ọ̀rọ̀, èyí, mo, mọ, mọ́, rẹ̀, di, ọ̀hún, bẹ́ẹ̀, tún, nǹkan, ara, nítorí, ṣùgbọ́n, lọ́wọ́, mú, dá, lórí, ẹni, ọwọ́, bíi, jù, pẹ̀lú, ọjọ́, mi, sílẹ̀, iṣẹ́, bó, ohun, kó, un, dé, báyìí, pa, níbẹ̀, ibi, wà, ká, tàbí, láàrin, yẹ, gbọ́ , làwọn, bẹ̀rẹ̀, má, ìyẹn, kú, ìgbà, kọjá, lóun, eré, ọnà, pọ̀, àsìkò, jáde, gẹ́gẹ́, jọ, á, i, sọ̀rọ̀, fẹ́, irú, kúrò, bọ̀, o, le, lára, wáá, níbi, lẹ́yìn, wò, wo, nípa, ín, inú, méjì, nílé, yín, ọn, ńkọ́, ńlá, já, déédé, la, tẹ, rárá, ìdí, òpin, ṣeé, lọ́jọ́, lọ́sẹ̀, títí, wáyé, ò, padà, án, lójú, tẹ́lẹ̀, rò, gan, tàwọn, tán, rán, tòun, lọọ, ẹ̀yin, ọ̀dọ̀, wa, tóó, ṣẹ̀bọ, kọ́kọ́, ta, sá, n, yọ, dúró, hàn, ṣíṣẹ́, lọ́dún, kì, lẹ, ṣẹlẹ̀, pàápàá, nílùú, nìkan, níṣẹ́, síbẹ̀, nìyẹn, yẹn, kankan, bóyá, múra, fẹ́wọn, e, lọ́nà, yá, gbọ́dọ̀, lénu, wàhálà, wí, káàkiri, parí, síbi, kọ, mẹ́ta, ọ̀kan, kẹta, san, péré, dáadáa, láìpẹ́, wọlé, sùn, tilẹ̀, lélẹ̀, ẹnikẹ́ni, á, ẹnìkan, èmi, ọ̀pọ̀, mẹ́rin, peléke, nídìí, wẹ́wẹ́, ún, méjìlá, kínní, ú, sínú, sáré, kín, yàrá, fúnra, kọjú, díẹ̀, lòótọ́, nìyí, àtàwọn, bákan, méjèèjì, fẹ̀gbẹ́, àbí, kiri, torí, jẹ, káwọn, márún, jùlọ, ọ̀pọ̀lọpọ̀, síwájú, àgàgà, lásán, tọ, tètè, àwa, odidi, pàdé, tiẹ, tuntun, gba, sọ́dọ̀, kàn, yí, ọ, mìíràn, wọ̀nyí, ná, ọ́, ẹ́, afi.

Table 2: Stop list of Yoruba words without diacritics
ti, ni, won, awon, n, pe, to, si, o, ko, naa, se, lo, wa, ba, e, ki, kan, fi, yii, je, i, fun, a, mo, yoo, so, re, bi, le, ati, lati, gbe, ohun, maa, oun, ninu, ile, gba, omo, gbogbo, ri, oro, nigba, un, bee, de, ka, pa, owo, nitori, egbe, sugbon, lowo, on, bo, mu, tun, ara, di, lori, da, benikan, in, eni, fin, wo, eyi, jo, nnkan, mi, te, ise, pelu, oju, ilu, ede, ojo, gbo, yin, fee, lawon, bayii, bii, bere, ju, ebi, daa, ma, fe, ye, bu, iyen, jeun, gege, tabi, waa, gege, tabi, nibi, ta, leyin, lara, an, meji, iru, nibe, nipa, jade, koja, titi, nla, pada, inu, tele, okan, tawon, tan, loju, too, yo, loun, emi, po, rara, gbodo, sa, han, koko, sele, paapaa, nikan, laarin, see, yen, niyen, ran, kuro, ya, meta, iroyin, gan, daadaa, sibe, toun, keji, Kankan, wonyi, laipe, eyin, kinni, die, kun, bakan, bawo, pari, atawon, sinu, kawon, dara, siwaju, opolopo, kiri, merin, pade, enikan, niyi, yato, boya, looto, tete, eleyii, kin, marun, lodun, saa, tuntun, rere, seni, sibi, fowo, koju, dide, yi

VI. Conclusion

It is imperative to create stopword list because of its importance in text preprocessing, particularly for a resource-scare language like Yoruba. Preprocessing is not only useful for Information retrieval, it is the main step shared among text mining, NLP, Automatic Speech Recognition (ASR) and many other applications. A corpus of 756, 039 words was used in this experiment drawn from religious texts, published articles, online and offline newspapers and research projects.

This paper used an entropy based method to create a general stopword list for Yoruba, a resource scare language. The stopwords met the four standard set for a proposed stoplist for Yoruba. Viz: It contained only words that are bad indexing terms, therefore they cannot be used as keywords. The stopwords constitute over 50% of the full text. The stopwords only perform syntactic function and they bear no meaning in themselves. The stopwords are general words that are not restricted to a particular field.

Future work is expected on domain-specific stoplists for the language and the employement of other metrics for creating stoplist. Furthermore, it is suggested that future work will use corpus of larger size.

Acknowledgement

I wish to acknowledge the support of Dr. Tunde Adegbola, my mentor and the Executive Director of African Languages Technology Initiative, Ibadan, Nigeria.

References

[1] R. Baeza-Yates and B. Rebiero-Neto. “Modern Information Retrieval.” Addison Wesley, London, England. Chapter 1, pp. 9-15. 2003.

[2] C. Hsin-His. “Terms and Query Operation”, Department of Computer Science and Information Engineering, National Taiwan University, Taiwan. Pp.2-14. 2009

[3] P. Sinka, and W. Corne, “Evolving Better Stoplists for Document Clustering and Web Intelligence.” Design and Application of Hybrid Intelligent Systems. Pages 1015-1023. IOS Press Amsterdam, The Netherlands 2003.

[4] F. Zou, L. Wang, X. Deng, S. Han and L. Wang (2006). “Automatic Construction of Chinese Stop Word List”. Proceedings of the 5th WSEAS International Conference on Applied Computer Science, Hangzhou, China, April 16-18, 2006 (pp1010-1015)

[5] I. El-Khair (2006). “Effects of Stop Words Elimination for Arabic Information Retrieval: A Comparative Study”. International Journal of Computing and Information Sciences. Vol 4. No 3. December 2006, Pp 119-133.

[6] T. Asubiaro. “An Analysis of the Structure of Index Terms for Yoruba Texts.” Master’s Degree Project submitted to Africa Regional Centre for Information Science, University of Ibadan, 2011.

[7] J. Rijsbergen, Information Retrieval, Second Edition, Department of Computer Science, University of Glasgow, Butterworths, London. 1979

[8] C. Fox, “A Stop List for General Text.” SIGIR FORUM, 24(4), December 1990.

[9] J. Savoy, “A Stemming Procedure and Stopword List for general French Corpora.” Journal of the American Society for Information Science, 50(10), 1999, 944-952.

[10] M. Davarpanah, M. Sanji, and M. Aramideh (2009). “Farsi Lexical Analysis and Stop word List.” Library Hi Tech, Vol. 27, No 2, 2009.

[11] A. Alajmi, M. Saad, and R. Darwish, (2012). “Toward an Arabic Stop-Words List Generation. International Journal of Computer Applications (0975-8887). Volume 46-No.8, May 2012.

[12] M. Choy, “Effective Listings of Function Stop Words for Twitter 2012.” Arxiv.org/pdf/1205.6396, retrieved on 23rd May 2013

[13] R. Tsz-Wai, B. He, and I. "Automatically Building a Stopword List for an Information Retrieval System." 5th Dutch-Belgium Information Retrieval Workshop (DIR)'05Utrecht, the Netherlands 2005.

[14] T. Kiso, M. Shimbo, and Y. Matsumoto, "HITS-based Seed Selection and Stop List Construction for Bootstrapping." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: short papers, pages 30–36, Portland, Oregon, June 19-24, 2011. C 2011 Association for Computational Linguistics.

[15] T. Adegbola and L. Odilinye, "Quantifying the effect of Corpus Size on the Quality of Automatic Diacritization of Yorùbá Texts." Africa Languages Technology Initiative. 2012.

[16] E. Shanon, "Mathematical Theory of Communication", Bell System Technical Jounal. 27:, 379-423 and 623-656. 1948.

[17] D. Harman, "An Experimental Study of Factors Important in Document Ranking" in Proceedings of the 1986 ACM Conference on Research and Developments in Information Retrieval, Pisa, 1986.