Research Journal of Computer and Information Technology Sciences _____
Vol. **1(1),** 2-5, February **(2013)** *Res. J. Computer & IT Sci.*

Part-of-Speech tagging of Yoruba Standard, Language of Niger-Congo family

Adedjouma Sèmiyou A.[1], John O. R. Aoga[1] and Mamoud A. Igue[2]
[1]Laboratory of Electro-technical, Telecommunications and Applied Computing of the Polytechnic School of Abomey-Calavi (EPAC), BENIN
[2]Faculty of Letters, Arts and Human Science, University of Abomey-Calavi (UAC), BENIN

Available online at: www.isca.in
Received 18th February 2013, revised 20th February 2013, accepted 25th February 2013

Abstract

The utilization of corpora is a critical phase of systems of Natural Language Processing (NLP) based on statistical methods. This point is crucial for less equipped and less computerized languages like African languages. This paper aimed to design a yoruba corpus. Yoruba is an African language of Niger-Congo family. It is spoken by more than thirty million people around the world and particularly in Nigeria and Benin. The main motivation of this work was to obtain training data for PoS taggers and to provide applications of Yoruba Language Processing (YLP) with a basic tool. The tagging was performed with SVMTool one of the Part-of-Speech taggers widely used. The preprocessing of the text general outline has been ensured by Perl scripts. The corpus with 312,562 words, formed from the Web, was annotated with an accuracy of 98.04%. This annotated corpus might be used in translation system.

Keywords: Annotated corpus, PoS tagger, Natural Language Processing, Yoruba.

Introduction

The language technology is crucial today to ensure access to information and opportunities for economic development. With about two thousand different languages, Africa is a multilingual continent. It presents significant challenges for researchers who seek to promote and use African languages in the areas of business, development, education and humanitarian. The Natural Language Processing (NLP) allows these challenges in the creation of application that use languages. It is well known that taggers are the essential elements in the development of any serious application in this field. But very few African languages have this tool. The following languages that make use of this tool are: Amharic, Wolof, Bantu[1,2,3].

Yoruba, one of the twelve (12) Edekiri group languages of the Niger-Congo family, is an African language. It has problems similar to others languages. However, we can notice the work on the possibility of a version of Microsoft Windows Vista in Yoruba; on the speech of the Yoruba, and online dictionaries Yoruba-English, English-Yoruba[4,5,6]. More specifically in Yoruba Language Processing (YLP) we can highlight the lexical database of Yoruba containing 450,000 words and the model of Yoruba lexical analyzer using an approach based on the rules relating to morphology[7,8]. However, there is not yet tagger for Yoruba.

Our purpose is to design an annotated corpus of Yoruba with tagger. We believe that this work will help by providing a stepping stone for YLP-based applications and a methodology for a similar work on others languages.

Research Methodology

The design of the yoruba corpus was done in two (02) parts. First, we made the choice of tools and then we moved on to the PoS Tagging.

Tools: The first tool of our work is yoruba language. Yoruba is one of the twelve (12) languages of the Edekiri sub-branch from the great family of Niger-Congo. It is natively spoken in southwestern part of Nigeria (the second largest ethnic group in number), in Benin and Togo by over 30 million people[9]. Since 2011, Yoruba has a unique alphabet containing thirty (30) letters composed of 12 vowels: i, e, ẹ, a, o, ọ, u, in, n, un, n, an and 18 consonants: b, d, f, g, gb, h, j, k, l, m, n, p, r, s, ṣ, t, w, y[10]. Yoruba is a tonal language. It has three (03) tones: high tone (represented by the acute accent), the low tone (represented by the grave accent) and the average tone (represented by the absence of accent).

The categories of words or part-of-speech are defined as classes of words in a sentence. They are used as labels in PoS Taggers. The most important are: verbs, adverbs, adjectives, nouns, pronouns and connecting elements. To perform the PoS Tagging, we first perform the text preprocessing (segmentation, removing duplicates) using scripts written in Perl (Perl 5).

Then, we focused on the PoS Taggers. In fact, PoS Taggers are NLP tools that could be considered neutral, because they are

able to adapt to any language, as long as the corpus and models are available. The property, which also makes these tools very robust, is that they use methods of machine learning. These methods allows in the first time, to learn the language based on some relevant examples of sentences which are provided and then they are precisely predict the category of a new word of the language. Here, we used SVMTool based on SVM (Support Vector Machine) unlike many other taggers based Hidden Markov Models (HMM)[11].

Indeed, the SVM is a statistical learning algorithm based on the determination of maximum margins between many classes in a given set [12]. In the PoS training, we use the sample equation-1 that represents the pairs (word, tag) that are not linearly separable.

$$S = \{(x_1, u_1), \quad \ldots, \quad (x_m, u_m)\} \tag{1}$$

With x_i the i[th] word in corpus and u_i the i[th] tag in corpus. m is the size of S.

So we used the equation-2 to determine the parameters of the system's learning.

$$\begin{cases} Minimize & \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\xi_i \quad (C=cte>0) \\ constraints & u_i h(x) \geq 1 - \xi_i \quad (i = 1, \cdots, m) \end{cases} \tag{2}$$

To evaluate the performance of the system, we used the F-measure, equation-3 combining the precision (P) equation-4 representing the probability of an object returned by the system being relevant and the recall (R) equation-5 representing the probability of a relevant object is returned by the system. Where V_P= True positive, F_P= False positive, V_N= True negative et F_N= False negative,

$$F - mesure = \frac{2 \times P \times R}{(R+P)} \tag{3}$$

$$P = \frac{V_P}{V_P + F_P} \tag{4}$$

$$R = \frac{V_P}{V_P + F_N} \tag{5}$$

Steps of designing annotated corpus: Figure-1 shows a block diagram of the steps taken to implement the annotated corpus. i. Choice of digital resources: we used the Yoruba bible because of the widespread of using of biblical texts in NLP and the seriousness in the spelling of words [1, 2, 3]; ii. Retrieving text and segmentation (basic corpus): the text retrieved was segmented using a Perl script; iii. Recovery of different lexemes contained in basic corpus (Corpus 1); iv. Creating a training corpus containing the elements of finite categories (corpus 0) such as personal pronouns, possessive pronouns, relative pronouns, interrogative pronouns, v. Utilization of corpus 0 to train (with SVMTlearn of SVMTool) the system that was used to annotate the corpus 1 (corpus 2) by SVMTagger of SVMTool. Thus, manual annotation was reduced because it remains the labeling of the unknown elements of the corpus; vi. Manual annotation of the unknown elements of the corpus 2 achieved through the usage of an English-Yoruba and Yoruba-English dictionary [13]. This dictionary has allowed us to manually annotate corpus and identify elements of the corpus which presents any ambiguities; vii. Utilization of corpus 2 corrected to annotate basic corpus; viii. Manual correction of basic corpus and generating repairing dictionary (R). This corpus is final aim of this manipulation; ix. Performance evaluation of the system with SVMTeval of SVMTool using equation-3.

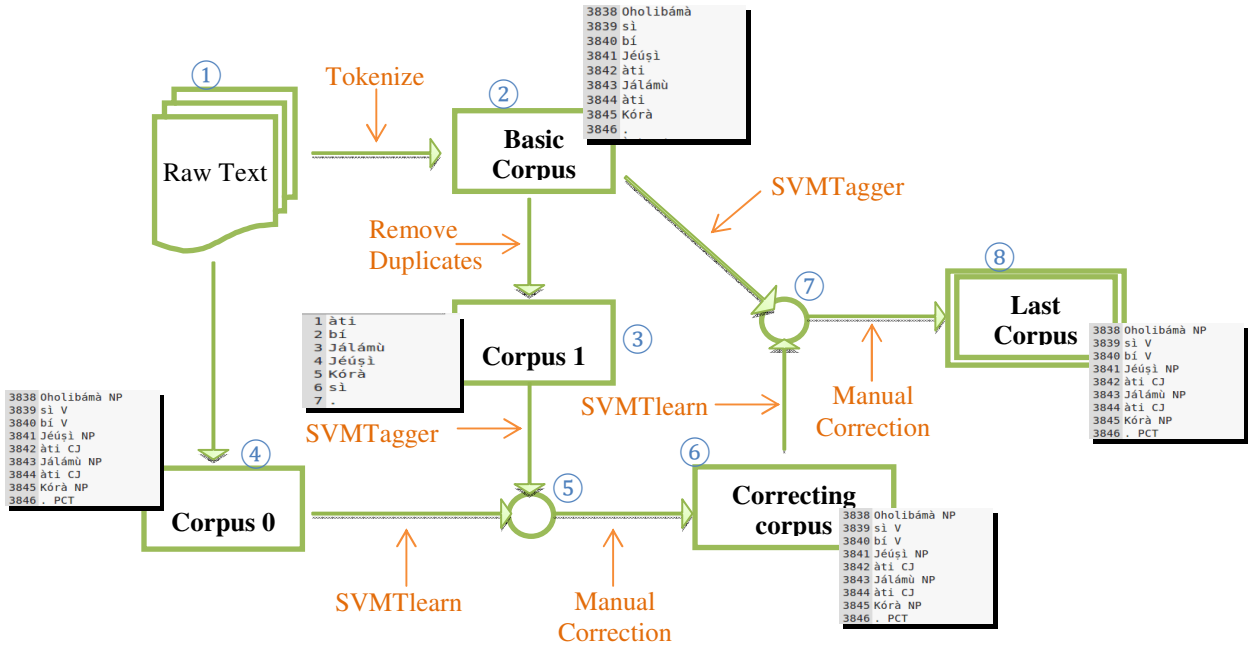

Figure-1
Steps of the realization of the annotated corpus

Results and Discussion

Results: The main result of our research is the creation of the annotated corpus with PoS Tagging for the Yoruba. This corpus contains 312,562 words.

To evaluate the performance of our system comparatively to the tagging in SVMTool, we considered 60% of the corpus to train the system and the remaining 40% were used for testing. Note that 40% of the corpus constitutes the text file and contain 5.5% of unknown words and 17.2% of ambiguous words.

We built our model on the combination of models 0, 1, 2 and 4 which are provided by SVMTool, this allowed us to use all possible strategies. The Management of ambiguities within known and unknown words was taken into account. Thus, the PoS tagging of yoruba, performed on the corpus with the configuration file SVMTool gave an average accuracy of 97.85% with a standard deviation of 0.33% for all strategies.

Table-1 shows for each strategy the overall accuracy, the accuracy on unknown words (UNK) and those of ambiguous words (AMBK). Thus, the strategy 2 gives the best performance which is 98.04%.

Table- 1
Details of each strategy applied to the annotated corpus

Strategies	AMBK (%)	UNK (%)	Accuracy (%)
0	93.16	85.97	98.02
1	93.3	86.78	98.03
2	93.9	87.00	98.04
4	93.55	82.09	97.89
5	93.17	86.98	98.03
6	93.56	82.08	97.89

(AMBK = ambiguous words, UNK = unknown words)

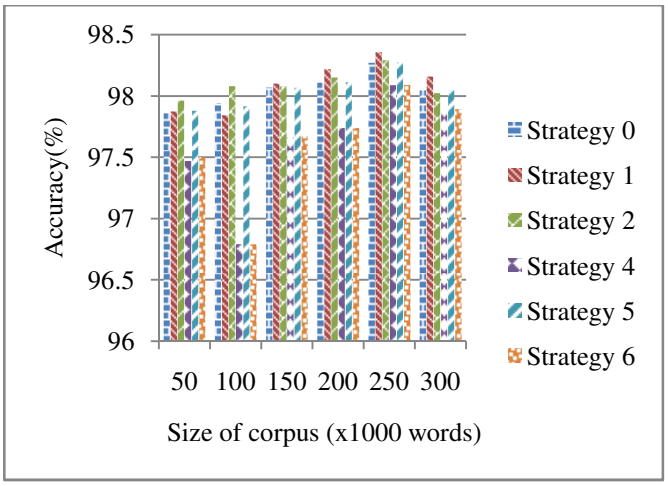

Figure-2
Variation of the precision of the label depending on the size of the annotated corpus

To evaluate the influence of size of the corpus on the accuracy, we evaluated the performance by gradually increasing to 50 000 words the size of the corpus starting from 50,000 to 312,562 words. The results of this evaluation are presented in Figure-2. We note that the development of the corpus size is not proportional to the accuracy.

Results and Discussion

The accuracies obtained on a small corpus indicate the relevance of the training examples and the accuracy of the selected features. But the adding of words in annotated corpus will not necessarily increase the accuracy because as we noted the evolution of accuracy is not proportional to the size of the corpus.

But this performance is quite interesting especially as it shows among other things that the SVM can provide better performance in labeling as long as we take account of over learning.

It is also to be noted that the PoS Tagging of the same corpus must be evaluate with others PoS Taggers. This allows to evaluate the effectiveness and robustness of the model and system. It also requires a careful error analysis.

Conclusion

Researches in Natural Language Processing allow a certain evolution of languages. This paper is included in the development of YLP-based Applications. Where the aim was to develop a Yoruba annotated corpus. This corpus was designed and evaluated through the SVM Tool and was found to be 98.04% accurate.

However, an evaluation of the annotated corpus on other taggers is possible. The study of the causes of errors and the identification of factors that contribute to the development of precision are many challenges to be overcome.

In future work, we will try to make a reliable yoruba speech synthesis using this corpus. A dependency parser Yoruba and a design of a bilingual corpus are also possible.

Acknowledgement

We thank Mr. IGUE Mamoud who helps us for the yoruba language.

References

1. Gamback B., Olsson F., Argaw A.A. and Asker L., Methods for Amharic part-of-speech tagging, *AfLaT*, Athens, Greece 104-111 **(2009)**

2. Dione C.M.B., Kuhn J. and Zarrie S., Design and development of part-of-speech-tagging resources for wolof (Niger-Congo, spoken in Senegal), *LREC'l0*, 1-8 **(2010)**

3. De Pauw G., De Schryver G.M. and van De Loo J., Resource-Light Bantu Part-of-speech Tagging, *SALTMILSI AfLaT*, 85-92 **(2012)**

4. Adegbola T., Owolabi K. and Odejobi T., Localising for Yoruba: Experience, challenges and future direction, *Proc. of HLT*, Alexandrie, Egypte, 7-10 **(2011)**

5. Odejobi O., Design of a text markup system for Yoruba text to speech synthesis applications, *Proc. of HLT for development*, Alexandrie, Egypte, 74-80 **(2011)**

6. Smith P. and Onayemi A., Yoruba Dictionary, *Ed. Bis Bus International*, http://www.yorubadictionary.com/, **(2003)**

7. Awoyale Y., Global Yoruba Lexical Database, *LDC*, 1-49 **(2008)**

8. Aladesote I., Olaseni O.E., Adetunmbi A.O. and Akinbohun F., A Computational Model Of Yoruba Morphology Lexical Analyzer, *Proc. of IJCL*, **2(1)**, 37-47 **(2011)**

9. Igue A. M., Grammaire Yoruba de base abrégée, *CASAS*, 1-49 **(2009)**

10. Adeniyi H., Yusuff A., Adesanya A., Olomu O., Igue A. M., Fadoro O., Fakeye F. and Bada M., Une orthographe standard et unifiée pour le Yoruba (Nigéria, République du Bénin et Togo), *CASAS and CBAAC*, 1-20 **(2011)**

11. Gimenez J. and Marquez L., SVM Tool: Technical Manual v1.3, *TALP Research Center*, LSI Department, Barcelone, 1-50 **(2006)**

12. Conuejols A. and Miclet L., Apprentissage artificiel : Concepts et algorithmes, *Ed. Eyrolles*, **2(1)**, 279-310 **(2003)**

13. Upplc, A Dictionary of The Yoruba Language, *Ed. University Press PLC IBADAN*, 0-239 **(2011)**