Journal of Software Engineering and Applications, 2016, 9, 516-523
http://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

Morpho-Syntactic Tagging of Text in "Baoule" Language Based on Hidden Markov Models (HMM)

Hyacinthe Konan[1], Bi Tra Gooré[2], Raymond Gbégbé[2], Olivier Asseu[1,2*]

[1]Ecole Supérieure Africaine des TICs (ESATIC), Abidjan, Côte d'Ivoire
[2]Institut National Polytechnique Félix Houphouët Boigny (INP-HB), Yamoussoukro, Côte d'Ivoire
Email: *oasseu@yahoo.fr

How to cite this paper: Konan, H., Gooré, B.T., Gbégbé, R. and Asseu, O. (2016) Morpho-Syntactic Tagging of Text in "Baoule" Language Based on Hidden Markov Models (HMM). *Journal of Software Engineering and Applications*, **9**, 516-523.
http://dx.doi.org/10.4236/jsea.2016.910034

Received: August 30, 2016
Accepted: October 22, 2016
Published: October 25, 2016

Abstract

The label text is a very important tool for the automatic processing of language. It is used in several applications such as morphological and syntactic text analysis, indexing, retrieval, finished networks deterministic (in which all combinations of words that are accepted by the grammar are listed) or by statistical grammars (e.g., an n-gram in which the probabilities of sequences of n words in a specific order are given), etc. In this article, we developed a morphosyntactic labeling system language "Baoule" using hidden Markov models. This will allow us to build a tagged reference corpus and represent major grammatical rules faced "Baoule" language in general. To estimate the parameters of this model, we used a training corpus manually labeled using a set of morpho-syntactic labels. We then proceed to an improvement of the system through the re-estimation procedure parameters of this model.

Keywords

Corpus, the Set of Tags, the Morpho-Syntactic Tagging, "Baoule" Language, Hidden Markov Model

1. Introduction

Each language has its own syntax. That of language "Baoule" is not that of the French and vice versa. In this article we are trying to answer the following question: How to bring out the structure of a given sentence to recognize and understand its contents? Indeed a sentence has meaning only when it is syntactically and semantically correct. The sentence will therefore be considered recognized. The syntactic analysis puts additional strain on the recognition system so that the studied paths correspond to words in the lexicon "Baoule" (lexical decoding), and for which, words are in proper sequence as specified by a sentence pattern. Such a model of sentence may again be represented by a

deterministic finite network, or by a statistical grammar [1]. For some tasks (command, control processes), one word of a finite set must be recognized and so the grammar is either trivial or useless [2]. For other applications (e.g., sequences of numbers), very simple grammars are often sufficient (for example, a figure can be discussed and followed by a number) [3]. Finally, there are tasks that the grammar is a dominant factor. It significantly improves the performance of recognition. The Semantic Analysis adds additional stress to all the recognition search path. One of the ways the semantic constraints are used is carried out by means of a dynamic model [4]. Depending on the condition of recognition, some syntactically correct input channels are eliminated from consideration.

This again serves to make easier the recognition task and leads to a better system performance. In Côte d'Ivoire, the French as official language is not always spoken by the entire population. Some local languages like the "Bambara" (Malinké) and especially "Baoule" language emerge, but fail to address the concerns of people who today have to do with the evolution of digital technologies without always under- standing or speaking the conveying languages. Our research work offer goes beyond what is currently available and will allow a person speaking only "Baoule" language to receive and understand "Baoule" language communication expressed in French.

2. Research Question

What is the probability that a sentence in "Baoulé" language is recognized correctly? The linguistic model we propose to build in this section will help us answer this question.

2.1. System Overview

Syntactic categories

Consider the following labels representing the syntactic categories in language "Baoule": N = name; V = Word; P = preposition; ADV = adverb; ADJ = adjective; D = Determinant; etc.

We want to build a system that will input a sequence of words $PH = w_1, w_2, \cdots, w_n$ and will output a sequence of labels $ET = et_1, et_2, \cdots, et_n$ (**Figure 1**).

Input is: S = "le professeur parle"

Output is: D N V (le/D professeur/N parle/V)

Some elements of "Baoule" grammar

The semantic categories of time and appearance match in the conjugation of "Baoulé" to different morphological phenomena. The grammatical expression mode only involves the tone. Expression of aspectuality involves affixes, as time has no direct expression in the context of the combination (see [1] for more detail on the elements of grammar).

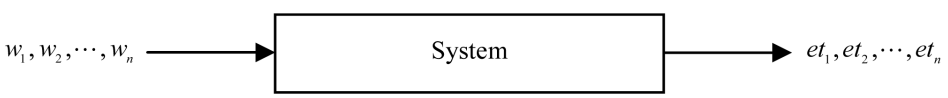

Figure 1. System using the linguistic model.

2.2. The Part-of-Speech Tagging

A label corpus is a corpus in which are associated to text segments (usually words) other information of any kind be it morphological, syntactic, semantic, prosodic, critical, etc. [2] [3]. In particular, in the community of automatic natural language processing, when talking of tag corpus it is most often referred to a document in which each word has a morphosyntactic tag and a single. The automatic labeling morphosyntactic is a process that is usually done in three stages [4] [5]: the segmentation of text into tokens, the a priori labeling disambiguation which assigns to each lexical unit and depending on its context, relevant morphosyntactic tag. The size of the label set and the size of the training corpus are important factors for good performance of the labeling system [6] [7]. In general, there are two methods for part-of-speech tagging: rule-based method [7] [8] and the probabilistic method. In this article we have used the second approach.

3. Methodology

The choice of the most likely label at a given point is in relation to the history of the last labels which have just been assigned. In general this history is limited to one or two previous labels. This method assumes that we have a training corpus which must be of sufficient size to allow a reliable estimate of probabilities [9].

3.1. Hidden Markov Model (HMM) Taggers

We have an input sentence $PH = w_1, w_2, \cdots, w_n$ (w is the ith word in the sentence).

We have a tag sequence $ET = et_1, et_2, \cdots, et_n$ (et_i is the ith tag in the sentence).

We'll use an HMM to define $p(w_1, w_2, \cdots, w_n, et_1, et_2, \cdots, et_n)$ for any sentence $w_1, w_2, \cdots, w_n$ and tag sequence $ET = et_1, et_2, \cdots, et_n$ of same length.

Then the most likely tag sequence for ET is

$$e_1^* \cdots e_T^* = \arg\max_{et_1 et_2 \cdots et_n} p(w_1, w_2, \cdots, w_n, et_1, et_2, \cdots, et_n)$$

3.2. Trigram Hidden Markov Models (Trigram HMMs)

Basic definition

For any sentence $PH = w_1, w_2, \cdots, w_n$ (where $w_i \in V$ for $i = 1, \cdots, n$) and any tag sequence

$$ET = et_1, et_2, \cdots, et_n$$

(where $et_i \in S$ for $i = 1, \cdots, n$) and $et_{n+1} = \text{STOP}$, the joint probability of the sentence and tag sequence is:

$$p(w_1, w_2, \cdots, w_n, et_1, et_2, \cdots, et_n) = \prod_{i=1}^{n+1} q(et_i \mid et_{i-2}, et_{i-1}) \prod_{i=1}^{n} e(w_i \mid et_i)$$

where we have assumed that $w_0 = w_{-1} = *$.

Parameters of the model:

$q(s \mid u, v)$ for any $s \in S \cup \{\text{STOP}\}, u, v \in S \cup \{*\}$ (Trigram)

$e(w \mid s)$ for any $s \in S, w \in V$ (Emission Parameter)

Example:

If we have $n=3$, $w_1 \cdots w_3$ equal to the sentence "**le professeur parle**", and $et_1 \cdots et_4$ equal to the tag sequence **DNV** STOP, then

$$\begin{aligned} & p(w_1, w_2, \cdots, w_n, et_1, et_2, \ldots, et_n) \\ &= p(\mathbf{le}, \mathbf{professeur}, \mathbf{parle}, \mathbf{D}, \mathbf{N}, \mathbf{V}, \text{STOP}) \\ &= q(\mathbf{D} \mid *, *) \times q(\mathbf{N} \mid *, \mathbf{D}) \times q(\mathbf{V} \mid \mathbf{D}, \mathbf{N}) \times q(STOP \mid \mathbf{N}, \mathbf{V}) \\ &\quad \times e(\mathbf{le} \mid \mathbf{D}) \times e(\mathbf{professeur} \mid \mathbf{N}) \times e(\mathbf{parle} \mid \mathbf{V}) \end{aligned}$$

STOP is a special tag the t terminates the sequence.

We take $et_0 = et_{-1} = *$, where $*$ is a special "padding" symbol.

Why the Nane "HIDDEN MARKOV MODEL"

$$p(x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_n) = q(\text{STOP} \mid y_{n-1}, y_n) \prod_{i=1}^{n} q(y_i \mid y_{i-2}, y_{i-1}) \prod_{i=1}^{n} e(x_i \mid y_i)$$

$q(\text{STOP} \mid y_{n-1}, y_n) \prod_{i=1}^{n} q(y_i \mid y_{i-2}, y_{i-1})$ →Markov chain

$\prod_{i=1}^{n} e(x_i \mid y_i)$ →Are observed

Parameter estimation

Learning is a necessary operation to a pattern recognition system (in particular the labeling system); it can estimate the parameters of the model. Improper or inadequate learning decreases the performance of the labeling system. To prepare the training corpus, we proceed by successive approximations. A first training corpus, relatively short, makes it possible to label a much larger corpus. This is corrected, allowing to re-estimate the probabilities, and thus serves to second learning, and so on. In general there are three estimation methods of these parameters:

- The estimation by maximum likelihood (Maximum Likelihood Estimation), it is carried by the Baum-Welch algorithm [10] or the Viterbi algorithm [11].
- The maximum estimate by post [12].
- The estimate by maximum of mutual information [13] [14].

In our case we have used the maximum likelihood estimate because it is the most used and easiest to compute.

$$q(V_t \mid D, J) = \lambda_1 \times \frac{\text{Count}(D, J, V_t)}{\text{Count}(D, J)} + \lambda_2 \times \frac{\text{Count}(J, V_t)}{\text{Count}(J)} + \lambda_3 \times \frac{\text{Count}(V_t)}{\text{Count}( )}$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 1 \quad \text{and for all } i, \quad \lambda_i \geq 0$$

$$q(\text{base} \mid V_t) = \frac{\text{Count}(V_t, \text{base})}{\text{Count}(V_t)}$$

$$\frac{\text{Count}(D, J, V_t)}{\text{Count}(D, J)} \quad \text{(Trigram)}$$

$$\frac{\text{Count}(J, V_t)}{\text{Count}(J)} \quad \text{(Bigram)}$$

$$\frac{\text{Count}(V_t)}{\text{Count}(\ )} \quad \text{(Unigram)}$$

$e(x \mid y) = 0$ for all y if x is never seen in the training data.

3.3. The Viterbi Algorithm

Problem: For an input $w_1 \cdots w_n$ find

$$\arg\max_{et_1 et_2 \cdots et_n} p(w_1, w_2, \cdots, w_n, et_1, et_2, \cdots, et_{n+1})$$

where the arg max in taken over all sequences $et_1, \cdots, et_{n+1}$ such that $et_1 \in S$ for $i = 1, \cdots, n$ and $et_{n+1} = \text{STOP}$.

We assume that p again takes the form

$$p(w_1, w_2, \cdots, w_n, et_1, et_2, \cdots, et_n) = \prod_{i=1}^{n+1} q(et_i \mid et_{i-2}, et_{i-1}) \prod_{i=1}^{n} e(w_i \mid et_i)$$

Recall that we have assumed in this definition that $et_0 = et_{-1} = *$ and $et_{i-1} = \text{STOP}$.

Algorithm:

- Define n to be the length of the sentence $(w_1, w_2, \cdots, w_n)$
- Define S_k for $k = -1, \cdots, n$ to be the set of possible tag at position k:

$$S_{-1} = S_0 = \{*\}; \quad S_k = S \quad \text{for} \quad k \in \{1, \cdots, n\} \quad \text{(for example } S = \{\text{D, N, V, P}\}\text{)}$$

- Define

$$r(et_{-1}, et_0, et_1, et_2, \cdots, et_k) = \prod_{i=1}^{n+1} q(et_i \mid et_{i-2}, et_{i-1}) \prod_{i=1}^{n} e(w_i \mid et_i)$$

- Define a dynamic programming table $\pi(k, u, v)$ = maximum probability of a tag sequence ending in tag u, v at position k that is:

$$\pi(k, u, v) = \max_{(et_{-1}, et_0, et_1, \cdots, et_k): et_{k-1} = u, et_k = v} r(et_{-1}, et_0, et_1, \cdots, et_k)$$

Example

$$S = \{\text{D, N, V, P}\}$$

$$\pi(7, P, D)$$

		D	**D**	**D**	**D**	**D**			
		N	**N**	**N**	**N**	**N**			
		V	**V**	**V**	**V**	**V**			
*	*	**P**	**P**	**P**	**P**	**P**	**P**	**D**	
		the	teacher	call	the	student	with	the	telephone
		1	2	3	4	5	6	7	8

A Recursive Definition

Base case $\pi(0, *, *) = 1$

Recursive Definition

For any $k \in \{1, \cdots, n\}$; for any $u \in S_{k-1}$ and $v \in S_k$:

$$\pi(k,u,v) = \max_{w \in S_{k-2}} \left(\pi(k-1,w,u) \times q(v \mid w,u) \times e(x_k \mid v)\right)$$

4. Experimentation

Learning Data

The experimental work was carried out in three steps:

- Setting the label set and learning corpus construction.
- Estimate the parameters of the hidden Markov model.
- Automatic labeling and re-estimation of parameters of the hidden Markov model.

The definition of the set of morphosyntactic tags is particularly delicate; this phase is carried out in collaboration with linguists. This set of labels consists of several morpho-syntactic labels. The training corpus consists of a set of sentences representing the major morphological and syntactic rules used in "Baoule" language in general.

Results

The error rate is measured on two sets (Table 1):

- Set 1 consists of the same phrases in the training set but without labels,
- Set 2 consists of phrases (without labels) different from the training set.

Note that in the case of unvowelized texts the error rate increases in relation with vowelized texts, because of the increase in ambiguity (a word can take several labels). For the remaining errors, they are due to lack of training data (there are words and transitions between labels that are not represented in the training corpus).

5. Conclusions

In analyzing the results, we noticed that the majority of labeling errors are mainly due to lack of learning problem or insufficient data. In our case there are two types of problems of lack of data:

- one or more words, part of the sentence to be labeled by this system, do not exist in the lexicon, *i.e.* we do not have an estimate observation probabilities of the words in all states.
- one or more tags have no predecessors in the sentence to be labeled automatically, *i.e.* we do not have an estimate of the transition probabilities of these labels to all other system labels.

In the continuation of our work, we shall proceed to two solutions to address these two problems.

The first is to introduce a kind of morphological analysis based on morphological forms of words to be able to identify the labels of unknown words. The second is to in-

Table 1. The automatic labeling error rate.

	Set 1	Set 2
Vowelized texts	1.82%	2.3%
Unvowelized texts	2.7%	3.5%

troduce basic syntactic rules that define the possible transitions between different labels.

That said, and given that nowhere exists to date, tagged corpus of the "Baoulé" language, it was for us, through this research to fill this gap.

References

[1] Tymian, J., Kouadio, J. and Loucou, J.-N. (2003) Language "Baoulé"-French Dictionary. NEI Abidjan.

[2] Veronis, J. (2000) Corpus Automatic Annotation: Overview and Technics State of the Art. Languages Engineering, Paris, HERMES Sciences Europe, 111-128.

[3] Vergne, J. and Giguet, E. (1998) Theoretical Perspectives on "Tagging". *Processing of 5th Conference on Natural Language Automatic* (*TALN*98), Paris, 10-12 June 1998, c

[4] Nguyen, T.M.H., Romary, L. and Vu, X.L. (2003) Vietnamese Morpho-Syntaxic Texts Tagging: A Case Study. *Processing of 5th Conference on Natural Language Automatic* (*TALN* 2003), Batz-sur-Mer, 11-14 June 2003.

[5] Paroubek, P. and Rajman, M. (2000) Morpho-Syntaxic Tagging. Languages Engineering, Paris, HERMES Sciences Europe, 131-150.

[6] Chanod, J.-P. and Tapanainen, P. (1995) Tagging French—Comparing a Statistical and a Constraint Based Method. *Proceeding of the 7th Conference of the European Chapter of the Association for Computational Linguistics* (*EACL*.95), Dublin, 149-156.

[7] De Loupy, C. (1995) The Eric Brill Tagging Method. *Revue TAL*, **36**, 37-46.

[8] Brill, E. (1992) A Simple Rule-Based Part of Speech Tagger. *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento, April, 152-155. http://dx.doi.org/10.3115/974499.974526

[9] Habert, B., Nazarenko, A. and Salem, A. (1997) The Corpus Linguistics. Armand Colin/ Masson, Paris.

[10] Baum, L. (1972) An Inequality and Association Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. *Inequality*, **3**.

[11] Celux, G. and Clairambault, J. (1992) Estimation Markov Hidden Chains: Problems and Methods. CNRS, Thematic Days on Markovian Approaches on Images and Signal, September.

[12] Bahl, L.R., Brown, P.F., de Souza, P.V. and Mercer, R.L. (1986) Maximum Mutual Information Estimation in Hidden Markov Model Parameters for Speech Recognition. *Proceedings of ICASSP*, Tokyo, 49-52.

[13] Rice, J. (2006) Mathematical Statistics and Data analysis. Thomson Learning, 511-540.

[14] Fornay, D.R. (1973) The Viterbi Algorithm. *Proceedings of the IEEE*, **61**, 268-278. http://dx.doi.org/10.1109/proc.1973.9030

Annexe

The Viterbi Algorithm

Input: $x_1 \cdots x_n$, parameters $q(s \mid u,v)$ and $e(x \mid s)$

Initialisation: Set $\pi(0,*,*)=1$

Definition: $S_{-1}=S_0=\{*\}, S_k=S$ for $k \in \{1,\cdots,n\}$

Algorithm:

BEGIN

FOR $k=1$ to n **DO**

FOR $u \in S_{k-1}, v \in S_k$ **DO**

$$\pi(k,u,v)=\max_{w \in S_{k-2}}\left(\pi(k-1,w,u)\times q(v \mid w,u)\times e(x_k \mid v)\right)$$

END

END

RETURN $\max_{u \in S_{n-1}, v \in S_n}\left(\pi(n,u,v)\times q(STOP \mid u,v)\right)$

END

The Viterbi Algorithm with Backpointers

Input: a sentence $x_1 \cdots x_n$, parameters $q(s \mid u,v)$ and $e(x \mid s)$

Initialisation: Set $\pi(0,*,*)=1$

Definition: $S_{-1}=S_0=\{*\}, S_k=S$ for $k \in \{1,\cdots,n\}$

Algorithm:

BEGIN

FOR $k=1$ **TO** n **DO**

FOR $u \in S_{k-1}, v \in S_k$ DO

$$\pi(k,u,v)=\max_{w \in S_{k-2}}\left(\pi(k-1,w,u)\times q(v \mid w,u)\times e(x_k \mid v)\right)$$

$$bp(k,u,v)=\operatorname{argmax}_{w \in S_{k-2}}\left(\pi(k-1,w,u)\times q(v \mid w,u)\times e(x_k \mid v)\right)$$

END

END

Set $(y_{n-1},y_n)=\operatorname{argmax}_{(u,v)}\left(\pi(n,u,v)\times q(STOP \mid u,v)\right)$

FOR $k=n-2$ TO 1 **DO**

$$y_k=bp(k+2,y_{k+1},y_{k+2})$$

END

RETURN The tag sequence $y_1,\cdots,y_n$

END