A MORPHOLOGICAL ANALYSIS AND SMOOTHING TECHNIQUES TO IMPROVE A STATISTICAL POS TAGGER FOR ARABIC LANGUAGE

AZZEDINE MAZROUI and NABIL ABABOU
Faculty of sciences, University Mohamed I Oujda, Morocco.
azze.mezroui@gmail.com; nabilaabou@gmail.com

ABSTRACT

In this paper, we have developed a new Part-of-Speech Tagger based on the morphological Analyzer Alkhalil Morpho Sys [12] for an analysis out of context and statistical approach using a hidden Markov model to identify the likely tag in context. We also use the Absolute discounting method to smooth the estimation of emission probabilities. Most existing statistical systems assign to each word the possible tags from the training corpus and this affects their performance. In this paper, we propose a method which assigns to the analyzed word the tags provided by the Alkhalil_Morpho_Sys Analyzer when the word is not included in the training data. We prove that these considerations will greatly reduce the error rate of the Part of Speech Tagger system.

Keywords: Natural Language Processing, Part–of-Speech tagger, Morphological Analysis, Smoothing, Training set, Testing set.

I. INTRODUCTION

Part-of-speech (POS) tagging is an essential tool in many natural language processing applications such as word sense disambiguation, parsing, question answering and machine translation.

Much work has been done in this area. The best known are based on the Markov models or symbolic systems based on rules or neural networks. These systems are able to produce only 3-4% error in English language [7], but this is not the case in the Arabic language [1,5].

Despite the various researches on the Arabic language, there are few free and generic tools for the Arabic language that could be useful to the research community. Our work focuses on building a new annotating Arabic system. This construction is based on the morphological Analyzer Alkhalil Morpho Sys [12] and on the hidden Markov models and using the Nemlar corpus in training and testing phases [3].

II. LITERATURE REVIEW

The used approaches to develop the POS tagging can be classified into three categories: the statistical approaches, the approaches based on rules and the hybrid approaches.

II.1 STATISTICAL APPROACHES

The statistical approaches [6,7,13] use a training corpus to choose the most probable tag for a given word. Most of statistical methods are based on the Markov models of order one or two.

II.2. APPROACHES BASED ON RULES

These approaches consist of developing a knowledge base of rules written by linguists to define precisely how and where to assign the various POS tags. For instance, a word following a determiner and an adjective must be a noun [2, 4,15].

III. HYBRID APPROACHES

A Hybrid approaches [9,14] combine a method based on rules and a statistical approach to assign the best tag for a given word taking into account its context. It chooses the most likely tag using a training corpus and then applies a set of rules to see whether the tag should be changed or not. It saves any new rules learnt in this process, for future use.

III. MORPHOLOGICAL ANALYZER ALKHALIL MORPHO SYS

In the next section, we will use the morphological Analyzer Alkhalil Morpho Sys. We recall its main properties.

For a given word, this Analyzer is used to determine all the possible morphological and syntactic tags of the word. It doesn’t use the context of the word. It gives for a given world the following information:
- its possible vowelized form;
- its possible proclitics and enclitics;
- its possible roots (only for derivative nouns and verbs);
- its possible syntactic natures;
- its possible vowelized patterns (only for derivative nouns and verbs);
- its possible lemmas and their corresponding patterns;
- its possible syntactic forms.

IV. METHOD DESCRIPTION

The developed system goes through two steps as shown in Figure 1 below.
We give in what follows a detailed description of these steps.

**IV.1 First step: Possible tags for a word W**

Given a word W of a sentence Ph, we identify in this step the possible tags for W by performing the following two stages:

1. First stage: using a training corpus, we identify all possible tags of W.
2. Second stage: we use the morphological Analyzer Alkhali Morpho Sys to determine the possible tags of W.

The tags of the word W out of context are those obtained from these two stages. If the word W doesn’t belong to the training corpus and is not analyzed by Alkhali Analyzer, we use the information of its proclitic and enclitic to find the most likely Tag.

**IV.2 Second step: Statistical analysis**

After identifying the possible tags for each word of the sentence, the statistical step provides the most likely tag for each word of the sentence. It is based on the hidden Markov model, the smoothing techniques and the Viterbi algorithm [10]. In what follows, we give a brief overview of these mathematical concepts.

**IV.2.1. The Hidden Markov Models**

A hidden Markov model (HMM) is a statistical tool used to model a double Markov process one of which is observable and the other is hidden.

From the information on the observed data, we can estimate these hidden states. Indeed, the observation frequency depends on the state of the process in which they are issued.

Let O = \{o1, ..., on\} be a finite set of observations and E = \{e1, ..., em\} be a finite set of hidden states (unobserved).

Definition: A first-order HMM is a double process (Xp,Yp)1≤i≤1 where:

- (Xp)1 is a homogeneous Markov chain with values in the set of the hidden states E, i.e.:
  \[ \Pr(X_{i+1} = e_j / X_i = e_i, ..., X_1 = e_0) = \Pr(e_i) \]
  where:
  \[ a_{ij} = \Pr(e_i / e_j) \]  
  aij is the transition probability from the state e_i to the state e_j and A = (a_{ij}) is the transition matrix.

- (Yp)1 is an observable process taking values in the observations set O where:
  \[ \Pr(Y_i = o_k / X_i = s_i, ..., X_1 = s_0) = \Pr(s_i) \]
  \[ = \Pr(Y_i = o_k / X_i = s_j) = b_{ik} \]  
  bi(k) is the probability of observing o_k given the state sj and B = (b_{ik}) is the emission matrix.

In the following, we assume that the observations set consists of Arabic words while the hidden states set consists of Arabic Tags.

Let Ph be an observed sentence consisting of the words w1, w2, ..., wn, and \( \mathbb{E} = \{e_1, e_2, ..., e_n\} \) be the set of Arabic tags. Our goal is to find the most likely sequence of tags \( (e_1, e_2, ..., e_n) \) of the sequence of words \( (w_1, w_2, ..., w_n) \). This goal can be formulated as follows:

\[ (e_1, e_2, ..., e_n) = \arg\max_{e_1, e_2, ..., e_n} \Pr(e_1, ..., e_n) / \Pr(w_1, ..., w_n) \]  
where \( \mathbb{E} \) is the set of possible tags of the word w obtained in the first step.

**IV.2.2. The Viterbi Algorithm**

We can easily prove that (3) is equivalent to:

\[ (e_1', ..., e_n') = \arg\max_{e_1, e_2, ..., e_n} \Pr(e_1', ..., e_n') \]  
(4)

The Viterbi algorithm is well suited for finding the most likely path. It is based on the following functions:

\[ \phi(t, e_t') = \max_{e_{\theta(0)}', ..., e_{\theta(t)}'} \Pr(w_1, ..., w_t, e_t' \mid e_{\theta(0)}', ..., e_{\theta(t)}') \times \Pr(e_{\theta(0)}', ..., e_{\theta(t)}') \]  
\[ \psi(t, e_t') = \arg\max_{e_{\theta(0)}', ..., e_{\theta(t)}'} \phi(t, e_{\theta(0)}', ..., e_{\theta(t)}') \]  

\( \psi(t, e_t') \) is the probability of the most likely partial path until time t, and ending at the tag \( e_{\theta(t)}' \) (\( e_{\theta(t)}' \) belongs to the possible tags of the word w obtained in the morphological step).
ψ(t,e* \text{t}) allows storing at time \text{t} the tag e* \text{t} that achieves the optimal path reaching this tag.

Applying the Markov properties (1) and (2), we show that the expression of \( \phi \) is equivalent to:

\[
\phi(t,e^k) = \max_{e^j \in E^t} \mathcal{Pr} \left( e^j \middle| e^i \right) \mathcal{Pr} \left( e^i \right) \quad (5)
\]

This equation will allow us to find, recursively, the values of \( \phi \).

IV.2.3. SMOOTHING METHOD

To identify the most likely tags of the sentence words, the parameters of the statistical model, i.e. the transition matrix \( A = (a_{ij}) \) and the emission matrix \( B = (b(i)) \) will be estimated from a training corpus. The estimation method used is based on the maximum likelihood [8]. Thus, if \( w \) is an unvowelized word (observed state) and \( e \) and \( e^j \) are two tags (hidden states), we note:

- \( c(e,e^j) \): the number of occurrences in the training corpus of the transition from the hidden state \( e \) to the hidden state \( e^j \).
- \( c(w,e) \): the number of occurrences in the training corpus of the hidden state \( e \).
- \( c(w,e^j) \): the number of occurrences in the training corpus of the unvowelized word \( w \) with the tag \( e^j \) in the training corpus.

Then, the coefficients \( a_{ij} \) and \( b(i) \) will be approximated using the following equations:

\[
a_{ij} = \frac{c(e,e^j)}{c(e^i)} \quad 1 \leq i \leq N \quad 1 \leq j \leq N \quad (6)
\]

\[
b(i) = \frac{c(w,e^j)}{c(e^j)} \quad 1 \leq t \leq M \quad 1 \leq i \leq N \quad (7)
\]

Since in some cases the correct tag \( e^j \) of the word \( w \) is not included in the training data (i.e., \( c(w,e^j) = 0 \)), the Viterbi algorithm can miss the optimal path. To circumvent this problem, we conducted a smoothing technique for estimating the parameters \( b(i) \). We have tested several smoothing methods and the best performances are obtained with the Absolute discounting smoothing method [11].

For a given word \( w \) and a possible tag \( e^j \) obtained by the first step, the estimation of \( b(i) = P(w|e^j) \) is given by:

\[
b(i,t) = \left\{ \begin{array}{ll}
\frac{c(w,e^j) - d}{c(e^j)} & \text{if } c(w,e^j) \neq 0 \\
\frac{N - d}{c(e^j)} & \text{otherwise}
\end{array} \right.
\]

where, \( N \) = the number of words that appear in the corpus with the tag \( e^i \), \( z = \) the number of words that are not annotated in the corpus with the tag \( e^i \) and for which the Alkhalil Analyzer generates this tag, \( d = \) a value which is chosen between zero and one.

We give an example to illustrate the effect of the smoothing process.

Example: we assume that we have a small corpus \( V = \{ w_1, w_2, w_3, w_4, w_5, w_6, w_7 \} \) in which only the words \( w_1, w_2, w_3 \) appear with the noun tag, and Alkhalil gives the noun tag for words \( w_4, w_5 \).

Table1: The estimation of emission probabilities without the smoothing process

<table>
<thead>
<tr>
<th>( w_i/noun )</th>
<th>Frequency in Corpus</th>
<th>Alkhalil</th>
<th>Estimation of ( P(w_i/noun) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w_1/noun )</td>
<td>3</td>
<td>3/10 = 0.3</td>
<td></td>
</tr>
<tr>
<td>( w_2/noun )</td>
<td>3</td>
<td>3/10 = 0.3</td>
<td></td>
</tr>
<tr>
<td>( w_3/noun )</td>
<td>4</td>
<td>4/10 = 0.4</td>
<td></td>
</tr>
<tr>
<td>( w_4/noun )</td>
<td>0</td>
<td>True</td>
<td></td>
</tr>
<tr>
<td>( w_5/noun )</td>
<td>0</td>
<td>True</td>
<td></td>
</tr>
<tr>
<td>( w_6/noun )</td>
<td>0</td>
<td>False</td>
<td></td>
</tr>
<tr>
<td>( w_7/noun )</td>
<td>0</td>
<td>False</td>
<td></td>
</tr>
</tbody>
</table>

This table represents the estimation of emission probabilities before using the Absolute discounting smoothing method. After smoothing with \( d = 0.1 \), we get the values shown in the following table:

Table2: The estimation of emission probabilities after smoothing

<table>
<thead>
<tr>
<th>( w_i/noun )</th>
<th>Estimation of ( P(w_i/noun) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w_1/noun )</td>
<td>( (3/10)-(0.1/10) = 0.29 )</td>
</tr>
<tr>
<td>( w_2/noun )</td>
<td>( (3/10)-(0.1/10) = 0.29 )</td>
</tr>
<tr>
<td>( w_3/noun )</td>
<td>( (4/10)-(0.1/10) = 0.39 )</td>
</tr>
<tr>
<td>( w_4/noun )</td>
<td>( (0.1/10) \times 3/2 = 0.15 )</td>
</tr>
<tr>
<td>( w_5/noun )</td>
<td>( (0.1/10) \times 3/2 = 0.15 )</td>
</tr>
<tr>
<td>( w_6/noun )</td>
<td>0</td>
</tr>
<tr>
<td>( w_7/noun )</td>
<td>0</td>
</tr>
</tbody>
</table>

We see clearly that there was a redistribution of probabilities over all events.

V. SET OF TAGS

Our system uses Alkhalil Analyzer in the first phase of the system and the Nemlar corpus in the training and the testing phases. Since the tag set used in the Nemlar corpus is not identical to that used by the Alkhalil Analyzer, we have to choose a common set of tags between the Nemlar corpus and Alkhalil Analyzer. The following table shows the tag set of our system.

Table3: The set of tags used by our system

<table>
<thead>
<tr>
<th>Category</th>
<th>Arabic Term</th>
<th>English Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>أفعال</td>
<td>فعل</td>
<td>VERB</td>
</tr>
<tr>
<td>اسم عام</td>
<td>common noun</td>
<td>COMMON NOUN</td>
</tr>
<tr>
<td>اسم علم</td>
<td>common noun</td>
<td>COMMON NOUN</td>
</tr>
<tr>
<td>اسم موصول</td>
<td>familiar noun</td>
<td>FAMILIAR NOUN</td>
</tr>
<tr>
<td>اسم إشارة</td>
<td>pronoun</td>
<td>PRONOUN</td>
</tr>
<tr>
<td>الأدوات</td>
<td>tools</td>
<td>TOOLS</td>
</tr>
<tr>
<td>جملة</td>
<td>sentence</td>
<td>SENTENCE</td>
</tr>
<tr>
<td>جمل محدودة</td>
<td>finite sentence</td>
<td>FINITE SENTENCE</td>
</tr>
<tr>
<td>جمل غير محدودة</td>
<td>non-finite sentence</td>
<td>NON-FINITE SENTENCE</td>
</tr>
<tr>
<td>أدوات</td>
<td>tools</td>
<td>TOOLS</td>
</tr>
<tr>
<td>المرکبات</td>
<td>compounds</td>
<td>COMPOUNDS</td>
</tr>
</tbody>
</table>

VI. TESTING AND EVALUATION

We evaluate the contribution of the use of the Alkhalil Analyzer in the first phase and that of the smoothing method in the statistical phase. Indeed, we initially
calculate the accuracy of the system using just the training corpus to identify the possible tags of words and the formulas (6) and (7) to estimate the transition and emission probabilities. Then, we integrate the Alkhalil Analyzer in the first phase of the system and we use the smoothing method to estimate the emission probabilities and subsequently calculate the accuracy of the system.

To evaluate the system, we used Nemlar corpus [3]. It contains 500,000 words previously annotated. We used 90% of this corpus randomly selected as training corpus and the remaining 10% as testing corpus. The accuracy of the system is the ratio between the number of words of the testing corpus well labelled and the size of the test corpus. We present in Table 4 the obtained results.

Table 4: The values of accuracies for both tests

<table>
<thead>
<tr>
<th></th>
<th>System without Alkhalil Analyser and smoothing method</th>
<th>System with Alkhalil Analyser and smoothing method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>86.59%</td>
<td>94.63%</td>
</tr>
<tr>
<td>Error rate</td>
<td>13.41%</td>
<td>5.37%</td>
</tr>
</tbody>
</table>

We note that the use of the Alkhalil Analyzer and the smoothing method has allowed a significant improvement in system performance since the recognition rate increased by 8 points to exceed 94%.

VII. CONCLUSION

The results show that the use of the Alkhalil Analyzer and smoothing method improves significantly the performances of our system. We expect in the future to apply our approach to a large corpus.

References


