Morpho-Syntactic Tagging System Based on the Patterns Words for Arabic Texts

Abdelhamid El-Jihad[1], Abdellah Yousfi[2], and Aouragh Si-Lhoussain[3]
[1]Institute for Studies and Research on arabization, Rabat, Morocco
[2]University Mohamad V Suissi, Rabat, Morocco
[3]University Mohamad I-Oujda, Morocco

Abstract: *Text tagging is a very important tool for various applications in natural language processing, namely the morphological and syntactic analysis of texts, indexation and information retrieval, "vocalization" of Arabic texts, and probabilistic language model (n-class model). However, these systems based on the lexemes of limited size, are unable to treat unknown words consequently. To overcome this problem, we developed in this paper, a new system based on the patterns of unknown words and the hidden Markov model. The experiments are carried out in the set of labeled texts, the set of 3800 patterns, and the 52 tags of morpho-syntactic nature, to estimate the parameters of the new model HMM.*

Keywords: *Hidden markov model, morpho-syntactic tagging, Arabic text, and pattern.*

Received September 22, 2008; accepted May 17, 2009

1. Introduction

Most speech taggers research have been done specifically for Indo-European languages and a very little work for the Arabic language. Among these latters, we can mention:

- Khoja [11] has developed an Arabic part-of-speech tagger that uses 131 tags derived from traditional Arabic grammatical theory.
- Diab [4] has developed a Support Vector Machine (SVM) based approach to automatically tokenize part-of-speech tag and annotate base phrase in Arabic text.
- And others [5, 7, 17].

In this paper, we have developed a new system based on the patterns of unknown words and the Hidden Markov Model (HMM). The originality of this work is to overcome the deficiency observed in other systems that do not treat these unknown words. The experiments are carried using tagged texts, a set of 3800 patterns, and 52 tags of morpho-syntactic nature, to estimate the parameters of the new HMM model.

The automatic tagging of texts is a process that consists of assigning morphological, syntactic, semantic, prosodic, critical, morpho-syntactic information to segment of texts (generally a word) with [14, 15]. It consists of three steps [12, 13]:

- Segmentation of the text into lexemes.
- Tagging that assigns for each identified lexeme the whole of the possible morpho-syntactic tags.
- Disambiguation: in trivial cases there will be only one tag per word.

Generally, there are two main approaches to part-of-speech tagging: rule-based tagging and probabilistic. Among the difficulties which arise in the systems of tagging are the unknown words. In this paper we developed an approach to solve the problem of the unknown words by using the notion of word patterns. This approach is integrated into the morpho-syntactic system of tagging based on the model of hidden Markov, developed at the IERA [5].

1.1. Arabic Word Patterns

The Arabic language has a particular characteristic, namely verbs and derived nouns can be classified into patterns. The construction of a word pattern is given by the procedure described in [1] which extracts the radical letters composing its root from the word, then replaced as follows: the first radical letter is replaced by "ف", the second letter by "ع", the third by "ل" and the fourth by "ل".
Example:

- The patterns word of "تَدَحْرَجَ" is "تَفَعْلَلَ".
- The patterns word of "كَتَبْتُ" is "فَعَلْتُ".

In Arabic, every word has many patterns[1]. For the research of the pattern word, we developed a measure D, which measures the degree of similarity between the word and each pattern.

[1] For example the word "نَقُولُ" has two patterns :
- نَقُولُ → قَالَ/يَقُولُ
- فَعُولُ → نَقَلَ/يَنْقُلُ

$F = \{f_1, f_2, ..., f_N\}$ is the set of all patterns and *w* is a word. The set of patterns of *w* is given by:

$$F(w) = \{f_i \ / \ D(w, f_i) \geq 0\} \quad (1)$$

1.2. Tagging by Probabilistic Method

The choice of the most probable tag of a word is made in comparison with the history of the last tags which have been just affected. This history is limited to one or two previous tags. The probabilistic method supposes that one has a sufficient training corpus to allow a reliable estimation of the probabilities [8]. Let: $ph = w_1 ... w_p$ a sentence of words: $w_1, ..., w_p$, in the vocabulary *V*. $E = \{et_1, ..., et_N\}$ the set of morpho-syntax tags. The morpho-syntactic tagging of the *ph* by tags in *E* which is based on the probabilistic approach consist to find a set of tags $et_1^*, ..., et_p^*$ associated with the sentence *ph*, such as:

$$et_1^*, ..., et_p^* = \arg \max_{et_1, ..., et_p} \Pr(w_1, ..., w_p, et_1, ..., et_p) \quad (2)$$

The problem which arises in this formulas is words that do not exist in *V*. To solve the equation 2 by tagging into account of this problem, we adapted the Hidden Markov Model by introducing the patterns notion of unknown words.

1.3. Morpho-Syntactic Tagging by Using Word Patterns

The HMM model of order 1 by considering the word patterns, is a process with:

- X_t is a Markov chain of order 1, with value is in a finished set of states $Q = \{q_1, ..., q_N\}$, X_t checks:

$$\begin{aligned} &\Pr(X_{t+1} = q_j \ / \ X_1 = q_1, ..., X_t = q_i) \\ &= \Pr(X_{t+1} = q_j \ / \ X_t = q_i) = a_{ij} \\ &\Pr(X_1 = q_i) = \pi_i \quad i = 1, ..., N \end{aligned} \quad (3)$$

  a_{ij} is the transition probability between q_i and q_j., π_i is the probability that q_i is an initial state.

- Y_t is an observable process with values in a measurable unit *Y*, Y_t checks:

$$\begin{aligned} &\Pr(Y_t = y_t \ / \ X_1 = q_1, ..., X_t = q_i, Y_1 = y_1, ..., Y_{t-1} = y_{t-1}) = \\ &= \Pr(Y_t = y_t \ / \ X_t = q_i) = b_i(y_t) = b_{it} \end{aligned} \quad (4)$$

  b_{it} is the emission probability of the observation y_t from state q_i.

- Z_t is an observable process with values in a measurable set *Z*, Z_t checks:

$$\begin{aligned} &\Pr(Z_t = z_t \ / \ X_1 = q_1, ..., X_t = q_i, Z_1 = z_1, ..., Z_{t-1} = z_{t-1}) = \\ &= \Pr(Z_t = z_t \ / \ X_t = q_i) = d_i(z_t) = d_{it} \end{aligned} \quad (5)$$

  d_{it} is the emission probability of the observation z_t starting from state q_i .

In the continuation one, we will suppose that the process: (*X*, $Y_{t\,t}$, Z_t,) is HMM of order 1:

- $X_t = et_{it}$ representing the morpho-syntactic tag, with value is in *E*.
- $Y_t = w_t$ representing the words of the vocabulary $V = \{w_1, ..., w_L\}$,
- $Z_t = f_t$ representing the patterns words.

Note: this model is defined by a parameter vector □ = *(□, A, B, D)* where:

- $\Pi = \{\pi_1, ..., \pi_N\}$ The set of the initial probability.
- A = *(a*$(a_{ij})_{1 □ \ i,j □ \ N}$: the matrix of the transition probabilities.
- B = $(b_{it})_{1 □ i □ N \ and \ 1 □ t □ L}$: the matrix of the emission probabilities of the words from the states.
- D = $(d_{it})_{1 □ i □ N \ and \ 1 □ t □ L}$: the matrix of the emission probabilities of the patterns words from the states.

1.4. Learning Procedure (Parameters of Estimation)

Training is a necessary operation for a tagging system which makes to estimate the parameters of the model □ = *(□, A, B, D).* An incorrect or insufficient training decreases the performance of the tagging system. In general there are three methods to estimate these parameters [16]:

- Estimation by the likelihood maximum which is carried out by the algorithm of Baum-Welch [2] or Viterbi algorithm [3].
- Estimation by the maximum a posteriori [9].
- Estimation by a mutual information maximum [10].

In our case, we used the estimation by the Likelihood Maximum. If we take the training set $R = \{ph_1, ..., ph_k\}$ composed by the tagged sentences $ph_1, ..., ph_k$. . The Formulas of estimate the parameters of □ = *(□, A, B, D)*, are given by:

$$a_{ij} = \frac{\sum_{n=1}^{k} \text{number of transition } et_i \ et_j \text{ in the sentence } ph_n}{\sum_{n=1}^{k} \text{number of state } et_i \text{ in } ph_n}$$

$$\pi_i = \frac{\sum_{n=1}^{k} \delta[et_i \text{ is initial state}]}{k}$$

$$b_{it} = \frac{\sum_{n=1}^{k} \text{number of word } w_t \text{ are the tag } et_i \text{ in } ph_n}{\sum_{n=1}^{k} \text{number of state } et_i \text{ in } ph_n}$$

$$d_{it} \frac{\sum_{n=1}^{k} \text{number of the pattern } f_t \text{ are the tag } et_i \text{ in } ph_n}{\sum_{n=1}^{k} \text{nuumber of state } et_i \text{ in } ph_n}$$

with:

$$\delta[x] = \begin{cases} 1 & \text{if the evenement } x \text{ is true} \\ 0 & \text{else} \end{cases}$$

1.5. Automatic Tagging by the Adapted Viterbi Algorithme

For a faster calculation of the equation, we have adapted the Viterbi algorithm [6] to solve this equation 2. We note by:

$$\delta_t(et_j) = \max_{et_{i_1} \ldots et_{i_t}} \Pr(w_1 \ldots w_t, et_{i_1} \ldots et_{i_t}) \tag{6}$$

$$\textit{with} \quad et_{i_t} = et_j$$

To solve the problem of the unknown words, we have introduced the process of the patterns of these words into the formula 6. This formula becomes [16]:

$$\delta_t(et_j) = \begin{cases} \max\limits_{et_i} \delta_{t-1}(et_i).a_{ij}.b_j(w_t) & if \quad w_t \in V \\ \max\limits_{et_i, f_k \in F(w_t)} \delta_{t-1}(et_i).a_{ij}.d_j(f_k) & else \end{cases} \tag{7}$$

$F(w_t)$: The set of all possible patterns of w_t. We calculate this formula for all the values *t = 1..., T* and *j=1...,N*. At the end, the optimal path is obtained by using a recursive calculation of this formula.

1.6. Experimental Results

The Experimental work was completed in four major steps:

- Step of definition of the set of tag and construction of the corpus of training. The definition of our own morpho-syntactic set of tag was particularly delicate; this phase was carried out in collaboration with linguists to satisfy the need for the projects under development at IERA. This set of tag consists of 52 tags of morpho-syntactic nature which are illustrated in Table1:

Table 1. Example of some morpho-syntactic tags used in our system.

Significations	Tags
Prefix	س.ب
Active Participle	إ.ف
Genetive Particle	ح.ج
Exception Particle	ح.إ
Proper Name	إ.ع
Deiclic Particle	إ.إ
Exaggeration Noun	إ.غ

The training corpus is composed of a whole of sentences representing the major morphological and syntactic rules used in Arabic. This corpus was tagged manually by linguists.

أَكْرَمْتُ/ف.م.م.م ال/س.ب مُجْتَهِدِينَ/إ.ف ./.
أَحْسَنْتُ/ف.م.م.م إِلَى/ح.ج ال/س.ب مُجْتَهِدِينَ/إ.ف ./.
جَاءَ/ف.م.م.م ال/س.ب مُسَافِرُونَ/إ.ف إِلَّا/ح.إ سَعِيدًا/إ.ع ./.
قَلَّمَا/ف.ج فَعَلْتُ/ف.م.م.م هَذَا/إ.إ ./.

Figure 1. Example of an extract from our learning corpus.

Note: the present system enables to tag the Arabic vowelled texts. To tag the Arabic unvowelled texts, an unvowelled training corpora is used.

- Step of construction of the data base of the forms of the words. This base consists of 3800 patterns, and is generated by a morphological generator of verbs developed at the IERA. This base has been enriched by the patterns of derived names as shown in Table 2.
- Step of estimate the parameters of adapted hidden Markov model.

Automatic step of tagging and reestimation of the parameters of hidden Markov model. To carry out these two last steps, we have developed an application in C programming language which is based on three modules:

- Module of determination of the form of a given word.
- Module which makes it possible to carry out the phase of training.

Table 2. Example of an extract from the database patterns words.

فاعل	إ.ف
متفاعل	إ.ف
إفعال	م.ص
تفعيل	م.ص
مفعول	إ.م
مفعل	إ.م
فعلت	ف.م.م.م
أفعل	ف.ض.م.م

- Automatic module of tagging of rough corpus. The latter is corrected manually to be used for a reappraisal of the parameters of the model of adapted hidden Markov. The programs are evaluated on the basis of version of text. Results the error rate is measured on a whole of test containing 500 sentences and it is shown in Table 3.

Table 3. The error rate on the entire test to the old system.

	Ensemble Test
Old System (HMM)	3%

The error rate is measured on a test set containing 500 sentences. The latter are a sequence of words involved in the old system vocabulary [5]. 2% of these errors are the result of a bad assignment of tags by the system. 1% comes from problems of unpresented transition between tags within training set.

Table 4.The error rate on the entire test for the new system.

	Ensemble Test
New System (Adapted HMM)	1.78%

For the new system, the error rate is measured on a test set consisting of 500 sentences containing unknown word (99.14% of the sentences containing unknown word have been correctly tagged). 0.92% of these errors come from problems of unpresented transitions between tags in the training set. The remaining of mistakes is the result of the incorrect tagging.

2. Conclusions and Perspectives

By analyzing the results, we have note that the introduction of the concept of the patterns words has decreased the effect of the problem of unknown words. Moreover this new system has successfully labelled 1.78% sentences containing these unknown words.

As for the perspectives of this work, we intend to introduce syntactic rules in our system to address the problem of transitions.

References

[1] Alghalayni M., *Collection des Leçons Arabe*, Librairie Moderne, 2000.

[2] Baum L., "An Inequality and Association Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes," *Computer Journal of Inequality*, vol. 3, no. 3, pp. 145-147, 1972.

[3] Celux G. and Clairambault J., "Estimation de Chaînes de Markov Cachées: Méthodes et Problèmes," *Journées Mathématiques Sur Les Approches Markoviennes en Signal et Images*, CNRS-Paris, vol. 6, no. 3, pp. 5-19, 1992.

[4] Diab M., Kadri H., and Daniel J., "Automatic Tagging of Arabic Text: from Raw Text to Base Phrase Chunks," *in Proceedings of HLT-NAACL*, pp. 44-49, 2004.

[5] EL-Jihad A. and Yousfi A., "Etiquetage Morpho-Syntaxique des Textes Arabes par Modèle de Markov Caché," *in Proceedings of Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, France, pp. 649-654, 2005.

[6] Fornay D., "The Viterbi Algorithm," *in Proceedings of IEEE*, pp. 268-278, 1973.

[7] Habash N. and Rambow O., "Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop," *in Proceedings of the 43rd Meeting of the Association for Computational Linguistics,* pp. 258-263, 2005.

[8] Habert B., Nazarenko A., and Salem A., *Les Linguistiques de Corpus*, Armand colin/Masson, Paris, 1997.

[9] John R., *Mathematical Statistics and Data Analysis*, Duxbury Press, California, 2007.

[10] Kapadia S., Valtchev V., and Young S., "MMI Training for Continuous Phoneme Recognition on the TIMIT Database," *in Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, pp. 491-494, 1993.

[11] Khoja S., "APT: Arabic Part-of-Speech Tagger," *in Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, Pennsylvania, 2001.

[12] Paroubek P., and Martin R., "Etiquetage Morpho-Syntaxique," *Ingénierie des langues,* Paris, pp.131-150, 2000.

[13] Huyen T., Laurent R., and Xuan L., "Une Etude de Cas Pour L'étiquetage Morpho-Syntaxique de Textes Vietnamiens," *in Proceedings of Traitement Automatique des Langues Naturelles*, pp. 123-126, 2003.

[14] Vergne J. and Emmanuel G., "Regards Théoriques sur le Tagging," *in Proceedings of Traitement Automatique des Langues Naturelles*, France, pp. 22-31, 1998.

[15] Veronis J., "Annotation Automatique De Corpus: Panorama et Etat de la Technique," *in Proceedings of Ingénierie des langues*, Paris, pp. 111-128, 2000.

[16] Yousfi A., "Introduction de la Vitesse d'élocution dans un Modèle de Reconnaissance Automatique de la Parole," *Thèse de Doctorat*, Université Mohamed Premier, Maroc, 115 p, 2001.

[17] Zribi C., Torjmen A., and Ben Ahmed M., "An Efficient Multi-Agent System Combining POS-Taggers for Arabic Texts," *in Proceedings of Lecture Notes in Computer Science*, pp. 121-131, 2006.

Abdelhamid El-Jihad is currently working in the Department of Arabic Language Processing, Institute for Studies and Research on Arabization, Vth Mohamed University Suissi, Rabat, Morocco. He holds his PhD degree in electronic and computer science from the Faculty of Sciences, Vth Mohamed University Agdal, Rabat, Morocco, 1993. His research interests are in the area of Arabic computational linguistic.

Abdellah Yousfi is currently working in Vth Mohamed University Souissi, Rabat, Morocco. His research interests are in the area of automatic labeling of Arabic texts, morphological analysis, and information retrieval. He holds his PhD degree in automatic speech recognition in Faculty of Science, First Mohammad Univesity, Oujda, Morocco, 2001. His research interests include automatic labeling of Arabic texts, morphological analysis, and information retrieval.

Aouragh Si-Lhoussain holds his PhD degree on natural language processing in Faculty of Science, Oujda, Morocco, 2009. His research interests are in the area of statistical language model, part-of-speech tagger.