Probabilistic Approaches to Natural Language

Papers from the 1992 Fall Symposium
Technical Report FS-92-04

AAAI Press
American Association for Artificial Intelligence

Table of Contents

A nonstationary hidden Markov model with a hard capture of observations: application to the problem of morphological ambiguities*

Djamel Bouchaffra and Jacques Rouault
Centre de Recherche en Informatique appliquée aux Sciences Sociales B.P. 47, 38040 Grenoble Cedex 9, FRANCE.
E-mail : djamel@criss.fr
Phone : (+33) 76.82.56.94.
Fax : (+33) 76.82.56.75.

Abstract

This correspondence is concerned with the problem of morphological ambiguities using a Markov process. The problem here is to eliminate interferent solutions that might be derived from a morphological analysis. We start by using a Markov chain with one long sequence of transitions. In this model the states are the morphological features and a sequence corresponds to the transition of a word form from one feature to another. After having observed an inadequacy of this model, one will explore a nonstationary hidden Markov process. Among the main advantages of this latter model we have the possibility to assign a type to a text given some training samples. Therefore, a recognition of "style" or a creation of a new one might be developed.

1. Introduction

1.1. Automatic analysis of natural language

This work lies within a textual analysis system in natural language discourse (french in our case). In most systems used today, the analysis process is divided into *levels*, starting from morphology (first level) through syntax, and semantics to pragmatics. These levels are sequentially activated, without backtracking, originating in the morphological phase and ending in the pragmatic phase. Therefore, the i-th level knows only the results of preceding levels; in particular, the morphological analysis works without any reference to the other levels. This means that each word in the text *(a form)* is analyzed autonomously out of context. Hence, for each form, one is obliged to consider all possible analyses.

* Much of the research on which this paper is based was carried out in order to be used in the framework of MMI^2 (A Multi-Modal Interface for Man Machine Interaction). This interface is part of a project partially funded by the Commission of the European Communities ESPRIT program.

Example: let's consider the sequence of the two forms "cut" and "down":
- "cut" can be given 3 analyses: verb, noun, adjective;
- "down" can be a verb, an adverb or a noun.
- The number of possible combinations based upon the independence of the analysis of one form in relation with the others implies that the phrase "cut down" is liable to *nine* interpretations; whatever its context.

These multiple solutions are transmitted to syntactic parsing which doesn't eliminate them either. In fact, as a syntactic parser generates its own interferent analyses, often from interferent morphology analyses, the problems with which we are confronted are far from being solved.

In order to provide a solution to these problems, we have recourse to statistical methods. Thus, the result of the morphological analysis is filtered when using a Markov model.

1.2. Morphological analysis

A morphological analyser must be able to cut up a word form into smaller components and to interpret this action. The easiest segmentation of a word form consists in separating word terminations (flexional endings) from the rest of the word form called *basis*. We have then got *a flexional morphology*. A more accurate cutting up consists in splitting up the basis into affixes *(prefixes, suffixes)* and *root*. This is then called *derivational morphology*.

The interpretation consists in associating the segmentation of a word form with a set of information particulary including:
- the general morphological class: verb, noun-adjective, preposition, ...
- the values of relevant morphological variables: number, gender, tense,...

Therefore, an interpretation is a class plus values of variables; such a combination is called *a feature*. Note that a word form is associated with several features in case there are multiple solutions.

1.3. Why statistical procedures ?

Because of the independence of the analysis levels, it is difficult to provide contextual linguistic rules. This is one of the reasons why we fall back on statistical methods. These latter methods possess another advantage: they reflect simultaneously language properties, eg., the impossibility to obtain a determinant followed directly by a verb, and properties of the analysed corpus, eg., a large number of nominal sentences.

Some researchers used Bayesian approaches to solve the problem of morphological ambiguities. However, these methods have a clear conceptual framework and powerful representations, but must still be knowledge-engineered, rather than trained. Very often in

the application of those methods researchers have not a good observation of the individuals of the population, *because the observation is a relative notion*. Therefore, we have difficulty in observing possible transitions of these individuals. The way of "capturing" the individuals depends on the environment encountered.

2. A morphological features Markov chain

2.1. The semantic of the model

Let (f_i) (i = 1 to m) be the states or morphological features, we have only one individual (n =1) for each transition time $t \in \{1,2,...,T\}$. A first-order m-state Markov chain is defined by an m*m state transition probability matrix P, and an m*1 initial probability vector Π, where

$P = \{P_{fifj}\}$, $P_{fifj} = Prob[e_{t+1} = f_j / e_t = f_i]$, i,j = 1,2,...,m,

$\Pi = \{\Pi_{fi}\}$, $\Pi_{fi} = Prob[e_1 = f_i]$, i = 1,2,...,m,

$E = \{e_t\}$ is a morphological features sequence, t = 1,2,...,T.

The parameters m and T are respectively the number of states and the length of state sequence. By definition, we have

$$\sum_{j=1}^{j=m} P_{f_i f_j} = 1 \quad \text{for } i = 1,2,...,m \quad \text{and} \quad \sum_{k=1}^{k=m} \Pi_{f_k} = 1.$$

The probability associated to a realization E of this Markov chain is

$$Prob\,[E / P,\Pi] = \Pi_{e_1} * \prod_{t=2}^{t=T} P_{e_{t-1}e_t}.$$

2.2. Estimation of transition probabilities

As pointed out by Bartlett in Anderson and Goodman [1], the asymptotic theory must be considered with respect to the variable "number of times of observing the word form in a single sequence of transitions" instead of the variable "number of individuals in a state when T is fixed". However, this asymptotic theory was considered because the number of times of observing the word form increases ($T \to \infty$). Furthermore, we cannot investigate the stationarity properties of the Markov process since we only have one word form (one individual) at each transition time. Therefore, we assumed stationarity. Thus, if N_{fifj} is the number of times that the observed word form was in the feature f_i at time

(t - 1) and in the feature fj at time t for all t ∈ {1..T}, then the estimates of the transition probabilities are

$$P_{f_i f_j} = \frac{N_{f_i f_j}}{N_{f_i +}},$$

where N_{fi+} is the number of times that the word form was in state f_i. The estimated transition probabilities are evaluated on one training sample. We removed the morphological ambiguities by choosing the sequence E of higher probability.

3. A Markov model with hidden states and observations

The inadequacy of the previous model to remove certain morphological ambiguities has led us to believe that some unknown hidden states govern the distribution of the morphological features. Instead of passing from one morphological feature to another we might transit from a hidden state to another. Besides, in section 2 we focused only on the surface of one random sample, i.e., an observation was a morphological feature. As pointed out in [4], this latter entity cannot be extracted without a context effect in a sample. In order to consider this context effect, we have chosen criteria like "the nature of the feature", "its successor feature", "its predecessor feature", "its position in a sentence" and "the position of this sentence in the text". An observation o_i is then a "known hidden vector" whose components are values of the criteria presented here. However, one can explore other criteria.

Definition 3.1. A hidden Markov model (HMM) is a Markov chain whose states cannot be observed directly but only through a sequence of observation vectors.

An HMM is represented by the state transition probability P, the initial state probability vector Π and a (T*K) matrix V whose elements are the conditional densities $v_i(o_t)$ = density of observation o_t given $e_t = i$, K is the number of states. Our aim is the determination of the optimal model estimate $\vartheta^* = [\Pi^*, P^*, V^*]$ given a certain number of samples, this is the training problem.

Theorem 3.2. *The probability of a sample $S = \{o_1, o_2, ..., o_T\}$ given a model ϑ can be written as:*

$$Pr(S/\vartheta) = \sum_{E} \Pi_{e_1} v_{e_1}(o_1) * \prod_{t=2}^{t=T} P_{e_{t-1}e_t} v_{e_t}(o_t).$$

Proof. For a fixed state sequence $E = e_1,e_2,...,e_T$, the probability of the observation sequence $S = o_1,o_2,...o_T$ is $Prob(S / E,\vartheta) = v_{e1}(o_1)*v_{e2}(o_2)*...*v_{eT}(o_T)$. The probability of a state sequence is: $Prob(E / \vartheta) = \Pi_{e1} P_{e1e2}*P_{e2e3}*...*P_{e(T-1)eT}$. Using the formula: $Prob(S , E / \vartheta) = Prob(S / E,\vartheta) * Prob(E / \vartheta)$ and summing this joint probability over all possible state sequences E, one demonstrates the theorem.

The interpretation of the previous equation is: initially à time t = 1, the system is in state e_1 with probability Π_1 and we observe o_1 with probability $v_{e1}(o_1)$. The system then makes a transition to state e_2 with probability P_{e1e2} and we observe o_2 with probability $v_{e2}(o_2)$. This process continues until the last transition from state e_{T-1} to state e_T with probability P_{eT-1eT} and then we observe o_T with probability $v_{eT}(o_T)$.

In oder to determine one of the estimate of the model $\vartheta = [\Pi,P,V]$, one can use the maximum likelihood criterion for a certain family S_i ($i \in \{1,2,...,L\}$) of training samples. Some methods of choosing representative samples of fixed length are presented in [2]. The problem can be expressed mathematically as:

$$\max_{\vartheta_i} f(S_1,S_2,S_3,...,S_L/ \vartheta) = \max_{\vartheta_i} \left\{ \prod_{j=1}^{j=L} \left[\sum_{E} \Pi_{e_1} * v_{e_1}(o_1^j) * \prod_{t=2}^{t=T} P_{e_{t-1}e_t} * v_{e_t}(o_t^j) \right] \right\}.$$

There is no known method to solve this problem analytically, that is the reason why we use iterative procedures. We start by determining first the optimal path for each sample. An optimal path E* is the one which is associated to the higher probability of the sample. Using the well-known Viterbi algorithm, one can determine this optimal path. The different steps for finding the single best state sequence in the viterbi algorithm are:

---

*step1: **initialization***

$$\delta_1(i) = \Pi_i v_i(o_1), \quad (1 \le i \le K)$$

$$\Psi_1(i) = 0,$$

---

*step2: **recursion***

for $(2 > t > T)$ and $(1 \le j \le K)$,

$$\delta_t(j) = \max_{1 \le i \le K} [\delta_{t-1}(i) P_{ij}] v_j(o_t),$$

$$\Psi_t(j) = \operatorname{argmax}_{1 \le i \le K} [\delta_{t-1}(i) P_{ij}] ,$$

*step3: **termination***

$$P^* = \max_{1 \le i \le K} [\delta_T(i)]$$

$$e_T^* = \operatorname*{argmax}_{1 \le i \le K} [\delta_T(i)]$$

*step4: **state sequence backtracking***

for t = T-1, T-2,...,1

$$e_T^* = \Psi_{t+1}\, e_{t+1}^*.$$

P^* is the state-optimized likelihood function and $E^* = \{e_1^*, e_2^*, ..., e_T^*\}$ is the optimal state sequence. Instead of tracking all possible paths, one successively tracks only the optimal paths E_i^* of all samples. Thus, this can be written as:

$$g(o_1,o_2,o_3,...,o_T ; E^*, \vartheta) = \max_{E} \left\{ \Pi_{e_1}^* \, v_{e_1}(o_1) * \prod_{t=2}^{t=T} P_{e_{t-1}e_t}^* \, v_{e_t}(o_t) \right\},$$

this computation has to be done for all the samples. Among all the ϑ_i, $i \in \{1,...,L\}$ associated to optimal paths, we decide to choose as best model estimate the one which maximizes the probability associated to a sample. It can be written as:

$$\vartheta^* = \arg \left\{ \max_{\vartheta_i} \left\{ g(o_1^i, o_2^i, ..., o_T^i ; E^*, \vartheta_i) \right\} \right\}, \quad i \in \{1,2,3,...,L\}.$$

4. The different steps of the method

We present an iterative method which enables us to obtain an estimator of the model ϑ. This method is suitable for direct computation.

First step: one has to cluster the sample with respect to the chosen criteria, two possibilities are offered : *a classification* or *a segmentation*. In this latter procedure, The user may structure the states, operating in this way, the states appears like known hidden states. However, in a classification the system structures its own states according to a suitable norm. Thus, the states appears like unknown hidden ones. The clusters formed by one of the two procedures represent the first states of the model, they form *the first training path.*

Second step: one estimates the transition probabilities using the following equations and the probability of each training vector for each state i.e., $v_i(o_t)$, this is the first model ϑ_1.

$$\Pi_i = \frac{\text{number of times the observation } o_1 \text{ belongs to the state i}}{\text{number of training paths}},$$

$$P_{ij}(t) = \frac{\text{number of times } \{(o_{t-1} \text{ belongs to i}) \text{ and } (o_t \text{ belongs to j})\}}{\text{number of times the observation } o_{t-1} \text{ belongs to i}},$$

the previous estimation formula can be written as

$$P_{ij}(t) = \frac{N_{ij}(t)}{N_i(t-1)} = \frac{N_{ij}(t)}{N_{i+}(t)},$$

$$v_i(o_t) = \frac{\text{the expected number of times of being in state i and observing } o_t}{\text{the expected number of times of being in state i}},$$

for i,j = 1,2,...,K and t = 1,2,...,T, where $N_{ij}(t)$ is the number of transitions from state i at time (t - 1) to state j at time t and $N_i(t - 1)$ is the number of times the state i is visited at time (t - 1).

Third step: one computes $f(o_1,o_2,...,o_T ; \vartheta_1)$ and determines the next training path or clustering capable to increase $f(o_1,o_2,...,o_T ; \vartheta_1)$. We apply the second step to this training path. This procedure is repeated until we reach the maximum value of the previous function. At this optimal value, we have E_1^* and ϑ_1 of the first sample. This step uses Viterbi algorithm.

This algorithm is applied to a family of samples of the same text, so we obtain a family of E_i^* and ϑ_i. As mentioned previously, one decides reasonably to choose the model ϑ^* whose probability associated to a sample is maximum. This last model makes the sample the most representative, i.e., *we have a good observation in some sense*. This optimal model estimate is considered as *a type of the text processed.*

5. Test for first-order stationarity

As outlined by Anderson and Goodman [1], the following test can be used to determine whether the Markov chain is first-order stationary or not. Thus, we have to test

the null hypothesis

$H: P_{ij}(t) = P_{ij} \quad (t = 1,2,...,T).$

The likelihood ratio with respect to the null and alternate hypothesis is

$$\lambda = \prod_{t=1}^{t=T} \prod_{i=1}^{i=K} \prod_{j=1}^{j=K} \frac{P_{ij}^{N_{ij}(t)}}{P_{ij}^{N_{ij}(t)}(t)}.$$

We now determine the confidence region of the test. In fact, the expression - 2 log λ is distributed as a Chi-square distribution with (T - 1)*K*(K-1) degrees of freedom when the null hypothesis is true. As the distribution of the statistic S = -2 log λ is χ^2, one can compute a β point (β = 95, 99.95% etc.) as the threshold S_β. The test is formulated as: if $S < S_\beta$, the null hypothesis is accepted, i.e., the Markov chain is first-order stationary. Otherwise, the null hypothesis is rejected at 100% - β level of significance, i.e., the chain is not a first-order stationary, and one decides in favour of the nonstationary model.

6. How to solve the morphological ambiguities

This is the most important phase of our application. Let's consider an example of nine possible paths encountered in a text. Among these paths, the system has to choose the most likely according to the probability measure, see figure 1. Our decision of choosing the most likely path comes from the optimal model ϑ^* obtained in the training phase. We show in this example how to remove the morphological ambiguities.

If the optimal state sequence obtained in the training phase is the one which corresponds to the figure 2., then one for example can choose between the two following paths of the figure 1:

Path 1: $s_1 \; s_2 \; s_3 \; s_4 \; s_5 \; s_6 \; s_7$.

path 2: $s_1 \; s_2 \; s'_3 \; s_4 \; s_5 \; s'_6 \; s_7$.

One computes the probabilities of these two realizations of the observations o_i (i =1,...,7) using the formula:

$$Pr\,(o_1,o_2,...,o_7 / \vartheta^*) = \Pi_{e_1} * v_{e_1}(o_1) * \prod_{t=2}^{t=7} P_{e_{t-1}e_t} * v_{e_t}(o_t).$$

The figure 2. shows that each s_i belongs to a state e_j and using the optimal model $\vartheta^* = [\Pi,P,V]$ one can computes the probability of a path. Our decision to remove the morphological ambiguities is to choose the path with the highest probability.

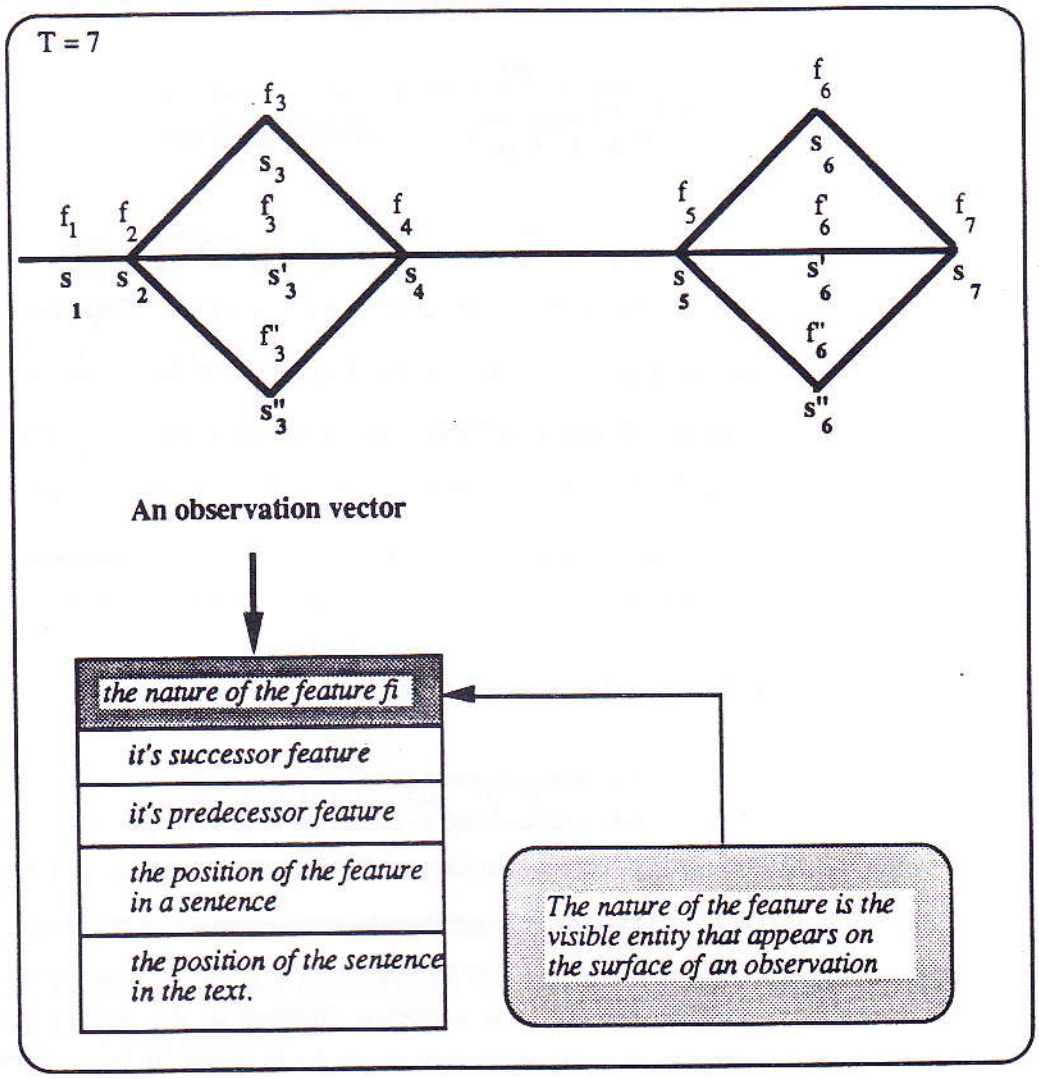

Fig.1.

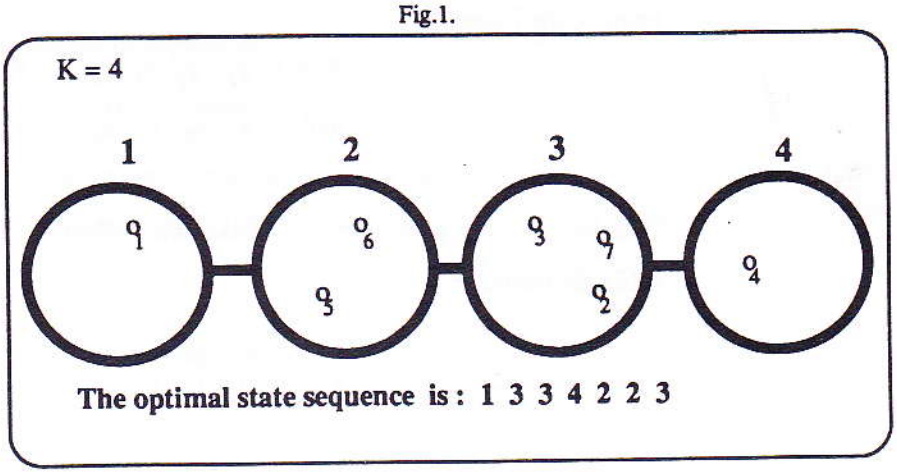

Fig. 2.

7. Conclusion

We have presented a new approach for solving the morphological ambiguities using a hidden Markov model. This method may also be applied to other analysis levels as syntax. The main advantage of the method is the possibility to assign many different classes of criteria (fuzzy or completely known) to the training vectors and investigating many samples. Furthermore, we can define a "distance" between any sample and a family of types of texts called models. One may choose the model which gives the higher probability of this sample and conclude that the sample belongs to this specific type of texts. We can also develop a proximity measure between two models ϑ^*_1 and ϑ^*_2 through representative samples. However, some precautions must be taken in the choice of the distance used between the training vectors in the cluster process. In fact, the value of the probability associated to a sample may depend on this norm and therefore, the choice of the best model estimate can be affected.

So far, we supposed that the criteria described the observations and the states are completely known (hard observation). Very often, when we want to make deep investigations, fuzziness or uncertainty due to some criteria or states are encountered, what to do in this case? How can we cluster the observations according to any uncertainty measure? What is the optimal path and the best estimate model according to this uncertainty measure? We are working in order to propose solutions to those questions in the case of a probabilistic [3] and a fuzzy logic.

References

[1] T.W., Anderson and L.A., Goodman, "Statistical inference about Markov chains", Ann. Math.Statist. 28, 89-110 (1957).

[2] D., Bouchaffra, *Echantillonnage et analyse multivariée de paramètres linguistiques*, Thèse de Doctorat (P.H.D), Université Pierre Mendès France, to appear in (1992).

[3] D., Bouchaffra, "A relation between isometries and the relative consistency concept in probabilistic logic", 13th IMACS World Congress on Computation and applied Mathematics, Dublin, Ireland, July 22-26, (1991), selected papers in *AI, Expert Systems and Symbolic Computing for Scientific Computation*, edit. John Rice and Elias N. Houstis, published by Elsevier, North-Holland.

[4] J., Rouault, *Linguistique automatique: Applications documentaires*, Editions Peter Lang SA, Berne (1987).