algorithms

MDPI

Communication

Equivalence between LC-CRF and HMM, and Discriminative Computing of HMM-Based MPM and MAP

Elie Azeraf [1], Emmanuel Monfrini [2] and Wojciech Pieczynski [2,*]

[1] Watson Department, IBM GSB France, avenue de l'Europe,92270 Bois-Colombes, France
[2] SAMOVAR, Telecom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France
* Correspondence: wojciech.pieczynski@telecom-sudparis.eu

Abstract: Practitioners have used hidden Markov models (HMMs) in different problems for about sixty years. Moreover, conditional random fields (CRFs) are an alternative to HMMs and appear in the literature as different and somewhat concurrent models. We propose two contributions: First, we show that the basic linear-chain CRFs (LC-CRFs), considered as different from HMMs, are in fact equivalent to HMMs in the sense that for each LC-CRF there exists an HMM—that we specify—whose posterior distribution is identical to the given LC-CRF. Second, we show that it is possible to reformulate the generative Bayesian classifiers maximum posterior mode (MPM) and maximum a posteriori (MAP), used in HMMs, as discriminative ones. The last point is of importance in many fields, especially in natural language processing (NLP), as it shows that in some situations dropping HMMs in favor of CRFs is not necessary.

Keywords: hidden Markov model; linear chain conditional random field; Bayesian classifier; discriminative classifier; generative classifier; maximum posterior mode; maximum a posteriori

check for updates

Citation: Azeraf, E.; Monfrini, E.; Pieczynski, W. Equivalence between LC-CRF and HMM, and Discriminative Computing of HMM-Based MPM and MAP. *Algorithms* **2023**, *16*, 173. https://doi.org/10.3390/a16030173

Academic Editors: Eugene Semenkin, Todor Ganchev and Predrag S. Stanimirovic

Received: 13 January 2023
Revised: 8 March 2023
Accepted: 15 March 2023
Published: 21 March 2023

1. Introduction

Let $Z_{1:N} = (Z_1, \ldots, Z_N)$ be a stochastic sequence, with $Z_n = (X_n, Y_n)$. The random variables $X_1, \ldots, X_N$ take their values in a finite set Λ, while $Y_1, \ldots, Y_N$ take their values either in a discrete or continuous set Ω. Realizations of $X_{1:N} = (X_1, \ldots, X_N)$ are hidden while realizations of $Y_{1:N} = (Y_1, \ldots, Y_N)$ are observed, and the problem we deal with is to estimate $X_{1:N} = x_{1:N}$ from $Y_{1:N} = y_{1:N}$. We deal with Bayesian methods of estimation, which requires some probabilistic model. Probabilistic model is a distribution—or a family of distributions—which is denoted by $p(z_{1:N})$, or $p(x_{1:N}, y_{1:N})$. We are interested in the case of dependent $Z_1, \ldots, Z_N$. The simplest model taking into account this dependence is the well-known hidden Markov model (HMM) [1–5], whose distribution is given with

$$p(x_{1:N}, y_{1:N}) = p(x_1)p(y_1|x_1)\prod_{n=1}^{N-1} p(x_{n+1}|x_n)p(y_{n+1}|x_{n+1}) \tag{1}$$

In the whole paper, we consider HMMs such that $p(x_{n+1}|x_n)$ and $p(y_{n+1}|x_{n+1})$ in Equation (1) are non-null. HMMs allow recursive fast computation of Bayesian estimators, called "classifiers" in this paper, and recalled below. In spite of their simplicity, HMMs are very robust and provide quite satisfactory results in many applications.

Moreover, conditional random fields (CRFs) [6,7] also allow estimating $X_{1:N} = x_{1:N}$ from $Y_{1:N} = y_{1:N}$. Their definition is different from the definition of HMMs in that in CRFs, one directly considers $p(x_{1:N}|y_{1:N})$, and neither $p(x_{1:N}, y_{1:N})$ nor $p(y_{1:N}|x_{1:N})$ is needed to perform the estimation. The distribution of general LC-CRFs is written as such:

$$p(x_{1:N}|y_{1:N}) = p(x_1|y_{1:N})\prod_{n=1}^{N-1} p(x_{n+1}|x_n, y_{1:N}) \tag{2}$$

In this paper, we consider the following basic LC-CRF:

$$p(x_{1:N}|y_{1:N}) = \frac{1}{\kappa(y_{1:N})} exp\left[\sum_{n=1}^{N-1} V_n(x_n, x_{n+1}) + \sum_{n=1}^{N} U_n(x_n, y_n)\right] \tag{3}$$

with $\kappa(y_{1:N})$ the normalizing constant.

Authors usually consider the two families HMMs and LC-CRFs as different [6–13]. They classify the former in the category of "generative models", while they classify the latter in the category of "discriminative" models.

In this paper, we investigate the relationship between HMMs (Equation (1)) and LC-CRFs (Equation (3)). We immediately see that if $p(x_{1:N}, y_{1:N})$ is of the form (Equation (1)), then $p(x_{1:N}|y_{1:N})$ is of the form (Equation (3)). Indeed, $p(x_{n+1}|x_n)$ and $p(y_{n+1}|x_{n+1})$ in Equation (1) being non-zero, one can take $V_1(x_1, x_2) = Log[p(x_1, x_2)]$, $V_n(x_n, x_{n+1}) = Log[p(x_{n+1}|x_n)]$ for $n = 2, \ldots, N-1$, $U_n(x_n, y_n) = Log[p(y_n|x_n)]$, for $n = 1, \ldots, N$, and $\kappa(y_{1:N}) = p(y_{1:N})$. Thus, the posterior distribution $p(x_{1:N}|y_{1:N})$ of each HMM (Equation (1)) with the positivity conditions assumed above is a LC-CRF (Equation (3)). Our first contribution is to show the converse: for a given LC-CRF $p(x_{1:N}|y_{1:N})$ of form (Equation (3)), there is an HMM $q(x_{1:N}, y_{1:N})$ of form (Equation (3)) such that $q(x_{1:N}|y_{1:N}) = p(x_{1:N}|y_{1:N})$. Moreover, we give exact computation of the HMM parameters $q(x_1)$, $q(x_{n+1}|x_n)$, $q(y_{n+1}|x_{n+1})$, from the LC-CRF parameters $V_n(x_n, x_{n+1})$, $U_n(x_n, y_n)$. Such a general result, without additional constraints on LC-CRF, is new.

Our second contribution is related to the computation of Bayesian classifiers. In some situations, CRFs are preferred over HMMs because the Bayesian classifiers based on CRFs are computed without considering $p(y_n|x_n)$ distributions, which are difficult to model. This is particularly the case with automatic natural language processing (NLP). Indeed, when considering an HMM (Equation (1)) and calculating the Bayesian classifiers "maximum posterior margins" (MPM) and "maximum posterior" (MAP) in the standard way, we use $p(y_n|x_n)$ and this is the reason why LC-CRFs are preferred over HMMs. In this paper, we show that the HMM-based MPM and MAP can also be computed without using $p(y_n|x_n)$. Also recall that each Bayesian "discriminative" classifier based on LC-CRF (Equation (3)) is identical to the "generative" Bayesian classifier based on an HMM (Equation (1)), since the posterior distribution $p(x_{1:N}|y_{1:N})$ of an HMM gives the one of a LC-CRF. Indeed, Bayesian classifiers only depend on the posterior distribution. Thus, the distinction between "generative" classifiers and "discriminative" classifiers is misleading, they are all "discriminative", but they can be computed in a "generative" way, using $p(y_n|x_n)$, or in a discriminative manner, without using $p(y_n|x_n)$. Thus, we can say that our second contribution is to show that the HMM-based MPM and MAP, usually computed in a generative manner, can also be computed in a discriminative manner. This is important because it shows that abandoning HMMs in favor of CRFs in the aforementioned situations is not justified. Indeed, attached to the first contribution, it shows that the use of the MPM or MAP based on HMM (Equation (1)) is as interesting as the use of the MPM or MAP based on LC-CRF (Equation (3)).

Let us give some more technical details on the two contributions.

1. We establish an equivalence between HMMs (Equation (1)) and basic linear-chain CRFs (Equation (3)), which completes the results presented in [14].

Let us notice that wanting to compare the two models directly is somewhat misleading. Indeed, HMMs and CRFs are defined with distributions on different spaces. To be precise, we adopt the following definition:

Definition 1. *Let $X_1, \ldots, X_N$, $Y_1, \ldots, Y_N$ be the two stochastic sequences defined above.*

(i) We will call "model" a distribution $p(x_{1:N}, y_{1:N})$;

(ii) We will call "conditional model" a distribution $p(x_{1:N}|y_{1:N})$;

(iii) We will say that a model $p(x_{1:N}, y_{1:N})$ is "equivalent" to a conditional model $q(x_{1:N}|y_{1:N})$ if there exists a distribution $r(y_{1:N})$ such that $p(x_{1:N}, y_{1:N}) = q(x_{1:N}|y_{1:N})r(y_{1:N})$;

(iv) *We will say that a family of models A is "equivalent" to a family of conditional models B if for each model $p(x_{1:N}, y_{1:N})$ in A there exists an equivalent conditional model $q(x_{1:N}|y_{1:N})$ in B.*

According to Definition 1, HMMs are particular "models", while CRFs are particular "conditional models". Then a particular HMM model cannot be equal to a particular conditional CRF model, but it can be equivalent to the latter.

Our contribution is to show that the family of LC-CRFs (Equation (3)) is equivalent to the family of HMMs (Equation (1)). In addition, we specify, for each LC-CRF $q(x_{1:N}|y_{1:N})$, a particular HMM $p(x_{1:N}, y_{1:N})$ such that $p(x_{1:N}|y_{1:N}) = q(x_{1:N}|y_{1:N})$.

Finally, the core of our first contribution is the following. Let $p(x_{1:N}, y_{1:N}, \theta)$ be an HMM (Equation (1)), with parameters θ. Taking $r(y_{1:N}) = p(y_{1:N}, \theta)$, it is immediate to see that $p(x_{1:N}|y_{1:N}, \theta)$ is an equivalent CRF. The converse is not immediate. Is a given CRF $p(x_{1:N}|y_{1:N}, \theta)$ equivalent to a certain HMM? If yes, can we find $r(y_{1:N})$ such that $p(x_{1:N}|y_{1:N}, \theta)r(y_{1:N})$ is an HMM? Moreover, can we give its (Equation (1)) form? Answering to these questions in a simple linear-chain CRF case is our first contribution. More precisely, we show that the family of LC-CRFs (Equation (3)) is equivalent to the family of HMMs (Equation (1)), and we specify, for each LC-CRF $p(x_{1:N}|y_{1:N})$, a particular HMM $q(x_{1:N}, y_{1:N})$ given in the form (Equation (1)), such that $p(x_{1:N}|y_{1:N}) = q(x_{1:N}|y_{1:N})$.

2. We show that the "generative" estimators MPM and MAP in HMM are computable in a "discriminative" manner, exactly as in LC-CRF.

One of the interests of HMMs and CRFs is that in both of them there exist Bayesian classifiers, which allow estimating $x_{1:N}$ from $y_{1:N}$ in a reasonable computer time. As examples, let us consider the "maximum of posterior margins" (MPM) defined with:

$$[g(y_{1:N}) = \hat{x}_{1:N} = (\hat{x}_1, \ldots, \hat{x}_N)] \iff [\forall n = 1, \ldots, N, p(\hat{x}_n|y_{1:N}) = \underset{x_n}{sup}(p(x_n|y_{1:N}))] \quad (4)$$

and the "maximum a posteriori" (MAP), defined with

$$[g(y_{1:N}) = \hat{x}_{1:N}] \iff [p(\hat{x}_{1:N}|y_{1:N}) = \underset{x_{1:N}}{sup}(p(x_{1:N}|y_{1:N}))] \quad (5)$$

Note that likely to any other Bayesian classifier, MPM and MAP are independent from $p(y_{1:N})$. This means that in any generative model $p(x_{1:N}, y_{1:N})$, any related Bayesian classifier is strictly the same as the one related to the equivalent (in the meaning of Definition 1) CRF model $p(x_{1:N}|y_{1:N})$. We see that the distinction between "generative" and "discriminative" classifiers is not justified: all Bayesian classifiers are discriminative. However, in HMM the related MPM and MAP classifiers are computed using $p(y_n|x_n)$, while this is not the case in LC-CRF. We show that both MPM and MAP in HMM can also be computed in a "discriminative" way, without using $p(y_n|x_n)$. Thus, the feasibility of using MPM and MAP in HMM is strictly the same as that of their use in LC-CRF, which is our second contribution. One of the consequences is that the use of MPM and MAP in the two families HMMs and LC-CRFs presents exactly the same interest, in particular in NLP. This shows that abandoning HMMs in favor of LC-CRFs in NLP because of the "generative" nature [6–9,15–18] of their related Bayesian classifiers was not justified.

2. Related Works

Concerning the first contribution, several authors noticed similarities between LC-CRFs and HMMs in different previous works. Our first remark is to notice that trying to compare the two families directly is somewhat incorrect, as they are defined on different spaces. In particular, they cannot be equal. It is well-known that the posterior distribution $p(x_{1:N}|y_{1:N})$ of an HMM $p(x_{1:N}, y_{1:N})$ is a LC-CRF. Conversely, showing that for a given CRF $p(x_{1:N}|y_{1:N})$ it is possible to find an HMM $q(x_{1:N}, y_{1:N})$ such that $q(x_{1:N}|y_{1:N}) = p(x_{1:N}|y_{1:N})$ is more difficult and at our knowledge, there is no general results, except [14], published. However, let us mention [19] where authors show a similar result to ours assuming an additional constraint on the LC-CRF considered. In [7], authors comment similarities and differences between LC-CRFs and HMMs considered here; however, the problem of searching an HMM equivalent to a given LC-CRF is not addressed. In this paper we show how to compute, from the LC-CRF $p(x_{1:N}|y_{1:N})$ given with Equation (3), an HMM $q(x_{1:N}, y_{1:N})$ verifying $q(x_{1:N}|y_{1:N}) = p(x_{1:N}|y_{1:N})$, without any additional constraints. Concerning the second contribution, the authors generally distinguish between "discriminative" classifiers, linked to discriminative models, and "generative" classifiers, linked to generative models. As mentioned above, our contribution is to show that the MPMs and MAPs based on generative HMMs can also be considered as discriminative classifiers. To our knowledge, there is no work on such conversions, except [20].

3. Equivalence between HMMs and Simple Linear-Chain CRFs

We will use the following Lemma:

Lemma 1. *Let $W_{1:N} = (W_1, \ldots, W_N)$ be random sequence, taking its values in a finite set Δ. Then*

(i) *$W_{1:N}$ is a Markov chain if and only if (iff) there exist $N-1$ functions $\varphi_1, \ldots, \varphi_{N-1}$ from Δ^2 to R^+ such that*

$$p(w_1, \ldots, w_N) \propto \varphi_1(w_1, w_2) \ldots \varphi_{N-1}(w_{N-1}, w_N) \tag{6}$$

where "$\propto$" means "proportional to";

(ii) *For the HMM defined with $\varphi_1, \ldots, \varphi_{N-1}$ verifying Equation (6), $p(w_1)$ and $p(w_{n+1}|w_n)$ are given with*

$$p(w_1) = \frac{\beta_1(w_1)}{\sum_{w_1} \beta_1(w_1)}; p(w_{n+1}|w_n) = \frac{\varphi_n(w_n, w_{n+1})\beta_{n+1}(w_{n+1})}{\beta_n(w_n)} \tag{7}$$

where $\beta_1(w_1), \ldots, \beta_N(w_N)$ are defined with the following backward recursion:

$$\beta_N(w_N) = 1,\ \beta_n(w_n) = \textstyle\sum_{w_{n+1}} \varphi_n(w_n, w_{n+1})\beta_{n+1}(w_{n+1}) \tag{8}$$

Proof of Lemma.

1. Let $W_{1:N}$ be Markov: $p(w_1, \ldots, w_N) = p(w_1)p(w_2|w_1)p(w_3|w_2) \ldots p(w_N|w_{N-1})$. Then (Equation (6)) is verified by $\varphi_1(w_1, w_2) = p(w_1)p(w_2|w_1)$, $\varphi_2(w_2, w_3) = p(w_3|w_2)$, $\ldots$, $\varphi_{N-1}(w_{N-1}, w_N) = p(w_N|w_{N-1})$.
2. Conversely, let $p(w_1, \ldots, w_N)$ verifies (Equation (6)). Thus $p(w_1, \ldots, w_N) = K\varphi_1(w_1, w_2) \ldots \varphi_{N-1}(w_{N-1}, w_N)$ with K constant. This implies that for each $n = 1, \ldots, N-1$ we have

$$\begin{gathered} p(w_{n+1}|w_1, \ldots, w_n) = \tfrac{p(w_1, \ldots, w_n, w_{n+1})}{p(w_1, \ldots, w_n)} = \\ \tfrac{\sum_{(w_{n+2}, \ldots, w_N,)} \varphi_1(w_1, w_2) \ldots \varphi_n(w_n, w_{n+1})\varphi_{n+1}(w_{n+1}, w_{n+2}) \ldots \varphi_{N-1}(w_{N-1}, w_N)}{\sum_{(w_{n+1}, w_{n+2}, \ldots, w_N,)} \varphi_1(w_1, w_2) \ldots \varphi_n(w_n, w_{n+1})\varphi_{n+1}(w_{n+1}, w_{n+2}) \ldots \varphi_{N-1}(w_{N-1}, w_N)} = \\ \tfrac{\varphi_n(w_n, w_{n+1})\sum_{(w_{n+2}, \ldots, w_N,)} \varphi_{n+1}(w_{n+1}, w_{n+2}) \ldots \varphi_{N-1}(w_{N-1}, w_N)}{\sum_{(w_{n+1}, w_{n+2}, \ldots, w_N,)} \varphi_n(w_n, w_{n+1})\varphi_{n+1}(w_{n+1}, w_{n+2}) \ldots \varphi_{N-1}(w_{N-1}, w_N)} = p(w_{n+1}|w_n) \end{gathered} \tag{9}$$

which shows that $p(w_1, \ldots, w_N)$ is Markov.

Moreover, let us set $\beta_n(w_n) = \sum_{(w_{n+1}, w_{n+2}, \ldots, w_N)} \varphi_n(w_n, w_{n+1}) \ldots \varphi_{N-1}(w_{N-1}, w_N)$ for $n = 1, \ldots, N-1$. On the one hand, we see that $\beta_n(w_n) = \sum_{w_{n+1}} \varphi_n(w_n, w_{n+1})\beta_{n+1}(w_{n+1})$. On the other hand, according to (Equation (9)) we have $p(w_{n+1}|w_n) = \frac{\varphi_n(w_n, w_{n+1})\beta_{n+1}(w_{n+1})}{\beta_n(w_n)}$. As $p(w_1) = \frac{\beta_1(w_1)}{\sum_{w_1} \beta_1(w_1)}$, (Equation (7)) and (Equation (8)) are verified, which ends the proof. □

Proposition 1 below shows that the LC-CRF defined with (Equation (3)) is equivalent to an HMM defined with Equation (1). In addition, $p(x_1)$, $p(x_{n+1}|x_n)$, and $p(y_n|x_n)$ in Equation (1) defining an equivalent HMM are computed from $V_n(x_n, x_{n+1})$ and $U_n(x_n, y_n)$. To the best of our knowledge, except some first weaker results in [19], these results are new.

Proposition 1. *Let $Z_{1:N} = (Z_1, \ldots, Z_N)$ be a stochastic sequence, with$Z_n = (X_n, Y_n)$. Each (X_n, Y_n) takes its values in $\Lambda \times \Omega$, with Λ and Ω finite. If $Z_{1:N}$ is a LC-CRF with the distribution $p(x_{1:N}|y_{1:N})$ defined by*

$$p(x_{1:N}|y_{1:N}) = \frac{1}{\kappa(y_{1:N})} exp\left[\sum_{n=1}^{N-1} V_n(x_n, x_{n+1}) + \sum_{n=1}^{N} U_n(x_n, y_n)\right] \tag{10}$$

then (*Equation (10)*) *is the posterior distribution of the HMM*

$$q(x_{1:N}, y_{1:N}) = q_1(x_1)q(y_1|x_1)\prod_{n=1}^{N-1} q(x_{n+1}|x_n)q(y_{n+1}|x_{n+1}) \tag{11}$$

with

$$q(x_1, y_1) = \frac{\beta_1(x_1, y_1)}{\sum_{(x_1, y_1)} \beta_1(x_1, y_1)} \tag{12}$$

and, for $n = 1, \ldots, N-1$:

$$q(x_{n+1}|x_n) = \frac{\psi(x_{n+1})exp[V_{\ n}(x_n, x_{n+1})]}{\sum_{x_{n+1}} \psi(x_{n+1})exp[V_{\ n}(x_n, x_{n+1})]} \tag{13}$$

$$q(y_{n+1}|x_{n+1}) = \frac{exp[U_{n+1}(x_{n+1}, y_{n+1})]\beta_{n+1}(x_{n+1}, y_{n+1})}{\psi(x_{n+1})} \tag{14}$$

where

$$\psi(x_{n+1}) = \sum_{y_{n+1}} exp[U_{n+1}(x_{n+1}, y_{n+1})]\beta_{n+1}(x_{n+1}, y_{n+1}) \tag{15}$$

and $\beta_1(x_1, y_1), \ldots, \beta_N(x_N, y_N)$are given by the backward recursion

$$\beta_N(x_N, y_N) = 1$$

$$\beta_n(x_n, y_n) = \sum_{(x_{n+1}, y_{n+1})} \exp[V_{\ n}(x_n, x_{n+1}) + U_{n+1}(x_{n+1}, y_{n+1})]\beta_{n+1}(x_{n+1}, y_{n+1}) \tag{16}$$

Proof of Proposition 1. Let us consider functions $\varphi_1, \ldots, \varphi_N$ defined on $[\Lambda \times \Omega]^2$ by

$$\varphi_1(x_1, y_1, x_2, y_2) = exp[V_1(x_1, x_2) + U_1(x_1, y_1) + U_2(x_2, y_2)] \tag{17}$$

$$\varphi_n(x_n, y_n, x_{n+1}, y_{n+1}) = \exp[V_{\ n}(x_n, x_{n+1}) + U_{n+1}(x_{n+1}, y_{n+1})], \text{ for } n = 2, \ldots, N-1 \tag{18}$$

According to the lemma, they define a Markov chain $Z_{1:N} = (Z_1, \ldots, Z_N)$, with $Z_n = (X_n, Y_n)$. Let us denote its distribution by $q(z_{\ 1:N}) = q(x_{1:N}, y_{1:N})$. As $q(x_{1:N}, y_{1:N}) = Kexp\left[\sum_{n=1}^{N-1} V_n(x_n, x_{n+1}) + \sum_{n=1}^{N} U_n(x_n, y_n)\right]$ with K constant, we have $q(x_{1:N}|y_{1:N}) = p(x_{1:N}|y_{1:N})$. Let us show that $q(x_{1:N}, y_{1:N})$ verifies (Equations (11)–(16)). According to the lemma, considering $\beta_1(x_1, y_1), \ldots, \beta_N(x_N, y_N)$ defined with (Equation (16)), and $\psi(x_{n+1})$ defined with (Equation (15)), we have

$$\begin{gathered} q(x_{n+1}, y_{n+1}|x_n, y_n) = \frac{\varphi_n(x_n, y_n, x_{n+1}, y_{n+1})\beta_{n+1}(x_{n+1}, y_{n+1})}{\sum_{(x_{n+1}, y_{n+1})} \varphi_n(x_n, y_n, x_{n+1}, y_{n+1})\beta_{n+1}(x_{n+1}, y_{n+1})} = \\ \frac{\exp[V_{\ n}(x_n, x_{n+1}) + U_{n+1}(x_{n+1}, y_{n+1})]\beta_{n+1}(x_{n+1}, y_{n+1})}{\sum_{(x_{n+1}, y_{n+1})} \exp[V_{\ n}(x_n, x_{n+1}) + U_{n+1}(x_{n+1}, y_{n+1})]\beta_{n+1}(x_{n+1}, y_{n+1})} = \end{gathered} \tag{19}$$

$$\frac{\exp[V_n(x_n,x_{n+1})]\exp[U_{n+1}(x_{n+1},y_{n+1})]\beta_{n+1}(x_{n+1},y_{n+1})}{\sum_{x_{n+1}}\psi(x_{n+1})\exp[V_n(x_n,x_{n+1})]} =$$
$$\left[\frac{\exp[V_n(x_n,x_{n+1})]}{\sum_{x_{n+1}}\psi(x_{n+1})\exp[V_n(x_n,x_{n+1})]}\right][\exp[U_{n+1}(x_{n+1},y_{n+1})]\beta_{n+1}(x_{n+1},y_{n+1})] =$$
$$\left[\frac{\psi(x_{n+1})\exp[V_n(x_n,x_{n+1})]}{\sum_{x_{n+1}}\psi(x_{n+1})\exp[V_n(x_n,x_{n+1})]}\right]\left[\frac{\exp[U_{n+1}(x_{n+1},y_{n+1})]\beta_{n+1}(x_{n+1},y_{n+1})}{\psi(x_{n+1})}\right] =$$
$$q(x_{n+1}|x_n)q(y_{n+1}|x_{n+1})$$

(12) being directly implied by the lemma, this ends the proof. □

4. Discriminative Classifiers in Generative HMMs

One of the interests of HMMs and some CRFs with hidden discrete finite data lies in possibilities of analytic fast computation of Bayesian classifiers. As examples of classic Bayesian classifiers, let us consider MPM (Equation (4)) and MAP (Equation (5)). However, in some domains such as NLP, CRFs are preferred to HMMs for the following reasons. As HMM is a generative model, MPM and MAP used in HMM are also called "generative", and people consider that HMM-based MPM and MAP require the knowledge of $p(y_n|x_n)$. Then people consider it as improper to use the HMM-based MPM and MAP in situations where the distributions $p(y_n|x_n)$ are hard to handle. We show that this reason is not valid. More precisely, we show two points:

(i) First, we notice that regardless of the distribution $p(x_{1:N}, y_{1:N})$, all Bayesian classifiers are independent from $p(y_{1:N})$, so that the distinction between « generative » and « discriminative » classifiers is misleading: they are all discriminative;

(ii) Second, we show "discriminative" computation of MPM and MAP in HMMs is not intrinsic to HMMs but is due to its particular classic parameterization Equation (1). In other words, changing the parametrization, it is possible to compute the HMM-based MPM and MAP without using $p(y_{1:N}|x_{1:N})$ or $p(y_{1:N})$.

The first point is rather immediate: we note that Bayesian classifier g_L is defined by a loss function $L : \Omega^2 \to R^+$ through

$$[g_L(y_{1:N}) = \hat{x}_{1:N}] \Longleftrightarrow [E[L(g_L(y_{1:N}), X_{1:N})|y_{1:N}] = \underset{x_{1:N}}{inf} E[L(x_{1:N}, X_{1:N})|y_{1:N}]] \tag{20}$$

it is thus immediate to notice that $g_L(y_{1:N})$ only depends on $p(x_{1:N}|y_{1:N})$. This implies that it is the same in a generative model or its equivalent (within the meaning of Definition 1) discriminative model.

We show (ii) by separately considering the MPM and MAP cases.

4.1. Discriminative Computing of HMM-Based MPM

To show (ii), let us consider Equation (1) with $p(y_n|x_n) = \frac{p(y_n)p(x_n|y_n)}{p(x_n)}$. It becomes,

$$p(x_{1:N}, y_{1:N}) = p(x_1|y_1)\prod_{n=2}^{T} p(x_n|x_{n-1})\frac{p(x_n|y_n)}{p(x_n)}\prod_{n=1}^{N} p(y_n) \tag{21}$$

We see that (Equation (21)) is of the form $p(x_{1:N}, y_{1:N}) = h(x_{1:N}, y_{1:N})\prod_{n=1}^{N} p(y_n)$, where $h(x_{1:N}, y_{1:N})$ does not depend on $p(y_1), \ldots, p(y_N)$. This implies that $h(x_n, y_{1:N}) = \sum_{(x_{1:n-1}, x_{n+1:N})} h(x_{1:N}, y_{1:N})$ does not depend on $p(y_1), \ldots, p(y_N)$ either. Then $p(x_n|y_{1:N}) = \frac{p(x_n,y_{1:N})}{p(y_{1:N})} = \frac{h(x_n,y_{1:N})\prod_{n=1}^{N} p(y_n)}{[\sum_{x_n} h(x_n,y_{1:N})]\prod_{n=1}^{N} p(y_n)} = \frac{h(x_n,y_{1:N})}{[\sum_{x_n} h(x_n,y_{1:N})]}$, so that

$$p(x_n|y_{1:N}) = \frac{\sum_{(x_{1:n-1},x_{n+1:N})} p(x_1|y_1)\prod_{t=2}^{N} p(x_n|x_{n-1})\frac{p(x_n|y_n)}{p(x_n)}}{\sum_{(x_{1:N})} p(x_1|y_1)\prod_{t=2}^{N} p(x_n|x_{n-1})\frac{p(x_n|y_n)}{p(x_n)}} \tag{22}$$

neither depends on $p(y_1), \ldots, p(y_N)$. Thus, the HMM-based classifier MPM also verifies the "discriminative classifier" definition.

How to compute $p(x_n|y_{1:N})$? It is classically computable using "forward" probabilities $\alpha_n(x_n)$ and "backward" ones $\beta_n(x_n)$ defined with

$$\alpha_n(x_n) = p(x_n, y_{1:n}) \tag{23}$$

$$\beta_n(x_n) = p(y_{n+1:N}|x_n) \tag{24}$$

then

$$p(x_n|y_{1:N}) = \frac{\alpha_n(x_n)\beta_n(x_n)}{\sum_{x_n} \alpha_n(x_n)\beta_n(x_n)} \tag{25}$$

with all $\alpha_n(x_n)$ and $\beta_n(x_n)$ computed using the following forward and backward recursions [21]:

$$\alpha_1(x_1) = p(x_1)p(y_1|x_1);\ \alpha_{n+1}(x_{n+1}) = \sum\nolimits_{x_n} p(x_{n+1}|x_n)p(y_{n+1}|x_{n+1})\alpha_n(x_n) \tag{26}$$

$$\beta_N(x_N) = 1;\ \beta_n(x_n) = \sum\nolimits_{x_{n+1}} p(x_{n+1}|x_n)p(y_{n+1}|x_{n+1})\beta_{n+1}(x_{n+1}) \tag{27}$$

Setting $p(y_n|x_n) = \frac{p(y_n)p(x_n|y_n)}{p(x_n)}$ and recalling that $p(x_n|y_{1:N})$ does not depend on $p(y_1), \ldots, p(y_N)$, we can arbitrarily modify them. Let us consider the uniform distribution over Ω, so that $p(y_1) = \ldots = p(y_N) = \frac{1}{\#\Omega} = c$. Then (Equation (26)) and (Equation (27)) become

$$\alpha^*_{n+1}(x_{n+1}) = \sum\nolimits_{x_n} p(x_{n+1}|x_n)\frac{cp(x_{n+1}|y_{n+1})}{p(x_{n+1})}\alpha^*_n(x_n) \tag{28}$$

$$\beta^*_n(x_n) = \sum\nolimits_{x_{n+1}} p(x_{n+1}|x_n)\frac{cp(x_{n+1}|y_{n+1})}{p(x_{n+1})}\beta^*_{n+1}(x_{n+1}) \tag{29}$$

and we still have

$$p(x_n|y_{1:N}) = \frac{\alpha^*_n(x_n)\beta^*_n(x_n)}{\sum_{x_n} \alpha^*_n\beta^*_n(x_n)} \tag{30}$$

Finally, we see that $p(x_n|y_{1:N})$ is independent from c, so that we can take $c = 1$. Then we can state the following proposition.

Proposition 2. *Let $X_1, \ldots, X_N, Y_1, \ldots, Y_N$ be an HMM Equation (1). Let us define "discriminative forward" quantities $\alpha^D_1(x_1), \ldots, \alpha^D_N(x_N)$, and "discriminative backward" ones $\beta^D_1(x_1), \ldots, \beta^D_N(x_N)$by the following forward and backward recursions:*

$$\alpha^D_1(x_1) = p(x_1|y_1); \alpha^D_{n+1}(x_{n+1}) = \sum\nolimits_{x_n} p(x_{n+1}|x_n)\frac{p(x_{n+1}|y_{n+1})}{p(x_{n+1})}\alpha^D_n(x_n) \tag{31}$$

$$\beta^D_N(x_N) = 1; \beta^D_n(x_n) = \sum\nolimits_{x_{n+1}} p(x_{n+1}|x_n)\frac{p(x_{n+1}|y_{n+1})}{p(x_{n+1})}\beta^D_{n+1}(x_{n+1}) \tag{32}$$

then

$$p(x_n|y_{1:N}) = \frac{\alpha^D_n(x_n)\beta^D_n(x_n)}{\sum_{x_n} \alpha^D_n(x_n)\beta^D_n(x_n)} \tag{33}$$

Consequently, we can compute the MPM classifier in a discriminative manner, only using $p(x_1), \ldots, p(x_N), p(x_2|x_1), \ldots, p(x_N|x_{N-1})$, and $p(x_1|y_1), \ldots, p(x_N|y_N)$.

Note that this result is similar to the result in [21], with a different proof.

Remark 1. *Let us notice that according to Equations (31) and (32), it is possible to compute $\alpha^D_n(x_n)$and$\beta^D_n(x_n)$ by a very slight adaptation of classic computing programs giving classic $\alpha_n(x_n) = p(x_n, y_{1:n})$ and $\beta_n(x_n) = p(y_{n+1:N}|x_n)$ with recursions (Equations (26) and (27)).*

All we have to do is to replace$p(y_{n+1}|x_{n+1})$ *with* $\frac{p(x_{n+1}|y_{n+1})}{p(x_{n+1})}$. *Of course,* $\alpha_n^D(x_n) \neq p(x_n, y_{1:n})$ *and*$\beta_n^D(x_n) \neq p(y_{n+1:N}|x_n)$, *but (Equation (33)) holds and thus* $p(x_n|y_{1:N})$ *is computable allowing MPM.*

Remark 2. *We see that we can compute the MPM in HMM only using* $p(x_1), \ldots, p(x_N)$, $p(x_2|x_1), \ldots, p(x_N|x_{N-1})$, *and* $p(x_1|y_1), \ldots, p(x_N|y_N)$. *This means that in supervised classification, where we have a learn sample, we can use any parametrization to estimate them. For example, we can model them with logistic regression, as currently done in CRFs. It is of importance to note that such a parametrization is unusual; however, what is important is that the model remains the same.*

Remark 3. *When using the MPM to deal with a concrete problem, Proposition 2 implies that talking about a comparison between LC-CRFs and HMMs is somewhat incorrect. Indeed, there is only one model. However, there are two different parameterizations, and this can produce two different results. Indeed, the estimation of the parameters can give two models of the same nature (posterior distribution of an HMM) but unequally situated with respect to the optimal model, the best suited to the data. It would therefore be more correct to speak of a comparison of two parametrizations of HMM, each associated with its own parameter estimator. The same is true concerning the MAP discussed in the next paragraph.*

4.2. Discriminative Computing of HMM-Based Map: Discriminative Viterbi

Let $X_1, \ldots, X_N$, $Y_1, \ldots, Y_N$ be an HMM (Equation (1)). The Bayesian MAP classifier (Equation (5)) based on such HMM is computed with the following Viterbi algorithm [22]. For each $n = 1, \ldots, N$, and each x_n, let $x_{1:n-1}^{max}(x_n) = \left(x_1^{max}, \ldots, x_{n-1}^{max}\right)(x_n)$ be the path x_1^{max}, $\ldots$, x_{n-1}^{max} verifying

$$p(x_{1:n-1}^{max}(x_n), x_n, y_{1:n}) = \underset{x_{1:n-1}}{sup}\, p(x_{1:n-1}, x_n, y_{1:n}) \tag{34}$$

We see that $x_{1:n-1}^{max}(x_n)$ is a path maximizing $p(x_{1:n-1}, x_n|y_{1:n})$ over all paths ending in x_n. Then having the paths $x_{1:n-1}^{max}(x_n)$ and the probabilities $p\left(x_{1:n-1}^{max}(x_n), x_n, y_{1:n}\right)$ for each x_n, one determines, for each x_{n+1}, the paths $x_{1:n}^{max}(x_{n+1})$ and the probabilities $p\left(x_{1:n}^{max}(x_{n+1}), x_{n+1}, y_{1:n+1}\right) = p\left(x_{1:n-1}^{max}(x_n^{max}), x_n^{max}, x_{n+1}, y_{1:n+1}\right)$, searching x_n^{max} with

$$\begin{gathered} p\left(x_{1:n-1}^{max}(x_n^{max}), x_n^{max}, x_{n+1}, y_{1:n+1})\right) = \\ \sup_{x_n}\left[p\left(x_{1:n-1}^{max}(x_n), x_n, y_{1:n}\right)p(x_{n+1}|x_n)p(y_{n+1}|x_{n+1})\right] \end{gathered} \tag{35}$$

Setting in Equation (35) $p(y_{n+1}|x_{n+1}) = \frac{p(y_{n+1})p(x_{n+1}|y_{n+1})}{p(x_{n+1})}$, we see that x_n^{max} which verifies Equation (35) is the same that x_n^{max} which maximizes $p\left(x_{1:n-1}^{max}(x_n), x_n, y_{1:n}\right)p(x_{n+1}|x_n)\frac{p(x_{n+1}|y_{n+1})}{p(x_{n+1})}$, so that we can suppress $p(y_{n+1})$. In other words, we can replace Equation (35) with

$$\begin{gathered} p\left(x_{1:n-1}^{max}(x_n^{max}), x_n^{max}, x_{n+1}, y_{1:n+1})\right) = \\ \sup_{x_n}[p(x_{1:n-1}^{max}(x_n), x_n, y_{1:n})p(x_{n+1}|x_n)\frac{p(x_{n+1}|y_{n+1})}{p(x_{n+1})}] \end{gathered} \tag{36}$$

Finally, we propose the following discriminative version of the Viterbi algorithm:

- Set $x_1^{max} = \underset{x_n}{argmax}[p(x_1|y_1)]$;
- For each $n = 1, \ldots, N-1$, and each x_{n+1}, apply (3.17) to find a path $x_{1:n}^{max}(x_{n+1})$ from the paths $x_{1:n-1}^{max}(x_n)$ (for all x_n), and the probabilities $p\left(x_{1:n}^{max}(x_{n+1}), x_{n+1}, y_{1:n+1}\right)$ (for all x_{n+1});
- End setting $x_{1:N}^{max} = \underset{x_N}{argmax}\left[p\left(x_{1:N-1}^{max}(x_N), x_N, y_{1:N}\right)\right]$.

As with MPM above, we see that we can find $x_{1:N}^{max}$ with the only use of $p(x_1), \ldots,$ $p(x_N)$, $p(x_2|x_1), \ldots, p(x_N|x_{N-1})$, and $p(x_1|y_1), \ldots, p(x_N|y_N)$, exactly as in CRF case. As above, it appears that dropping HMMs in some NLP tasks on the grounds that MAP is a "generative" classifier, is not justified. In particular, in supervised stationary framework, distributions $p(x_n)$, $p(x_{n+1}|x_n)$, and $p(x_n|y_n)$ can be estimated in the same way as in LC-CRFs case.

5. Discussion and Conclusions

We have proposed two results. First, we have shown that the basic LC-CRF (Equation (3)) is equivalent to a classical HMM (Equation (1)) in that one can find an HMM whose posterior distribution is exactly the given LC-CRF. More precisely, we specified the way to calculate the parameters $p(x_1)$, $p(x_{n+1}|x_n)$, and $p(y_n|x_n)$ which define Equation (1) from the parameters $V_n(x_n, x_{n+1})$, $U_n(x_n, y_n)$ which define Equation (3). Second, noting that all Bayesian classifiers are discriminative in the sense that they do not depend on the observation distribution, we showed that classifiers based on HMMs, usually considered as "generative", can also be considered as discriminative classifiers. More specifically, we have proposed discriminative methods for computing classic maximum posterior mode (MPM) and maximum a posteriori (MAP) classifiers based on HMMs. The first result shows that LC-CRFs are as general as classical HMMs. The second shows that at the application level, HMMs offer strictly the same processing power, at least with regard to MPM and MAP, as LC-CRFs.

The practical interest of our contributions is as follows. Until now, some authors considered LC-CRFs and HMMs to be equivalent without presenting rigorous proof, and others considered LC-CRFs to be more general [19]. Partly because of this uncertainty, CRFs have often been preferred over HMMs. We have proved, at least in the particular framework considered, that the two models were equivalent, and the abandonment of HMMs in favor of CRFs was not always justified. In other words, faced with a particular application, there is no reason to choose CRFs systematically. However, we also cannot say that HMMs should be chosen. Finally, our contribution is likely to encourage practitioners to consider both models, or rather, according to Remark 3, both parametrizations of the same model, on an equal footing.

We considered basic LC-CRFs and HMMs, which are a limited, yet widely used framework. LC-CRFs and HMMs can be extended in different directions. The general question to study is to search an extension of HMM which would be equivalent to the general CRF (Equation (2)). This seems to be a hard problem. However, one can study some existing extensions of HMMs and wonder what kind of CRFs they would be equivalent to. For example, HMMs have been extended to "pairwise Markov models" (PMMs [23]), and the question is therefore what kind of CRFs would be equivalent to PMMs? Another kind of extension consists of adding a latent process $U_{1:N}$ to the pair $(X_{1:N}, Y_{1:N})$. In the case of CRFs this leads to hidden CRFs (HCRFs [24]), and in the case of HMMs this leads to triplet Markov models (TMMs [25]). Finally, CRFs were generalized to semi-Markov CRFs [26], and HMMs were generalized to hidden semi-Markov models, with explicit distribution of the exact sojourn time in a given state [27], or with explicit distribution of the minimal sojourn time in a given state [28]. The study of the relations between these different extensions of CRFs and HMMs is an interesting perspective for further investigations.

Author Contributions: Conceptualization, E.A., E.M., and W.P.; methodology, E.A., E.M., and W.P.; validation, E.A., E.M., and W.P.; investigation, E.A., E.M., and W.P.; writing—original draft preparation, W.P.; writing—review and editing, E.A., E.M., and W.P; supervision, W.P.; project administration, W.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly funded by the French Government Agency ASSOCIATION NATIONALE RECHERCHE TECHNOLOGIES (ANRT), as part of the CIFRE ("Conventions industrielles de formation par la recherché") scholarship.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Stratonovich, R.L. Conditional Markov Processes. In *Non-Linear Transformations of Stochastic Processes*; Pergamon Press: Oxford, UK, 1965; pp. 427–453.
2. Baum, L.E.; Petrie, T. Statistical Inference for Probabilistic Functions of Finite state Markov Chains. *Ann. Math. Stat.* **1966**, *37*, 1554–1563. [CrossRef]
3. Rabiner, L.; Juang, B. An Introduction to Hidden Markov Models. *IEEE ASSP Mag.* **1986**, *3*, 4–16. [CrossRef]
4. Ephraim, Y. Hidden Markov Processes. *IEEE Trans. Inf. Theory* **2002**, *48*, 1518–1569. [CrossRef]
5. Cappé, O.; Moulines, E.; Ryden, T. *Inference in Hidden Markov Models*; Springer Series in Statistics; Springer: Berlin/Heidelberg, Germany, 2005.
6. Lafferty, J.; McCallum, A.; Pereira, F.C. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001.
7. Sutton, C.; McCallum, A. An Introduction to Conditional Random Fields. *Found. Trends Mach. Learn.* **2012**, *4*, 267–373. [CrossRef]
8. Jurafsky, D.; Martin, J.H. *Speech and Language Processing*; Kindle edition; Prentice Hall Series in Artificial Intelligence; Prentice Hall: Hoboken, NJ, USA, 2014.
9. Ng, A.; Jordan, M. On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes. *Adv. Neural Inf. Process. Syst.* **2001**, *14*, 841–848.
10. He, H.; Liu, Z.; Jiao, R.; Yan, G. A Novel Nonintrusive Load Monitoring Approach based on Linear-Chain Conditional Random Fields. *Energies* **2019**, *12*, 1797. [CrossRef]
11. Condori, G.C.; Castro-Gutierrez, E.; Casas, L.A. Virtual Rehabilitation Using Sequential Learning Algorithms. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 639–645. [CrossRef]
12. Fang, M.; Kodamana, H.; Huang, B.; Sammaknejad, N. A Novel Approach to Process Operating Mode Diagnosis using Conditional Random Fields in the Presence of Missing Data. *Comput. Chem. Eng.* **2018**, *111*, 149–163. [CrossRef]
13. Saa, J.F.D.; Cetin, M. A Latent Discriminative model-based Approach for Classification of Imaginary Motor tasks from EEG data. *J. Neural Eng.* **2012**, *9*, 026020. [CrossRef] [PubMed]
14. Azeraf, E.; Monfrini, E.; Pieczynski, W. On Equivalence between Linear-Chain Conditional Random Fields and Hidden Markov Chains. In Proceedings of the International Conference on Agents and Artificial Intelligence, Virtual, 3–5 February 2022.
15. Liliana, D.Y.; Basaruddin, C. A Review on Conditional Random Fields as a Sequential Classifier in Machine Learning. In Proceedings of the International Conference on Electrical Engineering and Computer Science (ICECOS), Palembang, Indonesia, 22–23 August 2017; pp. 143–148.
16. Ayogu, I.I.; Adetunmbi, A.O.; Ojokoh, B.A.; Oluwadare, S.A. A Comparative Study of Hidden Markov Model and Conditional Random Fields on a Yorùba Part-of-Speech Tagging task. In Proceedings of the IEEE International Conference on Computing Networking and Informatics (ICCNI), Lagos, Nigeria, 29–31 October 2017; pp. 1–6.
17. McCallum, A.; Freitag, D.; Pereira, F.C. Maximum Entropy Markov Models for Information Extraction and Segmentation. *ICML* **2000**, *17*, 591–598.
18. Song, S.L.; Zhang, N.; Huang, H.T. Named Entity Recognition based on Conditional Random Fields. *Clust. Comput. J. Netw. Softw. Tools Appl.* **2019**, *22*, S5195–S5206. [CrossRef]
19. Heigold, G.; Ney, H.; Lehnen, P.; Gass, T.; Schluter, R. Equivalence of Generative and Log-Linear Models. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *19*, 1138–1148. [CrossRef]
20. Azeraf, E.; Monfrini, E.; Vignon, E.; Pieczynski, W. Hidden Markov Chains, Entropic Forward-Backward, and Part-Of-Speech Tagging. *arXiv* **2020**, arXiv:2005.10629.
21. Baum, L.E.; Petrie, T.; Soules, G.; Weiss, N. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Ann. Math. Stat.* **1970**, *41*, 164–171. [CrossRef]
22. Viterbi, A. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Trans. Inf. Theory* **1967**, *13*, 260–269. [CrossRef]
23. Pieczynski, W. Pairwise Markov Chains. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 634–639. [CrossRef]
24. Quattoni, A.; Wang, S.B.; Morency, L.-P.; Collins, M.; Darrell, T. Hidden Conditional Random Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1848–1852. [CrossRef] [PubMed]
25. Pieczynski, W.; Hulard, C.; Veit, T. Triplet Markov Chains in hidden signal restoration. In Proceedings of the SPIE's International Symposium on Remote Sensing, Crete, Greece, 22–27 September 2002.

26. Sarawagi, S.; Cohen, W. Semi-Markov conditional random fields for information extraction. *Adv. Neural Inf. Process. Syst.* **2004**, *17*.
27. Yu, S.-Z. Hidden semi-Markov models. *Artif. Intell.* **2010**, *174*, 215–243. [CrossRef]
28. Li, H.; Derrode, S.; Pieczynski, W. Adaptive on-line lower limb locomotion activity recognition of healthy individuals using semi-Markov model and single wearable inertial sensor. *Sensors* **2019**, *19*, 4242. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.