# ICFHR 2010 Tutorial

# Sequence Modeling: from Hidden Markov Models to Structured Output Prediction (Duration: 3 Hrs.)

## Speakers

Thierry Artières, Professor
Université Pierre et Marie Curie (UPMC)
LIP6, 104 Av. Président Kennedy, 75016, Paris, France
Tel +33-1 44 27 72 20, Fax +33-1 44 27 70 00
E-mail Thierry.Artieres@lip6.fr; Web http://www-connex.lip6.fr/~artieres/

**Thierry Artières** earned a PhD at Paris Sud University in 1995. He joined University of Cergy Pontoise as assistant Professor in 1996 and moved to the computer Science Lab of Pierre et Marie Curie University in Paris (LIP6-UPMC) in 2001 where he is now full Professor since 2007. His teaching covers pattern recognition, data mining and statistical machine learning. He belongs to the PASCAL European network of excellence on machine learning. He is author or coauthor of about 50 international journal and conference papers (IEEE. Trans. on PAMI, Pattern Recognition, ICML, ECML, IJCAI, ECAI, ICDAR, ICFHR). He co-organized IWFHR 2006 at La Baule and KDD 2009 in Paris. He joined a number of program committees and is a frequent reviewer for a number of journals and conferences in data mining, pattern recognition and machine learning. His main research focus are statistical machine learning, sequences and signal processing, signal labeling, Hidden Markov models and Conditional Random Fields, Hybrid Systems. He worked on the development of Markovian systems for signal and sequence labeling tasks with application on various data (speech, on-line and off-line handwriting, eye movements, accelerometers signals, navigation logs) using multi-stream HMMs, segmental and trajectory HMMs, hierarchical HMMs and conditional models such as Conditional Random Fields and its variants. He was also interested in methods to improve the discriminative power of generative systems (such as HMMs and Gaussian Mixtures) through the use of a discriminative criterion for learning GMMs and HMMs or via the development of hybrid systems that combine Neural Networks, Support Vector Machines and HMMs.

## Abstract

Gaussian Hidden Markov Models (GHMMs) have been widely used for automatic speech recognition [Rabiner 1990] and for off-line and on-line handwriting recognition. GHMMs are traditionally learnt to maximize the joint likelihood of observation sequences and of state sequences (MLE) via an EM algorithm. Training is performed based on a partially labelled training set which consists in a set of observation sequences together with the corresponding character sequences. This is the usual setting where one never gets the complete sequence of states corresponding to an observation sequence in the training stage. In test, segmentation is performed through Viterbi decoding that maps an observation sequence into a state sequence. Based on the underlying semantics of the states (e.g. passing through the three states of the left-right HMM corresponding to a particular character means this character has been recognized), the sequence of states translates into a sequence of labels. This typical use of HMMs is very popular since it is both simple and efficient, and it scales well with large corpus.

However, such a learning strategy does not focus on what one is primarily concerned with, namely minimizing the classification (or the segmentation) error rate. Two main directions have been explored to improve the discriminative power of HMM based systems. The first strategy consists in defining an appropriate discriminative criterion and to derive an optimization algorithm able to optimize HMM parameters with respect to this criterion. A number of attempts have been made in this direction. First studies were performed in the speech recognition field, and proposed to optimize a probabilistic criterion such as the Maximum A Posteriori criterion (MAP), Conditional Maximum Likelihood (CML), or information theoretic criterion such as the Maximum Mutual Information (MMI) criterion [Woodland 2002]. Similar ideas have been applied in the handwriting recognition field [Zhang 2007]. Another approach focused on the definition of a criterion that would be more related to the error rate. A first attempt was to use the Minimum Classification Error (MCE) [Juang 1992]. More recently, different authors proposed to exploit a large margin criterion, building on the success of support vector machine for vector data [Jiang 2006, 2007, Sha 2007] (see [Yu 2007] for a review). Up to now such large margin approaches seem to significantly outperform many previously proposed discriminant methods such as Conditional Maximum Likelihood (CML) and Minimum Classification Error (MCE) (Cf. [Sha 2007]). The second strategy consists in investigating slightly different Markovian models whose nature is intrinsically discriminative. One may distinguish here between probabilistic models and non probabilistic ones. Among probabilistic models one should note Maximum Entropy Markov Networks and Conditional Random Fields (CRFs). These models are conditional models that aim at learning the conditional distribution over state sequence given observation sequence, rather than the joint distribution. During training, one finds model's parameters that maximize the conditional likelihood over a supervised training set consisting of input-output pairs. CRFs are an instance of Markov networks where the distribution of state sequence is conditioned on the observation sequence. CRFs avoid the well-known label-bias problem of MEMMs and of HMMs by considering a global normalization [Laferty, 2001]. To deal with partially labelled datasets such as those used in handwriting, a number of works have focused on using hidden variables in the CRF framework. A first investigation dates back 2004 with the work from [Quattoni 2005] who proposed Hidden Conditional Random Fields (HCRF) and the work from [Sarawagi 2005] extending CRF to semi-Markov CRFs. Few other studies on Hidden Conditional Random Fields have been proposed since then in the speech community [Gunawardana 2005; Reiter 2007; Yu 2009]. Non probabilistic models have also been proposed for dealing efficiently with sequences and structured data. They are related to what is called structured output prediction in the machine learning community [Tsochantaridis 2004, Sarawagi 2008]. Among non-probabilistic models Hidden Markov Support Vector Machines (HMSVM) [Altun 2003] and Maximum Margin Markov Networks (M3N) [Taskar et al., 2004] play a central role. While CRF achieves discriminative training by defining a conditional probability and by using maximum conditional likelihood as training criterion, large margin methods such as M3N and HMSVM directly focus on the definition of a discriminative function exploiting the same Markov structure as CRF. Parameters are then learned through the optimization of a large margin criterion. This tutorial aims to provide an overview of existing methods in the two families mentioned above, training generative models discriminatively and designing discriminative models. It will provide the technical basis for understanding the strengths and weaknesses of these methods and to identify potential implementation difficulties.

## Topics to be Covered

1. **Introduction : generative vs. discriminative models for sequences**
2. **Hidden Markov Models for sequences**
   - **Fundamentals and notations**
     - **o Learning and inference**
     - **o Designing models and systems**
   - **Extending Hidden Markov Models**
     - **o Segmental models**
     - **o Trajectory based models**
3. **Discriminative learning for HMMs**
   - **Probabilistic criterion**
     - **o Maximum A Posteriori**
   - **Information theoretic criterion**
     - **o Maximum Mutual Information**
   - **Classification error based criterion**
     - **o Minimum Classification Error**
     - **o Maximum Margin**
4. **Discriminative Markovian approaches**
   - **Structured output prediction**
     - **o Framework**
     - **o Primal and dual forms**
     - **o Special case of sequences**
   - **Conditional Random Fields (CRFs) and Maximum Margin Markov Networks (M3N)**
     - **o Markov Random Fields**
     - **o Conditional models**
       - **· Definition**
       - **· Pros and cons**
     - **o Learning and Inference**
       - **· Basic algorithms**
       - **· Towards scalability**
     - **o CRF as a generalization of HMMs**
   - **Variants of CRF and M3N**
     - **o Dealing with latent variable variant : Hidden CRF**
     - **o Segmental and trajectory variants**