

# Aspects of Semi-Supervised and Active Learning in Conditional Random Fields

Nataliya Sokolovska

LRI, CNRS UMR 8623 & INRIA Saclay,  
University Paris Sud, Orsay, France  
nataliya@lri.fr

**Abstract.** Conditional random fields are among the state-of-the art approaches to structured output prediction, and the model has been adopted for various real-world problems. The supervised classification is expensive, since it is usually expensive to produce labelled data. Unlabeled data are relatively cheap, but how to use it? Unlabeled data can be used to estimate marginal probability of observations, and we exploit this idea in our work.

Introduction of unlabeled data and of probability of observations into a purely discriminative model is a challenging task.

We consider an extrapolation of a recently proposed semi-supervised criterion to the model of conditional random fields, and show its drawbacks. We discuss alternative usage of the marginal probability and propose a pool-based active learning approach based on quota sampling. We carry out experiments on synthetic as well as on standard natural language data sets, and we show that the proposed quota sampling active learning method is efficient.

**Key words:** conditional random fields, probability of observations, active learning, semi-supervised learning

## 1 Introduction

In real-world applications (text, image, audio data processing) unlabeled data are plentiful and cheap. Labeled data, on the contrary, are usually rather expensive to gather. The problem how to exploit unlabeled instances is not recent and many proposals have been already made. Another problem is how to select training data. How to choose instances of high training utility is the active learning problem.

Intuitively, the information one can get from unlabeled data is the marginal probability of observations. In the asymptotic case, when we dispose infinitely many unlabeled instances, we can estimate the true marginal probability of observations. In real-world problems this is not feasible, since the number of observations is always limited. However, the probability of observations can be approximated.

Attention of the machine learning and data mining communities has been drawn to semi-supervised approaches (see [4] for an overview), especially by probabilistic semi-supervised classifiers. Logistic regression is a simple efficient discriminant model widely used for various applications. However, a number of real-world applications has sequential structure, e.g., natural language and biological applications. Conditional random fields [11] are a generalization of logistic regression, and therefore, a discriminative approach, which models sequential dependencies and allows to take a rich set of features into account.

Probabilistic generative models fare easily with unlabeled data, usually via the expectation-maximization algorithm [6,22]. Discriminative probabilistic models are reported to perform better than probabilistic generative models [17]. The introduction of unlabeled instances into discriminative models is much more challenging, since it is not straightforward how to integrate marginal probability of observations into a discriminative model.

Among the state-of-the art semi-supervised methods are combinations of generative and discriminative approaches in order to profit from both aspects, a better generalization error of a discriminative model and information extracted from unlabeled data, integrated into a generative model. A convex combination of a discriminative model and a generative model is considered e.g. by [2] and [9]. A Bayesian point of view for the hybrid approaches has been explored by [16] and [12]. The proposed hybrid method is based on the fact that parameters of a discriminative and of a generative models are related via their Bayesian distribution. However, the number of parameters to be estimated is usually doubled in the hybrid approaches, since the number of models increases.

The criterion of Bengio-Grandvalet [8] was probably the first attempt to introduce unlabeled data into a discriminant classifier. The criterion implements the idea that the classes have to be well-separated; conditional entropy over unlabeled instances is taken as a measure of overlap of classes. Among significant disadvantages of the criterion are its non-convexity and instability in cases where the number of labeled points is small.

In this paper, we discuss ideas how to introduce the marginal probability of observations into a purely discriminative model, into the model of conditional random fields (CRFs). It has been shown that the recently proposed semi-supervised discriminative criterion [26] is efficient under model misspecification and covariate shift scenarios for “simple” (i.e. without underlying structure) output tasks. We apply the semi-supervised approach to the criterion of conditional random fields and carry out experiments on structured output problems. We discuss the limits and drawbacks of the criterion.

We propose to integrate the marginal probability of observations into an active learning framework for the structured output prediction. We demonstrate on the standard natural language processing data sets that the proposed pool-based active learning approach based on quota sampling is efficient.

The paper is organized as follows: in Section 2 we consider the asymptotically optimal semi-supervised criterion [26], Section 3 introduces the model of conditional random fields [11], widely used for structured output prediction. In

Section 4 we discuss the application of the semi-supervised criterion to the model of conditional random fields. Section 5 discusses the limits of the semi-supervised discriminative criterion applied to the CRFs and introduces our approach of pool-based active learning based on quota sampling. Section 6 illustrates our experiments on synthetic as well as real-world applications. We discuss the state-of-the-art approaches and related work in Section 7. Concluding remarks and perspectives close the paper.

## 2 Semi-Supervised Discriminant Estimator

To start with, let us place in a context of classification without taking any structure into consideration. Let the observation variable  $X$  take its values in a finite set  $\mathcal{X}$ ;  $Y$  is the class variable which takes its values in  $\mathcal{Y}$ . We suppose  $\mathcal{Y}$  to be a finite set, and  $\{X_i, Y_i\}_{i=1}^n$  are observations and their labels available for the training.

Let  $g(y|x; \theta)$  be the conditional probability function, parameterized by  $\theta$ . Then the standard conditional maximum likelihood estimator is defined by

$$\hat{\theta}_n = \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ell(Y_i|X_i; \theta), \quad (1)$$

where  $\ell(y|x; \theta) = -\log g(y|x; \theta)$  denotes the negated conditional log-likelihood function.

The asymptotically optimal semi-supervised estimator  $\hat{\theta}_n^s$  proposed by [26] is defined by

$$\hat{\theta}_n^s = \arg \min_{\theta \in \Theta} \sum_{i=1}^n \frac{q(X_i)}{\sum_{j=1}^n \mathbb{1}\{X_j = X_i\}} \ell(Y_i|X_i; \theta), \quad (2)$$

where  $q(x)$  is the marginal probability of observations and can be considered as some prior knowledge. We suppose that infinitely many observations are available, and that the true value  $q(x)$  can be estimated. The semi-supervised estimator presented as eq. (2) is a weighted version of the usual conditional maximum likelihood estimator.

The semi-supervised estimator is shown to be asymptotically optimal and to be particularly efficient for the misspecified cases, that is if  $g(y|x; \theta_*) \neq \eta(y|x)$ , where  $\eta(y|x)$  is the true conditional probability that generated the data; in the following,  $\pi(y, x) = \eta(y|x)q(x)$ .

To be precise, the essential properties of the standard and weighted (semi-supervised) estimators consist in the following:

$$\sqrt{n} \left( \hat{\theta}_n - \theta_* \right) \xrightarrow{L} \mathcal{N} \left( 0, J^{-1}(\theta_*) I(\theta_*) J^{-1}(\theta_*) \right), \quad (3)$$

$$\sqrt{n} \left( \hat{\theta}_n^s - \theta_* \right) \xrightarrow{L} \mathcal{N} \left( 0, J^{-1}(\theta_*) H(\theta_*) J^{-1}(\theta_*) \right), \quad (4)$$

where

$$H(\theta_*) = E_q (V_\eta [\nabla_\theta \ell(Y|X; \theta_*)|X]), \quad (5)$$

$$I(\theta_*) = E_\pi \left[ \nabla_\theta \ell(Y|X; \theta_*) \{ \nabla_\theta \ell(Y|X; \theta_*) \}^T \right], \quad (6)$$

$$J(\theta_*) = E_\pi [\nabla_{\theta^T} \nabla_\theta \ell(Y|X; \theta_*)]. \quad (7)$$

The case of a covariate shift (observation variables are sometimes called explanatory variables or covariates) is a rather frequent situation in real-world applications. The covariate shift arises if  $q_0(x) \neq q_1(x)$ , where  $q_0(x)$  is determined by the sampling scheme and  $q_1(x)$  is determined by the population.

In the absence of covariate shift:

$$\lim_{n \rightarrow \infty} \frac{q_1(x_i)}{n^{-1} \sum_{j=1}^n \mathbb{1}\{x_j = x_i\}} \rightarrow 1. \quad (8)$$

With a covariate shift, we have:

$$\lim_{n \rightarrow \infty} \frac{q_1(x_i)}{n^{-1} \sum_{j=1}^n \mathbb{1}\{x_j = x_i\}} \rightarrow \frac{q_1(x_i)}{q_0(x_i)}. \quad (9)$$

The weighting scheme by the importance ratio is considered in [24].

The semi-supervised estimator of eq. (2) is shown to be asymptotically optimal under the covariate shift [25]. The advantage of the semi-supervised approach can be observed only when considering the scaled excess logarithmic risk

$$n(E_\pi[\ell(Y|X; \hat{\theta}_n)] - E_\pi[\ell(Y|X; \theta_*)]) \quad (10)$$

or the scaled squared error

$$n \|\hat{\theta}_n - \theta_*\|^2. \quad (11)$$

The true marginal probability of observations have to be provided to compute both the scaled excess logarithmic risk and the scaled squared error. However, as we have already mentioned, this is not possible for real-world applications.

### 3 Conditional Random Fields

Conditional random fields (CRF) [11,27] are a discriminative model based on the following probabilistic distribution

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\theta(\mathbf{x})} \exp \left\{ \sum_{t=1}^T \sum_{k=1}^K \theta_k f_k(y_{t-1}, y_t, x_t) \right\}, \quad (12)$$

where  $\mathbf{x} = (x_1, \dots, x_T)$  denotes the sequence of observations (input) and  $\mathbf{y} = (y_1, \dots, y_T)$  is the sequence of labels (output);  $\{f_k\}_{1 \leq k \leq K}$  is an arbitrary set of feature functions and  $\{\theta_k\}_{1 \leq k \leq K}$  are the associated real-valued parameter values. By convention,  $y_0$  refers to a particular (always observed) label which indicates the beginning of the sequence.

The CRF form presented as eq. (12) is usually referred to as linear-chain CRF, although  $y_t$  and  $x_t$  could be composed not only of the individual sequence tokens, but of sub-sequences ( $n$ -grams) of some fixed length or other localized characteristics.

We will denote by  $\mathcal{Y}$ ,  $\mathcal{X}$ , respectively, the sets in which  $y_t$  and  $x_t$  take their values. The normalization factor in eq. (12) is defined by

$$Z_\theta(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}^T} \exp \left\{ \sum_{t=1}^T \sum_{k=1}^K \theta_k f_k(y_{t-1}, y_t, x_t) \right\}. \quad (13)$$

One of the possible ways to define features is the combination of bigram  $\lambda_{y',y,x}$  and unigram  $\mu_{y,x}$  features

$$\begin{aligned} \sum_{k=1}^K \theta_k f_k(y_{t-1}, y_t, x_t) &= \sum_{y \in \mathcal{Y}, x \in \mathcal{X}} \mu_{y,x} \mathbb{1}\{y_t = y, x_t = x\} + \\ &\quad \sum_{y', y \in \mathcal{Y}^2, x \in \mathcal{X}} \lambda_{y',y,x} \mathbb{1}\{y_{t-1} = y', y_t = y, x_t = x\}, \end{aligned} \quad (14)$$

where  $\mathbb{1}(\text{test}) = 1$ , if the variables are observed jointly and 0 otherwise. We can rewrite equation (14) as  $\mu_{y_t, x_t} + \lambda_{y_{t-1}, y_t, x_t}$ , and we use this more compact representation in the following. We use this feature combination, unigram and bigram templates, in our experiments in Section 6.

Given  $N$  independent labeled sequences  $\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^N$ , the conditional maximum likelihood estimation is based on the minimization, with respect to  $\theta$ , of the negated log-likelihood

$$\begin{aligned} \ell(\mathcal{D}; \theta) &= - \sum_{i=1}^N \log p_\theta(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}) \\ &= \sum_{i=1}^N \left\{ \log Z_\theta(\mathbf{x}^{(i)}) - \sum_{t=1}^{T_i} \sum_{k=1}^K \theta_k f_k(y_{t-1}^{(i)}, y_t^{(i)}, x_t^{(i)}) \right\}, \end{aligned} \quad (15)$$

where  $T_i$  is the length of an observation  $\mathbf{x}^{(i)}$ .

Although  $\ell(\mathcal{D}; \theta)$  is a smooth convex function, it has to be optimized numerically, and standard gradient-based methods, such as a quasi-Newton approach, can be applied directly.

The gradient of  $\ell(\mathcal{D}; \theta)$  is given by

$$\begin{aligned} \frac{\partial \ell(\theta)}{\partial \theta_k} &= \sum_{i=1}^N \sum_{t=1}^{T_i} \mathbb{E}_{p_\theta(\mathbf{y} | \mathbf{x}^{(i)})} f_k(y_{t-1}, y_t, x_t^{(i)}) - \\ &\quad \sum_{i=1}^N \sum_{t=1}^{T_i} f_k(y_{t-1}^{(i)}, y_t^{(i)}, x_t^{(i)}), \end{aligned} \quad (16)$$

where  $\mathbb{E}_{p_\theta(\mathbf{y} | \mathbf{x}^{(i)})} f_k(y_{t-1}, y_t, x_t^{(i)})$  denotes the conditional expectation.

In our experiments, the log-likelihood is penalized by the  $L_2$  norm to avoid overfitting.

## 4 Semi-Supervised Conditional Random Fields

The semi-supervised criterion presented as eq. (2) applied to the conditional random fields criterion, referred later to as weighted CRFs, takes the form:

$$C(\theta) = \sum_{\mathbf{x} \in \mathcal{X}} -q(\mathbf{x}) \frac{1}{N_{\mathbf{x}}} \log p_{\theta}(\mathbf{y}|\mathbf{x}), \quad (17)$$

where  $p_{\theta}(\mathbf{y}|\mathbf{x})$  is defined by eq. (12), and  $N_{\mathbf{x}}$  is the number of times a sequence  $\mathbf{x}$  has been observed in the training corpus.

The marginal probability of observations  $q(\mathbf{x})$  has to be provided or approximated and introduced into the model of conditional random fields. In our case, the observations are sequences, what makes the task even more difficult.

If our data are artificial, generated by a hidden Markov model of the first order, then estimation of the probability of the observation sequences is straightforward. Following the standard notations [20], let  $A$  be the state transition probabilities,  $B$  be the observation probability matrix,  $p(y)$  be the initial state distribution,  $\mathbf{x} = (x_1, x_2, \dots, x_T)$  be an observation sequence of the length  $T$ . The probability of a series of observations, i.e., of a sequence is given by

$$\begin{aligned} q(\mathbf{x}) &= \sum_{\mathbf{Y}} p(\mathbf{x}, \mathbf{y}) \\ &= \sum_{\mathbf{Y}} p(y_1) b_{y_1}(x_1) a_{x_1, x_2} b_{y_2}(x_2) \dots a_{x_{T-1}, x_T} b_{y_T}(x_T). \end{aligned} \quad (18)$$

Usually in real-world problems, the structure of observations is unknown. It is not possible to compute the marginal probability of observations exactly, and it has to be estimated empirically.

Note, that  $N_{\mathbf{x}}$  equals 1 in a number of real-world applications, since each sequence is observed usually only once in a training corpus.

## 5 Motivation for Pool-Based Active Learning

The application of the semi-supervised discriminative estimator to real-world data sets does not always ameliorate the performance.

There are several reasons, why the performance of the criterion does not dominate the performance of the standard approach. The semi-supervised criterion performs better in the case of a misspecified model (the more a model is misspecified, the more efficient is the semi-supervised criterion compared to the standard, not weighted approach) and under a covariate shift. Usually, both scenarios are typical for a real data set. However, carrying out experiments on simulated data, we have noticed that the advantage of the semi-supervised approach is observed only when considering the scaled excess logarithmic risk, eq. (10), and the squared error, eq. (11). To compute these values, the true distribution of the observations has to be provided. In any real-world task the distribution is not available. From a number of experiments on the real data we

made a conclusion that although the marginal probability of observations can be efficiently approximated, the approximation is still not good enough to be used in the semi-supervised estimator instead of the true one.

However, we guess that even an approximation of the marginal distribution can be informative. We propose to use the probability of observations to sample a pre-defined pool of training instances (of a small size  $n$ ) to achieve the best possible generalization error. Our idea is close, in some sense, to [32], who considered an active learning approach based on self-training.

In the discriminative semi-supervised criterion presented as eq. (2), training instances with high probability can be automatically considered to be more important than those with low probability. In this sense, the discriminative semi-supervised approach is associated with stratified sampling. However, one of natural language phenomena consists in that rare events are as important as frequent events, and can therefore not be neglected.

We propose to apply the non-probabilistic quota sampling to select training sequences efficiently. In the method we propose, candidates for training instances are sorted according to their marginal probabilities and are divided into  $n$  groups, where  $n$  is the number of observations we use for the training procedure. We choose (randomly) one training instance from each group. Under quota sampling we mean here that we sample (uniformly) data instances from each frequency group. Therefore, we guarantee that we train our model taking frequent as well as rare dependencies into consideration.

We illustrate on standard natural language processing problems, in Sections 6.3 and 6.4 that the quota sampling (QS) pool-based active learning approach outperforms training procedures which choose their training instances randomly. In Section 6.2 we show that the proposed method outperforms a state-of-the art approach FuSAL.

## 6 Experiments on Artificial and Real Data Sets

In this section, we describe our experiments and provide our results on synthetic and two standard natural language processing problems, namely on NetTalk Phonetisation Task and on CoNLL 2003 challenge.

Section 6.1 illustrates limits of the semi-supervised criterion, even for an artificial data set. We demonstrate on real data that the proposed QS pool-based active learning is an efficient approach (Sections 6.3 and 6.4), in particular in case where the number of observations  $n$  is small.

The state-of-the art performance (mostly in the context of fully supervised learning) of the considered real data sets can be found, for instance, in [25].

We would like to underline that we are especially interested in cases where  $n$ , the number of observed instances, is small.

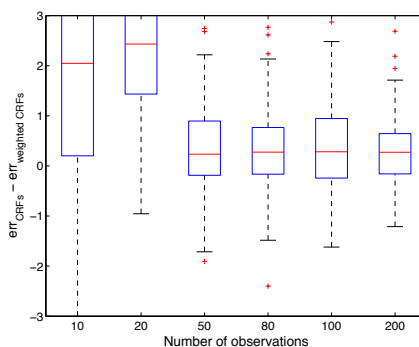
### 6.1 Weighted Conditional Random Fields Experiments

The synthetic sequential data are simulated with hidden Markov models of the first order. The observation alphabet contains 5 symbols, the size of the labels

alphabet is 6. All simulated sequences are of the same length which equals 5. The minimal achievable error is about 6%. The value of Bayes error is approximated by a percentage of errors obtained by decoding using the true values of the state transition and observation probability matrices.

Since we know the distribution which generates the data and the true parameters are available, we use the forward algorithm and eq. (18) to compute the marginal probability of observations  $q(\mathbf{x})$ .

The results of our experiments with the synthetic sequential data are illustrated by Figure 1. The size of the training corpus varies from 10 to 200 training instances. The percentage of error is always estimated on test data (test data contains 10000 instances). The number of Monte-Carlo replications in the experiment is 150. The boxplotted difference, which is shown on Figure 1, is positive, if the weighted CRFs performs better, i.e. has a lower error rate than the standard approach.



**Fig. 1.** Simulated data. Difference of error rates of standard and weighted conditional random fields by marginal probability.

The difference in performance of the standard and the semi-supervised CRFs is significant only for  $n = 10, 20$ , and not significant for larger  $n$ , even in the ideal situation, where we know the exact marginal probability of observations.

As to the real-world data experiments, where we dispose only the approximated values of the probability of observations, we consider the difference in performance to be not significant.

## 6.2 Fully Supervised Active Learning Approach (FuSAL)

We compare the performance of the proposed pool-based active learning to the one of a state-of-the art method called FuSAL (Fully Supervised Active Learning) introduced in [32]. Algorithm 1 describes the approach. A utility function we use in our experiments is the same as in [32]

$$u_{\theta}(\mathbf{x}) = 1 - p_{\theta}(\hat{\mathbf{y}}|\mathbf{x}), \quad (19)$$

where  $\hat{\mathbf{y}}$  is computed using the decoding Viterbi algorithm, and  $\theta$  corresponds to the current model. The intuition behind the utility function is to consider sequences for which the current model is least confident to be more important than other observations.

---

**Algorithm 1** General Active Learning Framework
 

---

$m$  – number of examples selected within one loop  
 $\mathcal{D}_l$  – set of labeled instances  
 $\mathcal{D}_u$  – set of unlabeled instances  
 $u_\theta(\mathbf{x})$  – utility function

**while** stopping criterion is not met **do**  
   train model  $M$  using  $\mathcal{D}_l$   
   estimate  $u_\theta(\mathbf{x}_i) \forall \mathbf{x}_i \in \mathcal{D}_u$   
   choose  $m$  examples whose  $u_\theta(\mathbf{x})$  is maximal  
   get labels for the  $m$  chosen instances  
   move the  $m$  labeled examples from  $\mathcal{D}_u$  to  $\mathcal{D}_l$   
**end while**

---

In our experiments, we add instances which are to be labeled one by one ( $m = 1$ ). The first instance is chosen randomly from the training corpus. The stopping criterion is the number  $n$  of observed sequences. If the cardinality of  $\mathcal{D}_l$  is equal to  $n$ , the stopping criterion is met.

### 6.3 Active Learning Experiments on Nettetalk Phonetisation Task

The original Nettetalk corpus has been introduced in [23]. The Nettetalk corpus we use in our experiments has been suggested for the Pascal Letter-to-Phoneme Conversion Challenge<sup>1</sup>. The English data set contains 16280 words aligned with their phonetical transcriptions. The alphabet of observation symbols includes 26 letters, and the number of phonemes, i.e., the number of labels, is 53 including the alignment symbol. The corpus is split into 10 parts. Each part includes 1628 sequences of observations and corresponding labels. One part, i.e., 1628 sequences, is used to test the performance. We use all available observation sequences of the corpus to estimate empirically the probability of observations  $q(\mathbf{x})$ .

To approximate  $q(\mathbf{x})$ , we follow the idea of  $n$ -gram linguistic models [7]. We let  $q(\mathbf{x}) = q(x_1, \dots, x_T) = \prod_t p(x_t | x_{t-1}, x_{t-2}, x_{t-3})$ , where

$$\frac{p(x_t | x_{t-1}, x_{t-2}, x_{t-3})}{C(x_t, x_{t-1}, x_{t-2}, x_{t-3}) / C(x_{t-1}, x_{t-2}, x_{t-3})}, \quad (20)$$

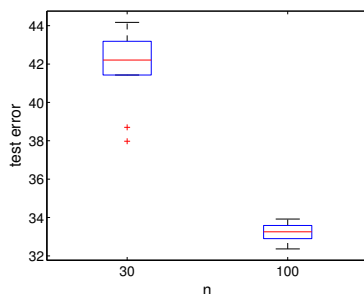
$C(\cdot)$  means counts estimated on all available observations.

The estimated  $q(\mathbf{x})$  are sorted into  $n$  frequency groups, and we sample one training instance from each frequency group. The training is performed with two

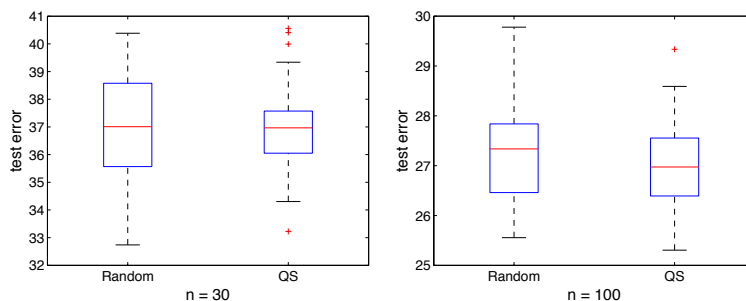
<sup>1</sup> <http://pascallin.ecs.soton.ac.uk/Challenges/PRONALSYL/>

types of features, bigram and unigram, as shown by eq. (14). The regularization parameter  $\rho$  of the penalty term  $\rho\|\theta\|^2$  is the same for all tested approaches, QS active learning, random sampling, and FuSAL, and is fixed to  $\rho = 0.1$  (the value is chosen by cross validation).

Figure 2 illustrates the performance of the FuSAL method on the Nettetalk data sets (50 Monte-Carlo replications). One of its obvious disadvantages and hence causes of its poor performance is that the method is not suitable for cases where  $n$  is small. To train the initial model, the method requires a number of labeled sequences, and if these sequences are not selected carefully, the training results in a model whose error rate on the testing set is large. It is not reasonable to compute the utility function, and therefore perform active learning based on a model which is not efficient.



**Fig. 2.** FuSAL performance (error rates). Nettetalk corpus,  $n = 30, 100$ .



**Fig. 3.** Nettetalk corpus. Comparison of error rates for  $n = 30$  and  $n = 100$ . The pool-based active learning based on quota sampling (QS) is more efficient than random choice of training sequences.

Figure 3 illustrates our results of the pool-based active learning approach compared to random sampling. We performed 50 Monte-Carlo replications. For a small number of observations,  $n = 30$  and  $n = 100$ , we noticed that the test

error and its variance are smaller if observations are chosen according to the proposed pool-based active learning method and not randomly.

It is easy to see that the QS approach outperforms the random sampling and FuSAL.

For the qualitative analysis of sequences selected by the proposed quota sampling method and the standard approach presented as Algorithm 1, see Tables 1 and 2 respectively. Note that applying the utility function, eq. (19), we tend to select sequences of similar morphological structure.

**Table 1.** Nettealk corpus. Sequences chosen by QS,  $n = 30$

ail	inconceivably	neat	superlative	chase
sworn	interstate	strain	unnaturally	fresco
secret	invertebrate	comrade	ennoble	haughtily
dribble	meditate	parasite	woodwork	meteoric
shoemaker	unstained	simpleton	soberly	snake
chloroform	aspire	babe	cheese	rise

**Table 2.** Nettealk corpus. Sequences chosen by FuSAL (Algorithm 1),  $n = 30$ ,  $m = 1$

hogshead	shepherdess	aggressiveness
misrepresentation	representation	representative
misapprehension	interdependence	superintendence
superintendent	misunderstanding	experimentation
standardization	interpretation	transcontinental
undenominational	unconstitutional	counterrevolution
indiscriminately	characteristically	internationally
characterization	instantaneously	enthusiastically
constitutional	conscientiously	incomprehensible
intermittently	instrumentality	correspondingly

#### 6.4 Active Learning Experiments on CoNLL 2003 Corpus

Named entity recognition consists in extracting groups of syntagmas that correspond to named entities (e.g., names of persons, organizations, places, etc.). The data used for our experiments are taken from the CoNLL 2003 challenge [31] and imply four distinct types of named entities. Labels have the form B-X or I-X, where B means “begin” and I means “inside” of a named entity X (note that the label B-PER is not present in the corpus). Words that are not included in any named entity, are labeled with O (outside). Overall, there are 8 labels.

At each position in the text, the input consists of three separate components, so we have three types of observations: a word (with 30290 distinct words in the

corpus), its part-of-speech (44), and syntactic (18) tags. The training set includes about 15000 sequences (phrases). Development set (Test A) and test set (Test B) include about 3500 sequences each.

We use all available sequences to estimate  $q(\mathbf{x})$ . We apply the same approach as for the Nettealk corpus, described in the previous section. However, for the CoNLL 2003 data set we use the Markovian dependency of the second, and not of the third, order. Since the data set has three types of observations, we have to take into consideration marginal probabilities of each type of observation. The probability  $q(\mathbf{x})$  is approximated by the product of marginal probabilities of its components  $p(\mathbf{x}_{\text{word}})p(\mathbf{x}_{\text{POS tag}})p(\mathbf{x}_{\text{synt. tag}})$ .

The training is carried out with two dependencies, unigram and bigram, extracted for each type of observation, i.e. our feature choice is as follows:

$$\begin{aligned} & \mu_{y_t, x_{\text{word}, t}} + \mu_{y_t, x_{\text{POS tag}, t}} + \mu_{y_t, x_{\text{synt. tag}, t}} \\ & + \lambda_{y_t, x_{\text{word}, t}} + \lambda_{y_t, x_{\text{POS tag}, t}} + \lambda_{y_t, x_{\text{synt. tag}, t}}. \end{aligned} \quad (21)$$

The regularization parameter  $\rho$  of the penalty term  $\rho\|\theta\|^2$  is chosen by cross validation and is the same for all tested sampling methods,  $\rho = 0.5$ .

Figure 4 illustrates the results of our experiments with FuSAL. Figure 5 shows the results for random sampling and QS. We carried out 50 Monte-Carlo replications for all methods. For a small number of observations,  $n = 10$  and  $n = 50$ , as illustrated on the figure, it is obvious that the error rate on the test data (both test A and test B sets) is much smaller while using the quota sampling active learning than choosing training instances arbitrary. FuSAL is less efficient as well.

Our experiments show that FuSAL is an acceptable active learning method if some initial, not very small,  $\mathcal{D}_l$  is provided and if a reasonable initial model  $M$  can be created.

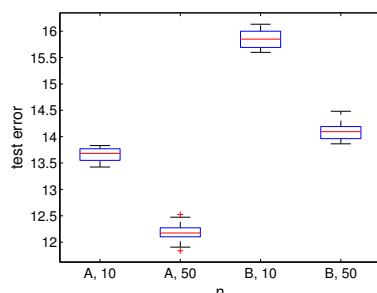
## 7 Related Work

Many proposals of semi-supervised methods have been recently made to sequence labeling. As to active learning for structured output prediction, there are much less published ideas.

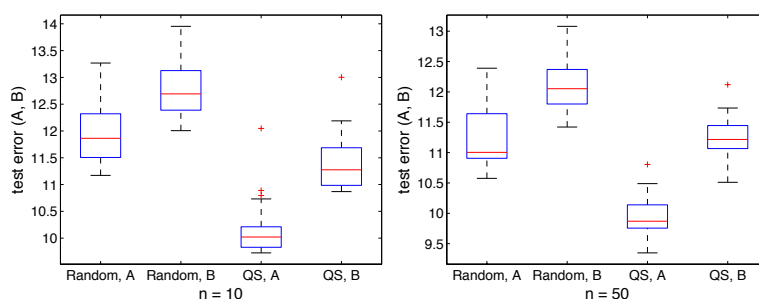
A maximum margin semi-supervised learning approaches for structured output prediction are described in [1] and [3]; [10], [15], and [13] discuss semi-supervised learning for conditional random fields.

The minimum entropy regularization approach of Grandvalet and Bengio [8] has been applied to conditional random fields by [10] :

$$\begin{aligned} & - \sum_{i=1}^{|\mathcal{D}_l|} \log p_{\theta}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}) + \frac{\|\theta\|^2}{2\sigma^2} - \\ & \rho \sum_{i=|\mathcal{D}_l|+1}^{|\mathcal{D}_l|+|\mathcal{D}_u|} \sum_{\mathbf{y}} p_{\theta}(\mathbf{y}|\mathbf{x}^{(i)}) \log p_{\theta}(\mathbf{y}|\mathbf{x}^{(i)}), \end{aligned} \quad (22)$$



**Fig. 4.** FuSAL performance (error rates). CoNLL 2003, for test A and test B sets,  $n = 10, 50$ .



**Fig. 5.** CoNLL 2003 data set. Comparison of error rates (for test A and test B sets) while training on  $n = 10$  and  $n = 50$  sequences. Active learning based on marginal probability (QS on the boxplots) is much more efficient than arbitrary choice of observations for training.

where  $\mathcal{D}_l$  are labeled data and  $\mathcal{D}_u$  are unlabeled instances;  $\sigma^2$  and  $\rho$  are parameters fixed usually by cross validation. The direct computation of the gradient of the entropy term of the criterion requires  $O(T^2|Y|^3)$  operations in comparison to  $O(T|Y|^2)$  of a standard forward-backward procedure. [13] proposed an efficient way (complexity of a standard forward-backward algorithm) to compute the gradient of the criterion presented in (22).

A hybrid semi-supervised model is proposed in [28]. The model combines discriminative and generative models, the parameters  $\Gamma = \{\{\gamma_i\}_{i=1}^I, \{\gamma_j\}_{j=I+1}^{I+J}\}$  are associated with  $I$  generative and  $J$  discriminative models. Unlabeled data are introduced into the generative models. The following criterion

$$p(\mathbf{y}|\mathbf{x}, \Lambda, \Theta, \Gamma) \propto \prod_i p_i^D(\mathbf{y}|\mathbf{x}, \lambda_i)^{\gamma_i} \prod_j p_j^G(\mathbf{x}, \mathbf{y}, \theta_j)^{\gamma_j} \quad (23)$$

contains three sets of parameters to be estimated,  $\Gamma$ ,  $\Lambda$ , and  $\Theta$ . The values of  $\Lambda$  are estimated on labeled data. An iterative optimization procedure runs until

convergence is used to adjust  $F$  (parameters of hybrid models) and parameters  $\Theta$  associated with discriminative components.

Recently [29] introduced a semi-supervised approach that is simpler than the one proposed in [28], since there are only two parameter vectors to be estimated. The parameter vector  $\Lambda$  is estimated on labeled data using a discriminative model, and  $\Theta$  on unlabeled data, using a generative approach. However, the number of parameters to be estimated is quite large.

The approach discussed in [28] was called a great step forward in hybrid models [5], since it combines models that take the underlying structure into account, namely hidden Markov models and conditional random fields. The approach of [29] has been recently applied to parsing problems by [30].

One of the recent works on semi-supervised learning applied to natural language processing is a trial to add incomplete annotations [33]. Ambiguous annotations are considered as candidate labels, and parameters are estimated by marginalizing out the unknown labels. The method is a particular case of hidden conditional random fields, introduced in [19].

The idea to introduce the knowledge of labels proportions, the method called “expectation regularization”, proposed in [14] for maximum entropy models, has been generalized in [15] for structured output prediction, using linear-chain CRFs. The approach was called generalized expectation. It was supposed that not only fully labeled instances can be used but labeled features as well.

Recently [32] proposed an approach that combines semi-supervised and active learning. The semi-supervised active learning method [32] which is actually self-training active learning approach, selects instances of high utility to be labeled and to be used for training. Estimation of utility of a given sequence is a problem in itself, since it can be done in many different ways.

It is discussed in [32] whether it is more reasonable to label manually only subsequences (e.g., features) of high utility instead of labeling entire sequences. A similar idea, an efficient learning of features from unlabeled data is considered in [18].

## 8 Conclusion

In this contribution, we addressed two problems, semi-supervised learning and active learning in discriminative models, more specifically, in conditional random fields. We demonstrated on the artificial data set that the considered discriminative semi-supervised method can be applied to conditional random fields. However, its application to real tasks is still an open problem, since an efficient approximation of the probability of observations, whose structure is complex and unknown, is still a challenge.

The proposed pool-based active learning method is based on the intuition that rare observations are not less important than frequent observations. In particular, this is the case in the domain of natural language processing. We have shown that selecting training instances using quota sampling is much more efficient in terms of error rates on test data than choosing them randomly. The

proposed approach is also more efficient than FuSAL, a state-of-the art method. Most of state-of-the art active learning methods (e.g., FuSAL) are based on the idea that there already exists a set of labeled instances, and are therefore not suitable for cases where the number of labeled points is very limited.

An important advantage of the proposed quota sampling approach is simplicity of implementation. The open issue is the theoretical analysis of the proposed quota sampling pool-based active learning approach, which is quite efficient on the real-world data sets.

## References

1. Y. Altun, D. McAllester, and M. Belkin. Maximum margin semi-supervised learning for structured variables. *NIPS*, 2005.
2. G. Bouchard and B. Triggs. The trade-off between generative and discriminative classifiers. *IASC*, 2004.
3. U. Brefeld and T. Scheffer. Semi-supervised learning for structured output variables. *ICML*, 2006.
4. O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.
5. Hal Daumé III. Semi-supervised or semi-unsupervised? In *NAACL Workshop on Semi-supervised Learning for NLP*, 2009.
6. A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of The Royal Statistical Society Series B*, 39(1):1–38, 1977.
7. J. T. Goodman. A bit of progress in language modeling. Technical Report MSR-TR-2001-72, Microsoft Research, Redmond, August 2001.
8. Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. *NIPS*, 2004.
9. A. Holub and P. Perona. A discriminative framework for modelling object classes. *CVPR*, 2005.
10. F. Jiao, S. Wang, C. H. Lee, R. Greiner, and D. Schuurmans. Semi-supervised conditional random fields for improved sequence segmentation and labeling. *ACL/COLING*, 2006.
11. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. *ICML*, 2001.
12. J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. *CVPR*, 2006.
13. G. Mann and A. McCallum. Efficient computation of entropy gradient for semi-supervised conditional random fields. *NAACL/HLT*, 2007.
14. G. Mann and A. McCallum. Simple, robust, scalable semi-supervised learning via expectation regularization. *ICML*, 2007.
15. G. Mann and A. McCallum. Generalized expectation criteria for semi-supervised learning of conditional random fields. *ACL*, 2008.
16. T. Minka. Discriminative models, not discriminative training. Technical Report TR-2005-144, Microsoft Cambridge, 2005.
17. A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *NIPS*, 2002.
18. Y. Qi, P. P. Kuksa, R. Collobert, K. Kavukcuoglu, and J. Weston. Semi-supervised sequence labelling with self-learned feature. *ICDM*, 2009.

19. A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. *NIPS*, 2004.
20. L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
21. H.J. Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11:363–371, 1965.
22. M. Seeger. Learning with labeled and unlabeled data. Technical report, University of Edinburgh, Institute for Adaptive and Neural Computation, 2002.
23. T. J. Sejnowski and C. R. Rosenberg. Parallel networks that learn to pronounce english text. *Complex Systems*, 1, 1987.
24. H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244, 2000.
25. N. Sokolovska. *Contributions to estimation of probabilistic discriminative models: semi-supervised learning and feature selection*. PhD thesis, TELECOM ParisTech, 2010.
26. N. Sokolovska, O. Cappé, and F. Yvon. The asymptotics of semi-supervised learning in discriminative probabilistic models. *ICML*, 2008.
27. C. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. In Lise Getoor and Ben Taskar, editors, *Introduction to Statistical Relational Learning*. The MIT Press, Cambridge, MA, 2006.
28. J. Suzuki, A. Fujino, and H. Isozaki. Semi-supervised structured output learning based on a hybrid generative and discriminative approach. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007.
29. J. Suzuki and H. Isozaki. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. *ACL*, 2008.
30. J. Suzuki, H. Isozaki, X. Carreras, and M. Collins. An empirical study of semi-supervised structured conditional models for dependency parsing. *EMNLP*, 2009.
31. E. F. Tjong Kim Sang and F. de Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *CoNLL*, 2003.
32. K. Tomanek and U. Hahn. Semi-supervised active learning for sequence labeling. *ACL and AFNLP*, 2009.
33. Y. Tsuboi, H. Kashima, S. Mori, H. Oda, and Y. Matsumoto. Training conditional random fields using incomplete annotations. *COLING*, 2008.