2001

Latent Maximum Entropy Principle for Statistical Language Modeling

Shaojun Wang  
Wright State University - Main Campus, shaojun.wang@wright.edu

Ronald Rosenfeld

Yunxin Zhao

Follow this and additional works at: https://corescholar.libraries.wright.edu/knoesis

Part of the Bioinformatics Commons, Communication Technology and New Media Commons, Databases and Information Systems Commons, OS and Networks Commons, and the Science and Technology Studies Commons

Repository Citation
https://corescholar.libraries.wright.edu/knoesis/281

This Conference Proceeding is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact corescholar@www.libraries.wright.edu, library-corescholar@wright.edu.
LATENT MAXIMUM ENTROPY PRINCIPLE FOR
STATISTICAL LANGUAGE MODELING

Shaojun Wang¹ Ronald Rosenfeld² Yunxin Zhao²
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213¹
Dept. of CECS, University of Missouri, Columbia, MO 65211²
swang,roni@cs.cmu.edu zhao@cecs.missouri.edu

ABSTRACT
In this paper, we describe a unified probabilistic framework for statistical
language modeling, latent maximum entropy principle. The salient feature
of this approach is that the hidden causal hierarchical dependency structure can be
encoded into the statistical model in a principled way by mixtures of exponential families
with a rich expressive power. We first show the problem formulation, solution, and cer-
tain convergence properties. We then describe how to use this machine learning technique
to model various aspects of natural language such as syntactic structure of sentence,
semantic information in document. Finally, we draw a conclusion and point out future research
directions.

1. INTRODUCTION
Markov chain (n-gram) source models for natural lan-
guage were first explored by Shannon in his monumental
paper [19] which led to the birth of information theory. N-
gram language models have been widely used in current
speech recognition systems to help resolve acoustic ambi-
guities by placing higher probabilities on more likely word
strings. While Markov chains are efficient at encoding lo-
cal word interactions, it has been long argued that natu-nal language has a deep structure (see for example [2, 17]),
and Markov chain is a completely inadequate model. How-
ever, very few approaches managed to propose a model that
can effectively exploit relevant syntactic structure [5] and
semantic information [1] of natural language and to out-
perform simple n-gram in perplexity. The difficulty lies
in the lack of a unified probabilistic framework to model
language, which can simultaneously take into account
the lexical information inherent in Markov chain models,
the hierarchical syntactic tree structure in stochastic branch-
ning processes [5, 14], the semantic content in bag-of-words
categorical mixture log-linear models [1, 9], and so on.
The most commonly used technique for combining vari-
ous statistical models is linear interpolation [5, 16]. Lin-
ear interpolation is simple, and easy to implement, and its
result is never worse than any of its components, but a
linear interpolated model makes suboptimal use of its com-
ponents and is generally inconsistent with its components,
and as the result, performance improvement is very limited.
Another approach is based on Jaynes’ maximum entropy
principle [10]. Compared with other approaches in statisti-
cal modeling, there are several advantages to this approach,
including no data fragmentation as in decision tree, no in-
dependence assumption as in naive Bayes, and automatic fea-
ture weights determination. The major weakness of current
maximum entropy approach is that it can only deal with
explicit features. In natural language, there are hidden hi-
erarchical structures which we do not observe directly, such
as semantic information [1] or syntactic structure [5]. Is
it possible to incorporate the hidden hierarchical structure
information which we believe into maximum entropy prin-
ciple framework? A previously proposed direct approach
[12, 14, 18] is to use the component models’ output informa-
tion and formulate it as certain constraints. This approach
achieved some improvement in perplexity and word error
rate reductions.

Motivated by the need of establishing a unified prob-
abilistic framework for natural language modeling, we have
recently proposed a latent maximum entropy (LME) prin-
ciple. The LME principle is beyond Jaynes’ original max-
imum entropy (ME) principle as it can handle latent vari-
ables. In the next section, we first present the latent max-
imum entropy principle, its problem formulation, solution,
and certain convergence properties. We then show how
to use this new principle for statistical language modeling
by mixtures of exponential families with a rich expressive
power.

2. LATENT MAXIMUM ENTROPY
PRINCIPLE
Let X ∈ X, say natural language, be the complete-data
with density p(X) and Y ∈ Y, say words, sentences, docu-
ments, etc be the observed incomplete data, and Y = Y(X)
is a many-to-one mapping from X to Y. The missing
data can be semantic content at document level, syntactic
structure at sentence level etc., see Fig. 1 for illustration.
Let p(Y) denote the density of Y and p(X|Y) the condi-
tional density of X given Y. Then p(Y) = ΣX|Yp(X|Y)
where X|Y = {X : X ∈ X, Y(X) = Y}, and p(X) =
p(Y)p(X|Y).

Figure 1. Natural language, observed incomplete data are
words, sentences, documents, missing data are syntactic
structure at sentence level, semantic content at document
level, where dark nodes denote missing information.

The problem of maximum entropy principle with latent
variables is to select a model ρ, from a set of allowed prob-
ability distributions to maximize the entropy

$$H(p) = - \sum_X p(X) \log p(X)$$

subject to

$$\sum_X p(X)f_i(X) = \sum_{Y=i} \bar{p}(y) \sum_{X \in \mathcal{X}(Y)} p(X|Y = y)f_i(X),$$

\[i = 1, \ldots, N \tag{2}\]

where \( \bar{p}(y) \) is the empirical distribution of a set of ob-
servable training samples \( y_1, \ldots, y_N \), and is thus given by

$$\bar{p}(y) = \frac{C(y)}{C}, \quad C(y) = \sum_{i=1}^N \delta(y, y_i)$$

is the occurrence count of \( y \) among the training samples, \( f_i(X) \), \( i = 1, \ldots, N \) are a set of features that correspond to weak learners in boost-
ing and to sufficient statistics in exponential models, and

$$p(X|Y = y)$$

encode the hierarchical dependency structure into the statis-
tical model. Note that some features
are functions of observable data \( Y \), say \( f_i(X) = f_i(Y) \). In such a case, the constraint is reduced to the common one, \( \sum_{Y \in Y} p(Y) f_i(Y) = \sum_{Y \in Y} p(Y = y) f_i(Y = y) \). There are no constraints on latent variables, and the maxent solution will assign equal probability on latent variables. If there is no missing data, then the problem is reduced to Jaynes’ model. Thus (2) is a more general description than MSE.

Note that due to the nonlinear mapping by \( p(X|Y) \), eq. (2) forms nonlinear constraints on \( p(X) \) and the feasible set is no longer convex. Even though the objective function (1) is concave, no unique optimal solution can be expected. In fact, minima and saddle points may exist.

In order to solve this problem, we consider log-linear models with incomplete data, since without missing data and higher-order terms, that is \( \sum_{X \in X(Y)} p(A) \) and the loglikelihood function of the observed data is

\[
L(\lambda) = \log \prod_{y \in Y} p_A(y)^{p_A(y)} = \sum_{y \in Y} p(y) \log p_A(y) \tag{3}
\]

Now we resort to the EM algorithm [8] to solve the maximization problem of eq. (3). Decompose \( L(\lambda) \) into two parts, that is

\[
L(\lambda) = \sum_{y \in Y} \bar{p}(y) \log p_A(y) = Q(\lambda, \lambda') - K(\lambda, \lambda') \tag{4}
\]

where \( Q(\lambda, \lambda') = \sum_{y \in Y} \bar{p}(y) \sum_{X \in X(Y)} p_X(X|Y)p_A(X|Y) \) is the conditional expected complete-data loglikelihood, and \( K(\lambda, \lambda') = \sum_{y \in Y} \bar{p}(y) \sum_{X \in X(Y)} p_X(X|Y) \log p_A(Y) \) is the conditional expected missing-data loglikelihood.

The EM algorithm maximizes \( L(\lambda) \) by iteratively maximizing \( Q(\lambda, \lambda') \) over \( \lambda \). The \( i\)th iteration \( \lambda^{(i)} \rightarrow \lambda^{(i+1)} \) of the EM algorithm is defined by an expectation, E step, which computes \( Q(\lambda, \lambda^{(i)}) \) as a function of \( \lambda \), followed by an maximization, M step, which finds \( \lambda = \lambda^{(i+1)} \) to maximize \( Q(\lambda, \lambda^{(i)}) \). Each iteration of EM increases \( L(\lambda) \), and very generally, if EM converges to \( \lambda^* \), then \( \lambda^* \) is a local maximum of \( L(\lambda) \) [8, 21].

For this particular log-linear model, we have

\[
Q(\lambda, \lambda^{(i)}) = \sum_{y \in Y} \bar{p}(y) \sum_{X \in X(Y)} p_{X^{(i)}}(X|Y = y) \log p_A(X) \tag{5}
\]

Surprisingly, maximizing \( Q(\lambda, \lambda^{(i)}) \) is equivalent to maximizing the dual function of the complete data maximum entropy problem as follows:

\[
\max_{p} H(p) = -\sum_{X} p(X) \log p(X) \tag{6}
\]

s.t. \( \sum_{X} p(X) f_i(X) = \sum_{y \in Y} \bar{p}(y) \sum_{X \in X(Y)} p_{X^{(i)}}(X|Y = y) f_i(X) \quad i = 1, \ldots, N \)

This is because

\[
Q(\lambda, \lambda^{(i)}) = -\log(Z_\lambda) + \sum_{i=1}^{N} \lambda_i (\sum_{y \in Y} \bar{p}(y)) \sum_{X \in X(Y)} p_{X^{(i)}}(X|Y = y) f_i(X)
\]

which is exactly the dual function of (6).

The generalized iterative scaling (GIS) [7] or improved iterative scaling (IIS) [3, 6] algorithms can be used to maximize \( Q(\lambda, \lambda^*) \). Usually only a few GIS (or IIS) steps are needed for the M step.

Thus the proposed EM algorithm for maximum entropy with latent variables (Latent-maxent) is

**Latent-maxent:**

**E step:** compute \( \sum_{y \in Y} p(y) \sum_{X \in X(Y)} p_{X^{(i)}}(X|Y = y) f_i(X) \quad i = 1, \ldots, N \)

**M step:** \( K \) iterations of full parallel update of parameter values \( \lambda_i, i = 1, \ldots, N \) by (GIS) or (IIS) algorithm.

![Figure 2. Latent-maxent, an EM procedure embedding an iterative scaling loop, where \( A^{(\lambda^{(i)}+1)}(Y) = A^{(\lambda^{(i)})}(Y) \) is the auxiliary function in IIS, \( s \) denotes the index of one cycle of full parallel update of \( \lambda_i \), \( i = 1, \ldots, N \) and \( K \) denotes the number of cycles of full parallel updates.](image)
it to statistical language modeling. We first describe various natural language models which aims at a specific linguistic phenomenon. Then we describe how to formulate them into the framework of latent maximum entropy principle.

3.1. Modeling Local Lexical Information

The commonly used n-gram model, or (n-1)th order Markov chain model, is constructed by assuming all histories with the same last n-1 words to belong to the same equivalence class. The maximum likelihood estimate of an n-gram probability given a training corpus is

$$p(w_1 \cdots w_n) = \frac{C(w_1 \cdots w_n)}{C_n}$$

where $C(w_1 \cdots w_n)$ is the occurrence count of the n-word string $w_1 \cdots w_n$, and $C_n$ is the count of total n-word strings in the corpus.

3.2. Modeling Syntactic Structure

There are two approaches to model syntactic structure in natural language. One approach uses the probability distribution of stochastic context-free grammars (SCFG) over strings of words [14]. The other uses a parser to uncover phrasal heads standing in an important relation to the current word [4, 5]. For brevity, we only demonstrate here the first approach.

Following [13], let $G$ be a context-free grammar consisting of a collection of rules $(A \rightarrow \alpha)$, where each $\alpha$ is a string of terminals and nonterminals. For each sentence $S \in L(G)$, the language of $G$, there is a corresponding set of parse trees $t$, each of which has $S = w_1 \cdots w_L$ as leaves. If we observe only $S$, then for an ambiguous grammar, the actual parse tree used to derive $S$ is hidden.

Suppose we have a joint distribution $p(S, T)$, the probability of deriving $S$ using the tree $T$. Then the marginal distribution

$$p(S) = \sum_T p(S, T) = \sum_T \prod_A p(A \rightarrow \alpha)^{C(A \rightarrow \alpha; T, S)}$$

(9)

gives a language model. In the equation above, $p(S, T)$ is the complete data density and $p(S)$ is the incomplete data density. $C(A \rightarrow \alpha; T, S)$ is the number of times that the rule $A \rightarrow \alpha$ appears in the parse tree $T$ for the sentence $S$. The probability parameters $p(A \rightarrow \alpha)$ are normalized so that $\sum_A p(A \rightarrow \alpha) = 1$.

The model is simplified by making the Markovian assumption that the probability with which a nonterminal is rewritten as a string $\alpha$ depends only on the nonterminal, and not on any surrounding context. This assumption leads to an efficient training algorithm. For convenience, we assume that the grammar is in Chomsky normal form. Thus, each rule is either of the form $A \rightarrow BC$ or $A \rightarrow w$. By EM algorithm, the parameters $p(A \rightarrow \alpha)$ can be estimated iteratively and the E step can be accomplished by inside-outside algorithm through dynamic programming in a parse chart.

3.3. Modeling Semantic Information

A document can be viewed as a collection of semantically homogeneous sentences. With a huge amount of documents on hand, the task of latent semantic analysis (LSA) in the context of information retrieval is to discover the compact semantic representations of high-dimensional categorical text data, which is beyond the lexical level of word occurrences, through the mapping of high-dimensional term-frequency (count) vectors in the vector space representation of documents to a lower dimensional representation in a so-called latent semantic space. Semantic relations between words and documents can then be easily defined in terms of their proximity in the semantic space[1, 9].

Following [9], a generative model of word-document co-occurrence by bag-of-words assumption is described as follows: (1) choose a document $d_n$ with probability $p(d_n)$, (2) select a semantic class $z_k$ with probability $p(z_k|d_n)$, (3) pick a word $w_m$ with probability $p(w_m|z_k)$. Since only pair of $(d_n, w_m)$ is being observed, as a result, the joint probability model is a mixture of log-linear model with the expression

$$p(d_n, w_m) = p(d_n) \sum_k p(w_m|z_k)p(z_k|d_n).$$

Typically the number of documents, words in the vocabulary, and latent class variables is on the order of 100,000, 10,000 and hundreds, respectively. Thus latent class variables function as bottleneck variables to constrain word occurrences in documents. Illustrations of latent semantic analysis in terms of a graphical model and dimensionality reduction are depicted in Figs. 3 and 4, respectively.
Semantic content as described by PLSA can be encoded by the constraints
\[
\sum_x p(x)\delta(d,n)\delta(w,m) = \sum_d \hat{p}(d) \sum_s \hat{p}(s|d) \tag{13}
\]
\[
\sum_{w \in V} \hat{p}(w|s) \sum_{x \in X|w(d)} p(z|w,d)\delta(z,v)\delta(d,n)\delta(w,m),
\]
i = 1, \ldots, K, n = 1, \ldots, N, m = 1, \ldots, M

The goal is to find \( p(X) \) which maximizes the entropy subject to the constraints (11-13).

Figure 5. A sample realization of the mixed chain/tree/stable graphical model, where the document has two sentences with lengths 5 and 6, respectively.

By information theoretic arguments, it can be shown that when each constraint is considered separately, the solution of latent maxent will be reduced to the individual models described in subsection (3.1-3.3).

The exponential form of the complete data density function is often called a Gibbsian field. For every Gibbsian field, there is an equivalent Markovian field. The proposed model is a rather complicated mixed chain/tree/stable graphical model (see Fig. 5 for illustration) for the case of the added lexical neighborhoods due to the n-gram, the distribution is no longer context-free, and the calculation of the right hand side of Eqn. (13) has to be performed by a variant of the inside-outside algorithm or Markov chain Monte Carlo simulation. Since the size of the configuration space is large, the feature expectations may need to be calculated by loopy belief propagation [15] or also by efficient Markov chain Monte Carlo methods.

4. CONCLUSION AND RESEARCH DIRECTIONS

We presented a latent maxent principle which is beyond Jaynes' original maximum entropy principle. LME provides a general statistical framework for incorporating arbitrary aspects of natural language into a parametric model. The parameters can be estimated in the sense of maximum likelihood, interactions among various aspects of language can be taken into account automatically and simultaneously, and the general model is reduced to a familiar model when aiming at a specific linguistic phenomenon.

We are currently implementing the latent maxent model using real text training data. Since the number of features is large, model complexity control and automatic feature selection is under investigation.

REFERENCES


