LATENT MAXIMUM ENTROPY PRINCIPLE FOR STATISTICAL LANGUAGE MODELING

Shaojun Wang[1] *Ronald Rosenfeld*[1] *Yunxin Zhao*[2]

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213[1]
Dept. of CECS, University of Missouri, Columbia, MO 65211[2]
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 certain convergence properties. We then discribe 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

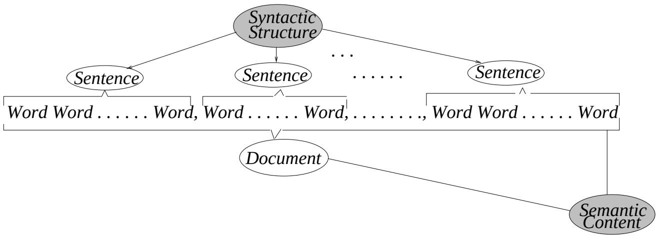
Markov chain (n-gram) source models for natural language were first explored by Shannon in his monumental paper [19] which led to the birth of information theory. N-gram language models have been widely used in current speech recognition systems to help resolve acoustic ambiguities by placing higher probabilities on more likely word strings. While Markov chains are efficient at encoding local word interactions, it has been long argued that natural language has a deep structure (see for example [2, 17]), and Markov chain is a completely inadequate model. However, 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 outperform simple n-gram in perplexity. The difficulty lies in *the lack of a unified probabilistic framework to encode language*, which can simultaneously take into account the lexical information inherent in Markov chain models, the hierarchical syntactic tree structure in stochastic branching 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 various statistical models is linear interpolation [5, 16]. Linear 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 components 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 statistical modeling, there are several advantages to this approach, including no data fragmentation as in decision tree, no independence assumption as in naive Bayes, and automatic feature 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 hierarchical strcutures 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 principle framework? A previously proposed direct approach [12, 14, 18] is to use the component models' output information 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 probabilistic framework for natural language modeling, we have recently proposed a latent maximum entropy (LME) principle. The LME principle is beyond Jaynes' original maximum entropy (ME) principle as it can handle latent variables. In the next section, we first present the latent maximum 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 $X \in \mathcal{X}$, say natural language, be the complete-data with density $p(X)$ and $Y \in \mathcal{Y}$, say words, sentences, documents, etc be the observed incomplete data, and $Y = Y(X)$ is a many-to-one mapping from $\mathcal{X}$ to $\mathcal{Y}$. 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 conditional density of X given Y. Then $p(Y) = \sum_{\mathcal{X}(Y)} p(X)$, where $\mathcal{X}(Y) = \{X : X \in \mathcal{X}, Y(X) = Y\}$, and $p(X) = p(Y)p(X|Y)$.



Syntactic Structure
Sentence
Sentence
. . .
. . . . . .
Sentence
Word Word . . . . . . Word, Word . . . . . . Word, ... , Word Word . . . . . . Word
Document
Semantic Content

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 p_* from a set of allowed probability distributions to maximize the entropy

$$H(p) = -\sum_X p(X) log p(X) \tag{1}$$

subject to

$$\sum_X p(X)f_i(X) = \sum_{y\in\mathcal{Y}} \tilde{p}(y) \sum_{X\in X(Y)} p(X|Y=y)f_i(X), \qquad i = 1, \cdots, N \tag{2}$$

where $\tilde{p}(y)$ is the empirical distribution of a set of observable training samples $y_1, \cdots, y_C$, and is thus given by $\tilde{p}(y) = \frac{C(y)}{C}$, $C(y) = \sum_{i=1}^{C} \delta(y, y_i)$ is the occurrence count of y among the training samples, $f_i(X), i = 1, \cdots, N$ are a set of features that correspond to weak learners in boosting and to sufficient statistics in exponential models, and $p(X|Y = y)$ encodes the hierarchical dependency structure into the statistical model. Note that some features are functions of observable data Y, say $f_j(X) = f_j(Y)$. In such a case, the constraint is reduced to the common one, $\sum_{Y \in \mathcal{Y}} p(Y) f_j(Y) = \sum_{y \in \mathcal{Y}} \tilde{p}(Y = y) f_j(Y = y)$. 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 ME.

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 term of p in eq. (2), the solution for p is a log-linear model. Define $p_\lambda(X) = Z_\lambda^{-1} e^{\sum_i \lambda_i f_i(X)}$. Then $p_\lambda(Y) = \sum_{X \in X(Y)} p_\lambda(X)$ and the loglikelihood function of the observed data is

$$L(\lambda) = log \prod_{y \in \mathcal{Y}} p_\lambda(y)^{\tilde{p}(y)} = \sum_{y \in \mathcal{Y}} \tilde{p}(y) log p_\lambda(y) \quad (3)$$

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

$$L(\lambda) = \sum_{y \in \mathcal{Y}} \tilde{p}(y) log p_\lambda(y) = Q(\lambda, \lambda') - K(\lambda, \lambda') \quad (4)$$

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

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

For this particular log-linear model, we have

$$Q(\lambda, \lambda^{(j)}) = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{X \in X(Y)} p_{\lambda^{(j)}}(X|Y = y) log p_\lambda(X) \quad (5)$$

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

$$\max_p H(p) = -\sum_X p(X) log p(X) \quad (6)$$

$$\text{s.t.} \sum_X p(X) f_i(X) = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{X \in X(Y)} p_{\lambda^{(j)}}(X|Y = y) f_i(X)$$

$$i = 1, \cdots, N$$

This is because

$$\begin{aligned} Q(\lambda, \lambda^{(j)}) &= -log(Z_\lambda) + \sum_{i=1}^{N} \lambda_i (\sum_{y \in \mathcal{Y}} \tilde{p}(y) \\ &\quad \sum_{X \in X(Y)} p_{\lambda^{(j)}}(X|Y = y) f_i(X) \end{aligned}$$

which is exactly the dual function of (6).

The generalized iteratiive 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 $\sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{X \in X(Y)} p_{\lambda^{(j)}}(X|Y = y)$ $f_i(X), i = 1, \cdots, N$;

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

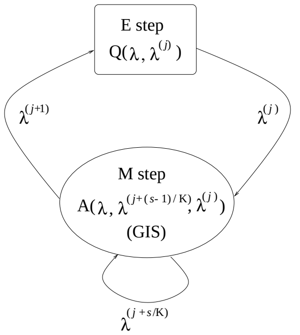

E step
$Q(\lambda, \lambda^{(j)})$
$\lambda^{(j+1)}$
$\lambda^{(j)}$
M step
$A(\lambda, \lambda^{(j+(s-1)/K)}, \lambda^{(j)})$
(GIS)
$\lambda^{(j+s/K)}$

Figure 2. *Latent-maxent, an EM procedure embedding an iterative scaling loop, where $A(\lambda^{(j+s/K)}, \lambda^{(j+(s-1)/K)}, \lambda^{(j)})$ is the auxiliary function in* IIS, *s denotes the index of one cycle of full parallel update of $\lambda_i, i = 1, \cdots, N$ and K denotes the number of cycles of full parallel updates.*

A natural interpretation of this iterative procedure is the following: If the right hand side of eq. (2) is constant, then the optimal soultion of $p_\lambda(X)$ is a log-linear model with parameters provided by GIS/IIS algorithms. Once $p_\lambda(X)$ is obtained, we could calculate the values of the right hand side of eq. (2). If this value matches the constant we assigned before, then by optimality condition, the extremum of the entropy subject to the required constraints is reached; otherwise, the EM procedure is iterated until meeting the constraints.

The convergence proof for the proposed latent maxent algorithm is quite similar to the GEM algorithm [21] and is omitted here. For details, see [20]. Here we simply state the result in the following theorem.

Theorem: The latent-maxent algorithm, an EM nested by iterative scaling, monotonically increases the likelihood function $L(\lambda)$. All limit points of any latent-maxent sequence $\{\lambda^{(j)}, j \geq 0\}$ belong to the set

$$\Gamma = \left\{ \lambda \in \Re^N : \frac{\partial L(\lambda)}{\partial \lambda} = 0 \right\} \quad (7)$$

and in the set Γ, the entropy $H(p_*(\lambda))$ achieves (local) maximum, and $L(\lambda) = Q(\lambda, \lambda) = -H(p_*(\lambda)), \forall \lambda \in \Gamma$.

3. LATENT MAXENT APPROACH FOR STATISTICAL LANGUAGE MODELING

Natural language is a composite, hierarchically organized code to represent messages. Simpler patterns at a lower level are combined in a well-defined manner to form more complex patterns at succeeding higher levels. The function of such a hierarchical structure is to constrain the ways in which the individual patterns at that level can be combined, thus building redundancy into the source code and making it robust to errors made by speakers. As a result, relatively few primitive patterns can be combined in a multilevel hierarchy according to a complex process to form a rich, robust information-bearing code.

The latent maximum entropy principle as discussed above can be used to describe natural language in a principled way by mixtures of exponential families with a rich expressive power. In this section, we discuss how to apply it to statistical language modeling. We first describe various 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

$$p(w_1 \cdots w_n) = \frac{C(w_1 \cdots w_n)}{C_n} \tag{8}$$

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 α is a string of terminals and nonterminals. For each sentence $S \in \mathcal{L}(G)$, the language of G, there is a corresponding set of parse trees t, each of which has $S = w_1 w_2 \cdots w_L$ 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

$$p(S) = \sum_T p(S,T) = \sum_T \prod_{A \rightarrow \alpha} p(A \rightarrow \alpha)^{C(A \rightarrow \alpha; T, S)} \tag{9}$$

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 $\sum_\alpha p(A \rightarrow \alpha) = 1$.

The model is simplified by making the Markovian assumption that the probability with which a nonterminal is rewritten as a string α 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 cooccurrences 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 $p(d_n, w_m) = p(d_n) \sum_{k=1}^{K} p(w_m|z_k) p(z_k|d_n)$. 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 demensionality reduction are depicted in Figs. 3 and 4, respectively.

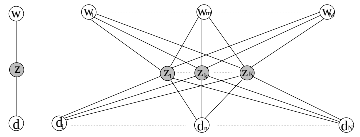


Figure 3. *Graphical representation of dependency of words, documents, and semantic content, where semantic nodes form a bottleneck, and dark nodes are not observable*

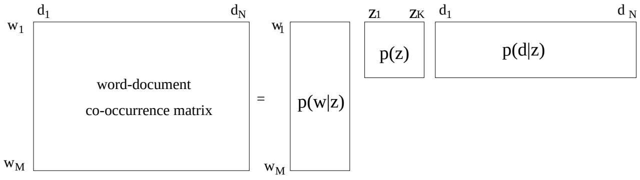


Figure 4. *Dimensionality reduction by probabilistic latent semantic analysis*

By assuming that the joint probability of (d_n, w_m) is a multinomial distribution, the likelihood function can be written as

$$L = \sum_{n=1}^{N} \sum_{m=1}^{M} C(d_n, w_m) log[p(d_n) \sum_{k=1}^{K} p(w_m|z_k) p(z_k|d_n)] \tag{10}$$

where $C(d_n, w_m)$ is the count of word w_m in document d_n. EM algorithm can be performed to estimate the parameters.

3.4. Modeling Various Aspects by LME

The various aspects of linguistic phenomena described above can be encoded into a unified probabilistic model by the latent maximum entropy principle. Define the complete data as $X = (S_1, T_1, \cdots, S_d, T_d, D, Z)$, where S_i is a sentence, $S_i = (W_{i_1} W_{i_2} \cdots W_{i_l})$, $W_{i_t} \in V$, T_i is a parse tree for S_i, Z is a semantic node, D is a document, and the observed data are $Y = (S_1, \cdots, S_d, D)$.

Explicit features such as Markov chain based n-grams can be modeled directly [16]. For example, for trigram model, we have

$$\begin{aligned} & \sum_X p(X) \delta(w_i w_j w_k) \qquad (11) \\ = & \sum_d \tilde{p}(d) \sum_s \tilde{p}(s|d) \sum_{w_i w_j w_k} \tilde{p}(w_i w_j w_k|s) \delta(w_i w_j w_k) \end{aligned}$$

Syntactic structure as described by SCFG can be encoded by the constraints

$$\begin{aligned} & \sum_X p(X) \delta(A \rightarrow \alpha) \qquad (12) \\ = & \sum_d \tilde{p}(d) \sum_s \tilde{p}(s|d) \sum_t p(t|s) \delta(A \rightarrow \alpha) \end{aligned}$$

Semantic content as described by PLSA can be encoded by the constraints

$$\sum_{X} p(X)\delta(d,n)\delta(w,m) = \sum_{d\in D}\tilde{p}(d)\sum_{s}\tilde{p}(s|d) \qquad (13)$$
$$\sum_{w\in V}\tilde{p}(w|s)\sum_{z\in X(w,d)} p(z|w,d)\delta(z,i)\delta(d,n)\delta(w,m),$$
$$i=1,\cdots,K, n=1,\cdots,N, m=1,\cdots,M$$

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

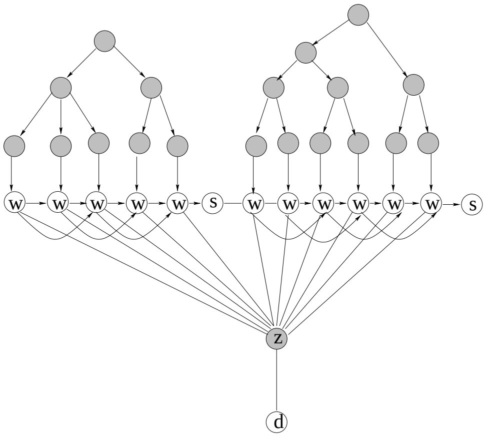


Figure 5. *A sample realization of the mixed chain/tree /table graphical model, where the document has two sentences with lengths 5 and 6, respectively.*

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/table graphical model (see Fig. 5 for illustration). Because 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 expections 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 maximum entropy 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

[1] J. Bellegarda, "Exploiting Latent Semantic Information in Statistical Language Modeling," *Proceedings of the IEEE*, Vol. 88, No. 8, pp. 1279-1296, August 2000

[2] J. Bellegarda, "Robustness in Statistical Language Modeling: Review and Perspectives," *Robustness in Languages and Speech Technology*, J. Junqua and G. van Noods (Editors), Kluwer Academic Publishers, 2001

[3] A. Berger, S. Della Pietra and V. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, Vol. 22, No. 1, pp. 39-71, 1996

[4] E. Charniak, "Immediate-Head Parsing for Language Models," *Proceedings of ACL2001*, pp. 116-123, 2001

[5] C. Chelba and F. Jelinek, "Structured Language Modeling," *Computer Speech and Language*, Vol. 14, No. 4, pp. 283-332, October 2000

[6] S. Della Pietra, V. Della Pietra and J. Lafferty, "Inducing Features of Random Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 4, pp. 380-393, April 1997

[7] J. Darroch and D. Ratchliff, "Generalized Iterative Scaling for Log-Linear Models," *The Annals of Mathematical Statistics*, Vol. 43, No. 5, pp. 1470-1480, 1972

[8] A. Dempster, N. Laird and D. Rubin, "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm," *Journal of Royal Statistical Society, Series B*, Vol. 39, pp 1-38, 1977

[9] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, Vol. 42, No. 1, pp.177-196, 2001

[10] E. Jaynes, *Papers on Probability, Statistics, and Statistical Physics*, edited by R. Rosenkrantz, D. Reidel Publishing Company, 1983

[11] F. Jelinek, J. Lafferty and R. Mercer, "Basic Methods of Probabilistic Context Free Grammars," *Speech Recognition and Understanding*, P. Laface and R. De Mori, eds., Springer, pp. 347-360, 1992

[12] S. Khudanpur and J. Wu, "Maximum Entropy Techniques for Exploiting Syntactic, Semantic and Collocational Dependencies in Language Modeling," *Computer Speech and Language*, Vol. 14, No. 4, pp. 355-372, October 2000

[13] J. Lafferty, Class Notes on Probabilistic Context-Free Grammars, 1999

[14] K. Mark, M. Miller and U. Grenander, "Constrained Stochastic Language Models," *Image Models And Their Speech Model Cousins*, S. Levinson and L. Shepp (Editors), Springer, 1996

[15] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan-Kaufmann, 1988

[16] R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," *Computer Speech and Language*, Vol. 10, pp. 187-228, July 1996

[17] R. Rosenfeld, "Two Decades of Statistical Language Modeling: Where Do We Go From Here?," *Proceedings of the IEEE*, Vol. 88, No. 8, pp. 1270-1278, 2000

[18] R. Rosenfeld, S. Chen and X. Zhu, "Whole Sentence Exponential Language Models: a Vehicle for Linguistic Statistical Integration," *Computer Speech and Language*, Vol. 15, No. 1, pp. 55-73, 2001

[19] C. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, Vol. 27, pp. 398-403, 1948

[20] S. Wang, R. Rosenfeld and Y. Zhao, "Latent Maximum Entropy Principle," submitted to *Neural Information Processing Systems*, Vancouver, 2001

[21] C. Wu, "On the Convergence Properties of the EM Algorithm," *Annals of Statistics*, Vol. 11, pp. 95-103, 1983