PARAMETER ESTIMATION FOR CONSTRAINED CONTEXT-FREE LANGUAGE MODELS

Kevin Mark, Michael Miller, Ulf Grenander[*]*, Steve Abney*[†]

Electronic Systems and Signals Research Laboratory
Washington University
St. Louis, Missouri 63130

ABSTRACT

A new language model incorporating both N-gram and context-free ideas is proposed. This constrained context-free model is specified by a stochastic context-free prior distribution with N-gram frequency constraints. The resulting distribution is a Markov random field. Algorithms for sampling from this distribution and estimating the parameters of the model are presented.

1. INTRODUCTION

This paper introduces the idea of N-gram constrained context-free language models. This class of language models merges two prevalent ideas in language modeling: N-grams and context-free grammars. In N-gram language models, the underlying probability distributions are Markov chains on the word string. N-gram models have advantages in their simplicity. Both parameter estimation and sampling from the distribution are simple tasks. A disadvantage of these models is their weak modeling of linguistic structure.

Context-free language models are instances of random branching processes. The major advantage of this class of models is its ability to capture linguistic structure. In the following section, notation for stochastic context-free language models and the probability of a word string under this model are presented. Section 3 reviews a parameter estimation algorithm for SCF language models.

Section 4 introduces the bigram-constrained context-free language model. This language model is seen to be a Markov random field. In Section 5, a random sampling algorithm is stated. In Section 6, the problem of parameter estimation in the constrained context-free language model is addressed.

2. STOCHASTIC CONTEXT-FREE GRAMMARS

A stochastic context-free grammar G is specified by the quintuple $< V_N, V_T, R, S, P >$ where V_N is a finite set of non-terminal symbols, V_T is a finite set of terminal symbols, R is a set of rewrite rules, S is a start symbol in V_N, and P is a parameter vector. If $r \in R$, then P_r is the probability of using the rewrite rule r.

For our experiments, we are using a 411 rule grammar which we will refer to as the Abney-2 grammar. The grammar has 158 syntactic variables, i.e., $|V_N| = 158$. The rules of the Abney-2 grammar are of the form $H \rightarrow G_1, G_2, \ldots G_k$ where $H, G_i \in V_N$ and $k = 1, 2, \ldots$. Hence, this grammar is not expressed in Chomsky Normal Form. We maintain this more general form for the purposes of linguistic analysis.

An important measure is the probability of a derivation tree T. Using ideas from the random branching process literature [2, 4], we specify a derivation tree T by its depth L and the counting statistics $z_l(i,k), l = 1, \ldots, L, i = 1, \ldots, |V_N|$, and $k = 1, \ldots, |R|$. The counting statistic $z_l(i,k)$ is the number of non-terminals $\sigma_i \in V_N$ rewritten at level l with rule $r_k \in R$. With these statistics the probability of a tree T is given by

$$\pi(T) = \prod_{l=1}^{L} \prod_{i=1}^{|V_N|} \prod_{k=1}^{|R|} P_{r_k}{}^{z_{l-1}(i,k)}. \qquad (1)$$

In this model, the probability of a word string $W_{1,N} = w_1 w_2 \ldots w_N$, $\beta(W_{1,N})$, is given by

$$\beta(W_{1,N}) = \sum_{T \in Parses(W_{1,N})} \pi(T) \qquad (2)$$

where $Parses(W_{1,N})$ is the set of parse trees for the given word string. For an unambiguous grammar, $Parses(W_{1,N})$ consists of a single parse.

[*] Division of Applied Mathematics, Brown University, Providence, Rhode Island 02904

[†] Bell Communications Research, Morristown, New Jersey 07962

3. PARAMETER ESTIMATION FOR SCFGS

An important problem in stochastic language models is the estimation of model parameters. In the parameter estimation problem for SCFGs, we observe a word string $W_{1,N}$ of terminal symbols. With this observation, we want to estimate the rule probabilities P. For a grammar in Chomsky Normal Form, the familiar Inside/Outside Algorithm is used to estimate P. However, the Abney-2 grammar is not in this normal form. Although the grammar could be easily converted to CNF, we prefer to retain its original form for linguistic relevance. Hence, we need an algorithm that can estimate the probabilities of rules in our more general form given above.

The algorithm that we have derived is a specific case of Kupiec's trellis-based algorithm [3]. Kupiec's algorithm estimates parameters for general recursive transition networks. In our case, we only have rules of the following two types:

1. $H \rightarrow G_1 G_2 \cdots G_k$ where $H, G_i \in V_N$ and $k = 1, 2, \ldots$
2. $H \rightarrow T$ where $H \in V_N$ and $T \in V_T$.

For this particular topology, we derived the following trellis-based algorithm.

Trellis-based algorithm

1. Compute inner probabilities $\alpha(i,j,\sigma) = \Pr[\sigma \overset{*}{\Rightarrow} W_{ij}]$ where $\sigma \in V_N$ and W_{ij} denotes the substring $w_i \ldots w_j$.

$$\begin{aligned}\alpha(i,i,\sigma) &= P^{\text{old}}_{\sigma\rightarrow w_i} + \sum_{\sigma_1:\sigma\rightarrow\sigma_1} P^{\text{old}}_{\sigma\rightarrow\sigma_1}\alpha(i,i,\sigma_1) \\ \alpha(i,j,\sigma) &= \sum_{\sigma_n:\sigma\rightarrow\ldots\sigma_n} \alpha_{nte}(i,j,\sigma_n,\sigma)\end{aligned}$$

$$\alpha_{nte}(i,j,\sigma_m,\sigma) = \left\{\begin{array}{l} P^{\text{old}}_{\sigma\rightarrow\sigma_m\ldots}\alpha(i,j,\sigma_m) \\ \qquad \text{if } \sigma \rightarrow \sigma_m \ldots \text{ or } m = 1 \\ \sum_{k=i+1}^{j-1} \alpha_{nte}(i,k,\sigma_{m-1},\sigma)\alpha(k,j,\sigma_m) \\ \qquad \text{if } \sigma \rightarrow \ldots \sigma_{m-1}\sigma_m \cdots \end{array}\right.$$

2. Compute outer probabilities $\beta(i,j,\sigma) = \Pr[S \overset{*}{\Rightarrow} W_{1,i-1}\ \sigma\ W_{j+1,N}]$ where $\sigma \in V_N$.

$$\beta(1,N,S) = 1.0$$

$$\begin{aligned}\beta(i,j,\sigma) = &\sum_{n\rightarrow\sigma\ldots} P^{\text{old}}_{n\rightarrow\sigma\ldots}\beta_{nte}(i,j,\sigma,n) \\ &+ \sum_{n\rightarrow\ldots p\sigma\ldots}\sum_{k=0}^{i-1}\alpha_{nte}(k,i,p,n)\beta_{nte}(k,j,\sigma,n)\end{aligned}$$

$$\beta_{nte}(i,j,\sigma_m,\sigma) = \left\{\begin{array}{l} \beta(i,j,\sigma) \\ \qquad \text{if } \sigma \rightarrow \ldots \sigma_m \\ \sum_{k=j+1}^{L} \alpha(j,k,\sigma_{m+1})\beta_{nte}(i,k,\sigma_{m+1},\sigma) \\ \qquad \text{if } \sigma \rightarrow \ldots \sigma_m\sigma_{m+1} \cdots \end{array}\right.$$

3. Re-estimate P.

$$P^{\text{new}}_{\sigma\rightarrow\sigma_1\sigma_2\ldots\sigma_n} = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\alpha_{nte}(i,j,\sigma_n,\sigma)\beta(i,j,\sigma)}{\sum_{i=1}^{N}\sum_{j=i}^{N}\alpha(i,j,\sigma)\beta(i,j,\sigma)}$$

$$P^{\text{new}}_{\sigma\rightarrow T} = \frac{\sum_{i:w_i=T}\alpha(i,i,\sigma)\beta(i,i,\sigma)}{\sum_{i=1}^{N}\sum_{j=i}^{N}\alpha(i,j,\sigma)\beta(i,j,\sigma)}$$

For CNF grammars, the trellis-based algorithm reduces to the Inside-Outside algorithm. We have tested the algorithm on both CNF grammars and non-CNF grammars. In either case, the estimated probabilities are asymptotically unbiased.

4. SCFGS WITH BIGRAM CONSTRAINTS

We now consider adding bigram relative frequencies as constraints on our stochastic context-free trees. The situation is shown in Figure 1. In this figure, a word string is shown with its bigram relationships and its underlying parse tree structure.

In this model, we assume a given prior context-free distribution as given by $\beta(W_{1,N})$ (Equation 2). This prior distribution may be obtained via the trellis-based estimation algorithm (Section 3) applied to a training text or, alternatively, from a hand-parsed training text. We are also given bigram relative frequencies,

$$h_{\sigma_i,\sigma_j}(W_{1,N}) = \sum_{k=1}^{N-1} 1_{\sigma_i,\sigma_j}(w_k, w_{k+1}) \qquad (3)$$

where $\sigma_i, \sigma_j \in V_T$.

Given this type of structure involving both hierarchical and bigram relationships, what probability distribution on word strings should we consider? The following theorem states the maximum entropy solution.

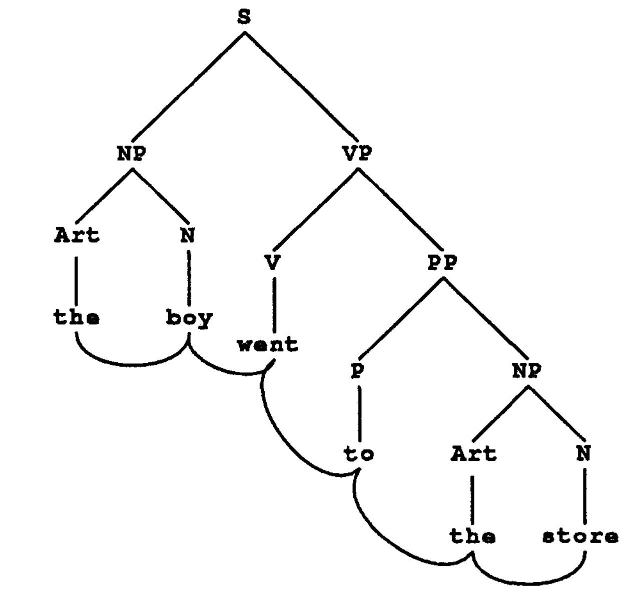

Figure 1: Stochastic context-free tree with bigram relationships.

Theorem 1 Let $c = W_{1,N}$ and $f(c) = \beta(W_{1,N})$. The distribution maximizing the generalized entropy

$$-\sum p(c) \log \frac{p(c)}{f(c)} \tag{4}$$

subject to the constraints $\{E[h_{\sigma_i,\sigma_j}(W_{1,N})] = H_{\sigma_i,\sigma_j}\}_{\sigma_i,\sigma_j \in V_T}$ is

$$\Pr(W_{1,N}) = p^*(c) = \tag{5}$$
$$Z^{-1} \exp\left(\sum_{\sigma_1 \in V_T} \sum_{\sigma_2 \in V_T} \alpha_{\sigma_1,\sigma_2} h_{\sigma_1,\sigma_2}(W_{1,N}) \right) \beta(W_{1,N})$$

where Z is the normalizing constant.

Remarks The specification of bigram constraints for $h(\cdot)$ is not necessary for the derivation of this theorem. The constraint function $h(\cdot)$ may be any function on the word string including general N-grams. Also, note that if the parameters $\alpha_{\sigma_1,\sigma_2}$ are all zero, then this distribution reduces to the unconstrained stochastic context-free model.

5. SIMULATION

For simulation purposes, we would like to be able to draw sample word strings from the maximum entropy distribution. The generation of such sentences for this language model cannot be done directly as in the unconstrained context-free model. In order to generate sentences, a random sampling algorithm is needed. A simple Metropolis-type algorithm is presented to sample from our distribution.

The distribution must first be expressed in Gibbs form:

$$\Pr(W_{1,N}) = \frac{1}{Z} e^{-E(W_{1,N})} \tag{6}$$

where

$$E(W_{1,N}) = -\sum_{\sigma_1 \in V_T} \sum_{\sigma_2 \in V_T} \alpha_{\sigma_1,\sigma_2} h_{\sigma_1,\sigma_2}(W_{1,N}) - \log \beta(W_{1,N}). \tag{7}$$

Given this 'energy' E, the following algorithm generates a sequence of samples, $\{W^1, W^2, W^3, \ldots\}$, from this distribution.

Random sampling algorithm

```
1. perturb $W^i$ to $W^{\text{new}}$
2. compute $\Delta E = E(W^{\text{new}}) - E(W^i)$
3. if $\Delta E \leq 0$ then
     $W^{i+1} \leftarrow W^{\text{new}}$
   else
     $W^{i+1} \leftarrow W^{\text{new}}$
       with probability $= \frac{P(W^{\text{new}})}{P(W)} = e^{-\Delta E}$
4. increment $i$ and repeat step 1.
```

In the first step, the perturbation of a word string is done as follows:

1. generate parses of the string W
2. choose one of these parses
3. choose a node in the parse tree
4. generate a subtree rooted at this node according to the prior rule probabilities
5. let the terminal sequence of the modified tree be the new word string W^{new}.

This method of perturbation satisfies the detailed balance conditions in random sampling.

Proposition Given a sequence of samples $\{W^1, W^2, W^3, \ldots\}$ generated with the random sampling algorithm above. The sequence converges weakly to the distribution $\Pr(W_{1,N})$.

6. PARAMETER ESTIMATION FOR THE CONSTRAINED CONTEXT-FREE MODEL

In the parameter estimation problem for the constrained context-free model, we are given an observed word string $W_{1,N}$ of terminal symbols and want to estimate the α parameters in the maximum entropy distribution, $\Pr(W_{1,N})$. One criterion in estimating these parameters is maximizing the likelihood given the observed data. Maximum likelihood estimation yields the following condition for the optimum (ML) estimates:

$$\left.\frac{\partial \Pr(W_{1,N})}{\partial \alpha_{\sigma_a,\sigma_b}}\right|_{\hat{\alpha}_{\sigma_a,\sigma_b}} = 0 \tag{8}$$

Evaluating the left hand side gives the following maximum likelihood condition

$$E_{\alpha_{\sigma_a,\sigma_b}}[h_{\sigma_a,\sigma_b}(W_{1,N})] = h_{\sigma_a,\sigma_b}(W_{1,N}) \tag{9}$$

One method to obtain the maximum likelihood estimates is given by Younes [5]. His estimation algorithm uses a random sampling algorithm to estimate the expected value of the constraints in a gradient descent framework. Another method is the pseudolikelihood approach which we consider here.

In the pseudolikelihood approach, an approximation to the likelihood is derived from local probabilities [1]. In our problem, these local probabilities are given by:

$$\Pr(w_i|w_1,\ldots,w_{i-1},w_{i+1},\ldots,w_N) = \frac{\exp(\alpha_{w_{i-1},w_i}+\alpha_{w_i,w_{i+1}})\beta(W_{1,N})}{\sum_{w_i'\in V_T}\exp(\alpha_{w_{i-1},w_i'}+\alpha_{w_i',w_{i+1}})\beta_i(W_{1,N},w_i')} \tag{10}$$

where .
$\beta_i(W_{1,N},w_i') = \sum_{T\in Parses(w_1,\ldots,w_{i-1},w_i',w_{i+1},\ldots,w_N)} \pi(T)$.

The pseudolikelihood $\tilde{\mathcal{L}}$ is given in terms of these local probabilities by

$$\tilde{\mathcal{L}} = \prod_{i=1}^{N} \Pr(w_i|w_1,\ldots,w_{i-1},w_{i+1},\ldots,w_N) \tag{11}$$

Maximizing the pseudolikelihood $\tilde{\mathcal{L}}$ is equivalent to maximizing the log-pseudolikelihood,

$$\log\tilde{\mathcal{L}} = N\log\beta(W_{1,N}) + 2\sum_{k=1}^{N-1}\alpha_{w_k,w_{k+1}} - \sum_{i=1}^{N}\log\left[\sum_{w_i'\in V_T} e^{\alpha_{w_{i-1},w_i'}+\alpha_{w_i',w_{i+1}}}\beta_i(W_{1,N},w_i')\right] \tag{12}$$

We can estimate the α parameters by maximizing the log-pseudolikelihood with respect to the α's. The algorithm that we use to do this is a gradient descent algorithm. The gradient descent algorithm is an iterative algorithm in which the parameters are updated by a factor of the gradient, i.e.,

$$\alpha^{(i+1)}_{\sigma_1,\sigma_2} = \alpha^{(i)}_{\sigma_1,\sigma_2} + \mu\frac{\partial\log\tilde{\mathcal{L}}}{\partial\alpha_{\sigma_1,\sigma_2}} \tag{13}$$

where μ is the step size and the gradient is given by

$$\frac{\partial\log\tilde{\mathcal{L}}}{\partial\alpha_{\sigma_1,\sigma_2}} = 2\sum_{k=1}^{N-1}\alpha_{w_k,w_{k+1}}1_{\sigma_1,\sigma_2}(w_k,w_{k+1}) - \sum_{i=1}^{N}\frac{\sum_{w_i'\in V_T}\frac{\partial}{\partial\alpha_{\sigma_1,\sigma_2}}e^{\alpha_{w_{i-1},w_i'}+\alpha_{w_i',w_{i+1}}}\beta_i(W_{1,N},w_i')}{\sum_{w_i'\in V_T}e^{\alpha_{w_{i-1},w_i'}+\alpha_{w_i',w_{i+1}}}\beta_i(W_{1,N},w_i')} \tag{14}$$

The gradient descent algorithm is sensitive to the choice of step size μ. This choice is typically made by trial and error.

7. CONCLUSION

This paper introduces a new class of language models based on Markov random field ideas. The proposed context-free language model with bigram constraints offers a rich linguistic structure. In order to facilitate exploring this structure, we have presented a random sampling algorithm and a parameter estimation algorithm. The work presented here is a beginning. Further work is being done in improving the efficiency of the algorithms and in investigating the correlation of bigram relative frequencies and estimated α parameters in the model.

References

1. Besag, J., "Spatial Interaction and the Statistical Analysis of Lattice Systems," *J. R. Statist. Soc. B*, Vol. 36, 1974, pp. 192-236.
2. Harris, T. E., *The Theory of Branching Processes*, Springer-Verlag, Berlin, 1963.
3. Kupiec, J., "A trellis-based algorithm for estimating the parameters of a hidden stochastic context-free grammar," 1991.
4. Miller, M. I., and O'Sullivan, J. A., "Entropies and Combinatorics of Random Branching Processes and Context-Free Languages," *IEEE Trans. on Information Theory*, March, 1992.
5. Younes, L., "Maximum likelihood estimation for Gibbsian fields," 1991.