Estimation of Consistent Probabilistic Context-free Grammars

Mark-Jan Nederhof
Max Planck Institute
for Psycholinguistics
P.O. Box 310
NL-6500 AH Nijmegen
The Netherlands
MarkJan.Nederhof@mpi.nl

Giorgio Satta
Dept. of Information Engineering
University of Padua
via Gradenigo, 6/A
I-35131 Padova
Italy
satta@dei.unipd.it

Abstract

We consider several empirical estimators for probabilistic context-free grammars, and show that the estimated grammars have the so-called consistency property, under the most general conditions. Our estimators include the widely applied expectation maximization method, used to estimate probabilistic context-free grammars on the basis of unannotated corpora. This solves a problem left open in the literature, since for this method the consistency property has been shown only under restrictive assumptions on the rules of the source grammar.

1 Introduction

Probabilistic context-free grammars are one of the most widely used formalisms in current work in statistical natural language parsing and stochastic language modeling. An important property for a probabilistic context-free grammar is that it be consistent, that is, the grammar should assign probability of one to the set of all finite strings or parse trees that it generates. In other words, the grammar should not lose probability mass with strings or trees of infinite length.

Several methods for the empirical estimation of probabilistic context-free grammars have been proposed in the literature, based on the optimization of some function on the probabilities of the observed data, such as the maximization of the likelihood of a tree bank or a corpus of unannotated sentences. It has been conjectured in (Wetherell, 1980) that these methods always provide probabilistic context-free grammars with the consistency property. A first result in this direction was presented in (Chaudhuri et al., 1983), by showing that a probabilistic context-free grammar estimated by maximizing the likelihood of a sample of parse trees is always consistent.

In later work by (Sánchez and Benedí, 1997) and (Chi and Geman, 1998), the result was independently extended to expectation maximization, which is an unsupervised method exploited to estimate probabilistic context-free grammars by finding local maxima of the likelihood of a sample of unannotated sentences. The proof in (Sánchez and Benedí, 1997) makes use of spectral analysis of expectation matrices, while the proof in (Chi and Geman, 1998) is based on a simpler counting argument. Both these proofs assume restrictions on the underlying context-free grammars. More specifically, in (Chi and Geman, 1998) empty rules and unary rules are not allowed, thus excluding infinite ambiguity, that is, the possibility that some string in the input sample has an infinite number of derivations in the grammar. The treatment of general form context-free grammars has been an open problem so far.

In this paper we consider several estimation methods for probabilistic context-free grammars, and we show that the resulting grammars have the consistency property. Our proofs are applicable under the most general conditions, and our results also include the expectation maximization method, thus solving the open problem discussed above. We use an alternative proof technique with respect to pre-

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 343–350,
New York, June 2006. ©2006 Association for Computational Linguistics

vious work, based on an already known renormalization construction for probabilistic context-free grammars, which has been used in the context of language modeling.

The structure of this paper is as follows. We provide some preliminary definitions in Section 2, followed in Section 3 by a brief overview of the estimation methods we investigate in this paper. In Section 4 we prove some properties of a renormalization technique for probabilistic context-free grammars, and use this property to show our main results in Section 5. Section 6 closes with some concluding remarks.

2 Preliminaries

In this paper we use mostly standard notation, as for instance in (Hopcroft and Ullman, 1979) and (Booth and Thompson, 1973), which we summarize below.

A **context-free grammar** (CFG) is a 4-tuple $G = (N, \Sigma, S, R)$ where N and Σ are finite disjoint sets of nonterminal and terminal symbols, respectively, $S \in N$ is the start symbol and R is a finite set of rules. Each rule has the form $A \to \alpha$, where $A \in N$ and $\alpha \in (\Sigma \cup N)^*$. We write V for set $\Sigma \cup N$.

Each CFG G is associated with a **left-most derive** relation $\Rightarrow$, defined on triples consisting of two strings $\gamma, \delta \in V^*$ and a rule $\pi \in R$. We write $\gamma \overset{\pi}{\Rightarrow} \delta$ if and only if $\gamma = uA\gamma'$ and $\delta = u\alpha\gamma'$, for some $u \in \Sigma^*$, $\gamma' \in V^*$, and $\pi = (A \to \alpha)$. A **left-most derivation** for G is a string $d = \pi_1 \cdots \pi_m$, $m \geq 0$, such that $\gamma_0 \overset{\pi_1}{\Rightarrow} \gamma_1 \overset{\pi_2}{\Rightarrow} \cdots \overset{\pi_m}{\Rightarrow} \gamma_m$, for some $\gamma_0, \ldots, \gamma_m \in V^*$; $d = \varepsilon$ (where ε denotes the empty string) is also a left-most derivation. In the remainder of this paper, we will let the term derivation always refer to left-most derivation. If $\gamma_0 \overset{\pi_1}{\Rightarrow} \cdots \overset{\pi_m}{\Rightarrow} \gamma_m$ for some $\gamma_0, \ldots, \gamma_m \in V^*$, then we say that $d = \pi_1 \cdots \pi_m$ **derives** γ_m from γ_0 and we write $\gamma_0 \overset{d}{\Rightarrow} \gamma_m$; $d = \varepsilon$ derives any $\gamma_0 \in V^*$ from itself.

A (left-most) derivation d such that $S \overset{d}{\Rightarrow} w$, $w \in \Sigma^*$, is called a **complete** derivation. If d is a complete derivation, we write $y(d)$ to denote the (unique) string $w \in \Sigma^*$ such that $S \overset{d}{\Rightarrow} w$. We define $D(G)$ to be the set of all complete derivations for G. The language generated by G is the set of all strings derived by complete derivations, i.e., $L(G) = \{y(d) \,|\, d \in D(G)\}$. It is well-known that there is a one-to-one correspondence between complete derivations and parse trees for strings in $L(G)$.

For $X \in V$ and $\alpha \in V^*$, we write $f(X, \alpha)$ to denote the number of occurrences of X in α. For $(A \to \alpha) \in R$ and a derivation d, $f(A \to \alpha, d)$ denotes the number of occurrences of $A \to \alpha$ in d. We let $f(A, d) = \sum_\alpha f(A \to \alpha, d)$.

A **probabilistic** CFG (PCFG) is a pair $\mathcal{G} = (G, p_G)$, where G is a CFG and p_G is a function from R to real numbers in the interval $[0, 1]$. We say that $\mathcal{G}$ is **proper** if, for every $A \in N$, we have

$$\sum_{A \to \alpha} p_G(A \to \alpha) \quad = \quad 1. \qquad (1)$$

Function p_G can be used to assign probabilities to derivations of the underlying CFG G, in the following way. For $d = \pi_1 \cdots \pi_m \in R^*$, $m \geq 0$, we define

$$p_G(d) \quad = \quad \prod_{i=1}^{m} p_G(\pi_i). \qquad (2)$$

Note that $p_G(\varepsilon) = 1$. The probability of a string $w \in \Sigma^*$ is defined as

$$p_G(w) \quad = \quad \sum_{y(d)=w} p_G(d). \qquad (3)$$

A PCFG is **consistent** if

$$\sum_{w} p_G(w) \quad = \quad 1. \qquad (4)$$

Consistency implies that the PCFG defines a probability distribution over both sets $D(G)$ and $L(G)$. If a PCFG is proper, then consistency means that no probability mass is lost in derivations of infinite length. All PCFGs in this paper are implicitly assumed to be proper, unless otherwise stated.

3 Estimation of PCFGs

In this section we give a brief overview of some estimation methods for PCFGs. These methods will be later investigated to show that they always provide consistent PCFGs.

In natural language processing applications, estimation of a PCFG is usually carried out on the basis of a tree bank, which in this paper we assume to be a **sample**, that is, a finite multiset, of complete derivations. Let $\mathcal{D}$ be such a sample, and let D be the underlying set of derivations. For $d \in D$, we let $f(d, \mathcal{D})$ be the multiplicity of d in $\mathcal{D}$, that is, the number of occurrences of d in $\mathcal{D}$. We define

$$f(A \to \alpha, \mathcal{D}) \; = \; \sum_{d \in D} f(d, \mathcal{D}) \cdot f(A \to \alpha, d), \qquad (5)$$

and let $f(A, \mathcal{D}) = \sum_\alpha f(A \to \alpha, \mathcal{D})$.

Consider a CFG $G = (N, \Sigma, R, S)$ defined by all and only the nonterminals, terminals and rules observed in D. The criterion of maximum likelihood estimation (MLE) prescribes the construction of a PCFG $\mathcal{G} = (G, p_G)$ such that p_G maximizes the likelihood of $\mathcal{D}$, defined as

$$p_G(\mathcal{D}) \; = \; \prod_{d \in D} p_G(d)^{f(d, \mathcal{D})}, \qquad (6)$$

subject to the properness conditions $\sum_\alpha p_G(A \to \alpha) = 1$ for each $A \in N$. The maximization problem above has a unique solution, provided by the estimator (see for instance (Chi and Geman, 1998))

$$p_G(A \to \alpha) \; = \; \frac{f(A \to \alpha, \mathcal{D})}{f(A, \mathcal{D})}. \qquad (7)$$

We refer to this as the supervised MLE method.

In applications in which a tree bank is not available, one might still use the MLE criterion to train a PCFG in an unsupervised way, on the basis of a sample of unannotated sentences, also called a corpus. Let us call $\mathcal{C}$ such a sample and C the underlying set of sentences. For $w \in C$, we let $f(w, \mathcal{C})$ be the multiplicity of w in $\mathcal{C}$.

Assume a CFG $G = (N, \Sigma, R, S)$ that is able to generate all of the sentences in C, and possibly more. The MLE criterion prescribes the construction of a PCFG $\mathcal{G} = (G, p_G)$ such that p_G maximizes the likelihood of $\mathcal{C}$, defined as

$$p_G(\mathcal{C}) \; = \; \prod_{w \in C} p_G(w)^{f(w, \mathcal{C})}, \qquad (8)$$

subject to the properness conditions as in the supervised case above. The above maximization problem provides a system of $|R|$ nonlinear equations (see (Chi and Geman, 1998))

$$p_G(A \to \alpha) \; = \; \frac{\sum_{w \in C} \; f(w, \mathcal{C}) \cdot E_{p_G(d \,|\, w)} \; f(A \to \alpha, d)}{\sum_{w \in C} \; f(w, \mathcal{C}) \cdot E_{p_G(d \,|\, w)} \; f(A, d)}, \qquad (9)$$

where E_p denotes an expectation computed under distribution p, and $p_G(d \,|\, w)$ is the probability of derivation d conditioned by sentence w (so that $p_G(d \,|\, w) > 0$ only if $y(d) = w$). The solution to the above system is not unique, because of the nonlinearity. Furthermore, each solution of (9) identifies a point where the curve in (8) has partial derivatives of zero, but this does not necessarily correspond to a local maximum, let alone an absolute maximum. (A point with partial derivatives of zero that is not a local maximum could be a local minimum or even a so-called saddle point.) In practice, this system is typically solved by means of an iterative algorithm called inside/outside (Charniak, 1993), which implements the expectation maximization (EM) method (Dempster et al., 1977). Starting with an initial function p_G that probabilistically extends G, a so-called growth transformation is computed, defined as

$$\overline{p}_G(A \to \alpha) \; = \; \frac{\sum_{w \in C} f(w, \mathcal{C}) \cdot \sum_{y(d)=w} \frac{p_G(d)}{p_G(w)} \cdot f(A \to \alpha, d)}{\sum_{w \in C} f(w, \mathcal{C}) \cdot \sum_{y(d)=w} \frac{p_G(d)}{p_G(w)} \cdot f(A, d)}. \qquad (10)$$

Following (Baum, 1972), one can show that $\overline{p}_G(\mathcal{C}) \geq p_G(\mathcal{C})$. Thus, by iterating the growth transformation above, we are guaranteed to reach a local maximum for (8), or possibly a saddle point. We refer to this as the unsupervised MLE method.

We now discuss a third estimation method for PCFGs, which was proposed in (Corazza and Satta, 2006). This method can be viewed as a generalization of the supervised MLE method to probability distributions defined over infinite sets of complete derivations. Let D be an infinite set of complete derivations using nonterminal symbols in N, start symbol $S \in N$ and terminal symbols in Σ. We assume that the set of rules that are observed in D is drawn from some finite set R. Let p_D be a probability distribution defined over D, that is, a function from set D to interval $[0, 1]$ such that $\sum_{d \in D} p_D(d) = 1$.

Consider the CFG $G = (N, \Sigma, R, S)$. Note that $D \subseteq D(G)$. We wish to extend G to some PCFG $\mathcal{G} = (G, p_G)$ in such a way that p_D is approximated by p_G (viewed as a distribution over complete derivations) as well as possible according to some criterion. One possible criterion is minimization of the **cross-entropy** between p_D and p_G, defined as the expectation, under distribution p_D, of the information of the derivations in D computed under distribution p_G, that is

$$H(p_D \,||\, p_G) \;=\; E_{p_D} \; \log \frac{1}{p_G(d)} \;=\; - \sum_{d \in D} \; p_D(d) \cdot \log p_G(d). \qquad (11)$$

We thus want to assign to the parameters $p_G(A \rightarrow \alpha)$, $A \rightarrow \alpha \in R$, the values that minimize (11), subject to the conditions $\sum_\alpha p_G(A \rightarrow \alpha) = 1$ for each $A \in N$. Note that minimization of the cross-entropy above is equivalent to minimization of the Kullback-Leibler distance between p_D and p_G. Also note that the likelihood of an infinite set of derivations would always be zero and therefore cannot be considered here.

The solution to the above minimization problem provides the estimator

$$p_G(A \rightarrow \alpha) \;=\; \frac{E_{p_D} \; f(A \rightarrow \alpha, d)}{E_{p_D} \; f(A, d)}. \qquad (12)$$

A proof of this result appears in (Corazza and Satta, 2006), and is briefly summarized in Appendix A, in order to make this paper self-contained. We call the above estimator the cross-entropy minimization method.

The cross-entropy minimization method can be viewed as a generalization of the supervised MLE method in (7), as shown in what follows. Let $\mathcal{D}$ and D be defined as for the supervised MLE method. We define a distribution over D as

$$p_{\mathcal{D}}(d) \;=\; \frac{f(d, \mathcal{D})}{|\mathcal{D}|}. \qquad (13)$$

Distribution $p_{\mathcal{D}}$ is usually called the **empirical distribution** associated with $\mathcal{D}$. Applying the estimator in (12) to $p_{\mathcal{D}}$, we obtain

$$\begin{aligned} p_G(A \rightarrow \alpha) \;=& \\ =& \; \frac{\sum_{d \in D} \; p_{\mathcal{D}}(d) \cdot f(A \rightarrow \alpha, d)}{\sum_{d \in D} \; p_{\mathcal{D}}(d) \cdot f(A, d)} \\ =& \; \frac{\sum_{d \in D} \; \frac{f(d,\mathcal{D})}{|\mathcal{D}|} \cdot f(A \rightarrow \alpha, d)}{\sum_{d \in D} \; \frac{f(d,\mathcal{D})}{|\mathcal{D}|} \cdot f(A, d)} \\ =& \; \frac{\sum_{d \in D} \; f(d, \mathcal{D}) \cdot f(A \rightarrow \alpha, d)}{\sum_{d \in D} \; f(d, \mathcal{D}) \cdot f(A, d)}. \qquad (14) \end{aligned}$$

This is the supervised MLE estimator in (7). This reminds us of the well-known fact that maximizing the likelihood of a (finite) sample through a PCFG distribution amounts to minimizing the cross-entropy between the empirical distribution of the sample and the PCFG distribution itself.

4 Renormalization

In this section we recall a renormalization technique for PCFGs that was used before in (Abney et al., 1999), (Chi, 1999) and (Nederhof and Satta, 2003) for different purposes, and is exploited in the next section to prove our main results. In the remainder of this section, we assume a fixed, not necessarily proper PCFG $\mathcal{G} = (G, p_G)$, with $G = (N, \Sigma, S, R)$.

We define the **renormalization** of $\mathcal{G}$ as the PCFG $\mathcal{R}(\mathcal{G}) = (G, p_{\mathcal{R}})$ with $p_{\mathcal{R}}$ specified by

$$\begin{aligned} p_{\mathcal{R}}(A \rightarrow \alpha) \;=& \\ & p_G(A \rightarrow \alpha) \cdot \frac{\sum_{d,w} \; p_G(\alpha \overset{d}{\Rightarrow} w)}{\sum_{d,w} \; p_G(A \overset{d}{\Rightarrow} w)}. \qquad (15) \end{aligned}$$

It is not difficult to see that $\mathcal{R}(\mathcal{G})$ is a proper PCFG. We now show an important property of $\mathcal{R}(\mathcal{G})$, discussed before in (Nederhof and Satta, 2003) in the context of so-called weighted context-free grammars.

Lemma 1 *For each derivation d with $A \overset{d}{\Rightarrow} w$, $A \in N$ and $w \in \Sigma^*$, we have*

$$p_{\mathcal{R}}(A \overset{d}{\Rightarrow} w) \;=\; \frac{p_G(A \overset{d}{\Rightarrow} w)}{\sum_{d',w'} \; p_G(A \overset{d'}{\Rightarrow} w')}. \qquad (16)$$

Proof. The proof is by induction on the length of d, written $|d|$. If $|d| = 1$ we must have $d = (A \rightarrow w)$, and thus $p_{\mathcal{R}}(d) = p_{\mathcal{R}}(A \rightarrow w)$. In this case, the statement of the lemma directly follows from (15).

Assume now $|d| > 1$ and let $\pi = (A \rightarrow \alpha)$ be the first rule used in d. Note that there must be at least one nonterminal symbol in α. We can then write α as $u_0 A_1 u_1 A_2 \cdots u_{q-1} A_q u_q$, for $q \geq 1$, $A_i \in N$, $1 \leq i \leq q$, and $u_j \in \Sigma^*$, $0 \leq j \leq q$. In words, $A_1, \ldots, A_q$ are all of the occurrences of nonterminals in α, as they appear from left to right. Consequently, we can write d in the form $d = \pi \cdot d_1 \cdots d_q$ for some derivations d_i, $1 \leq i \leq q$, with $A_i \overset{d_i}{\Rightarrow} w_i$, $|d_i| \geq 1$ and with

$w = u_0 w_1 u_1 w_2 \cdots u_{q-1} w_q u_q$. Below we use the fact that $p_{\mathcal{R}}(u_j \overset{\varepsilon}{\Rightarrow} u_j) = p_G(u_j \overset{\varepsilon}{\Rightarrow} u_j) = 1$ for each j with $0 \leq j \leq q$, and further using the definition of $p_{\mathcal{R}}$ and the inductive hypothesis, we can write

$$\begin{aligned}
& p_{\mathcal{R}}(A \overset{d}{\Rightarrow} w) \; = \\
& \quad = \; p_{\mathcal{R}}(A \to \alpha) \cdot \prod_{i=1}^{q} \, p_{\mathcal{R}}(A_i \overset{d_i}{\Rightarrow} w_i) \\
& \quad = \; p_G(A \to \alpha) \cdot \frac{\sum_{d',w'} \, p_G(\alpha \overset{d'}{\Rightarrow} w')}{\sum_{d',w'} \, p_G(A \overset{d'}{\Rightarrow} w')} \cdot \\
& \qquad\qquad \cdot \prod_{i=1}^{q} \, p_{\mathcal{R}}(A_i \overset{d_i}{\Rightarrow} w_i) \\
& \quad = \; p_G(A \to \alpha) \cdot \frac{\sum_{d',w'} \, p_G(\alpha \overset{d'}{\Rightarrow} w')}{\sum_{d',w'} \, p_G(A \overset{d'}{\Rightarrow} w')} \cdot \\
& \qquad\qquad \cdot \prod_{i=1}^{q} \frac{p_G(A_i \overset{d_i}{\Rightarrow} w_i)}{\sum_{d',w'} \, p_G(A_i \overset{d'}{\Rightarrow} w')} \\
& \quad = \; p_G(A \to \alpha) \cdot \frac{\sum_{d',w'} \, p_G(\alpha \overset{d'}{\Rightarrow} w')}{\sum_{d',w'} \, p_G(A \overset{d'}{\Rightarrow} w')} \cdot \\
& \qquad\qquad \cdot \frac{\prod_{i=1}^{q} \, p_G(A_i \overset{d_i}{\Rightarrow} w_i)}{\prod_{i=1}^{q} \sum_{d',w'} \, p_G(A_i \overset{d'}{\Rightarrow} w')} \\
& \quad = \; p_G(A \to \alpha) \cdot \frac{\sum_{d',w'} \, p_G(\alpha \overset{d'}{\Rightarrow} w')}{\sum_{d',w'} \, p_G(A \overset{d'}{\Rightarrow} w')} \cdot \\
& \qquad\qquad \cdot \frac{\prod_{i=1}^{q} \, p_G(A_i \overset{d_i}{\Rightarrow} w_i)}{\sum_{d',w'} \, p_G(\alpha \overset{d'}{\Rightarrow} w')} \\
& \quad = \; p_G(A \to \alpha) \cdot \frac{\prod_{i=1}^{q} \, p_G(A_i \overset{d_i}{\Rightarrow} w_i)}{\sum_{d',w'} \, p_G(A \overset{d'}{\Rightarrow} w')} \cdot \\
& \quad = \; \frac{p_G(A \overset{d}{\Rightarrow} w)}{\sum_{d',w'} \, p_G(A \overset{d'}{\Rightarrow} w')}. \qquad (17)
\end{aligned}$$

■

As an easy corollary of Lemma 1, we have that $\mathcal{R}(\mathcal{G})$ is a consistent PCFG, as we can write

$$\begin{aligned}
& \sum_{d,w} p_{\mathcal{R}}(S \overset{d}{\Rightarrow} w) \; = \\
& \quad = \; \sum_{d,w} \frac{p_G(S \overset{d}{\Rightarrow} w)}{\sum_{d',w'} \, p_G(S \overset{d'}{\Rightarrow} w')} \\
& \quad = \; \frac{\sum_{d,w} \, p_G(S \overset{d}{\Rightarrow} w)}{\sum_{d',w'} \, p_G(S \overset{d'}{\Rightarrow} w')} \; = \; 1. \qquad (18)
\end{aligned}$$

5 Consistency

In this section we prove the main results of this paper, namely that all of the estimation methods discussed in Section 3 always provide consistent PCFGs. We start with a technical lemma, central to our results, showing that a PCFG that minimizes the cross-entropy with a distribution over any set of derivations must be consistent.

Lemma 2 *Let $\mathcal{G} = (G, p_G)$ be a proper PCFG and let p_D be a probability distribution defined over some set $D \subseteq D(G)$. If $\mathcal{G}$ minimizes function $H(p_D \,||\, p_G)$, then $\mathcal{G}$ is consistent.*

Proof. Let $G = (N, \Sigma, S, R)$, and assume that $\mathcal{G}$ is not consistent. We establish a contradiction. Since $\mathcal{G}$ is not consistent, we must have $\sum_{d,w} p_G(S \overset{d}{\Rightarrow} w) < 1$. Let then $\mathcal{R}(\mathcal{G}) = (G, p_{\mathcal{R}})$ be the renormalization of $\mathcal{G}$, defined as in (15). For any derivation $S \overset{d}{\Rightarrow} w$, $w \in \Sigma^*$, with d in D, we can use Lemma 1 and write

$$\begin{aligned}
& p_{\mathcal{R}}(S \overset{d}{\Rightarrow} w) \; = \\
& \quad = \; \frac{1}{\sum_{d',w'} \, p_G(S \overset{d'}{\Rightarrow} w')} \cdot p_G(S \overset{d}{\Rightarrow} w) \\
& \quad > \; p_G(S \overset{d}{\Rightarrow} w). \qquad (19)
\end{aligned}$$

In words, every complete derivation d in D has a probability in $\mathcal{R}(\mathcal{G})$ that is strictly greater than in $\mathcal{G}$. But this means $H(p_D \,||\, p_{\mathcal{R}}) < H(p_D \,||\, p_G)$, against our hypothesis. Therefore, $\mathcal{G}$ is consistent and p_G is a probability distribution over set $D(G)$. Thus function $H(p_D \,||\, p_G)$ can be interpreted as the cross-entropy. (Observe that in the statement of the lemma we have avoided the term ‘cross-entropy’, since cross-entropies are only defined for probability distributions.) ■

Lemma 2 directly implies that the cross-entropy minimization method in (12) always provides a consistent PCFG, since it minimizes cross-entropy for a distribution defined over a subset of $D(G)$. We have already seen in Section 3 that the supervised MLE method is a special case of the cross-entropy minimization method. Thus we can also conclude that a PCFG trained with the supervised MLE method is always consistent. This provides an alternative proof of a property that was first shown in (Chaudhuri et al., 1983), as discussed in Section 1.

We now prove the same result for the unsupervised MLE method, without any restrictive assumption on the rules of our CFGs. This solves a problem that was left open in the literature (Chi and Geman, 1998); see again Section 1 for discussion. Let $\mathcal{C}$ and C be defined as in Section 3. We define the empirical distribution of $\mathcal{C}$ as

$$p_C(w) \;=\; \frac{f(w,\mathcal{C})}{|\mathcal{C}|}. \qquad (20)$$

Let $G = (N, \Sigma, S, R)$ be a CFG such that $C \subseteq L(G)$. Let $D(C)$ be the set of all complete derivations for G that generate sentences in C, that is, $D(C) = \{d \,|\, d \in D(G),\ y(d) \in C\}$.

Further, assume some probabilistic extension $\mathcal{G} = (G, p_G)$ of G, such that $p_G(d) > 0$ for every $d \in D(C)$. We define a distribution over $D(C)$ by

$$p_{D(C)}(d) \;=\; p_C(y(d)) \cdot \frac{p_G(d)}{p_G(y(d))}. \qquad (21)$$

It is not difficult to verify that

$$\sum_{d \in D(C)} p_{D(C)}(d) \;=\; 1. \qquad (22)$$

We now apply to $\mathcal{G}$ the estimator in (12), in order to obtain a new PCFG $\hat{\mathcal{G}} = (G, \hat{p}_G)$ that minimizes the cross-entropy between $p_{D(C)}$ and $\hat{p}_G$. According to Lemma 2, we have that $\hat{\mathcal{G}}$ is a consistent PCFG. Distribution $\hat{p}_G$ is specified by

$$\begin{aligned}
&\hat{p}_G(A \to \alpha) \;= \\
&= \frac{\sum_{d \in D(C)} p_{D(C)}(d) \cdot f(A \to \alpha, d)}{\sum_{d \in D(C)} p_{D(C)}(d) \cdot f(A, d)} \\
&= \frac{\sum_{d \in D(C)} \frac{f(y(d),\mathcal{C})}{|\mathcal{C}|} \cdot \frac{p_G(d)}{p_G(y(d))} \cdot f(A \to \alpha, d)}{\sum_{d \in D(C)} \frac{f(y(d),\mathcal{C})}{|\mathcal{C}|} \cdot \frac{p_G(d)}{p_G(y(d))} \cdot f(A, d)} \\
&= \frac{\sum_{w \in C} f(w,\mathcal{C}) \cdot \sum_{y(d)=w} \frac{p_G(d)}{p_G(w)} \cdot f(A \to \alpha, d)}{\sum_{w \in C} f(w,\mathcal{C}) \cdot \sum_{y(d)=w} \frac{p_G(d)}{p_G(w)} \cdot f(A, d)} \\
&= \frac{\sum_{w \in C} f(w,\mathcal{C}) \cdot E_{p_G(d \,|\, w)} f(A \to \alpha, d)}{\sum_{w \in C} f(w,\mathcal{C}) \cdot E_{p_G(d \,|\, w)} f(A, d)}. \qquad (23)
\end{aligned}$$

Since distribution p_G was arbitrarily chosen, subject to the only restriction that $p_G(d) > 0$ for every $d \in D(C)$, we have that (23) is the growth estimator (10) already discussed in Section 3. In fact, for each $w \in L(G)$ and $d \in D(G)$, we have $p_G(d \,|\, w) = \frac{p_G(d)}{p_G(w)}$. We conclude with the desired result, namely that a general form PCFG obtained at any iteration of the EM method for the unsupervised MLE is always consistent.

6 Conclusions

In this paper we have investigated a number of methods for the empirical estimation of probabilistic context-free grammars, and have shown that the resulting grammars have the so-called consistency property. This property guarantees that all the probability mass of the grammar is used for the finite strings it derives. Thus if the grammar is used in combination with other probabilistic models, as for instance in a speech processing system, consistency allows us to combine or compare scores from different modules in a sound way.

To obtain our results, we have used a novel proof technique that exploits an already known construction for the renormalization of probabilistic context-free grammars. Our proof technique seems more intuitive than arguments previously used in the literature to prove the consistency property, based on counting arguments or on spectral analysis. It is not difficult to see that our proof technique can also be used with probabilistic rewriting formalisms whose underlying derivations can be characterized by means of context-free rewriting. This is for instance the case with probabilistic tree-adjoining grammars (Schabes, 1992; Sarkar, 1998), for which consistency results have not yet been shown in the literature.

A Cross-entropy minimization

In order to make this paper self-contained, we sketch a proof of the claim in Section 3 that the estimator in (12) minimizes the cross entropy in (11). A full proof appears in (Corazza and Satta, 2006).

Let D, p_D and $G = (N, \Sigma, R, S)$ be defined as in Section 3. We want to find a proper PCFG $\mathcal{G} = (G, p_G)$ such that the cross-entropy $H(p_D \,||\, p_G)$ is minimal. We use Lagrange multipliers λ_A for each $A \in N$ and define the form

$$\nabla \;=\; \sum_{A \in N} \lambda_A \cdot (\sum_{\alpha} p_G(A \to \alpha) - 1) \;+$$

$$- \sum_{d \in D} p_D(d) \cdot \log p_G(d). \tag{24}$$

We now consider all the partial derivatives of ∇. For each $A \in N$ we have

$$\frac{\partial \nabla}{\partial \lambda_A} \; = \; \sum_{\alpha} p_G(A \to \alpha) - 1. \tag{25}$$

For each $(A \to \alpha) \in R$ we have

$$\begin{aligned}
&\frac{\partial \nabla}{\partial p_G(A \to \alpha)} \; = \\
&= \; \lambda_A - \frac{\partial}{\partial p_G(A \to \alpha)} \sum_{d \in D} p_D(d) \cdot \log p_G(d) \\
&= \; \lambda_A - \sum_{d \in D} p_D(d) \cdot \frac{\partial}{\partial p_G(A \to \alpha)} \log p_G(d) \\
&= \; \lambda_A - \sum_{d \in D} p_D(d) \cdot \frac{\partial}{\partial p_G(A \to \alpha)} \\
&\qquad \log \prod_{(B \to \beta) \in R} p_G(B \to \beta)^{f(B \to \beta, d)} \\
&= \; \lambda_A - \sum_{d \in D} p_D(d) \cdot \frac{\partial}{\partial p_G(A \to \alpha)} \\
&\qquad \sum_{(B \to \beta) \in R} f(B \to \beta, d) \cdot \log p_G(B \to \beta) \\
&= \; \lambda_A - \sum_{d \in D} p_D(d) \cdot \sum_{(B \to \beta) \in R} f(B \to \beta, d) \cdot \\
&\qquad \frac{\partial}{\partial p_G(A \to \alpha)} \log p_G(B \to \beta) \\
&= \; \lambda_A - \sum_{d \in D} p_D(d) \cdot f(A \to \alpha, d) \cdot \\
&\qquad \cdot \frac{1}{\ln(2)} \cdot \frac{1}{p_G(A \to \alpha)} \\
&= \; \lambda_A - \frac{1}{\ln(2)} \cdot \frac{1}{p_G(A \to \alpha)} \cdot \\
&\qquad \cdot \sum_{d \in D} p_D(d) \cdot f(A \to \alpha, d) \\
&= \; \lambda_A - \frac{1}{\ln(2)} \cdot \frac{1}{p_G(A \to \alpha)} \cdot \\
&\qquad \cdot E_{p_D} \, f(A \to \alpha, d).
\end{aligned} \tag{26}$$

By setting to zero all of the above partial derivatives, we obtain a system of $|N| + |R|$ equations, which we must solve. From $\frac{\partial \nabla}{\partial p_G(A \to \alpha)} = 0$ we obtain

$$\begin{aligned}
\lambda_A \cdot \ln(2) \cdot p_G(A \to \alpha) \; = \\
E_{p_D} f(A \to \alpha, d).
\end{aligned} \tag{27}$$

We sum over all strings α such that $(A \to \alpha) \in R$, deriving

$$\begin{aligned}
&\lambda_A \cdot \ln(2) \cdot \sum_{\alpha} p_G(A \to \alpha) \; = \\
&= \; \sum_{\alpha} E_{p_D} \, f(A \to \alpha, d) \\
&= \; \sum_{\alpha} \sum_{d \in D} p_D(d) \cdot f(A \to \alpha, d) \\
&= \; \sum_{d \in D} p_D(d) \cdot \sum_{\alpha} f(A \to \alpha, d) \\
&= \; \sum_{d \in D} p_D(d) \cdot f(A, d) \\
&= \; E_{p_D} \, f(A, d).
\end{aligned} \tag{28}$$

From each equation $\frac{\partial \nabla}{\partial \lambda_A} = 0$ we obtain $\sum_{\alpha} p_G(A \to \alpha) = 1$ for each $A \in N$ (our original constraints). Combining this with (28) we obtain

$$\lambda_A \cdot \ln(2) \; = \; E_{p_D} \, f(A, d). \tag{29}$$

Replacing (29) into (27) we obtain, for every rule $(A \to \alpha) \in R$,

$$p_G(A \to \alpha) \; = \; \frac{E_{p_D} \, f(A \to \alpha, d)}{E_{p_D} \, f(A, d)}. \tag{30}$$

This is the estimator introduced in Section 3.

References

S. Abney, D. McAllester, and F. Pereira. 1999. Relating probabilistic grammars and automata. In *37th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 542–549, Maryland, USA, June.

L. E. Baum. 1972. An inequality and associated maximization technique in statistical estimations of probabilistic functions of Markov processes. *Inequalities*, 3:1–8.

T.L. Booth and R.A. Thompson. 1973. Applying probabilistic measures to abstract languages. *IEEE Transactions on Computers*, C-22(5):442–450, May.

E. Charniak. 1993. *Statistical Language Learning*. MIT Press.

R. Chaudhuri, S. Pham, and O. N. Garcia. 1983. Solution of an open problem on probabilistic grammars. *IEEE Transactions on Computers*, 32(8):748–750.

Z. Chi and S. Geman. 1998. Estimation of probabilistic context-free grammars. *Computational Linguistics*, 24(2):299–305.

Z. Chi. 1999. Statistical properties of probabilistic context-free grammars. *Computational Linguistics*, 25(1):131–160.

A. Corazza and G. Satta. 2006. Cross-entropy and estimation of probabilistic context-free grammars. In *Proc. of HLT/NAACL 2006 Conference (this volume)*, New York.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B, 39:1–38.

J.E. Hopcroft and J.D. Ullman. 1979. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley.

M.-J. Nederhof and G. Satta. 2003. Probabilistic parsing as intersection. In *8th International Workshop on Parsing Technologies*, pages 137–148, LORIA, Nancy, France, April.

J.-A. Sánchez and J.-M. Benedí. 1997. Consistency of stochastic context-free grammars from probabilistic estimation based on growth transformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9):1052–1055, September.

A. Sarkar. 1998. Conditions on consistency of probabilistic tree adjoining grammars. In *Proc. of the 36th ACL*, pages 1164–1170, Montreal, Canada.

Y. Schabes. 1992. Stochastic lexicalized tree-adjoining grammars. In *Proc. of the 14th COLING*, pages 426–432, Nantes, France.

C. S. Wetherell. 1980. Probabilistic languages: A review and some open questions. *Computing Surveys*, 12(4):361–379.