Cross-Entropy and Estimation of Probabilistic Context-Free Grammars

Anna Corazza
Department of Physics
University "Federico II"
via Cinthia
I-80126 Napoli, Italy
corazza@na.infn.it

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

Abstract

We investigate the problem of training probabilistic context-free grammars on the basis of a distribution defined over an infinite set of trees, by minimizing the cross-entropy. This problem can be seen as a generalization of the well-known maximum likelihood estimator on (finite) tree banks. We prove an unexpected theoretical property of grammars that are trained in this way, namely, we show that the derivational entropy of the grammar takes the same value as the cross-entropy between the input distribution and the grammar itself. We show that the result also holds for the widely applied maximum likelihood estimator on tree banks.

1 Introduction

Probabilistic context-free grammars are able to describe hierarchical, tree-shaped structures underlying sentences, and are widely used in statistical natural language processing; see for instance (Collins, 2003) and references therein. Probabilistic context-free grammars seem also more suitable than finite-state devices for language modeling, and several language models based on these grammars have been recently proposed in the literature; see for instance (Chelba and Jelinek, 1998), (Charniak, 2001) and (Roark, 2001).

Empirical estimation of probabilistic context-free grammars is usually carried out on tree banks, that is, finite samples of parse trees, through the maximization of the likelihood of the sample itself. It is well-known that this method also minimizes the cross-entropy between the probability distribution induced by the tree bank, also called the empirical distribution, and the tree probability distribution induced by the estimated grammar.

In this paper we generalize the maximum likelihood method, proposing an estimation technique that works on any unrestricted tree distribution defined over an infinite set of trees. This generalization is theoretically appealing, and allows us to prove unexpected properties of the already mentioned maximum likelihood estimator for tree banks, that were not previously known in the literature on statistical natural language parsing. More specifically, we investigate the following information theoretic quantities

- the cross-entropy between the unrestricted tree distribution given as input and the tree distribution induced by the estimated probabilistic context-free grammar; and
- the derivational entropy of the estimated probabilistic context-free grammar.

These two quantities are usually unrelated. We show that these two quantities take the same value when the probabilistic context-free grammar is trained using the minimal cross-entropy criterion. We then translate back this property to the method of maximum likelihood estimation. Our general estimation method also has practical applications in cases one uses a probabilistic context-free grammar to approximate strictly more powerful rewriting systems,

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 335–342,
New York, June 2006. ©2006 Association for Computational Linguistics

as for instance probabilistic tree adjoining grammars (Schabes, 1992).

Not much is found in the literature about the estimation of probabilistic grammars from infinite distributions. This line of research was started in (Nederhof, 2005), investigating the problem of training an input probabilistic finite automaton from an infinite tree distribution specified by means of an input probabilistic context-free grammar. The problem we consider in this paper can then be seen as a generalization of the above problem, where the input model to be trained is a probabilistic context-free grammar and the input distribution is an unrestricted tree distribution. In (Chi, 1999) an estimator that maximizes the likelihood of a probability distribution defined over a finite set of trees is introduced, as a generalization of the maximum likelihood estimator. Again, the problems we consider here can be thought of as generalizations of such estimator to the case of distributions over infinite sets of trees or sentences.

The remainder of this paper is structured as follows. Section 2 introduces the basic notation and definitions and Section 3 discusses our new estimation method. Section 4 presents our main result, which is transferred in Section 5 to the method of maximum likelihood estimation. Section 6 discusses some simple examples, and Section 7 closes with some further discussion.

2 Preliminaries

Throughout this paper we use standard notation and definitions from the literature on formal languages and probabilistic grammars, which we briefly summarize below. We refer the reader to (Hopcroft and Ullman, 1979) and (Booth and Thompson, 1973) for a more precise presentation.

A **context-free grammar** (CFG) is a tuple $G = (N, \Sigma, R, S)$, where N is a finite set of nonterminal symbols, Σ is a finite set of terminal symbols disjoint from N, $S \in N$ 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 denote by $L(G)$ and $T(G)$ the set of all strings, resp., trees, generated by G. For $t \in T(G)$, the yield of t is denoted by $y(t)$.

For a nonterminal A and a string α, we write $f(A, \alpha)$ to denote the number of occurrences of A in α. For a rule $(A \to \alpha) \in R$ and a tree $t \in T(G)$, $f(A \to \alpha, t)$ denotes the number of occurrences of $A \to \alpha$ in t. We let $f(A, t) = \sum_\alpha f(A \to \alpha, t)$.

A **probabilistic** context-free grammar (PCFG) is a pair $\mathcal{G} = (G, p_G)$, with G a CFG and p_G a function from R to the real numbers in the interval $[0, 1]$. A PCFG is **proper** if for every $A \in N$ we have $\sum_\alpha p_G(A \to \alpha) = 1$. The probability of $t \in T(G)$ is the product of the probabilities of all rules in t, counted with their multiplicity, that is,

$$p_G(t) \quad = \quad \prod_{A \to \alpha} p_G(A \to \alpha)^{f(A \to \alpha, t)}. \qquad (1)$$

The probability of $w \in L(G)$ is the sum of the probabilities of all the trees that generate w, that is,

$$p_G(w) \quad = \quad \sum_{y(t) = w} p_G(t). \qquad (2)$$

A PCFG is **consistent** if $\sum_{t \in T(G)} p_G(t) = 1$.

In this paper we write $\log$ for logarithms in base 2 and $\ln$ for logarithms in the natural base e. We also assume $0 \cdot \log 0 = 0$. We write E_p to denote the expectation operator under distribution p. In case $\mathcal{G}$ is proper and consistent, we can define the **derivational entropy** of $\mathcal{G}$ as the expectation of the information of parse trees in $T(G)$, computed under distribution p_G, that is,

$$\begin{aligned} H_d(p_G) \quad &= \quad E_{p_G} \log \frac{1}{p_G(t)} \\ &= \quad - \sum_{t \in T(G)} p_G(t) \cdot \log p_G(t). \qquad (3) \end{aligned}$$

Similarly, for each $A \in N$ we also define the **nonterminal entropy** of A as

$$\begin{aligned} & H_A(p_G) = \\ & \quad = \quad E_{p_G} \log \frac{1}{p_G(A \to \alpha)} \\ & \quad = \quad -\sum_\alpha p_G(A \to \alpha) \cdot \log p_G(A \to \alpha). \qquad (4) \end{aligned}$$

3 Estimation based on cross-entropy

Let T be an infinite set of (finite) trees with internal nodes labeled by symbols in N, root nodes labeled by $S \in N$ and leaf nodes labeled by symbols in Σ. We assume that the set of rules that are observed in the trees in T is drawn from some finite set R. Let p_T be a probability distribution defined over T, that is, a function from T to set $[0, 1]$ such that $\sum_{t \in T} p_T(t) = 1$.

The **skeleton** CFG underlying T is defined as $G = (N, \Sigma, R, S)$. Note that we have $T \subseteq T(G)$ and, in the general case, there might be trees in $T(G)$ that do not appear in T. We wish anyway to approximate distribution p_T the best we can, by turning G into some proper PCFG $\mathcal{G} = (G, p_G)$ and setting parameters $p_G(A \rightarrow \alpha)$ appropriately, for each $(A \rightarrow \alpha) \in R$.

One possible criterion is to choose p_G in such a way that the cross-entropy between p_T and p_G is minimized, where we now view p_G as a probability distribution defined over $T(G)$. The **cross-entropy** between p_T and p_G is defined as the expectation under distribution p_T of the information, computed under distribution p_G, of the trees in $T(G)$

$$\begin{aligned} H(p_T \,||\, p_G) &= E_{p_T} \ \log \frac{1}{p_G(t)} \\ &= -\sum_{t \in T} p_T(t) \cdot \log p_G(t). \quad (5) \end{aligned}$$

Since $\mathcal{G}$ should be proper, the minimization of (5) is subject to the constraints $\sum_{\alpha} \ p_G(A \rightarrow \alpha) = 1$, for each $A \in N$.

To solve the minimization problem above, we use Lagrange multipliers λ_A for each $A \in N$ and define the form

$$\begin{aligned} \nabla &= \sum_{A \in N} \lambda_A \cdot (\sum_{\alpha} \ p_G(A \rightarrow \alpha) - 1) + \\ & \quad - \sum_{t \in T} p_T(t) \cdot \log p_G(t). \quad (6) \end{aligned}$$

We now view ∇ as a function of all the λ_A and the $p_G(A \rightarrow \alpha)$, and consider all the partial derivatives of ∇. For each $A \in N$ we have

$$\frac{\partial \nabla}{\partial \lambda_A} = \sum_{\alpha} \ p_G(A \rightarrow \alpha) - 1.$$

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

$$\begin{aligned} & \frac{\partial \nabla}{\partial p_G(A \rightarrow \alpha)} = \\ & \quad = \lambda_A - \frac{\partial}{\partial p_G(A \rightarrow \alpha)} \sum_{t \in T} p_T(t) \cdot \log p_G(t) \\ & \quad = \lambda_A - \sum_{t \in T} \ p_T(t) \cdot \frac{\partial}{\partial p_G(A \rightarrow \alpha)} \log p_G(t) \\ & \quad = \lambda_A - \sum_{t \in T} \ p_T(t) \cdot \frac{\partial}{\partial p_G(A \rightarrow \alpha)} \\ & \qquad \log \prod_{(B \rightarrow \beta) \in R} \ p_G(B \rightarrow \beta)^{f(B \rightarrow \beta, t)} \\ & \quad = \lambda_A - \sum_{t \in T} \ p_T(t) \cdot \frac{\partial}{\partial p_G(A \rightarrow \alpha)} \\ & \qquad \sum_{(B \rightarrow \beta) \in R} \ f(B \rightarrow \beta, t) \cdot \log p_G(B \rightarrow \beta) \\ & \quad = \lambda_A - \sum_{t \in T} \ p_T(t) \cdot \sum_{(B \rightarrow \beta) \in R} \ f(B \rightarrow \beta, t) \cdot \\ & \qquad \frac{\partial}{\partial p_G(A \rightarrow \alpha)} \log p_G(B \rightarrow \beta) \\ & \quad = \lambda_A - \sum_{t \in T} \ p_T(t) \cdot f(A \rightarrow \alpha, t) \cdot \\ & \qquad \cdot \frac{1}{\ln(2)} \cdot \frac{1}{p_G(A \rightarrow \alpha)} \\ & \quad = \lambda_A - \frac{1}{\ln(2)} \cdot \frac{1}{p_G(A \rightarrow \alpha)} \cdot \\ & \qquad \cdot \sum_{t \in T} \ p_T(t) \cdot f(A \rightarrow \alpha, t) \\ & \quad = \lambda_A - \frac{1}{\ln(2)} \cdot \frac{1}{p_G(A \rightarrow \alpha)} \cdot \\ & \qquad \cdot E_{p_T} \ f(A \rightarrow \alpha, t). \end{aligned}$$

We now need to solve a system of $|N| + |R|$ equations obtained by setting to zero all of the above partial derivatives. From each equation $\frac{\partial \nabla}{\partial p_G(A \rightarrow \alpha)} = 0$ we obtain

$$\begin{aligned} & \lambda_A \cdot \ln(2) \cdot p_G(A \rightarrow \alpha) = \\ & \quad = E_{p_T} \ f(A \rightarrow \alpha, t). \quad (7) \end{aligned}$$

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

$$\begin{aligned} & \lambda_A \cdot \ln(2) \cdot \sum_{\alpha} \ p_G(A \rightarrow \alpha) = \\ & \quad = \sum_{\alpha} \ E_{p_T} \ f(A \rightarrow \alpha, t) \\ & \quad = \sum_{\alpha} \sum_{t \in T} p_T(t) \cdot f(A \rightarrow \alpha, t) \\ & \quad = \sum_{t \in T} p_T(t) \cdot \sum_{\alpha} \ f(A \rightarrow \alpha, t) \\ & \quad = \sum_{t \in T} p_T(t) \cdot f(A, t) \\ & \quad = E_{p_T} \ f(A, t). \quad (8) \end{aligned}$$

From each equation $\frac{\partial \nabla}{\partial \lambda_A} = 0$ we obtain $\sum_\alpha p_G(A \rightarrow \alpha) = 1$ for each $A \in N$ (our original constraints). Combining with (8) we obtain

$$\lambda_A \cdot \ln(2) \;=\; E_{p_T}\; f(A,t). \tag{9}$$

Replacing (9) into (7) we obtain, for every rule $(A \rightarrow \alpha) \in R$,

$$p_G(A \rightarrow \alpha) \;=\; \frac{E_{p_T}\; f(A \rightarrow \alpha, t)}{E_{p_T}\; f(A,t)}. \tag{10}$$

The equations in (10) define the desired estimator for our PCFG, assigning to each rule $A \rightarrow \alpha$ a probability specified as the ratio between the expected number of $A \rightarrow \alpha$ and the expected number of A, under the distribution p_T. We remark here that the minimization of the cross-entropy above is equivalent to the minimization of the Kullback-Leibler distance between p_T and p_G, viewed as tree distributions. Also, note that the likelihood of an infinite set of derivations would always be zero and therefore cannot be considered here.

To be used in the next section, we now show that the PCFG $\mathcal{G}$ obtained as above is consistent. The line of our argument below follows a proof provided in (Chi and Geman, 1998) for the maximum likelihood estimator based on finite tree distributions. Without loss of generality, we assume that in $\mathcal{G}$ the start symbol S is never used in the right-hand side of a rule.

For each $A \in N$, let q_A be the probability that a derivation in $\mathcal{G}$ rooted in A fails to terminate. We can then write

$$q_A \;\leq\; \sum_{B \in N} q_B \cdot \sum_\alpha \; p_G(A \rightarrow \alpha) f(B, \alpha). \tag{11}$$

The inequality follows from the fact that the events considered in the right-hand side of (11) are not mutually exclusive. Combining (10) and (11) we obtain

$$q_A \cdot E_{p_T}\; f(A,t) \leq$$
$$\leq \; \sum_{B \in N} q_B \cdot \sum_\alpha \; E_{p_T} f(A \rightarrow \alpha, t) f(B, \alpha).$$

Summing over all nonterminals we have

$$\sum_{A \in N} \; q_A \cdot E_{p_T}\; f(A,t) \leq$$
$$\leq \; \sum_{B \in N} q_B \cdot \sum_{A \in N} \sum_\alpha \; E_{p_T}\; f(A \rightarrow \alpha, t) f(B, \alpha)$$
$$= \; \sum_{B \in N} q_B \cdot E_{p_T}\; f_c(B,t), \tag{12}$$

where $f_c(B,t)$ indicates the number of times a node labeled by nonterminal B appears in the derivation tree t as a child of some other node.

From our assumptions on the start symbol S, we have that S only appears at the root of the trees in $T(G)$. Then it is easy to see that, for every $A \neq S$, we have $E_{p_T} f_c(A,t) = E_{p_T} f(A,t)$, while $E_{p_T} f_c(S,t) = 0$ and $E_{p_T} f(S,t) = 1$. Using these relations in (12) we obtain

$$q_S \cdot E_{p_T}\; f(S,T) \;\leq\; q_S \cdot E_{p_T}\; f_c(S,T),$$

from which we conclude $q_S = 0$, thus implying the consistency of $\mathcal{G}$.

4 Cross-entropy and derivational entropy

In this section we present the main result of the paper. We show that, when $\mathcal{G} = (G, p_G)$ is estimated by minimizing the cross-entropy in (5), then such cross-entropy takes the same value as the derivational entropy of $\mathcal{G}$, defined in (3).

In (Nederhof and Satta, 2004) relations are derived for the exact computation of $H_d(p_G)$. For later use, we report these relations below, under the assumption that $\mathcal{G}$ is consistent (see Section 3). We have

$$H_d(p_G) \;=\; \sum_{A \in N} out_{\mathcal{G}}(A) \cdot H_A(p_G). \tag{13}$$

Quantities $H_A(p_G)$, $A \in N$, have been defined in (4). For each $A \in N$, quantity $out_{\mathcal{G}}(A)$ is the sum of the probabilities of all trees generated by $\mathcal{G}$, having root labeled by S and having a yield composed of terminal symbols with an unexpanded occurrence of nonterminal A. Again, we assume that symbol S does not appear in any of the right-hand sides of the rules in R. This means that S only appears at the root of the trees in $T(G)$. Under this condition, quantities $out_{\mathcal{G}}(A)$ can be exactly computed by solving the following system of linear equations (see also (Nederhof, 2005))

$$out_{\mathcal{G}}(S) \;=\; 1; \tag{14}$$

for each $A \neq S$

$$out_{\mathcal{G}}(A) =$$
$$= \; \sum_{B \rightarrow \beta} out_{\mathcal{G}}(B) \cdot f(A, \beta) \cdot p_G(B \rightarrow \beta) \tag{15}$$

We can now prove the equality

$$H_d(p_G) \;=\; H(p_T \,||\, p_G), \qquad (16)$$

where $\mathcal{G}$ is the PCFG estimated by minimizing the cross-entropy in (5), as described in Section 3.

We start from the definition of cross-entropy

$$\begin{aligned}
&H(p_T \,||\, p_G) = \\
&= -\sum_{t \in T} p_T(t) \cdot \log p_G(t) \\
&= -\sum_{t \in T} p_T(t) \cdot \log \prod_{A \to \alpha} p_G(A \to \alpha)^{f(A \to \alpha, t)} \\
&= -\sum_{t \in T} p_T(t) \cdot \\
&\quad \cdot \sum_{A \to \alpha} f(A \to \alpha, t) \cdot \log p_G(A \to \alpha) \\
&= -\sum_{A \to \alpha} \log p_G(A \to \alpha) \cdot \\
&\quad \cdot \sum_{t \in T} p_T(t) \cdot f(A \to \alpha, t) \\
&= -\sum_{A \to \alpha} \log p_G(A \to \alpha) \cdot \\
&\quad \cdot E_{p_T} \; f(A \to \alpha, t). \qquad (17)
\end{aligned}$$

From our estimator in (10) we can write

$$\begin{aligned}
&E_{p_T} \; f(A \to \alpha, t) = \\
&\quad = \; p_G(A \to \alpha) \cdot E_{p_T} \; f(A, t). \qquad (18)
\end{aligned}$$

Replacing (18) into (17) gives

$$\begin{aligned}
&H(p_T \,||\, p_G) = \\
&= \; -\sum_{A \to \alpha} \log p_G(A \to \alpha) \cdot \\
&\quad \cdot p_G(A \to \alpha) \cdot E_{p_T} \; f(A, t) \\
&= \; -\sum_{A \in N} E_{p_T} \; f(A, t) \cdot \\
&\quad \cdot \sum_{\alpha} p_G(A \to \alpha) \cdot \log p_G(A \to \alpha) \\
&= \; \sum_{A \in N} E_{p_T} \; f(A, t) \cdot H(p_G, A). \qquad (19)
\end{aligned}$$

Comparing (19) with (13) we see that, in order to prove the equality in (16), we need to show relations

$$E_{p_T} \; f(A, t) \;=\; out_{\mathcal{G}}(A), \qquad (20)$$

for every $A \in N$. We have already observed in Section 3 that, under our assumption on the start symbol S, we have

$$E_{p_T} \; f(S, t) \;=\; 1. \qquad (21)$$

We now observe that, for any $A \in N$ with $A \neq S$ and any $t \in T(G)$, we have

$$\begin{aligned}
&f(A, t) = \\
&\quad = \sum_{B \to \beta} f(B \to \beta, t) \cdot f(A, \beta). \qquad (22)
\end{aligned}$$

For each $A \in N$ with $A \neq S$ we can then write

$$\begin{aligned}
&E_{p_T} \; f(A, t) = \\
&= \sum_{t \in T} p_T(t) \cdot f(A, t) \\
&= \sum_{t \in T} p_T(t) \cdot \sum_{B \to \beta} f(B \to \beta, t) \cdot f(A, \beta) \\
&= \sum_{B \to \beta} \sum_{t \in T} p_T(t) \cdot f(B \to \beta, t) \cdot f(A, \beta) \\
&= \sum_{B \to \beta} E_{p_T} \; f(B \to \beta, t) \cdot f(A, \beta). \qquad (23)
\end{aligned}$$

Once more we use relation (18), which replaced in (23) provides

$$\begin{aligned}
&E_{p_T} \; f(A, t) = \\
&\quad = \sum_{B \to \beta} E_{p_T} \; f(B, t) \cdot \\
&\qquad \cdot f(A, \beta) \cdot p_G(B \to \beta). \qquad (24)
\end{aligned}$$

Notice that the linear system in (14) and (15) and the linear system in (21) and (24) are the same. Thus we conclude that quantities $E_{p_T} \; f(A, t)$ and $out_{\mathcal{G}}(A)$ are the same for each $A \in N$. This completes our proof of the equality in (16). Some examples will be discussed in Section 6.

Besides its theoretical significance, the equality in (16) can also be exploited in the computation of the cross-entropy in practical applications. In fact, cross-entropy is used as a measure of tightness in comparing different models. In case of estimation from an infinite distribution p_T, the definition of the cross-entropy $H(p_T \,||\, p_G)$ contains an infinite summation, which is problematic for the computation of such quantity. In standard practice, this problem is overcome by generating a finite sample $T^{(n)}$ of large size n, through the distribution p_T, and then computing the approximation (Manning and Schütze, 1999)

$$H(p_T \,||\, p_G) \;\sim\; -\frac{1}{n} \sum_{t \in T} f(t, T^{(n)}) \cdot \log p_G(t),$$

where $f(t, T^{(n)})$ indicates the multiplicity, that is, the number of occurrences, of t in $T^{(n)}$. However, in practical applications n must be very large in order to have a small error. Based on the results in this section, we can instead compute the exact value of $H(p_T \,||\, p_G)$ by computing the derivational entropy $H_d(p_G)$, using relation (13) and solving the linear system in (14) and (15), which takes cubic time in the number of nonterminals of the grammar.

5 Estimation based on likelihood

In natural language processing applications, the estimation of a PCFG is usually carried out on the basis of a finite sample of trees, called tree bank. The so-called maximum likelihood estimation (MLE) method is exploited, which maximizes the likelihood of the observed data. In this section we show that the MLE method is a special case of the estimation method presented in Section 3, and that the results of Section 4 also hold for the MLE method.

Let $\mathcal{T}$ be a tree sample, and let T be the underlying set of trees. For $t \in T$, we let $f(t, \mathcal{T})$ be the multiplicity of t in $\mathcal{T}$. We define

$$\begin{aligned} & f(A \to \alpha, \mathcal{T}) = \\ & \quad = \sum_{t \in T} f(t, \mathcal{T}) \cdot f(A \to \alpha, t), \end{aligned} \tag{25}$$

and let $f(A, \mathcal{T}) = \sum_\alpha f(A \to \alpha, \mathcal{T})$. We can induce from $\mathcal{T}$ a probability distribution $p_{\mathcal{T}}$, defined over T, by letting for each $t \in T$

$$p_{\mathcal{T}}(t) \quad = \quad \frac{f(t, \mathcal{T})}{|\mathcal{T}|}. \tag{26}$$

Note that $\sum_{t \in T} p_{\mathcal{T}}(t) = 1$. Distribution $p_{\mathcal{T}}$ is called the **empirical distribution** of $\mathcal{T}$.

Assume that the trees in T have internal nodes labeled by symbols in N, root nodes labeled by S and leaf nodes labeled by symbols in Σ. Let also R be the finite set of rules that are observed in $\mathcal{T}$. We define the skeleton CFG underlying T as $G = (N, \Sigma, R, S)$. In the MLE method we probabilistically extend the skeleton CFG G by means of a function p_G that maximizes the likelihood of $\mathcal{T}$, defined as

$$p_G(\mathcal{T}) \quad = \quad \prod_{t \in T} p_G(t)^{f(t, \mathcal{T})}, \tag{27}$$

subject to the usual properness conditions on p_G. Such maximization provides the estimator (see for instance (Chi and Geman, 1998))

$$p_G(A \to \alpha) \quad = \quad \frac{f(A \to \alpha, \mathcal{T})}{f(A, \mathcal{T})}. \tag{28}$$

Let us consider the estimator in (10). If we replace distribution p_T with our empirical distribution $p_{\mathcal{T}}$, we derive

$$\begin{aligned} & p_G(A \to \alpha) = \\ & \quad = \frac{E_{p_{\mathcal{T}}} \, f(A \to \alpha, t)}{E_{p_{\mathcal{T}}} \, f(A, t)} \\ & \quad = \frac{\sum_{t \in T} \frac{f(t, \mathcal{T})}{|\mathcal{T}|} \cdot f(A \to \alpha, t)}{\sum_{t \in T} \frac{f(t, \mathcal{T})}{|\mathcal{T}|} \cdot f(A, t)} \\ & \quad = \frac{\sum_{t \in T} \, f(t, \mathcal{T}) \cdot f(A \to \alpha, t)}{\sum_{t \in T} \, f(t, \mathcal{T}) \cdot f(A, t)} \\ & \quad = \frac{f(A \to \alpha, \mathcal{T})}{f(A, \mathcal{T})}. \end{aligned} \tag{29}$$

This is precisely the estimator in (28).

From relation (29) we conclude that the MLE method can be seen as a special case of the general estimator in Section 3, with the input distribution defined over a finite set of trees. We can also derive the well-known fact that, in the finite case, the maximization of the likelihood $p_G(\mathcal{T})$ corresponds to the minimization of the cross-entropy $H(p_{\mathcal{T}} \,||\, p_G)$.

Let now $\mathcal{G} = (G, p_G)$ be a PCFG trained on $\mathcal{T}$ using the MLE method. Again from relation (29) and Section 3 we have that $\mathcal{G}$ is consistent. This result has been firstly shown in (Chaudhuri et al., 1983) and later, with a different proof technique, in (Chi and Geman, 1998). We can then transfer the results of Section 4 to the supervised MLE method, showing the equality

$$H_d(p_G) \quad = \quad H(p_{\mathcal{T}} \,||\, p_G). \tag{30}$$

This result was not previously known in the literature on statistical parsing of natural language. Some examples will be discussed in Section 6.

6 Some examples

In this section we discuss a simple example with the aim of clarifying the theoretical results in the previous sections. For a real number q with $0 < q < 1$,

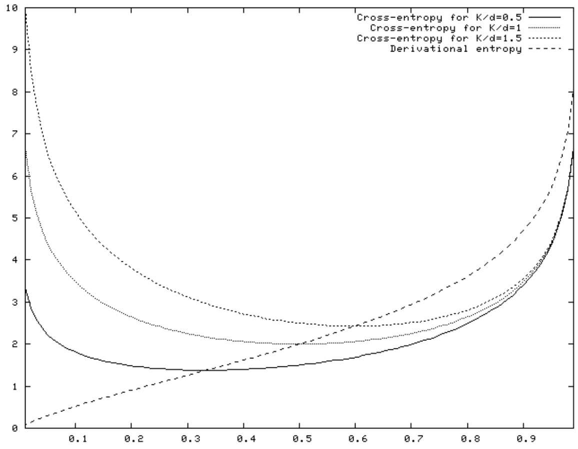

Figure 1: Derivational entropy of $\mathcal{G}_q$ and cross-entropies for three different corpora.

consider the CFG G defined by the two rules $S \rightarrow aS$ and $S \rightarrow a$, and let $\mathcal{G}_q = (G, p_{G,q})$ be the probabilistic extension of G with $p_{G,q}(S \rightarrow aS) = q$ and $p_{G,q}(S \rightarrow a) = 1 - q$. This grammar is unambiguous and consistent, and each tree t generated by G has probability $p_{G,q}(t) = q^i \cdot (1-q)$, where $i \geq 0$ is the number of occurrences of rule $S \rightarrow aS$ in t.

We use below the following well-known relations $(0 < r < 1)$

$$\sum_{i=0}^{+\infty} r^i = \frac{1}{1-r}, \tag{31}$$

$$\sum_{i=1}^{+\infty} i \cdot r^{i-1} = \frac{1}{(1-r)^2}. \tag{32}$$

The derivational entropy of $\mathcal{G}_q$ can be directly computed from its definition as

$$\begin{aligned} H_d(p_{G,q}) &= -\sum_{i=0}^{+\infty} q^i \cdot (1-q) \cdot \log\left(q^i \cdot (1-q)\right) \\ &= -(1-q)\sum_{i=0}^{+\infty} q^i \log q^i + \\ &\quad -(1-q) \cdot \log(1-q) \cdot \sum_{i=0}^{+\infty} q^i \\ &= -(1-q) \cdot \log q \cdot \\ &\quad \sum_{i=0}^{+\infty} i \cdot q^i - \log(1-q) \\ &= -\frac{q}{1-q} \cdot \log q - \log(1-q). \end{aligned} \tag{33}$$

See Figure 1 for a plot of $H_d(p_{G,q})$ as a function of q.

If a tree bank is given, composed of occurrences of trees generated by G, the value of q can be estimated by applying the MLE or, equivalently, by minimizing the cross-entropy. We consider here several tree banks, to exemplify the behaviour of the cross-entropy depending on the structure of the sample of trees. The first tree bank $\mathcal{T}$ contains a single tree t with a single occurrence of rule $S \rightarrow aS$ and a single occurrence of rule $S \rightarrow a$. We then have $p_{\mathcal{T}}(t) = 1$ and $p_{G,q}(t) = q \cdot (1-q)$. The cross-entropy between distributions $p_{\mathcal{T}}$ and $p_{G,q}$ is then

$$\begin{aligned} H(p_{\mathcal{T}}, p_{G,q}) &= -\log q \cdot (1-q) \\ &= -\log q - \log(1-q). \end{aligned} \tag{34}$$

The cross-entropy $H(p_{\mathcal{T}}, p_{G,q})$, viewed as a function of q, is a convex-$\cup$ function and is plotted in Figure 1 (line indicated by $\frac{K}{d} = 1$, see below). We can obtain its minimum by finding a zero for the first derivative

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}q} H(p_{\mathcal{T}}, p_{G,q}) &= -\frac{1}{q} + \frac{1}{1-q} \\ &= \frac{2q-1}{q \cdot (1-q)} = 0, \end{aligned} \tag{35}$$

which gives $q = 0.5$. Note from Figure 1 that the minimum of $H(p_{\mathcal{T}}, p_{G,q})$ crosses the line corresponding to the derivational entropy, as should be expected from the result in Section 4.

More in general, for integers $d > 0$ and $K > 0$, consider a tree sample $\mathcal{T}_{d,K}$ consisting of d trees t_i, $1 \leq i \leq d$. Each t_i contains $k_i \geq 0$ occurrences of rule $S \rightarrow aS$ and one occurrence of rule $S \rightarrow a$. Thus we have $p_{\mathcal{T}_{d,K}}(t_i) = \frac{1}{d}$ and $p_{G,q}(t_i) = q^{k_i} \cdot (1-q)$. We let $\sum_{i=1}^{d} k_i = K$. The cross-entropy is

$$\begin{aligned} & H(p_{\mathcal{T}_{d,K}}, p_{G,q}) = \\ &\quad = -\sum_{i=0}^{d} \frac{1}{d} \cdot \log q^{k_i} - \log(1-q) \\ &\quad = -\frac{K}{d} \log q - \log(1-q). \end{aligned} \tag{36}$$

In Figure 1 we plot $H(p_{\mathcal{T}_{d,K}}, p_{G,q})$ in the case $\frac{K}{d} = 0.5$ and in the case $\frac{K}{d} = 1.5$. Again, we have that these curves intersect with the curve corresponding to the derivational entropy $H_d(p_{G,q})$ at the points were they take their minimum values.

7 Conclusions

We have shown in this paper that, when a PCFG is estimated from some tree distribution by minimizing the cross-entropy, then the cross-entropy takes the same value as the derivational entropy of the PCFG itself. As a special case, this result holds for the maximum likelihood estimator, widely applied in statistical natural language parsing. The result also holds for the relative weighted frequency estimator introduced in (Chi, 1999) as a generalization of the maximum likelihood estimator, and for the estimator introduced in (Nederhof, 2005) already discussed in the introduction. In a journal version of the present paper, which is under submission, we have also extended the results of Section 4 to the unsupervised estimation of a PCFG from a distribution defined over an infinite set of (unannotated) sentences and, as a particular case, to the well-knonw inside-outside algorithm (Manning and Schütze, 1999).

In practical applications, the results of Section 4 can be exploited in the computation of model tightness. In fact, cross-entropy indicates how much the estimated model fits the observed data, and is commonly exploited in comparison of different models on the same data set. We can then use the given relation between cross-entropy and derivational entropy to compute one of these two quantities from the other. For instance, in the case of the MLE method we can choose between the computation of the derivational entropy and the cross-entropy, depending basically on the instance of the problem at hand. As already mentioned, the computation of the derivational entropy requires cubic time in the number of nonterminals of the grammar. If this number is large, direct computation of (5) on the corpus might be more efficient. On the other hand, if the corpus at hand is very large, one might opt for direct computation of (3).

Acknowledgements

Helpful comments from Zhiyi Chi, Alberto lavelli, Mark-Jan Nederhof and Khalil Simaan are gratefully acknowledged.

References

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. 2001. Immediate-head parsing for language models. In *39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 116–123, Toulouse, France, July.

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.

C. Chelba and F. Jelinek. 1998. Exploiting syntactic structure for language modeling. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 1, pages 225–231, Montreal, Quebec, Canada, August.

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.

M. Collins. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, pages 589–638.

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

C.D. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Massachusetts Institute of Technology.

M.-J. Nederhof and G. Satta. 2004. Kullback-Leibler distance between probabilistic context-free grammars and probabilistic finite automata. In *Proc. of the 20^{th} COLING*, volume 1, pages 71–77, Geneva, Switzerland.

M.-J. Nederhof. 2005. A general technique to train language models on language models. *Computational Linguistics*, 31(2):173–185.

B. Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2):249–276.

Y. Schabes. 1992. Stochastic lexicalized tree-adjoining grammars. In *Proc. of the fifteenth International Conference on Computational Linguistics*, volume 2, pages 426–432, Nantes, August.