On the Statistical Consistency of DOP Estimators

Detlef Prescher, Remko Scha, Khalil Sima'an and Andreas Zollmann

Institute for Logic, Language and Computation (ILLC)
Universiteit van Amsterdam

Abstract

Given a sample from an unknown probability distribution, a statistical estimator uses the sample in order to guess the unknown distribution. One desired property of an estimator is that its guess is increasingly likely to get arbitrarily close to the unknown distribution as the samples become larger. This property of an estimator is called *consistency*.

Data Oriented Parsing (DOP) employs *all fragments* of the trees in a training treebank, including the full parse-trees themselves, as the rewrite rules of a probabilistic tree-substitution grammar. Since the most popular DOP-estimator (DOP1) was shown to be inconsistent, there is an outstanding theoretical question concerning the possibility of DOP-estimators with reasonable statistical properties. This question constitutes the topic of the current paper.

First, we show that, contrary to common wisdom, any unbiased estimator for DOP is futile because it *will not* generalize over the training treebank. Subsequently, we show that a *consistent* estimator that generalizes over the treebank should involve a local smoothing technique. This exposes the relation between DOP and existing memory-based models that work with full memory and an analogical function such as k-nearest neighbor, which is known to implement backoff smoothing. Finally, we present a new consistent backoff-based estimator for DOP and discuss how it combines the memory-based preference for the longest match with the probabilistic preference for the most frequent match.

1 Introduction

The simplest probabilistic grammars (such as Stochastic Context-Free Grammars and Stochastic Tree-Adjoining Grammars) treat the different grammar rules as statistically independent. Since this results in sub-optimal predictions, more recent models allow the probabilities of rules to be conditioned on their local context, e.g. on the labels of parent- and sister-nodes, head-POS-tags or head-words (Black, Jelinek, Lafferty, Magerman, Mercer and Roukos 1993, Collins 1997, Johnson 1998, Charniak 1999, Collins and Duffy 2002).

Data-Oriented Parsing (DOP) (Scha 1990, Bod 1995) constitutes an early and radical example of this tendency. DOP treats *all* subtrees of an arbitrarily large treebank as accessible elements of a person's past linguistic experience, and assumes that all of them are potentially relevant for subsequent disambiguation decisions. DOP is thus reminiscent of non-probabilistic A.I. approaches like *Memory-Based Learning* (Stanfill and Waltz 1986, Daelemans 1995), which analyze new input by matching it with a store of earlier instances, using techniques such as *k-nearest-neighbor*.

As emphasized in (Scha 1990), DOP is a synthesis of these two research traditions: "(1) The analogy between the input sentence and the corpus should be constructed in the simplest possible way, i.e.: the number of constructions that is

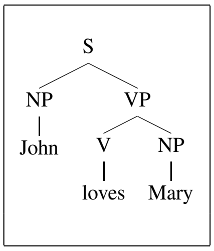

Figure 1: Treebank

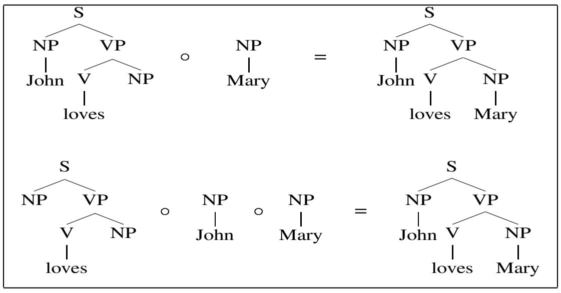

Figure 2: Two different derivations of the same parse

used to re-construct the sentence should be as small as possible. (2) More frequent constructions should be preferred above less frequent ones."

Non-probabilistic DOP systems which explicitly implement the first of these desiderata were developed only recently (De Pauw 2000a, De Pauw 2000b, Bod 2000b). Most work on DOP has employed Stochastic Tree Substitution Grammars, and has assumed that the preference for the simplest analysis will be a natural side-effect of choosing the right probability assignments. Various methods for estimating the subtree probabilities have been proposed for the DOP model. The motivation for these proposals, however, has been largely heuristic or practical. A fundamental question has thus remained unanswered: Is it possible to use statistically well-motivated estimators for treebank-grammars with such a large and redundant parameter space?

Background. Negative results about DOP estimators have been established already. Johnson (2002) investigated the DOP1 estimator (Bod 1995) which estimates the substitution probability of a subtree for a non-terminal node directly as the relative frequency of this subtree among all corpus subtrees with the same root-label. Johnson showed that this estimator is inconsistent and biased. A more fundamental problem was exposed by (Bonnema, Buying and Scha 1999), who provide an example where the Maximum-Likelihood Estimator completely overfits the DOP model to the training treebank, i.e. it assigns *zero* probability to *all* parses which do not occur in the training treebank.

It should be noted that implemented systems rarely deploy the full DOP model. If only for reasons of space and time, they restrict the subtree-set to subtrees satisfying certain criteria (maximum depth, minimal/maximal number of terminal/non-terminal leaves, etc.), thus loosing some properties of the full DOP model. In order to understand the essential features of DOP, however, it is interesting to investigate the unrestricted model.

Preview. This paper investigates the possibility of *consistent, non-overfitting* estimators for the *unrestricted* DOP model. (We thus exclude subtree selection on

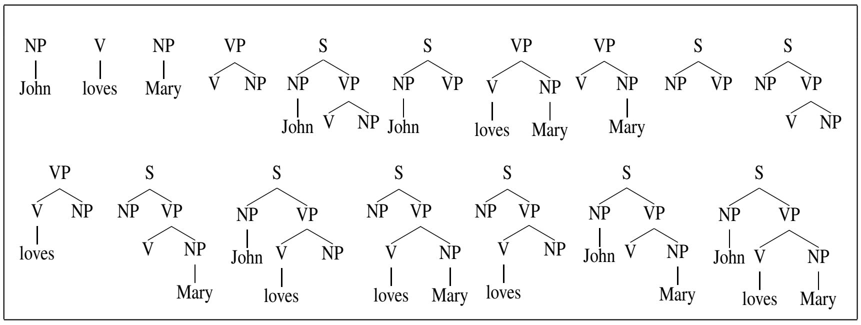

Figure 3: Fragments of the treebank in Figure 1

the basis of *a priori* criteria.) The paper proves that *any* unbiased estimator for the DOP model will yield a probability distribution that completely overfits the treebank. It identifies different ways to define a non-overfitting and consistent DOP estimator. The simplest solution is an estimator which uses smoothing, e.g. by *leaving-one-out* (Ney, Martin and Wessel 1997). We present such an estimator, and explain why it strikes a good balance between the probabilistic and the memory-based facets of DOP.

2 The DOP Model

Let be given a treebank TB, i.e. a finite sequence of utterance-parse pairs. Like other treebank models, DOP acquires from TB a finite set $\mathcal{F}$ of rewrite productions, called *subtrees* or *fragments*, together with their probability estimates. A connected subgraph of a treebank tree t is called a *subtree* iff it consists of one or more context-free productions[1] from t. The set $\mathcal{F}$ consists of *all* subtrees of the treebank trees. Figure 3 exemplifies the set of subtrees extracted from the treebank of Figure 1.

In DOP, the set of subtrees $\mathcal{F}$ is employed as a Stochastic Tree-Substitution Grammar (STSG), with the same start symbol, nonterminal and terminal sets as the treebank trees. A TSG is a rewrite system similar to a Context-Free Grammar (CFG), with the only difference that the productions of a TSG are subtrees of arbitrary depth. Like in CFGs, a (leftmost) derivation in a TSG starts with the start symbol S, and proceeds by replacing nonterminal symbols by subtrees using the (leftmost) substitution operation (denoted $\circ$). Given trees t_i and t_j, the rewriting $t_i \circ t_j$ is defined iff the leftmost nonterminal leaf-node μ of t_i carries the same label as the root node of t_j; the result is a tree consisting of t_i with t_j substituted at node μ. The derivation $(S \circ t_1 \circ \ldots \circ t_n)$ stands for a finite sequence of such left-associative substitutions, i.e. $(\ldots(S \circ t_1) \circ \ldots) \circ t_n)$. Multiple leftmost TSG-derivations may generate the same tree. (This constitutes an important conceptual

[1] Note that a non-leaf node labeled p and the sequence of its daughter nodes labeled $c_1, \cdots, c_n$, together constitute a graph that represents the context-free production: $p \longrightarrow c_1 \cdots c_n$.

and computational difference between TSG's and CFG's.) For instance, the parse in Figure 1 can be derived in at least two different ways as shown in Figure 2.

Given a specific TSG, i.e., given a specific set $\mathcal{F}$ of fragments and a procedure that re-combines these fragments, a Stochastic TSG (STSG) is defined on the basis of the following three notions:

Fragment probability: To each $t \in \mathcal{F}$, a real number $0 \leq \pi(t) \leq 1$ is assigned, such that for non-terminal label A, π induces a probability distribution on the set of fragments t whose root label R_t is A. I.e.:

$$\sum_{t\,:\,R_t=A} \pi(t) = 1$$

Derivation probability: The probability of a derivation d of a parse-tree is the product of its fragment probabilities:

$$p(d) \quad = \quad \prod_{t\in\mathcal{F}} \pi(t)^{f_t(d)} \tag{1}$$

Here $f_t(d)$ is the number of times the fragment t occurs in the derivation d.

Tree probability: The probability of a parse-tree x with a set of derivations $\mathcal{D}(x)$ is the sum of the probabilities of its derivations:

$$p(x) \quad = \quad \sum_{d\in\mathcal{D}(x)} p(d) \tag{2}$$

One of the simplest and most influential DOP grammars, called DOP1 (Bod 1995), employs $\pi(t) = \frac{count(t)}{count(R_t)}$ where $count(t)$ stands for the frequency count of subtree t in TB. This estimator was shown to be inconsistent (Johnson 2002). This raises the question whether and how $\pi(t\ )$ may be estimated in a consistent manner.

3 An Excursion into Estimation Theory

A statistics problem is a problem in which a *corpus* that has been generated in accordance with some unknown *probability distribution* is to be analyzed so that some inference about the unknown distribution can be made. In other words, in a statistics problem there is a choice between two or more probability distributions which might have generated the corpus. In practice, there are often an infinite number of different possible distributions – statisticians bundle these into one single *probability model* – which might have generated the corpus. By analyzing the corpus, an attempt is made to learn about the unknown distribution. On the basis of the corpus, an *estimation method* selects one *instance* of the probability model as its best guess about the original distribution.

This section provides the elements that are necessary for further discussion of DOP estimators. We start out by reviewing some definitions from Estimation Theory, inculding the properties of bias and consistency of estimators. Subsequently we state an important theorem concerning the relation between Maximum-Likelihood estimation, Relative-Entropy Estimation and relative frequency.

Corpora and Probability Models

Let $\mathcal{N}$ be the natural numbers (including zero), and let $\mathcal{X}$ be a countable set. Then, each function $f : \mathcal{X} \to \mathcal{N}$ is called a *corpus*, each $x \in \mathcal{X}$ is called a *type*, and each value of f is called a *type frequency*. The *corpus size* is defined as $|f| = \sum_{x \in \mathcal{X}} f(x)$. It is easy to check that these definitions cover the standard notion of the term *corpus* (used in Computational Linguistics) and of the term *sample* (used in Statistics).

Let $c = <x_1, \ldots, x_n> \in \mathcal{X}^n$ be a finite sequence of type instances from $\mathcal{X}$. Then, the *occurrence frequency* of a type x in c is defined as $c(x) = |\{ i \mid x_i = x\}|$. Clearly, $c(.)$ is a corpus in the sense of our definition above, since it has all the relevant properties: The type x does not occur in c if and only if $c(x) = 0$; in any other case, there are $c(x)$ tokens of x in c. Finally, the corpus size $|c|$ is identical to n, the number of tokens in c.

The *probability of a corpus* $f : \mathcal{X} \to \mathcal{N}$ under a probability mass function $p : \mathcal{X} \to [0, 1]$ is given by $L_p(f) = \prod_{x \in \mathcal{X}} p(x)^{f(x)}$.

A *probability model on* $\mathcal{X}$ is a non-empty set $\mathcal{M}$ of probability distributions on $\mathcal{X}$. The elements of $\mathcal{M}$ are called *instances*. The *unrestricted model* is the set $\mathcal{M}(\mathcal{X})$ of *all* probability distributions on $\mathcal{X}$, i.e.

$$\mathcal{M}(\mathcal{X}) = \left\{ p \colon \mathcal{X} \to [0,1] \;\middle|\; \sum_{x \in \mathcal{X}} p(x) = 1 \right\}$$

A probability model $\mathcal{M}$ is called *restricted* in all other cases: $\mathcal{M} \subseteq \mathcal{M}(\mathcal{X})$ and $\mathcal{M} \neq \mathcal{M}(\mathcal{X})$.

Note that in the earlier (more informal) literature on statistical aspects of DOP, the term ”probability model” was often used to indicate a specific estimation method, rather than the general DOP probability model in the sense defined here. The title of (Bonnema et al. 1999) is a case in point.

Estimators and Their Properties: Bias and Consistency

Let $\mathcal{C}_n$ be the set $\mathcal{X}^n = \{<x_1, ..., x_n> | x_i \in \mathcal{X}$ for all $i = 1, ..., n\}$, i.e, $\mathcal{C}_n$ comprises all corpora of size n, thereby distinguishing even between corpora which have the same occurrence frequencies but a different ordering of elements. If $\mathcal{M}$ is a probability model on $\mathcal{X}$, then each function $\mathrm{est}_n : \mathcal{C}_n \to \mathcal{M}$ is called an *estimator* for $\mathcal{M}$. Given a model instance $p \in \mathcal{M}$, the estimator's *expectation*[2] is calculated by $E_p(\mathrm{est}_n) = \sum_{f \in \mathcal{C}_n} L_p(f) \cdot \mathrm{est}_n(f)$. The estimator is called *unbiased* for $p \in \mathcal{M}$ iff

$$E_p(\mathrm{est}_n) = p \;.$$

A sequence of estimators est_n is called *asymptotically unbiased* for $p \in \mathcal{M}$ iff

$$\lim_{n \to \infty} E_p(\mathrm{est}_n) = p \;.$$

[2] The term *expectation* is justified because the corpus probabilities $L_p(f)$ form a probability distribution on $\mathcal{C}_n$, i.e., $L_p(\mathcal{C}_n) = \sum_{f \in \mathcal{C}_n} L_p(f) = 1$.

Moreover, a sequence of estimators est_n is called *consistent* for $p \in \mathcal{M}$ iff for all $x \in \mathcal{X}$ and for all $\epsilon > 0$

$$\lim_{n \to \infty} L_p\left(\{f \in \mathcal{C}_n : |\mathrm{est}_n(f)(x) - p(x)| > \epsilon\}\right) = 0 \ .$$

As Johnson (2002) noted, there is no unique definition of "consistency". The different definitions, however, share the intuition that an estimator is to be called consistent, if it outputs "in the limit" exactly the probability distribution $p \in \mathcal{M}$ that generated its input-corpus. For the sake of simplicity, we use a definition that operates on the types $x \in \mathcal{X}$, but that avoids the introduction of *loss functions*.

Maximum Likelihood Estimation

A *Maximum Likelihood estimate for* $\mathcal{M}$ *on* f is an instance $\hat{p} \in \mathcal{M}$ such that the corpus f is allocated a maximum probability, i.e., $L_{\hat{p}}\left(f\right) = \max_{p \in \mathcal{M}} L_p\left(f\right)$.

Relative-Frequency Estimation

The *relative-frequency estimate* on a non-empty and finite corpus f is defined by $\tilde{p}\colon \mathcal{X} \to [0,1]$ where $\tilde{p}(x) = |f|^{-1} \cdot f(x)$.

The following theorem clarifies the relation between relative-frequency estimation and Maximum Likelihood estimation.

Theorem 1 *The relative-frequency estimate $\tilde{p}$ is a Maximum Likelihood estimate for a (restricted or unrestricted) probability model $\mathcal{M}$ on f, if and only if $\tilde{p} \in \mathcal{M}$. In this case, $\tilde{p}$ is a unique ML estimate.*

Proof: Combine theorems 2 and 3 of the subsequent paragraph. **q.e.d.**

Relative-Entropy Estimation

The *relative entropy* $D(p||q)$ of two probability distributions p and q is defined by $D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$. Interestingly, ML estimation is equivalent to minimizing the relative entropy:

Theorem 2 *An instance $\hat{p}$ of a probability model $\mathcal{M}$ is a ML estimate for $\mathcal{M}$ on f, if and only if*

$$D(\tilde{p}||\hat{p}) = \min_{p \in \mathcal{M}} D(\tilde{p}||p) \ .$$

Proof: First, $D(\tilde{p}||p)$ is the difference between the *cross-entropy* $H(\tilde{p};p) = -\sum_{x \in \mathcal{X}} \tilde{p}(x) \log p(x)$ and the *entropy* $H(\tilde{p}) = -\sum_{x \in \mathcal{X}} \tilde{p}(x) \log \tilde{p}(x)$, i.e., $D(\tilde{p}||p) = H(\tilde{p};p) - H(\tilde{p})$. So minimizing $D(\tilde{p}||p)$ is equivalent to minimizing $H(\tilde{p};p)$. Finally, as $H(\tilde{p};p) = -\frac{1}{|f|} \log L_p\left(f\right)$, minimizing $D(\tilde{p}||p)$ is equivalent to maximizing $L_p\left(f\right)$ **q.e.d.**

Theorem 3 (Information Inequality) *Let p and q be two probability distributions. Then $D(p \mid\mid q) \geq 0$ with equality if and only if $p = q$.*

Proof: See, e.g., (Cover and Thomas 1991).

4 Estimation Theory and DOP

The probability model of a DOP grammar bundles specific probability distributions on the set $\mathcal{X}$ of derivable parse trees. Each model instance p is induced by a function π on the set $\mathcal{F}$ of tree fragments such that the equations (1) and (2) in section 2 are satisfied. I.e.:

$$\mathcal{M}_{\mathrm{DOP}} = \left\{ p \in \mathcal{M}(\mathcal{X}) \;\middle|\; \exists \pi : p(x) = \sum_{d \in \mathcal{D}(x)} \prod_{t \in \mathcal{F}} \pi(t)^{f_t(d)} \right\}$$

By contrast, the corresponding CFG's probability model is defined solely on the basis of fragments of depth one:

$$\mathcal{M}_{\mathrm{CFG}} = \left\{ p \in \mathcal{M}(\mathcal{X}) \;\middle|\; \exists \pi : p(x) = \prod_{r} \pi(r)^{f_r(x)} \right\}$$

(Here $f_r(x)$ is the number of occurrences of a CFG rule r in the full parse-tree $x \in \mathcal{X}$.)

A striking property of DOP's probability model is that it can model the relative frequencies of all trees in a given treebank:

Theorem 4 *Let f_{TB} be a treebank such that all trees have the same root label.*[3] *Moreover, let the treebank grammar generate at least one full parse-tree outside the treebank. Let $\mathcal{M}_{\mathrm{DOP}}$ and $\mathcal{M}_{\mathrm{CFG}}$ be the probability model of the DOP and the CFG grammar read-off from f_{TB}. Then the relative-frequency estimate $\tilde{p}$ on f_{TB}* ***is*** *an instance of $\mathcal{M}_{\mathrm{DOP}}$, but it* ***is not*** *an instance of $\mathcal{M}_{\mathrm{CFG}}$.*

Proof: For the CFG case: By definition, $\tilde{p}$ assigns positive probabilities to the trees in the treebank. Each instance of $\mathcal{M}_{\mathrm{CFG}}$, however, which assigns positive probabilities to the treebank trees, necessarily assigns positive probabilities to *all* full parse-trees $x \in \mathcal{X}$. In more detail, such a model instance $p \in \mathcal{M}_{\mathrm{CFG}}$ satisfies $p(x) > 0 \,\forall x \in \mathcal{X}$ as it can be shown that $p(x) = \prod_r \pi(r)^{f_r(x)}$ with $\pi(r) > 0 \,\forall r$:

> Assume that there is a rule r_0 with $\pi(r_0) = 0$. As the CFG is read off the treebank, the rule r_0 is read off a *treebank tree* x_0 (i.e. $f_{r_0}(x_0) > 0$). As $p \in \mathcal{M}_{\mathrm{CFG}}$ assigns positive probabilities to the treebank trees, it follows both
>
> $$p(x_0) > 0 \text{ and } p(x_0) = \prod_r \pi(r)^{f_r(x_0)} = \cdots \pi(r_0)^{f_{r_0}(x_0)} \cdots = 0$$
>
> Clearly, this is a contradiction, showing that the assumption is false.

So we conclude that $\tilde{p} \notin \mathcal{M}_{\mathrm{CFG}}$ as $\tilde{p}$ assigns probability zero to at least one full parse-tree outside the treebank.

For the DOP case: If S is the root label of the trees in the treebank, then define:

$$\pi(t) := \begin{cases} \tilde{p}(t) & \text{if } R_t = S \\ \text{arbitrarily} & \text{else} \end{cases}$$

[3] Every treebank tree can be augmented with a new root labeled with a fresh non-terminal symbol.

First of all, we choose arbitrary weights satisfying the side conditions for root labels $A \neq S$. Since S is unique, it is also guaranteed that $\sum_{t\,:\,R_t=S} \pi(t) = 1$. The argument is now as follows: In each derivation occurs at least one fragment with the root label S. By definition, however, a fragment rooted by S has a zero-assignment if it is not itself a full parse-tree of the treebank. Obviously, a derivation has a zero-assignment if it employs such a fragment. It follows especially that derivations combining two or more fragments have probability zero. So the DOP probabilities of full parse-trees x are calculated by $p(x) = \pi(x) = \tilde{p}(x)$. Therefore the relative-frequency estimate $\tilde{p}$ is an instance of the DOP model. **q.e.d.**

Hence, given a treebank, the relative frequency estimate of the trees in this treebank is a member of $\mathcal{M}_{\mathrm{DOP}}$ but is not guaranteed to be a member of $\mathcal{M}_{\mathrm{CFG}}$. When a probability distribution is not a member of a given probability model, it cannot be captured by an estimator over that model. Whereas DOP, in principle, can capture any relative frequency distribution over treebank trees (using a suitable estimator), Probabilistic CFGs read off the treebank cannot do so.

The fact that the DOP model always has the relative frequency over treebank trees as a member has a serious side effect: when the treebank is too small, any estimator that captures the relative frequency will risk overfitting the treebank, as the following theorem shows.

Theorem 5 *Let f_{TB}, $\mathcal{M}_{\mathrm{DOP}}$ and $\mathcal{M}_{\mathrm{CFG}}$ be given as in Theorem 4. Then the relative-frequency estimate $\tilde{p}$ on f_{TB}* ***is*** *the unique ML estimate for $\mathcal{M}_{\mathrm{DOP}}$ on f_{TB}, but* ***is not*** *a ML estimate for $\mathcal{M}_{\mathrm{CFG}}$ on f_{TB}.*

Proof: Combine theorems 1 and 4 **q.e.d.**

As a consequence, a serious overfitting problem is caused by the ML estimates for DOP models. In greater detail: an instance p of a probability model *completely overfits the treebank* if it assigns a probability of one to the treebank , i.e.,

$$\sum_{x\,:\,f_{\mathrm{TB}}(x)>0} p(x) = 1$$

or equivalently, if it assigns zero-probabilities to all full parse-trees outside the treebank:

$$p(x) = 0 \text{ for all } x \in \mathcal{X} \text{ with } f_{\mathrm{TB}}(x) = 0$$

Theorem 6 *Let f_{TB}, $\mathcal{M}_{\mathrm{DOP}}$ and $\mathcal{M}_{\mathrm{CFG}}$ be given as in Theorem 4. Then the ML estimate for $\mathcal{M}_{\mathrm{DOP}}$ on f_{TB} completely overfits the treebank, whereas a ML estimate for $\mathcal{M}_{\mathrm{CFG}}$ on f_{TB} does not.*

Proof: Apply Theorem 5 to DOP, and re-inspect the proof of Theorem 4 for the CFG case **q.e.d.**

In the following, we shall investigate whether there are other estimators for the DOP model that do not completely overfit the treebank, but that satisfy some of the good properties of the ML estimator. We start with a general result:

Theorem 7 *Let* $\mathrm{est}_n : \mathcal{C}_n \to \mathcal{M}$ *be an estimator. Let* $f_0 \in \mathcal{C}_n$ *be a corpus, and let* $x_0 \in \mathcal{X}$ *be a type outside the corpus such that*

$$\mathrm{est}_n(f_0)(x_0) > 0$$

Then the estimator is biased for all model instances $p \in \mathcal{M}$ *that assign a positive probability to the corpus but a zero-probability to the type outside the corpus, i.e.,* $L_p\,(f_0) > 0$ *and* $p(x_0) = 0$ *.*

Proof: Assume that the estimator is unbiased for a model instance $p \in \mathcal{M}$ satisfying $L_p\,(f_0) > 0$ and $p(x_0) = 0$. In what follows, we will show that this assumption leads to a contradiction. First, it follows by definition that

$$\sum_{f \in \mathcal{C}_n} L_p\,(f) \cdot \mathrm{est}_n(f) = p\ .$$

Next, let $\mathcal{X}_p = \{x \in \mathcal{X} \mid p(x) > 0\}$ be the *support* of the model instance p. Then

$$\sum_{x \in \mathcal{X}_p} \sum_{f \in \mathcal{C}_n} L_p\,(f) \cdot \mathrm{est}_n(f)(x) = \sum_{x \in \mathcal{X}_p} p(x)\ .$$

So,

$$\sum_{f \in \mathcal{C}_n} L_p\,(f) \cdot \sum_{x \in \mathcal{X}_p} \mathrm{est}_n(f)(x) = 1\ .$$

As $\sum_x \mathrm{est}_n(f)(x) \leq 1$ and $\sum_f L_p\,(f) = 1$, it follows

$$\sum_{x \in \mathcal{X}_p} \mathrm{est}_n(f)(x) = 1 \text{ whenever } L_p\,(f) > 0\ .$$

So especially $\sum_{x \in \mathcal{X}_p} \mathrm{est}_n(f_0)(x) = 1$ and thus $\mathrm{est}_n(f_0)(x) = 0$ for all $x \notin \mathcal{X}_p$. Finally, as $x_0 \notin \mathcal{X}_p$ and $\mathrm{est}_n(f_0)(x_0) > 0$, there is a contradiction **q.e.d.**

We now apply Theorem 7 to estimators for the CFG and the DOP model. So let $f_0 = f_{\mathrm{TB}}$ be a treebank and let $\mathcal{M}$ be the probability model of the CFG or of the DOP grammar that is read off from the treebank. Starting with the CFG case, Theorem 7 is not really useful: As mentioned several times, *there are no model instances* $p \in \mathcal{M}_{\mathrm{CFG}}$ *satisfying simultaneously* $L_p\,(f_{\mathrm{TB}}) > 0$ *and* $p(x_0) = 0$ *(for an arbitrary* x_0 *outside the treebank).* On the other hand, there is a surprising result for the DOP case:

Theorem 8 *Let* $\mathcal{M}_{\mathrm{DOP}}$ *be read off from a treebank* $f_{\mathrm{TB}} \in \mathcal{C}_n$*, and let all trees have the same root. Then each estimator* $\mathrm{est}_n : \mathcal{C}_n \to \mathcal{M}_{\mathrm{DOP}}$ *that does not completely overfit the treebank is biased for some instance* $p \in \mathcal{M}_{\mathrm{DOP}}$.

Proof: If the estimator does not completely overfit the treebank, there is a full parse-tree $x_0 \in \mathcal{X}$ outside the treebank, satisfying

$$\mathrm{est}_n(f_{\mathrm{TB}})(x_0) > 0$$

As $\tilde{p} \in \mathcal{M}_{\mathrm{DOP}}$ and $\tilde{p}(x_0) = 0$, it follows by Theorem 7 that $\mathrm{est}_n : \mathcal{C}_n \rightarrow \mathcal{M}_{\mathrm{DOP}}$ is biased for $\tilde{p}$ **q.e.d.**

It thus turns out that in the context of extremely rich models such as DOP, lack of bias is not a desirable property for an estimator: it precludes assigning probability mass to unseen events. In other words: in designing estimators for DOP, we should explicitly introduce bias towards predicting trees that were not observed in the corpus. Note, however, that we may still strive for consistency.

5 Approaches to Avoiding Overfitting

There are various ways to combat overfitting in learning, e.g. selecting model parameters and introducing prior preferences (Duda, Hart and Stork 2001). To consider the options for the DOP model, we start out from the formula that searches for the DOP model instance p^* that maximizes an objective function Φ over treebank TB and model instances p:

$$p^* \quad = \quad \arg \max_{p \in \mathcal{M}_{\mathrm{DOP}}} \Phi(p, \mathrm{TB}) \tag{3}$$

where $\mathcal{M}_{\mathrm{DOP}}$ is the space of all possible DOP probability distributions over parse trees. The MLE assignment is achieved when $\Phi(p, \mathrm{TB}) = L_p\,(\mathrm{TB})$, i.e., when Φ is the likelihood function. Due to the theorems in Section 4, we know that an unbiased instance of the DOP model – in particular the MLE estimate – results in complete overfitting. To avoid complete overfitting we may adapt formula (3) in two possible ways.

First of all, we may *constrain* $\mathcal{M}_{\mathrm{DOP}}$ such that "the overfitting instances" are excluded, and employ the MLE. Because here we are interested in exploring a memory-based DOP, i.e. the DOP model that employs all fragments, we will not consider this option any further in this paper.

Secondly, we may *change* Φ such that preference is given to all $p \in \mathcal{M}_{\mathrm{DOP}}$ for which $p(\mathrm{TB}) := \sum_{x \in \mathrm{TB}} p(x) < 1$, thereby reserving probability mass for parses not available in TB and distribute this mass among the subtrees in $\mathcal{F}$.

Hence, there are various options for building estimators that do not overfit DOP. But which ones are consistent, and which ones could be expected to exhibit good empirical results?

Starting from the MLE, which is a consistent estimator, it is attractive to work on constraining $\mathcal{M}_{\mathrm{DOP}}$. Although this approach could be interesting mathematically, it is not directly clear which constraints should be specified for reasonable estimators. Furthermore, within a memory-based framework it is hard to justify fixing a preference for some parametric form over another. It seems more intuitive to strive for preferences that rely to a lesser extent on mathematical abstraction. Hence, we do not pursue this path any further here.

We are left with the second option: allow Φ to exhibit preference for any $p \in \mathcal{M}_{\mathrm{DOP}}$ such that $p(\mathrm{TB}) < 1$, and where the mass $1 - p(\mathrm{TB})$ is reserved for unseen parses. Clearly, by deviating from the MLE, there is risk of inconsistency.

However, there exist various smoothing techniques (re-estimators) that are known to *preserve consistency*.

Smoothing as Estimation We consider resampling methods, including the Jackknife (Duda et al. 2001), held-out and leaving-one-out smoothing method (Ney et al. 1997), which apply a kind of repeated held-out estimation. Held-out estimation splits the treebank TB into two parts: the *extraction* part and the *held-out* part. The extraction part is used for obtaining the space of events $\{E\}$ (i.e. all subtrees $\mathcal{F}$ in the case of DOP), and partitions this space into equivalence classes according to frequency. The held-out part is used for obtaining (discounted) probability estimates for all events $e \in \{E\}$. Crucially, the probabilities are estimated *through MLE over the held-out part*, which reserves probability mass for unseen events. In the limit, when TB approaches infinity, under the condition that both the extraction and the held-out parts are allowed to approach infinity, the leaving-one-out will remain consistent, as it will approach the MLE itself. In fact, this is the reason why leaving-one-out is considered a reasonable smoothing tool for language models using Markov orders over word n-grams.

Consistent estimation for DOP by smoothing methods is in line with the findings of (Zavrel and Daelemans 1997) concerning the surprising similarity between smoothing by Katz backoff (Katz 1987) and by Memory-Based Learning (MBL) (Daelemans 1995).

As we shall see in the next section, the analogy between Memory-Based Learning and the estimation of DOP parameters by smoothing goes further than the surface impression suggests.

6 Consistent DOP Estimators

As mentioned earlier, consistency considerations imply that smoothing must be applied to parse probabilities (p) rather than subtree probabilities (π). Yet, direct smoothing of p will not solve the overfitting problem since it does not explain how the smoothing of p should result in estimates of the subtree probabilities (i.e. π). In this section we concentrate on how smoothing the parse probabilities p can be turned into smoothing the subtree probabilities π.

We present two new estimators and discuss their strengths and limitations. The first of these is the straightforward combination of Maximum-Likelihood with resampling methods, and the second is an application of backoff smoothing and resampling.

Maximum-Likelihood under Resampling: This estimator has four steps:

(1) The training treebank is split into a held-out and an extraction part,

(2) The set of all subtrees is extracted from the extraction part,

(3) Probabilities for these subtrees are obtained by Maximum-Likelihood Estimation over the held-out part, and

(4) Steps 1 to 3 are repeated to obtain averages.

This algorithm is consistent because repeated held-out estimation over MLE preserves consistency. As for Hidden Markov Models, MLE for DOP takes place under hidden derivations, and must be realized via an Inside-Outside algorithm (e.g. (Baker 1979, Bod 2000a)). However, there are no guarantees that Inside-Outside is consistent. In fact, this algorithm is not guaranteed to arrive at the MLE (Wu 1983). We argue, however, that the Inside-Outside algorithm will be consistent when a consistent function is found for the initial assignment, which is a circular situation. Moreover, this algorithm is extremely inefficient for DOP.

Backoff DOP: In DOP, because all treebank subtrees are included, a subtree $t \in \mathcal{F}$ of any depth > 1 can always be obtained by a derivation involving two[4] other subtrees $t_1, t_2 \in \mathcal{F}$ such that $t = t_1 \circ t_2$ (see Figure 4). By definition $p(t_1 \circ t_2 \mid R_{t_1}) = p(t_1 \mid R_{t_1}) p(t_2 \mid R_{t_2})$. Clearly, if this formula is an approximation of $p(t \mid R_t)$, then it must involve an independence assumption $p(t_2 \mid t_1) \approx p(t_2 \mid R_{t_2})$! This independence assumption is a so-called backoff that takes place from complex context t_1 to simpler context R_{t_2} (similar to the backoff from e.g. trigrams to bi-grams or uni-grams in Markov models). The backoff $p(t_2 \mid t_1) \approx p(t_2 \mid R_{t_2})$ represents an asymmetric relation between t and $t_1 \circ t_2$ (we say then that t_1 and t_2 *participate in a backoff of* t). This asymmetric relation can be used to impose a partial order (Sima'an and Buratto 2003) on the space of subtrees, just like Markov orders impose a partial order on the set $\cup_{k=1}^{n}\{\text{k-grams}\}$ (for some finite n). A subtree t_i participates in a lower (Markov) order than another subtree t_j iff one of two situations holds: (1) t_i participates in a backoff of t_j, or (2) there exists a subtree t_k that participates in a lower order than t_j such that t_i participates in a backoff of t_k. Note that, unlike an n-gram, a subtree can participate in multiple Markov orders and hence we will consider a unique copy of the subtree in each order in which it participates. Figure 5 exhibits examples.

Based on the backoff relation, a backoff estimator for DOP consists of the following steps:

(1) INITIALIZATION. Start out with the full training treebank trees themselves as the *current set of subtrees* $\mathcal{F}_{cu}$ and estimate the probabilities of every $t \in \mathcal{F}_{cu}$ by simple relative frequency,

(2) DISCOUNTING. Use e.g. leaving-one-out for discounting the probabilities of every $t \in \mathcal{F}_{cu}$, thereby reserving mass P_{rsv} for unseen events,

(3) BACKOFF. Define the set $\mathcal{F}_{bkf}$ to consist of all subtrees $t_1, t_2 \in \mathcal{F}$ such that $t_1 \circ t_2$ constitutes a backoff of some $t \in \mathcal{F}_{cu}$, and then employ the Katz backoff formula (Katz 1987) to distribute the reserved mass P_{rsv} to each $t_i \in \mathcal{F}_{bkf}$ (just like for n-gram backoff), and

[4] Note that a subtree can be generated through derivations involving more than two subtrees. However, note that all the derivations of a subtree can be simulated by a series of steps that each involve a pair of subtrees (break the subtree into two, break any of the two into two and so on). Hence, we adopt this simplifying assumption for the sake of arriving at a simpler model.

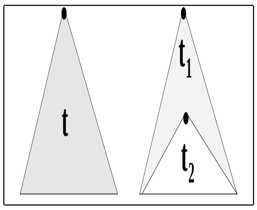

Figure 4: The backoff over subtrees. The subtree to the left (t) can be built up by substitution of t_2 at the suitable substitution-site in t_1 (right).

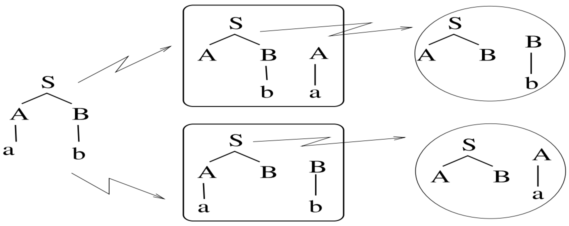

Figure 5: The Markov order imposed by the backoff relation between subtrees decreases from left to right. The left-most tree is backoffed to the boxed subtrees in the middle: two possible backoffs. Subsequently, the larger subtree in each of the two boxes is backed-off again to the encircled pairs of subtrees. Note the different copies of the same subtree in different Markov orders.

(4) RECURSION. Rename the set $\mathcal{F}_{bkf}$ as $\mathcal{F}_{cu}$ and repeat steps 2–4 until the set $\mathcal{F}_{bkf}$ is empty.

Backoff DOP estimation can be seen as a generalization of the known Katz backoff for word n-grams (Markov orders over word sequences). Backoff DOP is also consistent and non-overfitting as it smooths MLE over parses (initialization) using held-out estimation.

Backoff DOP is attractive because it provides a kind of *likelihood-weighted approach* to the empirically attractive shortest derivation DOP. By definition, the backoff mechanism gives preference to the shortest derivation as much as the training data allows, i.e. when the data is insufficient more mass is reserved for unseen parses, and longer derivations become more prominent. The more data there is, the shorter the derivations and the more the model starts to behave like a look up table. The less data, the more independence assumptions come into play (longer derivations) to achieve the desired coverage. In light of (Zavrel and Daelemans 1997), Backoff DOP is the probabilistic estimator that comes closest to the memory-based k-nearest backoff behavior.

These theoretical considerations are reinforced by an empirical result: A smoothing method similar to the one described here was employed in the BO-DOP1 model (Sima'an and Buratto 2003), obtaining good results on the OVIS corpus. (BO-DOP1 performed substantially better than DOP1 and than the model proposed in (Bonnema et al. 1999).) Since, however, BO-DOP1 uses DOP1 as the starting point for its back-off process, rather than the relative frequency of trees, we may surmise that it inherits DOP1's *inconsistency*. Therefore, unlike the backoff estimator described in this paper, BO-DOP1's theoretical status is dubious, and we may expect that it is not yet the optimal DOP estimator.

7 Conclusion

We proved that any unbiased parameter estimator (including Maximum-Likelihood) for a DOP model that employs all treebank subtrees will not generalize over the treebank. We also argued that smoothing techniques are reasonable for estimating these parameters. The latter result sheds a light on the tight relation between DOP and non-probabilistic memory-based models (k-nearest neighbor). We also presented two new consistent smoothing-based estimators and discussed their properties. Future work will concentrate on exploring the empirical aspects of the new estimators.

Acknowledgements: This work is supported by the Netherlands Organization for Scientific Research (NWO) project number 612.000.312 (`http://staff.science.uva.nl/˜simaan/LeStoGram.html`).

References

Baker, J. K.(1979), Trainable grammars for speech recognition, *Proc. of Spring Conference of the Acoustical Society of America*, pp. 547–550.

Black, E., Jelinek, F., Lafferty, J., Magerman, D., Mercer, R. and Roukos, S.(1993), Towards History-based Grammars: Using Richer Models for Probabilistic Parsing, *Proceedings of the 31st Annual Meeting of the ACL (ACL'93)*, Columbus, Ohio.

Bod, R.(1995), *Enriching Linguistics with Statistics: Performance models of Natural Language*, PhD dissertation. ILLC dissertation series 1995-14, University of Amsterdam.

Bod, R.(2000a), Combining semantic and syntactic structure for language modeling, *Proceedings ICSLP-2000*, Beijing, China.

Bod, R.(2000b), Parsing with the shortest derivation, *Proceedings of the 18th International Conference on Computational Linguistics (COLING'20 00)*, Saarbrcken, Germany.

Bonnema, R., Buying, P. and Scha, R.(1999), A new probability model for data oriented parsing, *in* P. Dekker (ed.), *Proceedings of the Twelfth Amsterdam Colloquium*, ILLC/Department of Philosophy, University of Amsterdam, Amsterdam, pp. 85–90.

Charniak, E.(1999), A maximum-entropy-inspired parser, *Report CS-99-12*, Providence, Rhode Island.

Collins, M.(1997), Three generative, lexicalized models for statistical parsing, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL) and the 8th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Madrid, Spain, pp. 16–23.

Collins, M. and Duffy, N.(2002), New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron, *Proc. of ACL 2002*.

Cover, T. M. and Thomas, J. A.(1991), *Elements of Information Theory*, Wiley, New York.

Daelemans, W.(1995), Memory-based lexical acquisition and processing, *in* P. Steffens (ed.), *Springer Lecture Notes in Artificial Intelligence*, Springer Lecture Notes in Artificial Intelligence no.898, Berlin: Springer-Verlag, pp. 85–98.

De Pauw, G.(2000a), Aspects of pattern-matching in dop, *Proceedings of the 18th International Conference of Computational Linguistics (COLING 2000)*, Saarbrücken.

De Pauw, G.(2000b), Probabilistische parsers - contextgevoeligheid en pattern-matching, APIL Report nr. 98, Universitaire Instelling Antwerpen.

Duda, R. O., Hart, P. E. and Stork, D. G.(2001), *Pattern Classification — 2nd ed*, Wiley–Interscience, New York.

Johnson, M.(1998), PCFG models of linguistic tree representations, *Computational Linguistics* **24**(4), 613–632.

Johnson, M.(2002), The DOP estimation method is biased and inconsistent, *Computational Linguistics* **28**(1), 71–76.

Katz, S.(1987), Estimation of probabilities from sparse data for the language model component of a speech recognizer, *IEEE Transactions on Acoustics, Speech and Signal Processing (ASSP)* **35(3)**, 400–401.

Ney, H., Martin, S. and Wessel, F.(1997), Statistical language modeling using leaving-one-out, *in* S. Young and G. Bloothooft (eds), *Corpus-based Methods in Language and Speech Processing*, Kluwer Academic, Dordrecht, pp. 174–207.

Scha, R.(1990), Taaltheorie en taaltechnologie; competence en performance, *in* Q. de Kort and G. Leerdam (eds), *Computertoepassingen in de Neerlandistiek, LVVN-jaarboek*, English translation as: Language Theory and Language Technology; Competence and Performance http://iaaa.nl/rs/LeerdamE.html, Almere, The Netherlands, pp. 7–22.

Sima'an, K. and Buratto, L.(2003), Backoff Parameter Estimation for the DOP Model, *in* H. B. N. Lavraĉ, D. Gamberger and L. Todorovski (eds), *Proceedings of the 14th European Conference on Machine Learning (ECML'03), Lecture Notes in Artificial Intelligence (LNAI 2837)*, Springer, Cavtat-Dubrovnik, Croatia, pp. 373–384.

Stanfill, C. and Waltz, D.(1986), Toward memory-based reasoning, *Communications of the ACM* **29**, 1213–1228.

Wu, C. F. J.(1983), On the convergence properties of the EM algorithm, *The Annals of Statistics* **11**(1), 95–103.

Zavrel, J. and Daelemans, W.(1997), Memory-based learning: Using similarity for smoothing, *in* P. R. Cohen and W. Wahlster (eds), *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Somerset, New Jersey, pp. 436–443.