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

A statistical estimator attempts to guess an unknown probability distribution by analyzing a sample from this distribution. One desirable property of an estimator is that its guess is increasingly likely to get arbitrarily close to the actual distribution as the sample size increases. This property 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

Probabilistic systems for syntactic disambiguation are usually based on non-redundant, linguistically inspired grammars. The simplest of these (such as Probabilistic Context-Free 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 2000, Collins and Duffy 2002).

Data-Oriented Parsing (DOP) (Scha 1990, Bod 1995) constitutes an alternative approach to this problem. DOP systems employ unconditional probabilities, but they achieve a fairly strong expressive power because they drop the assumption of non-redundancy. DOP treats *all* (arbitrarily large, and possibly overlapping) 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 classification techniques like *k-nearest-neighbor* and other forms of *Memory-Based Learning* (Stanfill and

Waltz 1986, Daelemans 1995), which analyze new input by matching it with a store of earlier instances. As emphasized in (Scha 1990), DOP should build a synthesis between the memory-based approach and the statistical approach: "(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 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 some 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. Bonnema et al. (1999) discuss Maximum Likelihood Estimation (MLE) and conclude that it cannot be used for DOP. They provide an example where the Maximum-Likelihood Estimator completely overfits the probability model to the training treebank, i.e. it assigns *zero* probability to *all* parses which do not occur in the treebank.

It should be noted that implemented systems rarely deploy the full DOP model. Limitations of space and time usually require that the subtree-set is restricted to subtrees which satisfy certain criteria (maximum depth, minimal/maximal number of terminal/non-terminal leaves, etc.). Thus, an implemented system may weaken or loose 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 the basis of *a priori* criteria.) The paper strengthens the existing negative results about DOP estimators. It proves that *any* unbiased estimator for the DOP model will yield a probability distribution that completely overfits the treebank. The constructive part of the paper then identifies different ways to define a non-overfitting and consistent DOP estimator. We briefly discuss an approach which uses held-out estimation, and then focus on an approach which combines MLE with smoothing by means of backoff. We point out that this approach strikes a good balance between the probabilistic and the memory-based facets of DOP.

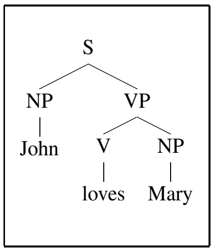

Figure 1: Treebank

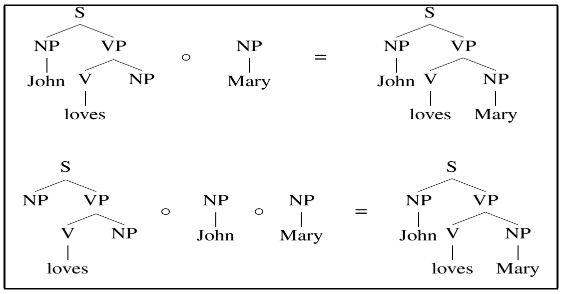

Figure 2: Two different derivations of the same parse

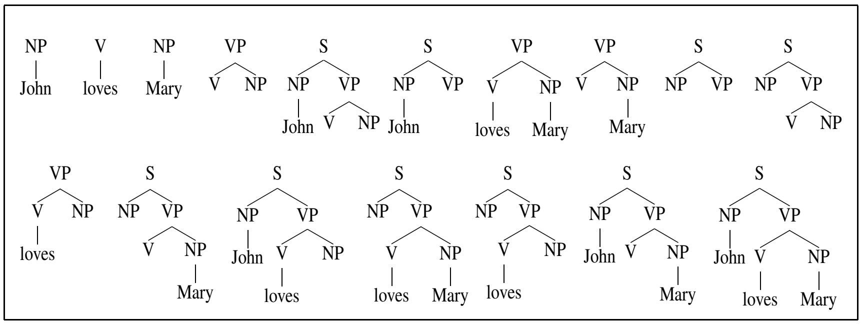

Figure 3: Fragments of the treebank in Figure 1

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 all 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

[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: $p \rightarrow c_1 \cdots c_n$.

(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. $(\ldots(S \circ t_1) \circ \ldots) \circ t_n)$. Note that multiple leftmost TSG-derivations may generate the same tree. (This constitutes an important conceptual 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 every 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 $p(d)$ of a derivation d is the product of its fragment probabilities:

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

where $f_t(d)$ denotes 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 $\pi(t) = \frac{count(t)}{\sum_{t'\,:\,R_{t'}=R_t} count(R_{t'})}$ 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 was 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, there is a choice between two or more probability distributions which might have generated the corpus. In fact, there is often an infinite number of different distributions which might have generated the corpus. (Statisticians bundle these into one single *probability model*.) 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, including 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 $f : \mathcal{X} \rightarrow \mathcal{N}$ 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 $|f| = \sum_{x \in \mathcal{X}} f(x)$. These definitions generalize the standard notion of the term *corpus* (used in Computational Linguistics) and of the term *sample* (used in Statistics). A "corpus" in the ordinary sense has of course a finite number of actual type-occurrences, and thus a finite value for $|f|$.

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* $f : \mathcal{X} \rightarrow \mathcal{N}$ given a probability distribution $p : \mathcal{X} \rightarrow [0, 1]$ is denoted as $L_p(f) = \prod_{x \in \mathcal{X}} p(x)^{f(x)}$. The disjunctive probability of a set of corpora F is denoted as $L_p(F) = \sum_{f \in F} L_p(f)$.

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

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

A probability model M is called *restricted* iff it is a proper subset of $\mathcal{M}(\mathcal{X})$.

Note that in earlier (less formal) discussions 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, Buying and Scha 1999) is a case in point.

Estimators and Their Properties: Bias and Consistency

Let $\mathcal{C}_n$ be the set $\mathcal{X}^n = \{<x_1, ..., x_n> |x_i \in \mathcal{X}\}$, i.e, $\mathcal{C}_n$ comprises all corpora of size n. If M is a probability model on $\mathcal{X}$, then each function $\mathrm{est}_n : \mathcal{C}_n \rightarrow M$ is called an *estimator* for M.

Given a model instance $p \in M$, the estimator's *expectation*[2] is calculated by

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

Note that the expectation is not a scalar but a distribution (because $\mathrm{est}_n(f)$ (for any $f \in \mathcal{C}_n$) is a distribution also). Hence, the sum $\sum_{f \in \mathcal{C}_n}$ is in fact a point-wise sum over distributions which also results in a distribution. We will not employ special notation for this situation to avoid complicating the formulae more than necessary.

The estimator is called *unbiased* for $p \in M$ iff

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

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

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

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 $\epsilon > 0$

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

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 generates its input-corpora. 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 M on f* is an instance $\hat{p} \in M$ which assigns a maximum probability to the corpus f, i.e., $L_{\hat{p}}\,(f) = \max_{p \in \mathcal{M}} L_p\,(f)$.

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 $\tilde{p}(x) = \frac{f(x)}{|f|}$.

The following theorems clarify the relation between the relative-frequency estimate $\tilde{p}$ and Maximum Likelihood Estimation.

Relative-Entropy Estimation

The *relative entropy* of probability distribution p with respect to probability distribution q is defined as $D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$. Interestingly, ML estimation is equivalent to finding the distribution p such that the relative entropy of $\tilde{p}$ with respect to p is minimized.

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

Theorem 1 *An instance $\hat{p}$ of a probability model M is a ML estimate for M on f, if and only if*

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

Proof: First, $D(\tilde{p}||p) = \sum_{x \in \mathcal{X}} \tilde{p}(x) \log \frac{\tilde{p}(x)}{p(x)} = \sum_{x \in \mathcal{X}} \tilde{p}(x) \log \tilde{p}(x) - \sum_{x \in \mathcal{X}} \tilde{p}(x) \log p(x)$. So minimizing $D(\tilde{p}||p)$ with respect to p is equivalent to minimizing $- \sum_{x \in \mathcal{X}} \tilde{p}(x) \log p(x)$. Finally, as this equals $-\frac{1}{|f|} \log L_p(f)$, minimizing $D(\tilde{p}||p)$ is equivalent to maximizing $L_p(f)$, **q.e.d.**

Theorem 2 (Information Inequality) *Let p and q be two probability distributions. Then $D(p \mid\mid q) \geq 0$; and $D(p \mid\mid q) = 0$ iff $p = q$.*

For a proof see, for instance, (Cover and Thomas 1991).

Theorem 3 *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: The proof is constructed by combining the theorems 1 and 2 mentioned above.

4 Estimation Theory and DOP

The probability model of a DOP grammar bundles 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.:

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

If we restrict the fragment-set $\mathcal{F}$ to fragments of depth 1, the DOP-grammar turns into a probabilistic CFG, and every tree has one unique derivation. The probability model of a CFG is thus given as:

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

(In formulas for CFG's we use r, for "rule", instead of t for "tree", to index the elements of the fragment set.)

A striking property of DOP's probability model is that it includes the probability distribution $\tilde{p}$ which assigns to every tree its relative frequency in the treebank TB (and which, therefore, assigns probability zero to all tree-types which do not occur in TB).

Theorem 4 *Let* 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* TB. *Then the relative-frequency estimate* $\tilde{p}$ *on* 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* derivable parse-trees $x \in \mathcal{X}$. In more detail, such a model instance $p \in \mathcal{M}_{\mathrm{CFG}}$ satisfies $p(x) > 0$ for all $x \in \mathcal{X}$ as it can be shown that $p(x) = \prod_r \pi(r)^{f_r(x)}$ with $\pi(r) > 0$ for all 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 that
>
> $$p(x_0) > 0 \quad \text{and} \quad p(x_0) = \prod_r \pi(r)^{f_r(x_0)} = \cdots \pi(r_0)^{f_{r_0}(x_0)} \cdots = 0$$
>
> Clearly, this is a contradiction, showing that the assumption is false.

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: As defined above, $\mathcal{M}_{\mathrm{DOP}}$ consists of all probability distributions p which (for some function π on $\mathcal{F}$) can be written as: $p(x) = \sum_{d \in \mathcal{D}(x)} \prod_{t \in \mathcal{F}} \pi(t)^{f_t(d)}$. We show that $\tilde{p}$ can indeed be specified in this fashion, if π is chosen as:

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

π assigns zero probability to all trees rooted with S, except the full parse trees from the corpus.Thus, every tree has only one derivation, and this derivation consists of one step. Therefore, $\sum_{d \in \mathcal{D}(x)} \prod_{t \in \mathcal{F}} \pi(t)^{f_t(d)} = \sum_{d=x} \prod_{t \in \{x\}} \pi(t)^1 = \tilde{p}$ **q.e.d.**

This theorem illustrates the expressive power of redundant Stochastic Tree-Substitution Grammars. Whereas DOP, in principle, can capture any frequency distribution over treebank trees, probabilistic CFGs cannot do this. This power has a serious disadvantage, however. As the following theorem shows, the Maximum Likelihood Estimator when applied to the DOP model will yield an overfitting probability distribution.

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

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

The proof is constructed by combining theorems 3 and 4.

To be able to formulate a corrollary of this theorem, we now introduce the notion of *complete overfitting*. An instance p of a probability model completely overfits the treebank iff it assigns probability zero 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* TB*,* $\mathcal{M}_{\mathrm{DOP}}$ *and* $\mathcal{M}_{\mathrm{CFG}}$ *be given as in Theorem 4. Then the Maximum Likelihood estimate for* $\mathcal{M}_{\mathrm{DOP}}$ *on* TB *completely overfits the treebank. This is not the case, however, for a Maximum Likelihood estimate for* $\mathcal{M}_{\mathrm{CFG}}$ *on* TB.

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

In the remainder of this paper, we 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* x_0 *(i.e.,* $L_p\,(f_0) > 0$ *and* $p(x_0) = 0$ *).*

Proof: Assume that est_n is unbiased for a model instance $p \in \mathcal{M}$ satisfying $L_p\,(f_0) > 0$ and $p(x_0) = 0$. 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 that

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

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 the DOP model.

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 $\mathrm{est}_n : \mathcal{C}_n \to \mathcal{M}_{\mathrm{DOP}}$ 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 aim at 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:

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

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., a DOP model that employs all fragments, we will not consider this option any further in this paper.

We therefore focus on 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 *Held-out estimation*, known from other areas of statistical NLP, and its close relative *leaving-one-out* (Ney, Martin and Wessel 1997) are smoothing methods that can be adapted to DOP. Held-out estimation avoids overfitting the training corpus by splitting it into two parts: the *extraction* part and the *held-out* part. The extraction part is used to obtain 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*. In the limit, when the size of the training corpus approaches infinity, under the condition that both the extraction and the held-out parts are allowed to approach infinity, the held-out estimator will approach the MLE itself, and is thus consistent. In fact, this is the reason why held-out estimation is considered a reasonable smoothing tool for language models using Markov orders over word sequences.

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 before, consistency considerations imply that smoothing must be applied to parse probabilities (p) rather than subtree probabilities (π). Yet, direct smoothing of p does not solve the overfitting problem, since it is not clear how it results in smoothed subtree probabilities π. This is the issue addressed in the current section.

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

Held-out DOP This estimator has four steps:

(1) The training treebank is split into a held-out part 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. Like in the case of Hidden Markov Models, MLE for DOP takes place under hidden derivations, and must be realized via an Inside-Outside algorithm; cf. also (Baker 1979, Bod 2000a)). However, there are no guarantees that Inside-Outside is consistent. In fact, it is not guaranteed to arrive at the MLE (Wu 1983).

In (Zollmann 2004), the maximum-likelihood problem in Step 3 is simplified by making assumptions on the multiple derivations per parse property of DOP. This way, Inside-Outside can be avoided, resulting in an efficient estimation of the parameters. A formal proof of consistency is provided together with experimental results showing improvements over the original DOP estimator.

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, $p(t_1 \circ t_2 \mid R_{t_1}) = p(t_1 \mid R_{t_1}) \cdot p(t_2 \mid R_{t_2})$. 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). As shown in (Sima'an and Buratto 2003), this asymmetric relation can be used to impose a partial order 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) <u>INITIALIZATION.</u> 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) <u>DISCOUNTING.</u> 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) <u>BACKOFF.</u> 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.

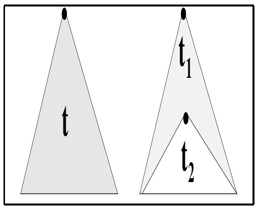

Figure 4: The backoff over subtrees. The subtree to the left (t) can be built up by substitution of t_2 at the suitable substitution-site in t_1 (right).

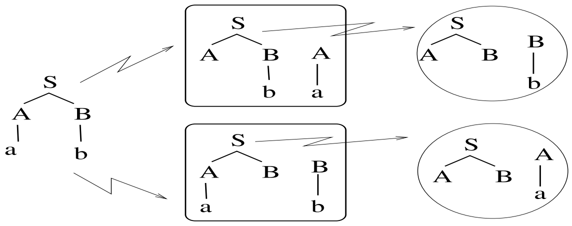

Figure 5: The Markov order imposed by the backoff relation between subtrees decreases from left to right. The left-most tree is backoffed to the boxed subtrees in the middle: two possible backoffs. Subsequently, the larger subtree in each of the two boxes is backed-off again to the encircled pairs of 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 wellknown Katz backoff for word n-grams (Markov chains over word sequences). Backoff DOP is also consistent and non-overfitting as it smooths MLE over parses 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 lookup 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 that resembles the one described here was presented 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).) The difference between the two methods is mainly in the Initialization step (1): BO-DOP1 uses DOP1 as the starting point for its backoff process, rather than the relative frequency of trees. Therefore, we may surmise that BO-DOP1 inherits DOP1's *inconsistency*. Therefore, unlike the backoff estimator described above, BO-DOP1's theoretical status is dubious, and we expect that it is not 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 the DOP parameters. Finally, we presented two new consistent smoothing-based estimators and discussed their properties.

Future work will concentrate on exploring the empirical aspects of the new estimators, and on alternative methods for restructuring and/or constraining the DOP model by data-driven algorithms.

Acknowledgements: This work is partially 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 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*, Vol. 3, Beijing, China, pp. 106–109.

Bod, R.(2000b), Parsing with the Shortest Derivation, *Proceedings COLING 2000*, Saarbrücken, Germany, pp. 69–75.

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, University of Amsterdam, pp. 85–90.

Charniak, E.(2000), A maximum entropy inspired parser, *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-00)*, Seattle, Washington, USA, pp. 132–139.

Collins, M.(1997), Three generative, lexicalized models for statistical parsing, *Proceedings ACL/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, *Proceedings ACL 2002*, pp. 263–270.

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 COLING 2000*, Saarbrücken, pp. 236–242.

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, 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, Almere, The Netherlands, pp. 7–22. English translation as: Language Theory and Language Technology; Competence and Performance. http://iaaa.nl/rs/LeerdamE.html.

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, *Proceedings ACL/EACL*, Madrid, pp. 436–443.

Zollmann, A.(2004), *A Consistent and Efficient Estimator for the Data-Oriented Parsing Model*, Master's thesis, ILLC, University of Amsterdam. Available at http://staff.science.uva.nl/∼azollman/publications.html.