Theoretical Evaluation of Estimation Methods for Data-Oriented Parsing

Willem Zuidema
Institute for Logic, Language and Computation
University of Amsterdam
Plantage Muidergracht 24, 1018 TV, Amsterdam, the Netherlands.
`jzuidema@science.uva.nl`

This is a corrected version of a paper published in Conference Companion/Proceedings 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06), pp. 183-186

Abstract

We analyze estimation methods for Data-Oriented Parsing, as well as the theoretical criteria used to evaluate them. We show that all current estimation methods are inconsistent in the "weight-distribution test", and argue that these results force us to rethink both the methods proposed and the criteria used.

1 Introduction

Stochastic Tree Substitution Grammars (henceforth, STSGs) are a simple generalization of Probabilistic Context Free Grammars, where the productive elements are not rewrite rules but elementary trees of arbitrary size. The increased flexibility allows STSGs to model a variety of syntactic and statistical dependencies, using relatively complex primitives but just a single and extremely simple global rule: substitution. STSGs can be seen as Stochastic Tree Adjoining Grammars without the adjunction operation.

STSGs are the underlying formalism of most instantiations of an approach to statistical parsing known as "Data-Oriented Parsing" (Scha, 1990; Bod, 1998). In this approach the subtrees of the trees in a tree bank are used as elementary trees of the grammar. In most DOP models the grammar used is an STSG with, in principle, all subtrees[1] of the trees in the tree bank as elementary trees. For disambiguation, the best parse tree is taken to be the most probable parse according to the weights of the grammar.

Several methods have been proposed to decide on the weights based on observed tree frequencies in a tree bank. The first such method is now known as "DOP1" (Bod, 1993). In combination with some heuristic constraints on the allowed subtrees, it has been remarkably successful on small tree banks. Despite this empirical success, (Johnson, 2002) argued that it is inadequate because it is *biased* and *inconsistent*. His criticism spearheaded a number of other methods, including (Bonnema et al., 1999; Bod, 2003; Sima'an and Buratto, 2003; Zollmann and Sima'an, 2005), and will be the starting point of our analysis. Here we show that alternative methods only partly remedy the problems with DOP1, leaving weight estimation as an important open problem.

2 Estimation Methods

The DOP model and STSG formalism are described in detail elsewhere, for instance in (Bod, 1998). The main difference with PCFGs is that multiple derivations, using elementary trees with a variety of sizes, can yield the same parse tree. The probability of a parse p is therefore given by: $P(p) = \sum_{d:\hat{d}=p} P(d)$, where $\hat{d}$ is the tree derived by derivation d, $P(d) = \prod_{t \in d} w(t)$ and $w(t)$ gives the weights of elementary trees t, which are combined in the derivation d (here treated as multiset).

2.1 DOP1

In Bod's original DOP implementation (Bod, 1993; Bod, 1998), henceforth DOP1, the weight of an elementary tree t is defined as its relative frequency (relative to other subtrees with the same root label) in the tree bank. That is, if $f_i = f(t_i)$ gives the frequency of subtree t_i in a corpus, and $r(t_i)$ is the root label of t_i, then the weight $w_i =$

[1] A subtree t' of a parse tree t is a tree such that every node i' in t' equals a node i in t, and i' either has no daughters or the same daughter nodes as i.

$w(t_i)$ of an elementary tree t_i is given by:

$$w_i = \frac{f_i}{\sum_{j:r(t_j)=r(t_i)}(f_j)}, \tag{1}$$

In his critique of this method, (Johnson, 2002) considers a situation where there is an STSG G (the *target grammar*) with a specific set of subtrees $(t_1 \dots t_N)$ and specific values of the weights $(w_1 \dots w_N)$. We can then evaluate an estimation procedure which produces a grammar G' (the *estimated grammar*), by looking at the difference between the weights of G and the expected weights of G'. This test for consistency is thus based on comparing the weight-distributions between target grammar and estimated grammar[2]. I will therefore refer to this test as the "weight-distribution test".

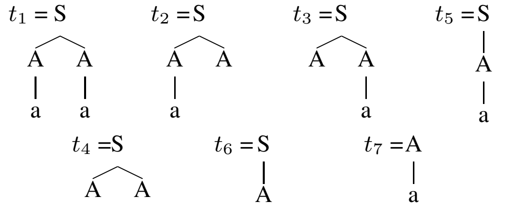

Figure 1: The example of (Johnson, 2002)

(Johnson, 2002) looks at an example grammar $G \in$ STSG with the subtrees as in figure 1. Johnson considers the case where the weights of all trees of the *target grammar* G are 0, except for w_7, which is necessarily 1, and w_4 and w_6 which are $w_4 = p$ and $w_6 = 1 - p$. He finds that the expected values of the weights w_4 and w_6 of the *estimated* grammar G' are:

$$\begin{aligned} \mathbf{E}[w'_4] &= \frac{p}{2+2p}, \quad (2) \\ \mathbf{E}[w'_6] &= \frac{1-p}{2+2p}, \quad (3) \end{aligned}$$

which are not equal to their target values for all values of p where $0 < p < 1$. This analysis thus already shows that DOP1 is unable to recover the true weights of the given STSG, and hence the inconsistency of the estimator with respect to the class of STSGs.

However, the analysis so far is not sufficient to distinguish DOP1 from alternative methods, because no possible estimation procedure can recover the true weights in the case considered. In the example there are only two complete trees that can be observed in the training data, corresponding to the trees t_1 and t_5. It is easy to see that when generating examples with the grammar in figure 1, the relative frequencies[3] $f_1 \dots f_4$ of the *subtrees* $t_1 \dots t_4$ must all be the same, and equal to the frequency of the complete tree t_1 which can be composed in the following ways from the subtrees in the original grammar:

$$t_1 = t_2 \circ t_7 = t_3 \circ t_7 = t_4 \circ t_7 \circ t_7. \tag{4}$$

It follows that the expected frequencies of each of these subtrees are:

$$\begin{aligned} \mathbf{E}[f_1] &= \mathbf{E}[f_2] = \mathbf{E}[f_3] = \mathbf{E}[f_4] \quad (5) \\ &= w_1 + w_2 w_7 + w_3 w_7 + w_4 w_7 w_7 \end{aligned}$$

Similarly, the other frequencies are given by:

$$\begin{aligned} \mathbf{E}[f_5] = \mathbf{E}[f_6] &= w_5 + w_6 w_7 \quad (6) \\ \mathbf{E}[f_7] &= 2\,(w_1 + w_2 w_7 + w_3 w_7 \\ &\quad + w_4 w_7 w_7) + w_5 + w_6 w_7 \\ &= 2\mathbf{E}[f_1] + \mathbf{E}[f_5]. \quad (7) \end{aligned}$$

From these equations it is immediately clear that, regardless of the amount of training data, the problem is simply *underdetermined*. The values of 6 weights $w_1 \dots w_6$ ($w_7 = 1$) given only 2 frequencies f_1 and f_5 (and the constraint that $\sum_{i=1}^{6}(f_i) = 1$) are not uniquely defined, and no possible estimation method will be able to reliably recover the true weights.

The relevant test, as argued by Johnson[4], is whether for all possible STSGs and in the limit of infinite data, the *expected* relative frequencies of trees given the estimated grammar, equal the *observed* relative frequencies. I will refer to this test as the "frequency-distribution test". As it turns out, the DOP1 method also fails this more lenient test. The easiest way to show this, using again figure 1, is as follows. The weights $w'_1 \dots w'_7$ of grammar G' will – by definition – be set to the relative frequencies of the corresponding subtrees:

$$w'_i = \begin{cases} \frac{f_i}{\sum_{j=1}^{6} f_j} & \textit{for } i = 1 \dots 6 \\ 1 & \textit{for } i = 7. \end{cases} \tag{8}$$

[2] More precisely, it is based on evaluating the estimator's behavior for any weight-distribution possible in the STSG model. (Prescher et al., 2003) give a more formal treatment of bias and consistency in the context of DOP.

[3] Throughout this paper I take frequencies f_i to be relative to the size of the corpus.

[4] In the published version of this paper I claim that (Johnson, 2002) only considered what I call the "weight-distribution test". This is incorrect; in fact, he explicitly states: "it is more natural to define bias and loss in terms of the probability distributions that the parameters specify, rather than in terms of the parameters themselves." (p.73) and does indeed show bias and inconsistency using these notions.

The grammar G' will thus produce the complete trees t_1 and t_5 with expected frequencies:

$$\begin{aligned}\mathbf{E}[f'_1] &= w'_1 + w'_2 w'_7 + w'_3 w'_7 + w'_4 w'_7 w'_7 \\ &= 4\frac{f_1}{\sum_{j=1}^{6} f_j} \qquad (9)\end{aligned}$$

$$\mathbf{E}[f'_5] = w'_5 + w'_6 w'_7 = 2\frac{f_5}{\sum_{j=1}^{6} f_j}. \qquad (10)$$

Now consider the two possible complete trees t_1 and t_5, and the fraction of their frequencies f_1/f_5. In the estimated grammar G' this fraction becomes:

$$\frac{\mathbf{E}[f'_1]}{\mathbf{E}[f'_5]} = \frac{4\frac{f_1}{\sum_{j=1}^{6} f_j}}{2\frac{f_5}{\sum_{j=1}^{6} f_j}} = \frac{2f_1}{f_5}. \qquad (11)$$

That is, in the limit of infinite data, the estimation procedure not only –understandably– fails to find the target grammar amongst the many grammars that could have produced the observed frequencies, it in fact chooses a grammar that could never have produced these observed frequencies at all (Johnson, 2002). This example shows the DOP1 method is biased and inconsistent for the STSG class in the frequency-distribution test[5].

2.2 Correction-factor approaches

Based on similar observations, (Bonnema et al., 1999; Bod, 2003) propose alternative estimation methods, which involve a correction factor to move probability mass from larger subtrees to smaller ones. For instance, Bonnema et al. replace equation (1) with:

$$w_i = 2^{-N(t_i)} \frac{f_i}{\sum_{j:r(t_j)=r(t_i)}(f_j)}, \qquad (12)$$

where $N(t_i)$ gives the number of internal nodes in t_i (such that $2^{-N(t_i)}$ is inversely proportional to the number of possible derivations of t_i). Similarly, (Bod, 2003) changes the way frequencies f_i are counted, with a similar effect. This approach solves the specific problem shown in equation (11). However, the following example shows that the correction-factor approaches cannot solve the more general problem.

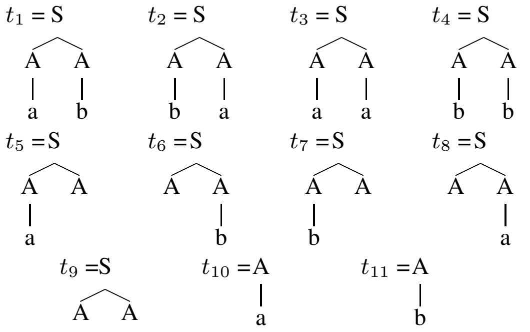

Figure 2: Counter-example to the correction-factor approaches

Consider the STSG in figure 2. The expected frequencies $f_1 \ldots f_4$ are here given by:

$$\begin{aligned}\mathbf{E}[f_1] &= w_1 + w_5 w_{11} + w_6 w_{10} + w_9 w_{10} w_{11} \\ \mathbf{E}[f_2] &= w_2 + w_7 w_{10} + w_8 w_{11} + w_9 w_{11} w_{10} \\ \mathbf{E}[f_3] &= w_3 + w_5 w_{10} + w_8 w_{10} + w_9 w_{10} w_{10} \\ \mathbf{E}[f_4] &= w_4 + w_6 w_{11} + w_7 w_{11} + w_9 w_{11} w_{11}\end{aligned} \qquad (13)$$

Frequencies $f_5 \ldots f_{11}$ are again simple combinations of the frequencies $f_1 \ldots f_4$. Observations of these frequencies therefore do not add any extra information, and the problem of finding the weights of the target grammar is in general again underdetermined. But consider the situation where $f_3 = f_4 = 0$ and $f_1 > 0$ and $f_2 > 0$. This constrains the possible solutions enormously. If we solve the following equations for $w_3 \ldots w_{11}$ with the constraint that probabilities with the same root label add up to 1: (i.e. $\sum_{i=1}^{9}(w_i) = 1$, $w_{10} + w_{11} = 1$):

$$\begin{aligned}w_3 + w_5 w_{10} + w_8 w_{10} + w_9 w_{10} w_{10} &= 0 \\ w_4 + w_6 w_{11} + w_7 w_{11} + w_9 w_{11} w_{11} &= 0,\end{aligned}$$

we find, in addition to the obvious $w_3 = w_4 = 0$, the following solutions: $w_{10} = w_6 = w_7 = w_9 = 0 \vee w_{11} = w_5 = w_8 = w_9 = 0 \vee w_5 = w_6 = w_7 = w_8 = w_9 = 0$. That is, if we observe no occurrences of trees t_3 and t_4 in the training sample, we know that at least one subtree in each derivation of these strings must have weight zero. However, any estimation method that uses the (relative) frequencies of subtrees and a (nonzero) correction factor that is based on the size of the subtrees, will give non-zero probabilities to all weights $w_5 \ldots w_{11}$ if $f_1 > 0$ and $f_2 > 0$, as we assumed. In other words, these weight estimation methods for STSGs are also *biased* and *inconsistent* in the frequency-distribution test.

[5] Note that there are settings of the weights $w_1 \ldots w_7$ that generate a frequency-distribution that could also have been generated with a PCFG. The example given applies to such distribution as well, and therefore also shows the inconsistency of the DOP1 method for PCFG distributions.

2.3 Shortest derivation estimators

Because the STSG formalism allows elementary trees of arbitrary size, every parse tree in a tree bank could in principle be incorporated in an STSG grammar. That is, we can define a trivial estimator with the following weights:

$$w_i = \begin{cases} f_i & \text{if } t_i \text{ is an observed parse tree} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

Such an estimator is not particularly interesting, because it does not generalize beyond the training data. It is a point to note, however, that this estimator is unbiased and consistent in the frequency-distribution test. (Prescher et al., 2003) prove that any unbiased estimator that uses the "all subtrees" representation has the same property, and conclude that lack of bias is not a desired property.

(Zollmann and Sima'an, 2005) propose an estimator based on held-out estimation. The training corpus is split into an estimation corpus EC and a held out corpus HC. The HC corpus is parsed by searching for the shortest derivation of each sentence, using only fragments from EC. The elementary trees of the estimated STSG are assigned weights according to their usage frequencies $u_1, \ldots, u_N$ in these shortest derivations:

$$w_i = \frac{u_i}{\sum_{j:r(t_j)=r(t_i)} u_j}. \tag{15}$$

This approach solves the problem with bias described above, while still allowing for consistency, as Zollmann & Sima'an prove. However, their proof only concerns consistency in the frequency-distribution test. As the corpus EC grows to be infinitely large, every parse tree in HC will also be found in EC, and the shortest derivation will therefore in the limit only involve a single elementary tree: the parse tree itself. Target STSGs with non-zero weights on smaller elementary trees will thus not be identified correctly, even with an infinitely large training set. In other words, the Zollmann & Sima'an method, and other methods that converge to the "complete parse tree" solution such as LS-DOP (Bod, 2003) and BackOff-DOP (Sima'an and Buratto, 2003), are inconsistent in the weight-distribution test.

3 Discussion & Conclusions

A desideratum for parameter estimation methods is that they converge to the correct parameters with infinitely many data – that is, we like an estimator to be consistent. The STSG formalism, however, allows for many different derivations of the same parse tree, and for many different grammars to generate the same frequency-distribution. Consistency in the weight-distribution test is therefore too stringent a criterion. We have shown that DOP1 and methods based on correction factors also fail the weaker frequency-distribution test.

However, the only current estimation methods that are consistent in the frequency-distribution test, have the linguistically undesirable property of converging to a distribution with all probability mass in complete parse trees. Although these method fail the weight-distribution test for the whole class of STSGs, we argued earlier that this test is not the appropriate test either. Both estimation methods for STSGs and the criteria for evaluating them, thus require thorough rethinking. In forthcoming work we therefore study yet another estimator, and the linguistically motivated evaluation criterion of convergence to a maximally general STSG consistent with the training data[6].

References

Rens Bod. 1993. Using an annotated corpus as a stochastic grammar. In *Proceedings EACL'93*, pp. 37–44.

Rens Bod. 1998. *Beyond Grammar: An experience-based theory of language*. CSLI, Stanford, CA.

Rens Bod. 2003. An efficient implementation of a new DOP model. In *Proceedings EACL'03*.

Remko Bonnema, Paul Buying, and Remko Scha. 1999. A new probability model for data oriented parsing. In Paul Dekker, editor, *Proceedings of the Twelfth Amsterdam Colloquium*. ILLC, University of Amsterdam.

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

D. Prescher, R. Scha, K. Sima'an, and A. Zollmann. 2003. On the statistical consistency of DOP estimators. In *Proceedings CLIN'03*, Antwerp, Belgium.

Remko Scha. 1990. Taaltheorie en taaltechnologie; competence en performance. In R. de Kort and G.L.J. Leerdam, eds, *Computertoepassingen in de Neerlandistiek*, pages 7–22. LVVN, Almere. http://iaaa.nl/rs/LeerdamE.html.

Khalil Sima'an and Luciano Buratto (2003). Backoff parameter estimation for the DOP model. In *Proceedings ECML'03*, pp. 373–384. Berlin: Springer Verlag.

Andreas Zollmann and Khalil Sima'an. 2005. A consistent and efficient estimator for data-oriented parsing. *Journal of Automata, Languages and Combinatorics*. In press.

[6] The author is funded by NWO, project nr. 612.066.405, and would like to thank the Mark Johnson, anonymous reviewers and several colleagues for comments.