Parameter Estimation and the Structure of the DOP Model

Detlef Prescher, Remko Scha and Khalil Sima'an
Institute for Logic, Language and Computation
University of Amsterdam

December 19, 2003

NWO Project LeStoGram, Oct 2003 - Sept 2006

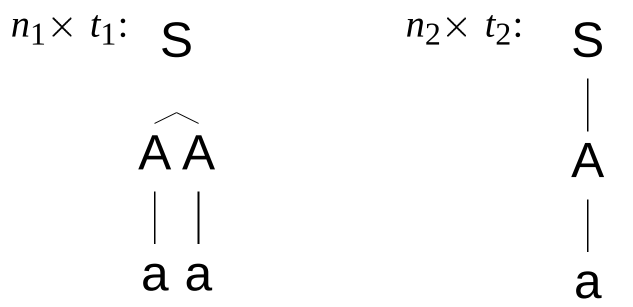
In this talk, we focus on Data-Oriented Parsing (DOP) and present first research results of the NWO project **Learning Stochastic Tree Grammars from Treebanks**. People involved: Prescher, Scha, and Sima'an (**PI**).

`http://staff.science.uva.nl/~simaan/LeStoGram.html`

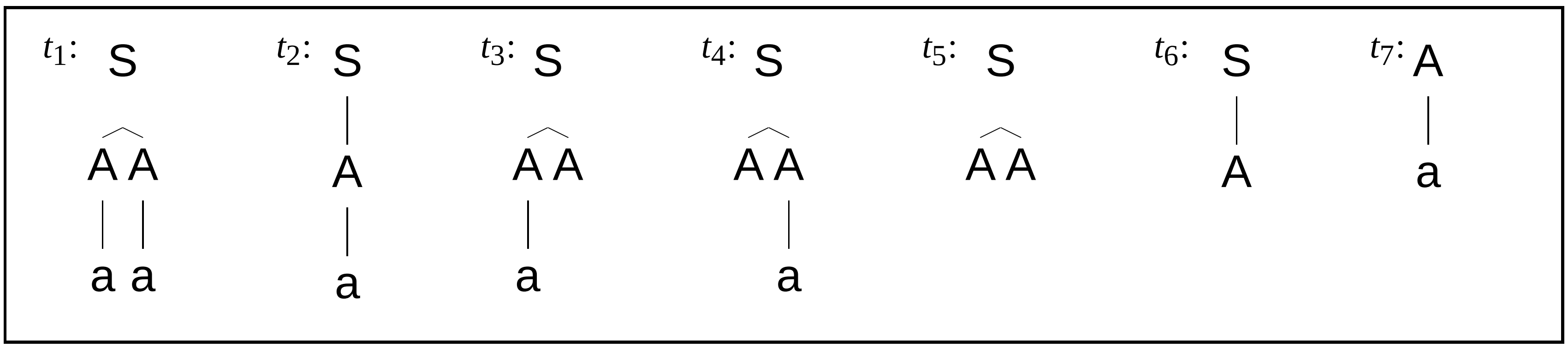
- Stochastic Tree-Grammars (STGs) generalize PCFGs: Extended contextual evidence is expressed in productions that are partial parse-trees. As a result, **STGs yield probability distributions of parse trees that can not be modeled by the underlying PCFGs**.
- Parsers based on STGs currently achieve **state-of-the-art performance**
- In the last few years, we observed an **increasing interest in investigating probabilistic versions of various STG formalisms** like Tree-Adjoining Grammars, Tree-Substitution Grammars, and Tree-Insertion Grammars.

The Symbolic Backbone of DOP

Treebank (Example by Johnson, 2002):



Tree fragments:



Tree derivations (Trees with hidden breakpoints):

$$D(t_1) = \{t_1,\ t_3 \circ t_7,\ t_4 \circ t_7,\ t_5 \circ t_7 \circ t_7\} \quad \text{and} \quad D(t_2) = \{t_2,\ t_6 \circ t_7\}$$

The Probability Model of DOP

Fragment probability: Each fragment is assigned a real number $\pi(t)$ such that π *induces probability distributions on the fragments having the same root label*

$$\pi(t) \geq 0 \quad \text{and} \quad \sum_{root(t)=A} \pi(t) = 1 \qquad (\text{for all } t \text{ and } A)$$

Derivation probability: The product of the derivation's fragment probabilities

$$p(d) = \prod_{t \in T(d)} \pi(t)$$

Tree probability: The sum of the tree's derivation probabilities

$$p(t) = \sum_{d \in D(t)} p(d)$$

Estimating an Instance of DOP's Probability Model

DOP's probability model: *Subset of the unrestricted probability model* $M(T)$ *on the set* T *of parse trees, parameterized by an assignment function* π

$$M_{\mathrm{DOP}} = \left\{ p \in M(T) \;\middle|\; p(t) = \sum_{d \in D(t)} \prod_{t \in T(d)} \pi(t) \text{ and } \pi\colon T \to [0,1] \right\}$$

DOP comes with a genuine idea for estimating the fragment probabilities: Data-driven calculation of $\pi(t)$ using a given treebank!

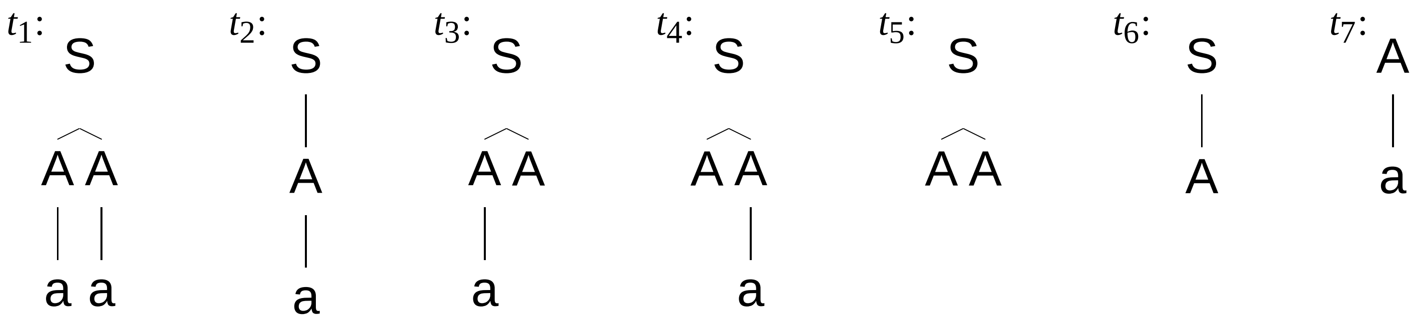
Challenge: Instances p can often be generated by multiple assignments π.

First instantiation by Bod (1992): DOP1 requests $\pi(t)$ being the *relative frequency of* t *in the sub-corpus of those fragments having the same root label*.

Still unsolved fundamental problem: Which instance of DOP's probability model is the best one? Which assignment $\pi\colon T \to [0,1]$ is the best choice?

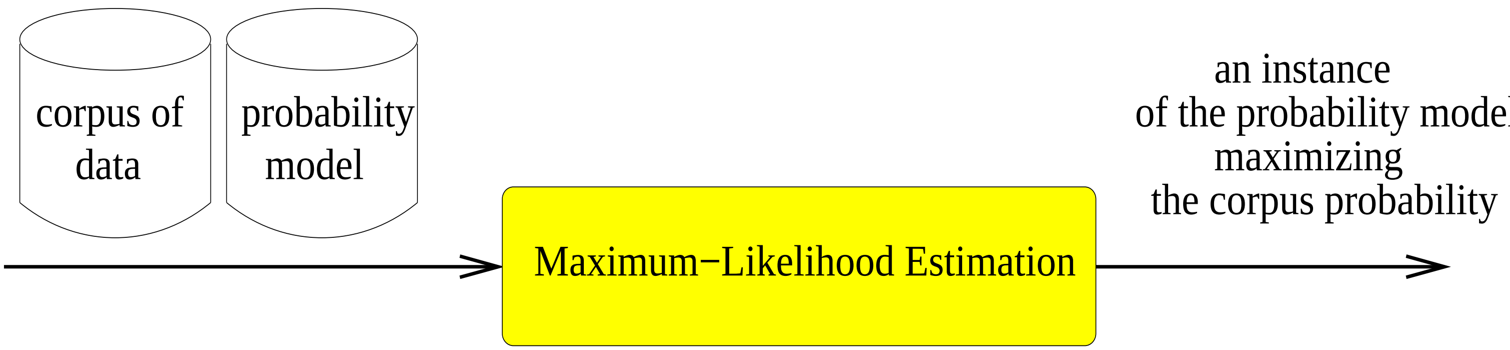
Johnson (2002): “DOP1 is biased and inconsistent”

Fragments



	t_1	t_2	t_3	t_4	t_5	t_6	t_7
$f_{\mathtt{DOP1}}$	n_1	$\boxed{n_2}$	n_1	n_1	n_1	$\boxed{n_2}$	n_2
$\pi_{\mathtt{DOP1}}$	$\frac{n_1}{4n_1+2n_2}$	$\boxed{\frac{n_2}{4n_1+2n_2}}$	$\frac{n_1}{4n_1+2n_2}$	$\frac{n_1}{4n_1+2n_2}$	$\frac{n_1}{4n_1+2n_2}$	$\boxed{\frac{n_2}{4n_1+2n_2}}$	1
$p_{\mathtt{DOP1}}$ ☹	$\frac{4n_1}{4n_1+2n_2}$	$\frac{2n_2}{4n_1+2n_2}$					
$p_{\mathtt{PCFG}}$ ☺	$\frac{n_1}{n_1+n_2}$	$\frac{n_2}{n_1+n_2}$			$\left(\frac{n_1}{n_1+n_2}\right)$	$\left(\frac{n_2}{n_1+n_2}\right)$	(1)

Maximum-Likelihood Estimation (General Theory)



Definition (Fisher, 1912): Maximum-likelihood estimates $\hat{p}$ of a model M on a corpus f satisfy

$$\hat{p} = \mathbf{arg}\max_{p \in M} \prod_{t \in T} p(t)^{f(t)}$$

MLE is the most widely used estimation method:

- Although there are theoretical problems concerning the existence, uniqueness and computability of ML estimates...
- ...for most practical problems, **MLE yields consistent estimates**.

Maximum-Likelihood Estimation for DOP

- In general, MLE is equivalent to minimizing the relative entropy $D_{\mathsf{KL}}(.||.)$ with respect to the relative frequency $\tilde{p} := |f|^{-1} \cdot f$ of the types in the corpus

$$\hat{p} = \mathbf{arg}\min_{p \in M} D_{\mathsf{KL}}(p||\tilde{p})$$

- For most treebanks (condition: no subtree of a tree is itself a treebank tree), **the relative-frequency estimate is an instance of DOP's probability model**

$$\tilde{p} \in M_{\mathsf{DOP}}$$

- **Theorem** The relative-frequency estimate $\tilde{p}$ is the unique ML estimate of M_{DOP} on each treebank f that satisfies the above weak condition.

The maximum-likelihood estimate of M_{DOP} over-fits the treebank. The consistency problem is solved at the cost of allocating all trees outside the treebank a zero probability...

Estimating DOP Instances: A Balancing Act

Starting with the comment by Johnson (2002) “DOP1 is biased and inconsistent”, we are arriving now at...

...the full problem: Sound estimation of an instance of DOP’s probability model seems to be a balancing act. When exploiting a given treebank, we have to look for those estimates

- that are **unbiased and consistent** (DOP1 ☹, MLE for M_{DOP} ☺)

and at the same time

- that **do not over-fit the treebank** (DOP1 ☺, MLE for M_{DOP} ☹)

Estimating DOP Instances: LeStoGram's Dual Vision

☺ **Vision "Empirical Research"**: Stick with DOP's probability model! Incorporate instead into MLE or into other sound estimation methods

- **smoothing techniques**: The aim is to learn accurately those DOP parameters that are insufficiently represented in the sparse treebank.
- **pruning techniques**: The aim is to drastically reduce DOP's huge parameter space by applying statistically sound parameter-selection procedures.

☺ **Vision "Fundamental Research"**: Stick with MLE! Augment instead DOP's standard probability-model M_{DOP} with constraints C

$$M_{\mathsf{DOP}}(C) = \{\ p \in M_{\mathsf{DOP}} \mid p \text{ satisfies the constraints } C\ \}$$

The aim is to find model constraints C such that *standard MLE of* $M_{DOP}(C)$ *results in consistent estimates that do not over-fit*. The ambition of this research effort is to **gain a better theoretical insight into DOP estimation**.

LeStoGram: Past research...

...related to our empirical-research vision

- The DOP fragments constitute a structured space of correlated events, which can be exploited via discounting and back-off smoothing.
  Reference: K. Sima'an and L. Buratto (2003). Back-off Parameter Estimation for the DOP Model. In *Proc. of the European Conf. on Machine Learning*.
- Split the given treebank: Read off the fragments from one part, but estimate the fragment probabilities using the other part.
  Reference: A. Zollmann (2003). *A Consistent Estimation Method for Data-Oriented Parsing*. Master thesis, ILLC, University of Amsterdam.

...related to our fundamental-research vision

- Initial experiments with constraints like: $\pi(t) > 0$ and $t > s \implies \pi(s) > 0$

LeStoGram: Future Research

- So far, there is no parameter-estimation method for DOP which is **known** to be unbiased and consistent, and which does not over-fit the treebank. However

  > “...the existing parameter-estimation methods for DOP are **known** to be inconsistent and biased towards either smaller or larger subtrees.”

  is an incorrect statement (Thanks to Rens Bod, personal communication).

- **Empirical research**: The success or failure of the (as we think: promising) experiments of Andreas Zollmann will be our starting point.

- **Fundamental research**: We are expecting great theoretical results by investigating constrained DOP models in the standard MLE framework.