Treebank Grammars and Other Infinite Parameter Models

Detlef Prescher
University of Amsterdam
prescher@science.uva.nl

December 2005

Overview

- Motivation
- Estimation / Estimators
  - Definition, Properties, Examples
- Grammar Estimation
  - for PCFGs (Maximum-Likelihood Estimation)
  - for DOP (first estimator: DOP1)
- More on DOP Estimation
  - Current Solutions, Current Problems
- Wrap-Up and Conclusion

Motivation

Johnson (2002). DOP1 is biased and inconsistent.

Learning Stochastic Tree Grammars from Treebanks (ongoing NWO project):
An unbiased DOP estimator over-fits the given tree-bank, i.e., bias is a desirable property for DOP. Moreover, there are consistent DOP estimators, which do not over-fit the given tree-bank.

People involved: Detlef Prescher, Remko Scha, Khalil Sima'an, and Andreas Zollmann.

This talk: Review of some of our discussions / Presentation of some research paths we did not follow yet...

Estimators

Definition 1. *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*

$$|f| = \sum_{x \in \mathcal{X}} f(x)$$

(Here, $\mathcal{N}$ is the set of all natural numbers, including 0.)

Estimators (continued)

Definition 2. *Let $\mathcal{X}$ be a countable set of types. A real-valued function $p\colon \mathcal{X} \to \mathcal{R}$ is called a* ***probability distribution on*** $\mathcal{X}$*, if p has two properties: First, p´s values are non-negative numbers*

$$p(x) \geq 0 \quad \text{for all } x \in \mathcal{X}$$

and second, p´s values sum to 1

$$\sum_{x \in \mathcal{X}} p(x) = 1$$

Estimators (continued)

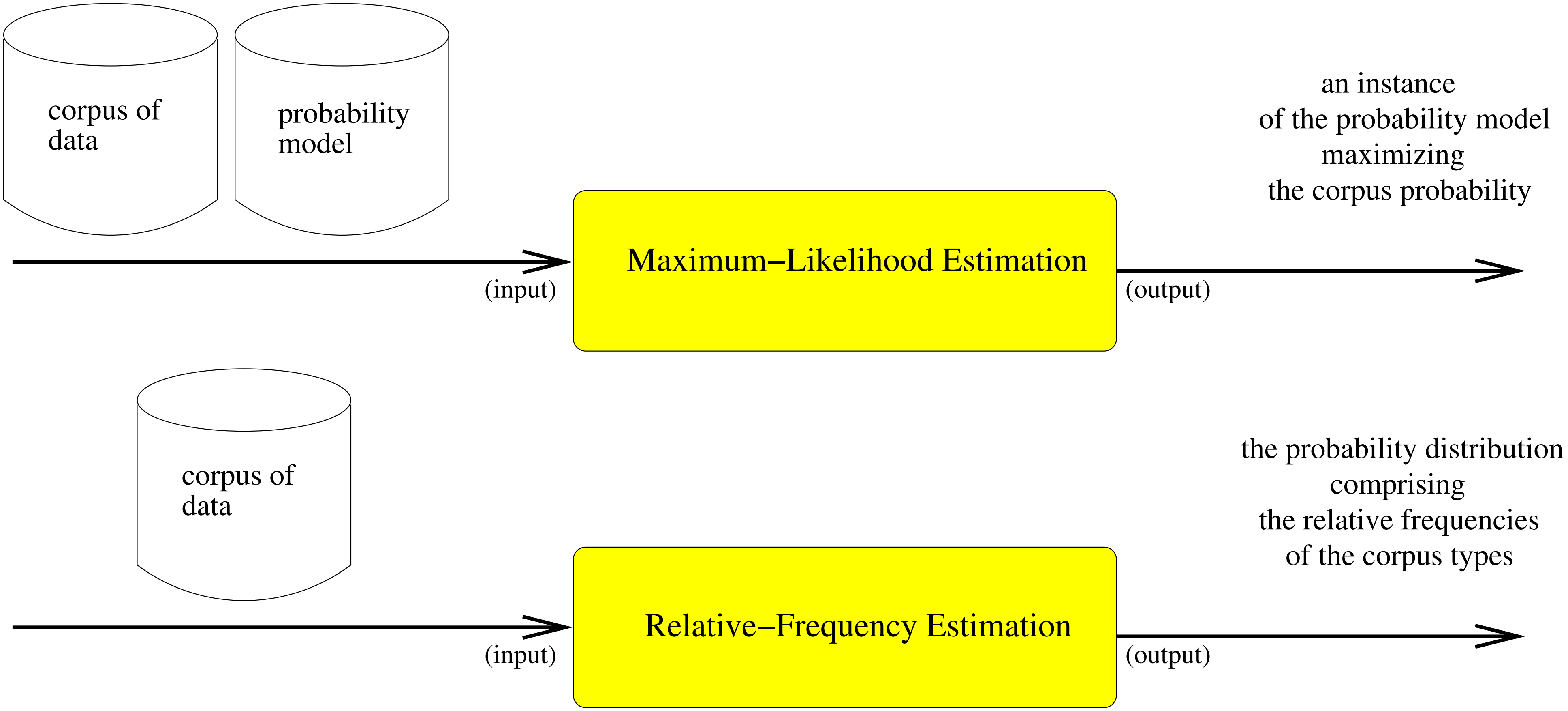
Definition 3. *A non-empty set $\mathcal{M}$ of probability distributions on a set $\mathcal{X}$ of types is called a* ***probability model on*** $\mathcal{X}$*. The elements of $\mathcal{M}$ are called* ***instances of the model*** $\mathcal{M}$*. The* ***unrestricted probability model*** *is the set $\mathcal{M}(\mathcal{X})$ of all probability distributions on the set of types*

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

A probability model $\mathcal{M}$ is called ***restricted*** *in all other cases*

$$\mathcal{M} \subseteq \mathcal{M}(\mathcal{X}) \quad \text{and} \quad \mathcal{M} \neq \mathcal{M}(\mathcal{X})$$

Estimators (continued)



corpus of data

probability model

(input)

Maximum–Likelihood Estimation

(output)

an instance
of the probability model
maximizing
the corpus probability

corpus of data

(input)

Relative–Frequency Estimation

(output)

the probability distribution
comprising
the relative frequencies
of the corpus types

Maximum-Likelihood Estimation and Relative-Frequency Estimation

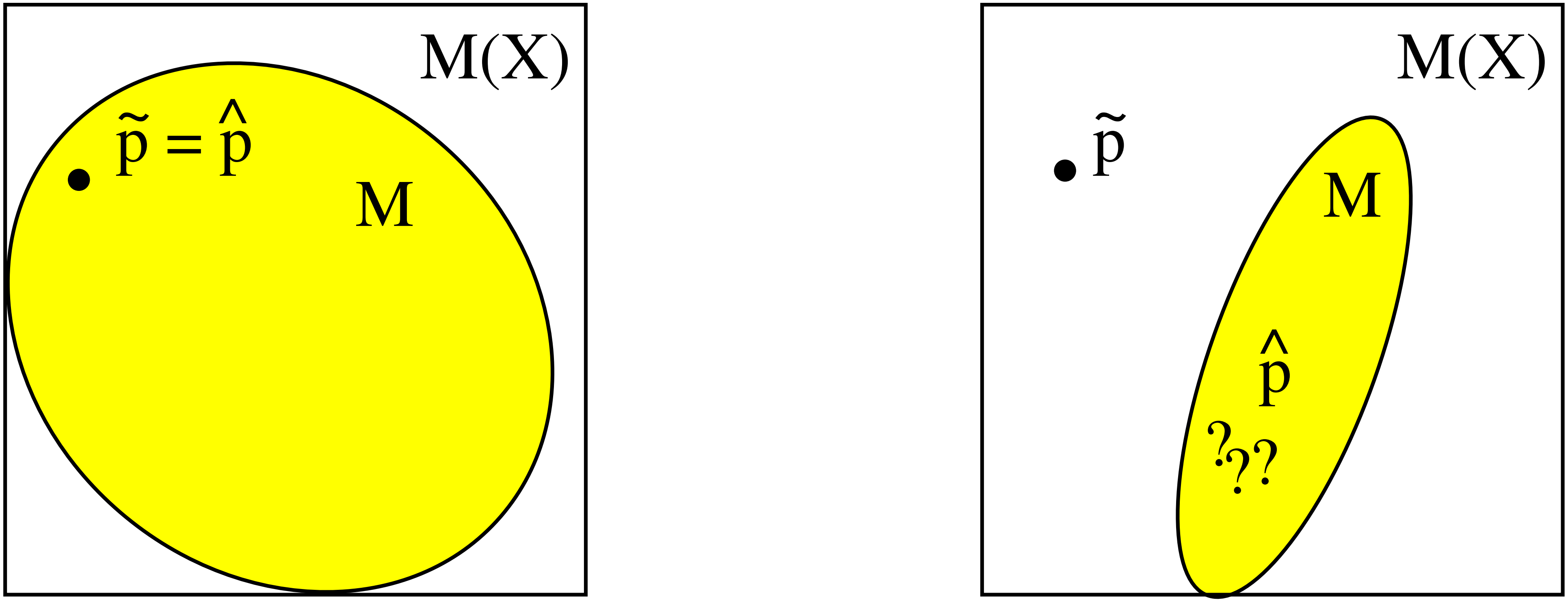
Estimators (continued)

Maximum-likelihood estimation (Fisher 1912), typically yields an excellent estimate if the given corpus is large:

- maximum-likelihood estimators fulfill the so-called **invariance principle**,
- under certain conditions which are typically satisfied in practical problems, they are **consistent estimators**,
- unlike the relative-frequency estimator, maximum-likelihood estimators typically **do not over-fit the given corpus** in practice.

Maximum-Likelihood Estimation is probably the most widely used estimation method

Estimators (continued)



In practice, Maximum-Likelihood Estimation and Relative-Frequency Estimation typically yield different estimates.

Estimators (continued)

Definition 4. *Let $\mathcal{X}$ be a countable set, let $\mathcal{C}_n$ be the set of all corpora $f\colon \mathcal{X} \to \mathcal{N}$ with size $|f| = n$, and let $\mathcal{M}$ be a probability model on $\mathcal{X}$. Then each function*

$$\mathrm{est}_n : \mathcal{C}_n \to \mathcal{M},\ f \mapsto p$$

is called an ***estimator*** *for $\mathcal{M}$.*

Estimators (continued)

Definition 5. *An estimator is called* ***unbiased*** *for a model instance $p \in \mathcal{M}$ iff*

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

Here, $E_p(\mathrm{est}_n) = \sum_{f \in \mathcal{C}_n} L_p(f) \cdot \mathrm{est}_n(f)$ is the estimator's expectation, calculated with the corpus probabilities $L_p(f)$. Note that these probabilities form a distribution on $\mathcal{C}_n$.

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

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

Mathematical properties of estimators

Grammar Estimation

Given / Setting:

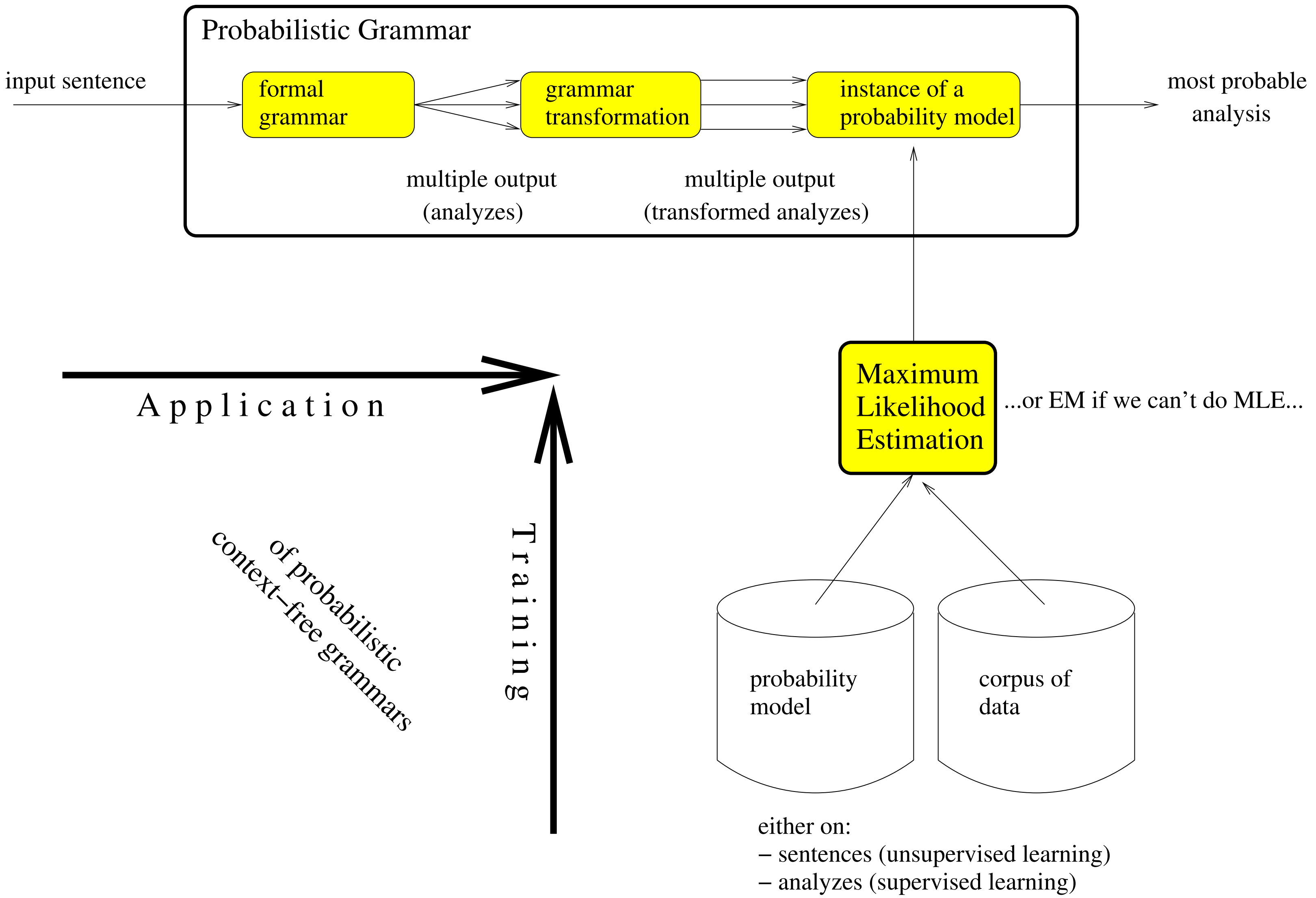
- a symbolic grammar
- a corpus of sentences
  (with or without grammatical analyzes)

Tasks:

- create a good probability model of the grammar
- select a good instance of the grammar's probability model
  (to be used as a language model or for disambiguation)

Probabilistic Modeling for Natural Language Grammars

Grammar Estimation



Probabilistic Grammar

input sentence

formal grammar

grammar transformation

instance of a probability model

most probable analysis

multiple output (analyzes)

multiple output (transformed analyzes)

Application

Training

of probabilistic context–free grammars

Maximum Likelihood Estimation

...or EM if we can't do MLE...

probability model

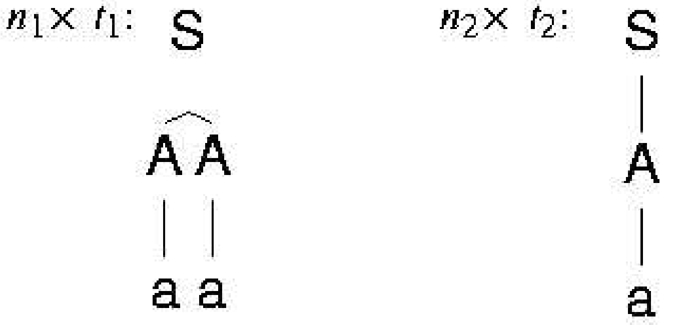
corpus of data

either on:
– sentences (unsupervised learning)
– analyzes (supervised learning)

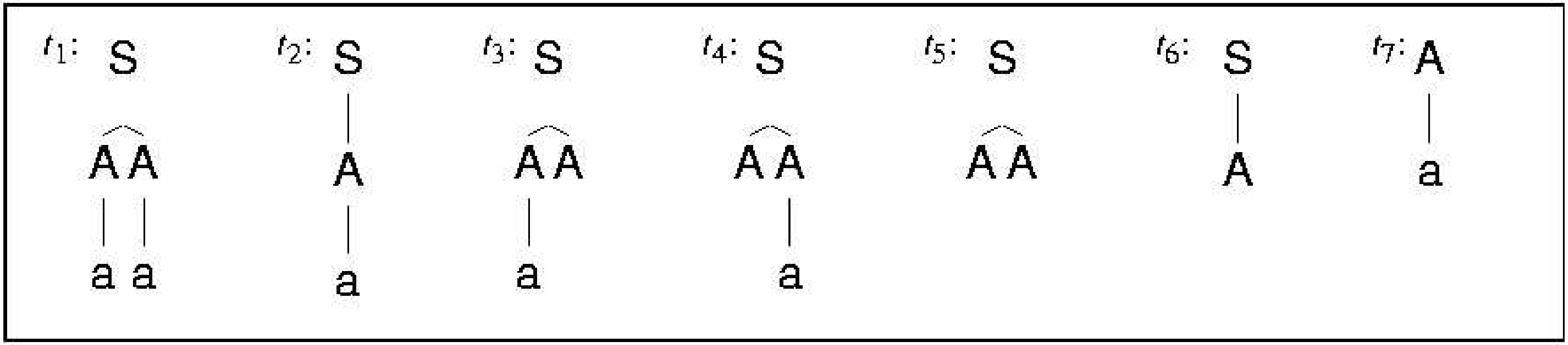
Grammar Estimation for PCFGs

Grammar Estimation (continued)

Treebank (Example by Johnson, 2002):



Tree fragments:



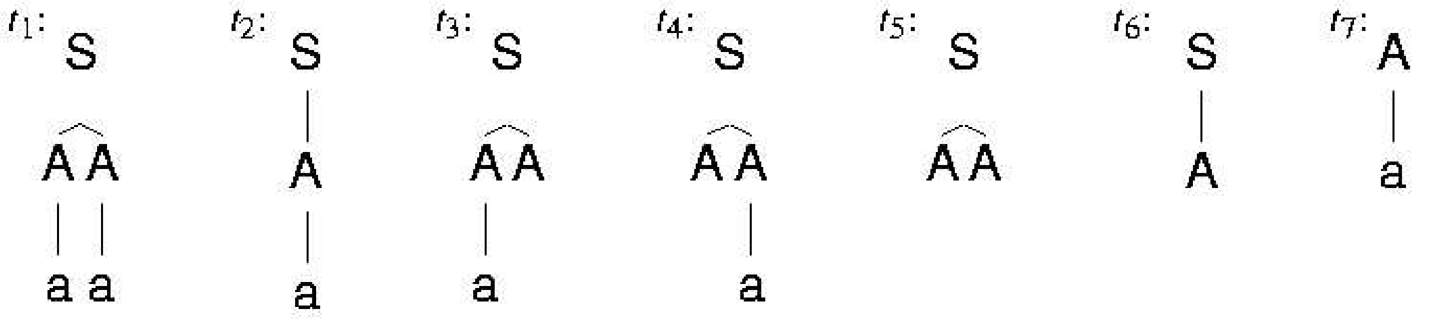
Tree derivations (Trees with hidden breakpoints):

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

Symbolic backbone of 'All-Fragment' DOP

Grammar Estimation (continued)

Fragments



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

DOP1 is biased and inconsistent

DOP Estimation

Research questions at the beginning of our project:

- Shall we search for a better estimation method (and stick with the 'All-Fragment' DOP)...
- ...or shall we search for a better / restricted DOP model (and stick with Maximum-Likelihood Estimation)?

Research context back then: We already knew at that time that Maximum-Likelihood Estimation results in a completely over-fitting instance of the standard DOP model, *which does not assign a positive probability to any tree outside the given treebank...*

DOP Estimation (continued)

Current solutions:

- **DOP Back-Off (Burrato and Sima'an 2003)**: Stick with the 'All-Fragments' DOP model, but smooth all fragment counts by backing off to smaller fragments and their counts...
- **DOP* (Zollmann and Sima'an 2005)**: Stick with Maximum-Likelihood Estimation by: (i) splitting the treebank into two parts, (ii) reading off selected fragments from one part, and (iii) estimating probabilities for these fragments by counting how often they occur in the other part...

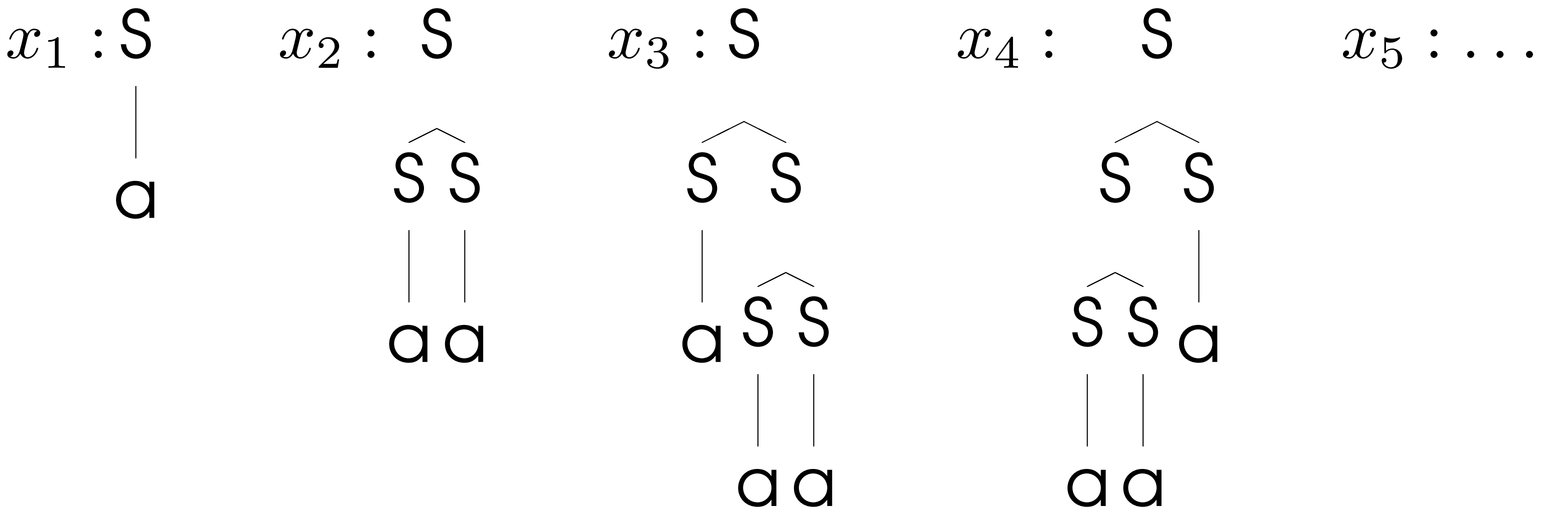
DOP Estimation (continued)

Current Problems: The original DOP model can be thought of as a kind of **Memory-Based-Learning** approach, where all trees and all tree fragments in the treebank comprise the memory of the DOP model:

- **Back-Off DOP** is not in this spirit, since its memory is based on back-off distributions. Look-up and update of this smoothed memory is considerably less simple than in Memory-Based Learning.
- Clearly, **DOP*** is not at all in the spirit of the ’All-Fragment’ DOP approach.
- (By contrast, **DOP1** could be easily modified to integrate new example trees on the fly.)

DOP Estimation (continued)

Current Problems (continued): In sharp contrast to PCFG estimation, the typical asymptotic behavior of DOP estimation is that *the symbolic backbone of DOP's probability model grows as the treebank grows*



In the limit of the treebank size, DOP risks learning an arbitrarily large grammar — even if the treebank is generated by a finite grammar.

Wrap-Up

- DOP1 is biased and inconsistent
- maximum-likelihood estimators for DOP are unbiased and consistent, but they over-fit
- all unbiased estimators for DOP over-fit
- DOP-Back-Off and DOP* are consistent and do not over-fit, but they do not fully satisfy important DOP aspects such as ‘trees and fragments and their probabilities should be easily accessible’ and ‘all fragments count’
- some symbolic aspects of DOP’s probability model do not even fit standard estimation theory

Conclusion

- stick with probabilistic modeling for DOP, but give up some aspects of the original DOP approach

  ⟶ smoothing as estimation
  ⟶ linguistically selected subsets of fragments
  ⟶ PCFGs only?!

- give up probabilistic modeling for DOP, thereby strengthening important aspects of the original DOP approach

  ⟶ Memory-Based Learning,
  ⟶ K-Nearest-Neighbors,
  ⟶ . . .

Thank you!