Consistent Estimators for Probabilistic Context-Free Grammars.

Alexander Clark and Nathanaël Fijalkow

Department of Philosophy
King's College London
alexander.clark@kcl.ac.uk

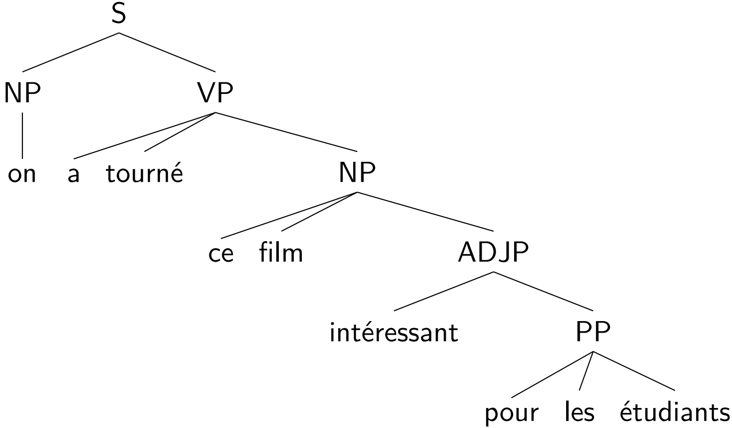
6 December 2018
Bordeaux

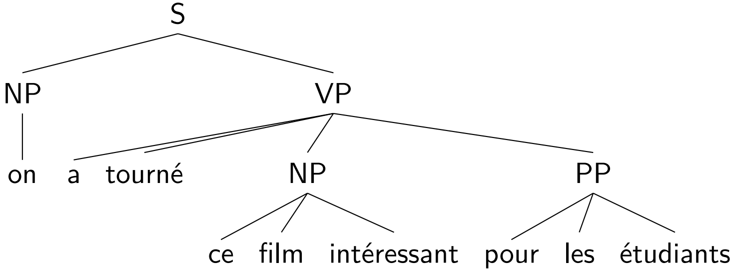
Outline

Motivation

- Natural languages – English, French etc – have *syntactic structure*

  [Levelt] "On a tourné ce film intéressant pour les étudiants"


S
NP
VP
on
a
tourné
NP
ce
film
ADJP
intéressant
PP
pour
les
étudiants


S
NP
VP
on
a
tourné
NP
PP
ce
film
intéressant
pour
les
étudiants

(Probabilistic) Context Free Grammars

Context-Free Grammars are the simplest model of hierarchical structure.

$$\langle \Sigma, V, S, P \rangle$$

- Σ is a set of terminal symbols (words)
- V is a set of nonterminal symbols (syntactic categories)
- S is a start symbol
- P is a set of productions which are one of :
  - $A \rightarrow a$, a is a terminal
  - $A \rightarrow BC$, $B, C \in V \setminus \{S\}$

  (using Chomsky Normal Form)

Probabilistic Context Free Grammars

Parameters $\theta : P \rightarrow [0, 1]$

$$\theta(A \rightarrow BC) = \frac{\mathbb{E}(A \rightarrow BC)}{\mathbb{E}(A)}$$

$$\theta(A \rightarrow a) = \frac{\mathbb{E}(A \rightarrow a)}{\mathbb{E}(A)}$$

Top-down generative process: start from S: Defines

- A distribution over *parse trees*
- and therefore a distribution over *strings*:
  Inside probabilities $\mathbb{P}(A \stackrel{*}{\Rightarrow} w)$
  Outside probabilities $\mathbb{P}(S \stackrel{*}{\Rightarrow} lAr)$

The Learning Problem

- We have a sequence of *strings* drawn i.i.d. from a distribution defined by a PCFG.
- We want to learn the grammar, and the parameters to arbitrary accuracy.

Motivation

First language acquisition:
Key question:

- Do the surface strings contain enough information to infer syntactic structure?
- Or must the learner rely on other sources of information (semantic, prosodic, innate ...)?

Weighted Context Free Grammars

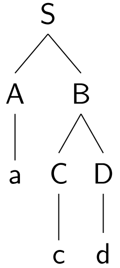
[Smith and Johnson(2007)]

Bottom up parameterisation

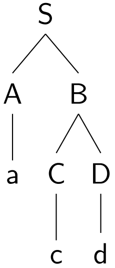
$$\theta(A \rightarrow BC) = \frac{\mathbb{E}(A \rightarrow BC)}{\mathbb{E}(B)\mathbb{E}(C)}$$

$$\theta(A \rightarrow a) = \mathbb{E}(A \rightarrow a)$$

Note that $\mathbb{E}(S) = 1$ so distribution is unchanged.



$$s(\tau) = \frac{\mathbb{E}(S \to AB)}{\mathbb{E}(S)} \cdot \frac{\mathbb{E}(B \to CD)}{\mathbb{E}(B)} \cdot \frac{\mathbb{E}(A \to a)}{\mathbb{E}(A)} \cdot \frac{\mathbb{E}(C \to c)}{\mathbb{E}(C)} \cdot \frac{\mathbb{E}(D \to d)}{\mathbb{E}(D)}$$



$$s(\tau) = \frac{\mathbb{E}(S \to AB)}{\mathbb{E}(S)} \cdot \frac{\mathbb{E}(B \to CD)}{\mathbb{E}(B)} \cdot \frac{\mathbb{E}(A \to a)}{\mathbb{E}(A)} \cdot \frac{\mathbb{E}(C \to c)}{\mathbb{E}(C)} \cdot \frac{\mathbb{E}(D \to d)}{\mathbb{E}(D)}$$

$$s(\tau) = \frac{1}{\mathbb{E}(S)} \cdot \frac{\mathbb{E}(S \to AB)}{\mathbb{E}(A)\mathbb{E}(B)} \cdot \frac{\mathbb{E}(B \to CD)}{\mathbb{E}(C)\mathbb{E}(D)} \cdot \mathbb{E}(A \to a) \cdot \mathbb{E}(C \to c) \cdot \mathbb{E}(D \to d)$$

Outline

Obvious problem

Infinitely many non isomorphic grammars define any non trivial language:
Consider the language

$$\{abc\}$$

Obvious problem

Infinitely many non isomorphic grammars define any non trivial language:
Consider the language

$$\{abc\}$$

We can't learn *all* PCFGs.

Anchored Context Free Grammars

Assume that for every nonterminal A there is a terminal a which occurs only in the production $A \to a$.
Reasonable assumption if $|\Sigma| \gg |V|$.
Implication (if a characterises A):

$$\mathbb{P}(lAr) = \frac{\mathbb{P}(lar)\mathbb{E}(A)}{\mathbb{E}(a)}$$

$$\theta(A \to a) = \frac{\mathbb{E}(a)}{\mathbb{E}(A)}$$

Basic Inequality with PCFGs

lexical rule

$$\mathbb{P}(lAr)\theta(A \to b) \leq \mathbb{P}(lbr)$$

Basic Inequality with PCFGs

lexical rule

$$\underbrace{\mathbb{P}(lAr)\theta(A \to b)}_{\text{sum over trees that use } A \to b} \leq \underbrace{\mathbb{P}(lbr)}_{\text{sum over all trees}}$$

Basic Inequality with PCFGs

lexical rule

$$\mathbb{P}(lAr)\theta(A \to b) \leq \mathbb{P}(lbr)$$

$$\theta(A \to b)\mathbb{E}(A) \leq \mathbb{E}(a) \min_{l,r} \frac{\mathbb{P}(lbr)}{\mathbb{P}(lar)}$$

Basic Inequality with PCFGs

lexical rule

$$\mathbb{P}(lAr)\theta(A \to b) \leq \mathbb{P}(lbr)$$

$$\underbrace{\theta(A \to b)\mathbb{E}(A)}_{\text{Bottom up parameters}} \quad \leq \quad \underbrace{\mathbb{E}(a) \min_{l,r} \frac{\mathbb{P}(lbr)}{\mathbb{P}(lar)}}_{\text{Properties defined by the distribution}}$$

Basic Inequality

binary rule

$$\mathbb{P}(lAr)\theta(A \to BC)\theta(B \to b)\theta(C \to c) \leq \mathbb{P}(lbcr)$$

Basic Inequality

binary rule

$$\underbrace{\mathbb{P}(lAr)\theta(A \rightarrow BC)\theta(B \rightarrow b)\theta(C \rightarrow c)}_{\text{sum over trees that use } A \rightarrow BC} \leq \underbrace{\mathbb{P}(lbcr)}_{\text{sum over all trees}}$$

Basic Inequality

binary rule

$$\mathbb{P}(lAr)\theta(A \to BC)\theta(B \to b)\theta(C \to c) \leq \mathbb{P}(lbcr)$$

$$\theta(A \to BC)\frac{\mathbb{E}(A)}{\mathbb{E}(B)\mathbb{E}(C)} \leq \frac{\mathbb{E}(a)}{\mathbb{E}(b)\mathbb{E}(c)} \min_{l,r} \frac{\mathbb{P}(lbcr)}{\mathbb{P}(lar)}$$

Basic Inequality

binary rule

$$\mathbb{P}(lAr)\theta(A \to BC)\theta(B \to b)\theta(C \to c) \leq \mathbb{P}(lbcr)$$

$$\underbrace{\theta(A \to BC)\frac{\mathbb{E}(A)}{\mathbb{E}(B)\mathbb{E}(C)}}_{\text{Bottom up parameters}} \leq \underbrace{\frac{\mathbb{E}(a)}{\mathbb{E}(b)\mathbb{E}(c)} \min_{l,r} \frac{\mathbb{P}(lbcr)}{\mathbb{P}(lar)}}_{\text{Properties defined by the distribution}}$$

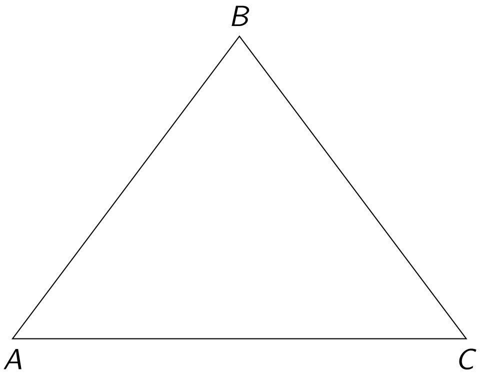
Ambiguity

Two further conditions:

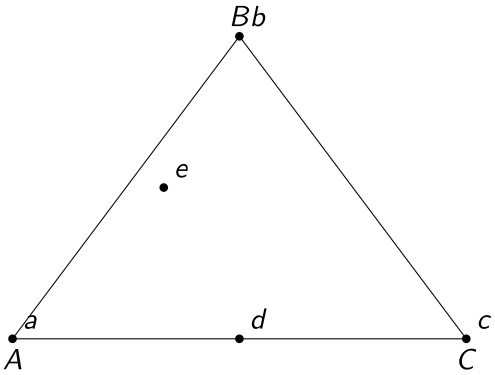
- Upwards monotonicity
- Downwards montonicity

Reasonable assumption if grammar is not excessively ambiguous: implies that we have equality in the inequalities above.

Picking characterising nonterminals

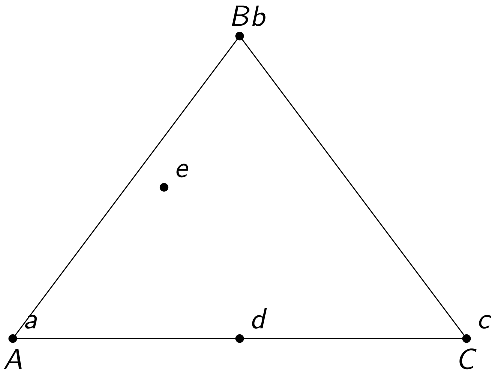


Picking characterising nonterminals



$$\frac{\mathbb{P}(lar)}{\mathbb{E}(a)} = \frac{\mathbb{P}(lAr)}{\mathbb{E}(A)}$$

Picking characterising nonterminals



$$\frac{\mathbb{P}(ldr)}{\mathbb{E}(d)} = \frac{1}{2}\frac{\mathbb{P}(lAr)}{\mathbb{E}(A)} + \frac{1}{2}\frac{\mathbb{P}(lCr)}{\mathbb{E}(C)}$$

Oracle probabilities

Assume for the moment that we have an oracle that will give us the true parameters: given a sample of strings we can recover directly the parameters:

-

Outline

Introduction

Grammar class

Parameter estimation

Paradigm

A consistent estimator (up to relabeling of nonterminals):
Input: $\{w_1, \ldots, w_m\}$
Output: as $m \to \infty$, $\hat{\theta}(A \to \alpha) \to \theta(A \to \alpha)$ in probability.
Not interested in the rate of convergence at the moment.

Plugin estimators

Naive approach:

- estimate the numerator and denominator separately and divide the estimates:
- minimize over observed frequent contexts of the denominator

$$\hat{\mathbb{E}}(a) = \frac{1}{N} \sum_{l,r} \#(lar)$$

$$\rho_N([[a]] \to [[b]][[c]]) = \frac{\hat{\mathbb{E}}(a)}{\hat{\mathbb{E}}(b)\hat{\mathbb{E}}(c)} \min_{l,r:c(lar)>\sqrt{N}} \frac{\#(lbcr)}{\#(lar)}$$

Ratio estimators

There are better ways of estimating these values:

Convergence of conditional KLD

If the estimates are close to the true values:

$$\varepsilon_{\mathsf{min}} < \log \frac{\hat{\theta}(A \to \alpha)}{\theta(A \to \alpha)} < \varepsilon_{\mathsf{max}}$$

then the conditional distribution of trees given strings is accurate too:

$$D\left(\mathbb{P}(\tau|w)\middle\|\hat{\mathbb{P}}(\tau|w)\right) \leq (2\mathbb{E}(|w|) - 1)(\varepsilon_{\mathsf{max}} - \varepsilon_{\mathsf{min}})$$

Normalisation

But the learned WCFG may even diverge and not define a distribution over trees at all.

- Standard normalisation techniques will maintain the conditional distribution gut give a very poor estimate of the joint distribution.

If we have a sample of strings we can use them to reestimate: Inside outside Algorithm

Conclusion

- Still a few gaps in the proof . . .
- Empirical work suggests that nearly all

Bibliography

Noah A Smith and Mark Johnson.
Weighted and probabilistic context-free grammars are equally expressive.
Computational Linguistics, 33(4):477–491, 2007.