Doubling DOP*

A comparison of Double-DOP and DOP*

Benno Kruit Sara Veldhoen

Supervised by:
Andreas van Cranenburg Khalil Sima'an

University of Amsterdam (UvA)

Project AI, January 2014

Data Oriented Parsing
Introduction to DOP
Bias and Consistency
Double-DOP and DOP*: a comparison
Introduction to Double-DOP and DOP*
Comparison
Experiments
Results
Analyzing grammars
Parsing Performance
Summary

Outline

Data Oriented Parsing
- Introduction to DOP
- Bias and Consistency

Double-DOP and DOP*: a comparison
- Introduction to Double-DOP and DOP*
- Comparison
- Experiments

Results
- Analyzing grammars
- Parsing Performance

Outline

Data Oriented Parsing
- Introduction to DOP
- Bias and Consistency

Double-DOP and DOP*: a comparison
- Introduction to Double-DOP and DOP*
- Comparison
- Experiments

Results
- Analyzing grammars
- Parsing Performance

Parsing

- input: sentence

  John Loves Mary

- output: constituent tree

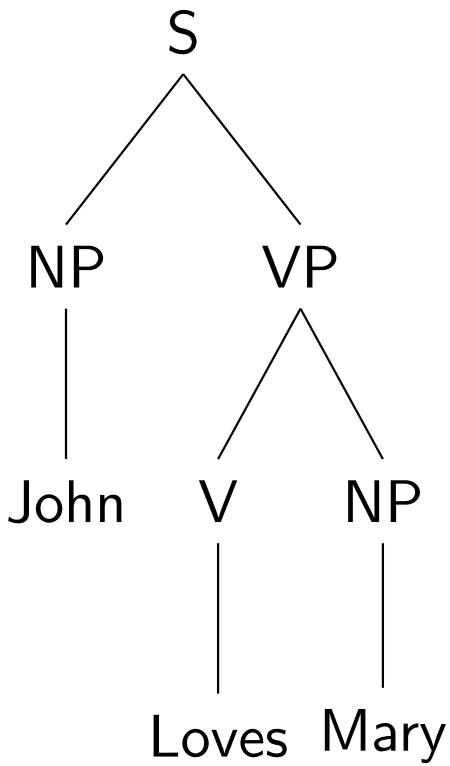
Data Oriented Parsing
Introduction to DOP
Bias and Consistency
Double-DOP and DOP*: a comparison
Introduction to Double-DOP and DOP*
Comparison
Experiments
Results
Analyzing grammars
Parsing Performance
Summary

Parsing

- input: sentence

  John Loves Mary

- output: constituent tree



Grammar

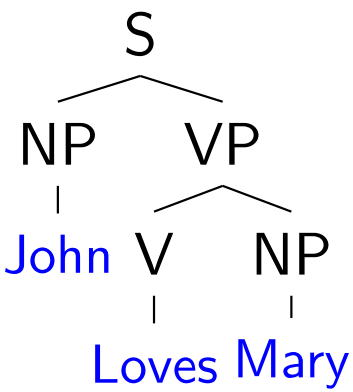
A grammar describes:

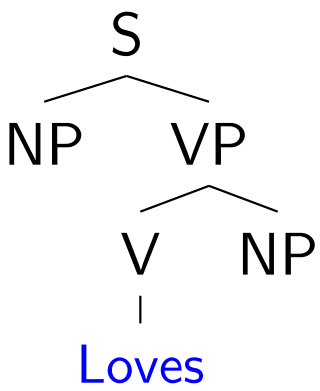
- how trees can be built
  - CFG's - elementary rules
  - TSG's - larger units: *fragments*
- how likely constructions are: *probabilistic* grammars
  - PCFG's - independence
  - PTSG's - derivations

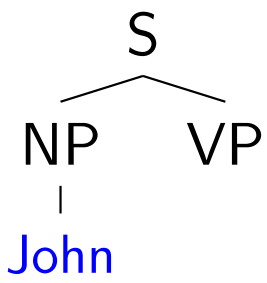
Grammar: CFG rules

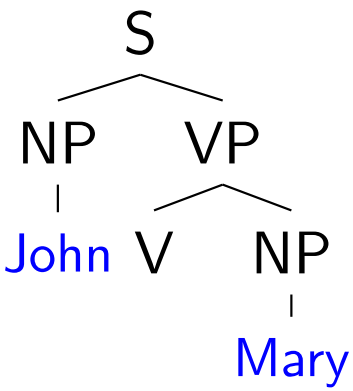
$S \rightarrow NP\ VP$
$VP \rightarrow V\ NP$
$NP \rightarrow John$
$NP \rightarrow Mary$
$V \rightarrow loves$

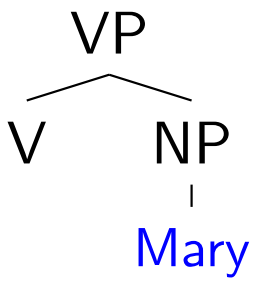
Grammar: Tree fragments











Etc...

Exponentially many

Outline

Data Oriented Parsing
- Introduction to DOP
- Bias and Consistency

Double-DOP and DOP*: a comparison
- Introduction to Double-DOP and DOP*
- Comparison
- Experiments

Results
- Analyzing grammars
- Parsing Performance

Consistency

- Assumption
  - Language is an infinite parse tree distribution
  - Treebank is a finite sample
- *Estimate* the true distribution
- Expected estimation should improve when the treebank grows $\rightarrow$ expected *loss* should decline
- **Consistency**: Expected loss becomes 0 when the sample size approaches ∞

Bias

- Assumption
  - An estimator should approach *any* distribution
  - Even finite distributions!
- If there's a distribution that doesn't match its expected estimate, the estimator is **biased**.
- What about unseen data?
- Bias is **good**

Outline

Data Oriented Parsing
- Introduction to DOP
- Bias and Consistency

Double-DOP and DOP*: a comparison
- Introduction to Double-DOP and DOP*
- Comparison
- Experiments

Results
- Analyzing grammars
- Parsing Performance

Double-DOP

- Extraction: Maximal Overlap
- Estimation: relative frequency
- Coverage: PCFG rules

DOP*

- Held-out estimation - *HC* and *EC*
- Extraction: Shortest derivations
- Estimation: relative frequency *in shortest derivations*
- Coverage: smoothing PCFG rules with probability p_{unkn}

Outline

Data Oriented Parsing
- Introduction to DOP
- Bias and Consistency

Double-DOP and DOP*: a comparison
- Introduction to Double-DOP and DOP*
- Comparison
- Experiments

Results
- Analyzing grammars
- Parsing Performance

Comparison

- Shortest derivations or Maximal overlap
- Held-out estimation or one vs. the rest

Outline

Data Oriented Parsing
- Introduction to DOP
- Bias and Consistency

Double-DOP and DOP*: a comparison
- Introduction to Double-DOP and DOP*
- Comparison
- Experiments

Results
- Analyzing grammars
- Parsing Performance

Outline

Data Oriented Parsing
- Introduction to DOP
- Bias and Consistency

Double-DOP and DOP*: a comparison
- Introduction to Double-DOP and DOP*
- Comparison
- Experiments

Results
- Analyzing grammars
- Parsing Performance

Maximal overlap $\leftrightarrow$ shortest derivation

Split $\leftrightarrow$ one vs. the rest

Data Oriented Parsing
Introduction to DOP
Bias and Consistency
Double-DOP and DOP*: a comparison
Introduction to Double-DOP and DOP*
Comparison
Experiments
Results
Analyzing grammars
Parsing Performance
Summary

Outline

Data Oriented Parsing
- Introduction to DOP
- Bias and Consistency

Double-DOP and DOP*: a comparison
- Introduction to Double-DOP and DOP*
- Comparison
- Experiments

Results
- Analyzing grammars
- Parsing Performance

F1 scores

Data Oriented Parsing
Introduction to DOP
Bias and Consistency
Double-DOP and DOP*: a comparison
Introduction to Double-DOP and DOP*
Comparison
Experiments
Results
Analyzing grammars
Parsing Performance
Summary

Summary

- So shortest derivations are not as useful as they seem
- And a split moves weight to large fragments
- Performance is not related to consistency

- Outlook
  - Why not?
  - Is it just grammar size?