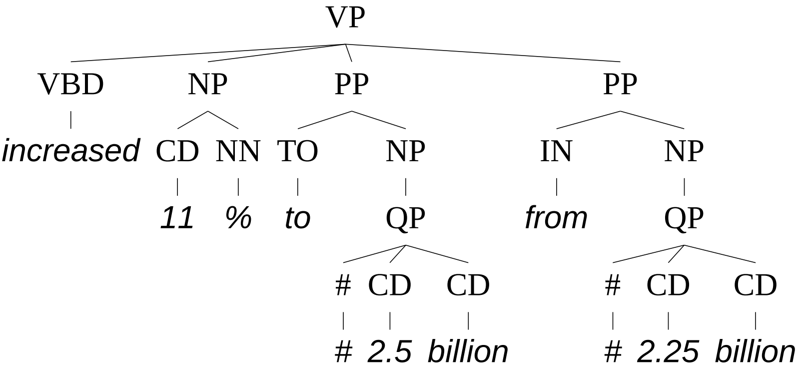
Non-Local Modeling with a Mixture of PCFGs

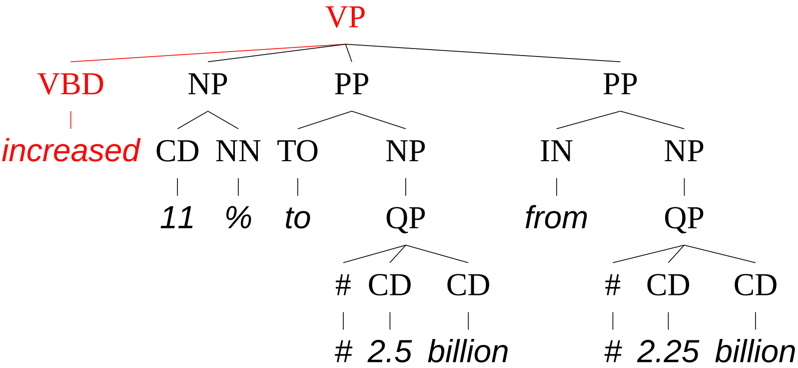
Slav Petrov, Leon Barrett and Dan Klein
University of California at Berkeley

CoNLL 2006

Empirical Motivation

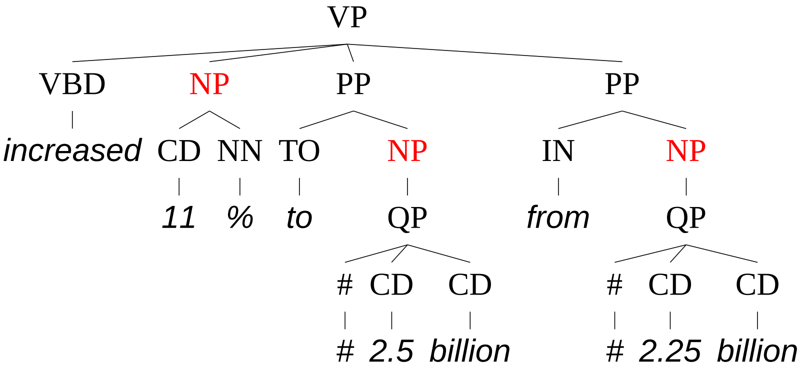


Empirical Motivation

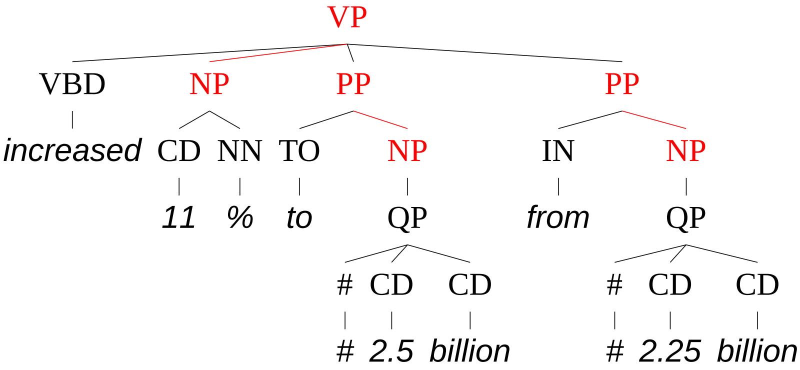


Verb Phrase Expansion: capture with lexicalization.
[Collins 1999, Charniak 2000]

Empirical Motivation

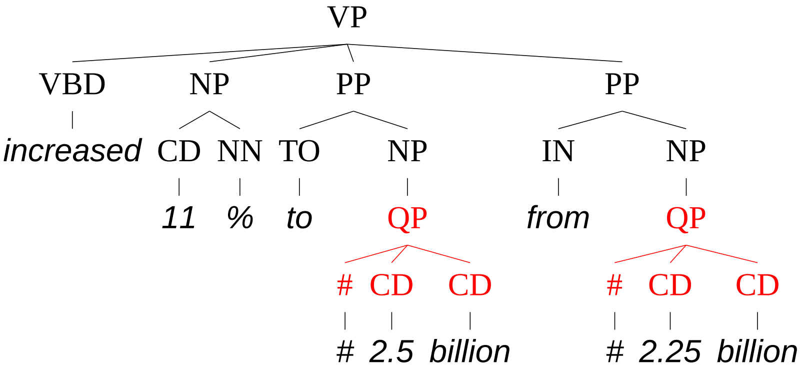


Empirical Motivation



Local Correlation: capture with parent annotation.
[Johnson 1998, Klein & Manning 2003]

Empirical Motivation



Non-Local Correlation.
[This work]

Correlations for QP $\rightarrow$ # CD CD

Rule	Score
QP $\rightarrow$ # CD CD	131.6
PRN $\rightarrow$ -LRB- ADJP -RRB	77.1
VP $\rightarrow$ VBD NP , PP PP	33.7
VP $\rightarrow$ VBD NP NP PP	28.4
PRN $\rightarrow$ -LRB- NP -RRB-	17.3
ADJP $\rightarrow$ QP	13.3
PP $\rightarrow$ IN NP ADVP	12.3
NP $\rightarrow$ NP PRN	12.3
VP $\rightarrow$ VBN PP PP PP	11.6
ADVP $\rightarrow$ NP RBR	10.1

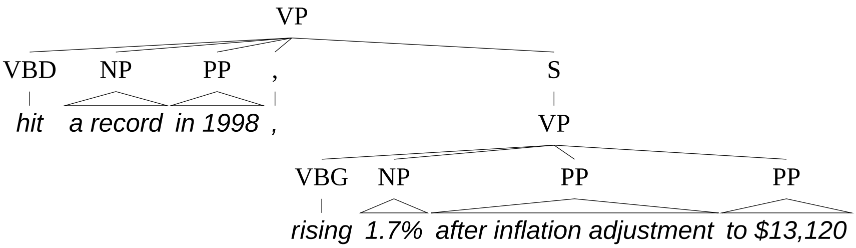
Correlations for QP → # CD CD

Rule	Score
QP → # CD CD	131.6
PRN → -LRB- ADJP -RRB	77.1
VP → VBD NP , PP PP	33.7
VP → VBD NP NP PP	28.4
PRN → -LRB- NP -RRB-	17.3
ADJP → QP	13.3
PP → IN NP ADVP	12.3
NP → NP PRN	12.3
VP → VBN PP PP PP	11.6
ADVP → NP RBR	10.1

Correlations for QP → # CD CD

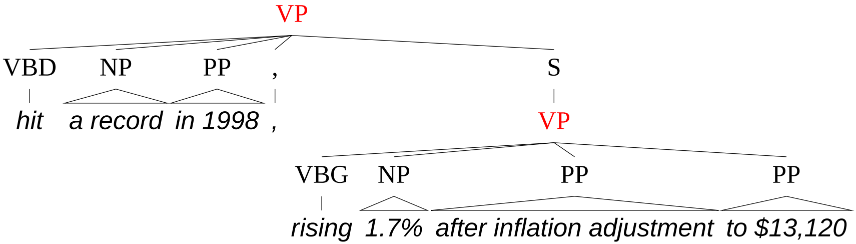
Rule	Score
QP → # CD CD	131.6
PRN → -LRB- ADJP -RRB	77.1
VP → VBD NP , PP PP	33.7
VP → VBD NP NP PP	28.4
PRN → -LRB- NP -RRB-	17.3
ADJP → QP	13.3
PP → IN NP ADVP	12.3
NP → NP PRN	12.3
VP → VBN PP PP PP	11.6
ADVP → NP RBR	10.1

Examples



Repeated formulaic structure in one grammar:
VP → VBD NP PP , S and VP → VBG NP PP PP.

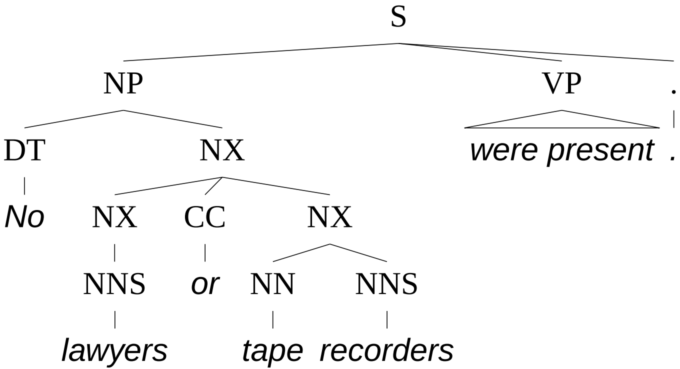
Examples



Repeated formulaic structure in one grammar:

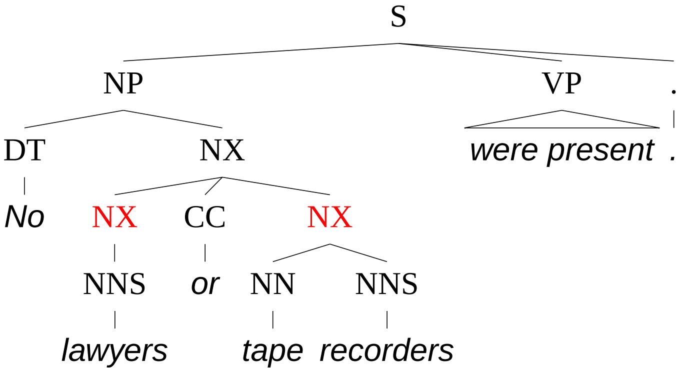
VP $\rightarrow$ VBD NP PP , S and VP $\rightarrow$ VBG NP PP PP.

Examples



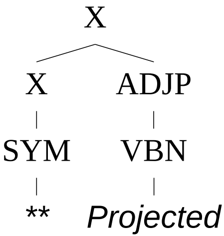
Sibling effects, though not parallel structure:
NX $\rightarrow$ NNS
and NX $\rightarrow$ NN NNS.

Examples



Sibling effects, though not parallel structure:
NX $\rightarrow$ NNS
and NX $\rightarrow$ NN NNS.

Examples



A special structure for footnotes:
ROOT $\rightarrow$ X
and X $\rightarrow$ SYM.

Motivation

Model non-local correlation that can stem from:

- Dialects,
- Priming effects,
- Genre,
- Stylistic conventions.

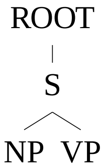
Single Grammar

ROOT

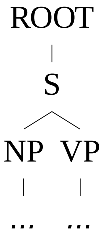
Single Grammar

ROOT
|
S

Single Grammar



Single Grammar



Mixture of PCFGs

ROOT-1 ROOT-2 ROOT-3 ROOT-4

Mixture of PCFGs

ROOT-1 ROOT-2 ROOT-3 ROOT-4

Mixture of PCFGs

ROOT-1 ROOT-2 ROOT-3 ROOT-4

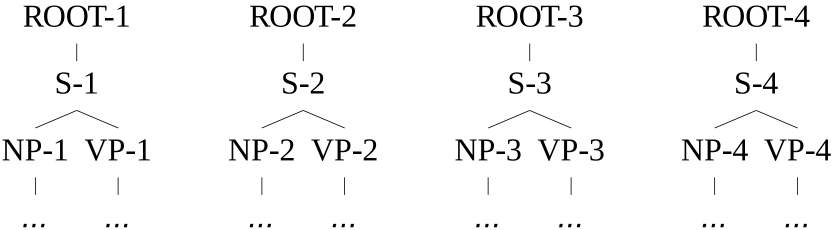
Mixture of PCFGs

ROOT-1 ROOT-2 ROOT-3 ROOT-4

Mixture of PCFGs

ROOT-1 ROOT-2 ROOT-3 ROOT-4

Mixture of PCFGs



Mixture of PCFGs

- Single grammar:

$$\mathrm{P}(T) = \prod_{X \to \alpha \in T} \mathrm{P}(\alpha | X).$$

Mixture of PCFGs

- Single grammar:

$$\mathrm{P}(T) = \prod_{X \rightarrow \alpha \in T} \mathrm{P}(\alpha | X).$$

- Single grammar from a mixture:

$$\mathrm{P}(T, i) = \mathrm{P}(i) \prod_{X \rightarrow \alpha \in T} \mathrm{P}(\alpha | X, i).$$

Mixture of PCFGs

- Single grammar:

$$P(T) = \prod_{X \rightarrow \alpha \in T} P(\alpha | X).$$

- Single grammar from a mixture:

$$P(T, i) = P(i) \prod_{X \rightarrow \alpha \in T} P(\alpha | X, i).$$

- Mixture of grammars:

$$P(T) = \sum_i P(T, i) = \sum_i P(i) \prod_{X \rightarrow \alpha \in T} P(\alpha | X, i).$$

Inference: Parsing

- Would like the *most probable parse*:

$$\mathrm{P}(T|S) \propto \sum_i \mathrm{P}(i)\mathrm{P}(T|i).$$

Inference: Parsing

- Would like the *most probable parse*:

$$P(T|S) \propto \sum_i P(i)P(T|i).$$

- Mixture of grammars:

$$\operatorname*{argmax}_T \sum_i P(T,i) = \operatorname*{argmax}_T \sum_i P(i) \prod_{X\rightarrow\alpha\in T} P(\alpha|X,i).$$

Inference: Parsing

- Would like the *most probable parse*:

$$\mathrm{P}(T|S) \propto \sum_i \mathrm{P}(i)\mathrm{P}(T|i).$$

- Mixture of grammars:

$$\operatorname*{argmax}_T \sum_i \mathrm{P}(T,i) = \operatorname*{argmax}_T \sum_i \mathrm{P}(i) \prod_{X \to \alpha \in T} \mathrm{P}(\alpha|X,i).$$

- Computing most probable parse is NP-hard.

Inference: Parsing

- Would like the *most probable parse*:

$$\mathrm{P}(T|S) \propto \sum_i \mathrm{P}(i)\mathrm{P}(T|i).$$

- Mixture of grammars:

$$\operatorname*{argmax}_T \sum_i \mathrm{P}(T,i) = \operatorname*{argmax}_T \sum_i \mathrm{P}(i) \prod_{X\to\alpha\in T} \mathrm{P}(\alpha|X,i).$$

- Computing most probable parse is NP-hard.
- Compute the *most probable derivation* instead.

Learning: Training

- Manually assign sentences to grammars, e.g. Brown corpus.
- Alternatively, use a standard Expectation-Maximization (EM) approach.

Learning: Training

- Manually assign sentences to grammars, e.g. Brown corpus.
- Alternatively, use a standard Expectation-Maximization (EM) approach.

E-Step:

- Fix model parameters and compute the posterior distributions of the latent variables.
- Component G of each sentence:

$$\mathrm{P}(i|T) = \frac{\mathrm{P}(T,i)}{\sum_j \mathrm{P}(T,j)}.$$

Learning: Training

M-Step:

- Given the posterior assignments find the maximum likelihood model parameters.
- Let $\mathbf{T} = \{T_1, T_2, \ldots\}$ be the training set. The M-Step updates are:
- Component prior:

$$\mathrm{P}(i) \leftarrow \frac{\sum_{T_k \in \mathbf{T}} \mathrm{P}(i|T_k)}{\sum_i \sum_{T_k \in \mathbf{T}} \mathrm{P}(i|T_k)} = \frac{\sum_{T_k \in \mathbf{T}} \mathrm{P}(i|T_k)}{k}.$$

- Estimate rule probabilities as for a single grammar but with fractional counts.

Hierarchical Estimation

- Pool common rules (e.g. NP $\rightarrow$ DT NN) in a *shared grammar* G_s.

Hierarchical Estimation

- Pool common rules (e.g. NP $\rightarrow$ DT NN) in a *shared grammar* G_s.
- Latent variable $L = \{\mathrm{s}, \mathrm{l}\}$ at each rewrite:

Hierarchical Estimation

- Pool common rules (e.g. NP $\rightarrow$ DT NN) in a *shared grammar* G_s.
- Latent variable $L = \{\text{s}, \text{l}\}$ at each rewrite:

$$\mathrm{P}(\alpha|X, i) = \lambda \mathrm{P}(\alpha|X, i, \ell = \text{l}) + (1 - \lambda)\mathrm{P}(\alpha|X, i, \ell = \text{s}),$$

Hierarchical Estimation

- Pool common rules (e.g. NP $\rightarrow$ DT NN) in a *shared grammar* G_s.
- Latent variable $L = \{\mathrm{s}, \mathrm{l}\}$ at each rewrite:

$$\mathrm{P}(\alpha|X, i) = \lambda\mathrm{P}(\alpha|X, i, \ell = \mathrm{l}) + (1 - \lambda)\mathrm{P}(\alpha|X, i, \ell = \mathrm{s}),$$

- Two kinds of hidden variables: the grammar G (for each sentence) and the level L (for each node).

E-Step

- Component G of each sentence as before:

$$\mathrm{P}(i|T) = \frac{\mathrm{P}(T,i)}{\sum_j \mathrm{P}(T,j)}.$$

- Hierarchy level L of each rewrite:

$$\mathrm{P}(\ell = \mathsf{l}|X \to \alpha, i, T) = \frac{\lambda\mathrm{P}(\alpha|X,\ell{=}\mathsf{l})}{\lambda\mathrm{P}(\alpha|X,i,\ell{=}\mathsf{l}) + (1-\lambda)\mathrm{P}(\alpha|X,\ell{=}\mathsf{s})}.$$

M-Step

- Component prior as before:

$$\mathrm{P}(i) \leftarrow \frac{\sum_{T_k \in \mathbf{T}} \mathrm{P}(i|T_k)}{\sum_i \sum_{T_k \in \mathbf{T}} \mathrm{P}(i|T_k)} = \frac{\sum_{T_k \in \mathbf{T}} \mathrm{P}(i|T_k)}{k}.$$

M-Step

- Component prior as before:

$$P(i) \leftarrow \frac{\sum_{T_k \in \mathbf{T}} P(i|T_k)}{\sum_i \sum_{T_k \in \mathbf{T}} P(i|T_k)} = \frac{\sum_{T_k \in \mathbf{T}} P(i|T_k)}{k}.$$

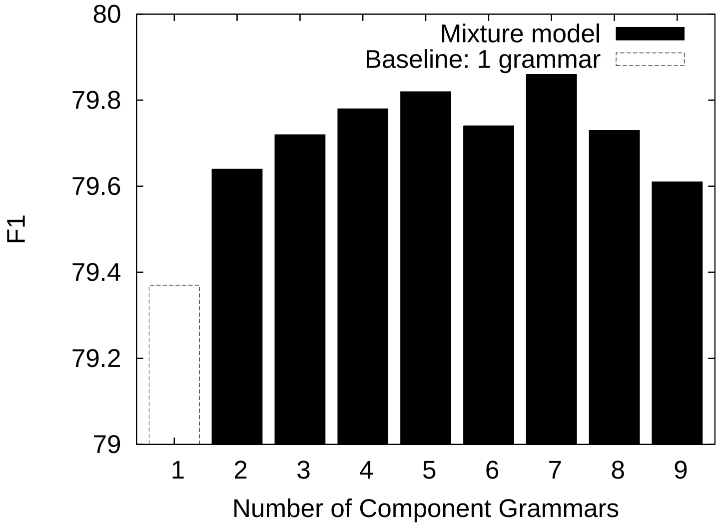
- Hierarchy Level:

$$P(l = \mathrm{l}) \leftarrow \frac{\sum_{T_k \in \mathbf{T}} \sum_{X \to \alpha \in T_k} P(\ell = \mathrm{l}|X \to \alpha)}{\sum_{T_k \in \mathbf{T}} |T_k|}.$$

Experimental Setup

- WSJ with standard setup:
  - Section 2-21 training set,
  - Section 22 validation set,
  - Section 23 test set.
- Baseline: Markovized grammar annotated with parent and sibling information (vertical order=2, horizontal order=1 [Klein & Manning 2003]).

Parsing Accuracy



Capturing Rule Correlations

- Mixture model captures non-local correlations.
- 10% reduction in total correlation error:
  - Estimate rule correlations from corpus.
  - Generate trees with grammar and estimate rule correlations.
  - Compute correlation difference.

Genre

- Brown corpus' genres are statistically coherent.
- Assign each genre to an individual grammar (no EM training):

$$F_1 = 79.48\text{, LL=-242332.}$$

- Initialize by genre then train with EM:

$$F_1 = 79.37\text{, LL=-242100.}$$

- EM with a random initialization:

$$F_1 = 79.16\text{, LL=-242459.}$$

- Model can capture variation between genres, but maximum training data likelihood does not necessarily give maximum accuracy.

Recent Development

"Learning Accurate, Compact, and Interpretable Tree Annotation", Petrov et al., ACL 2006:

- $F_1 = 90.2\%$.
- More flexible learning framework.
- Split and merge training to keep grammar compact.
- Similar in spirit to Klein & Manning 2003 and Matsuzaki et al. 2005.

Conclusions

- Examined rule correlations that may be found in natural language corpora, discovering non-local correlations not captured by traditional models.
- A Mixture of PCFGs can represent these non-local features and gives an improvement in parsing accuracy and data likelihood.
- This improvement is modest, however, primarily because local correlations are so much stronger than non-local ones.

Discussion

Thank you very much for your attention.

Questions?

{petrov, lbarrett, klein}@eecs.berkeley.edu