

**DOP models of the supervised and unsupervised
acquisition of constructions**

Jelle Zuidema

ILLC

University of Amsterdam

Plan of the talk

- Linguistic Motivation
- Data-Oriented Parsing as Probabilistic Construction Grammar
- Supervised Learning of Constructions
 - Problems with existing DOP estimators
 - New estimator: push-n-pull
 - Results
- Unsupervised Learning of Constructions

Probabilistic Linguistics

(Abney 1996, *Statistical Methods and Linguistics*)

- (1) a. #the a are of I word salad?
b. John saw Mary unambiguous?
- (2) a. a hectare is a hundred ares
b. As described in section I paragraph a ...
c. The a paragraph of I is hardly readable.
d. Typhoid Mary
e. the Russia house butler

Constructions

(Fillmore, Kay & O'Connor, 1988; Culicover & Nowak, 2004; Jackendoff, forthcoming)

Idioms

- (3) a. by and large
- b. lo and behold
- c. beat a dead horse
- d. make amends
- e. cast aspersions
- f. a flash in the pan

Constructions

VP constructions

- (4)
 - a. Pat sang/drank/sewed his heart out
 - b. *Pat sang the Marseillaise his heart out
 - c. Leslie talked/cooked/composed up a storm
 - d. *Leslie talked a storm up
 - e. *Leslie cooked eggs up a storm

- (5)
 - a. Elmer hobbled/laughed/joked his way to the bank.
 - b. Hermione slept/drank/sewed/programmed three whole evenings away.

Constructions

- (6)
- a. When is the next train *from Amsterdam to Paris*?
 - b. BA carried *more people than* cargo in 1987.
 - c. Lawson is *the closest thing* in London *to* a supply-side globalist.

Radical Construction Grammar

(Croft, Jackendoff)

- Constructions map conceptual, syntactic and phonological structures onto each other;
- Individual constructions carry a weight (frequency/recency information);
- Everything is a construction, from specific words to highly abstract word order principles;
- Constructions do not necessarily involve all three levels of conceptual, syntactic and phonological structure.

Stochastic Tree Substitution Grammars

<p style="text-align: center;"> NP PP-DIR PP-DIR IN NNP from Baltimore TO NNP to Oakland .2 .5 .5 .5 </p>	$\Pi = 0.0125$
<p style="text-align: center;"> NP PP-DIR PP-DIR IN NNP from Baltimore TO NNP to Oakland .2 .5 .5 </p>	$\Pi = 0.05$
<p style="text-align: center;"> NP PP-DIR PP-DIR IN NNP from Baltimore TO NNP to Oakland .1 </p>	$\Pi = 0.1$
$\Sigma = 0.1625$	

Stochastic Tree Substitution Grammars

An STSG is a 5-tuple $\langle V_n, V_t, S, T, w \rangle$

$$w : T \rightarrow [0, 1], \text{ such that } \forall r \sum_{t:r(t)=r} w(t) = 1$$

The probability of a derivation:

$$P(d = t_1 \circ \dots \circ t_n) = \prod_{i=1}^n (w(t_i))$$

The probability of a parse:

$$P(p) = \sum_{d:\hat{d}=p} (P(d))$$

Stochastic Tree Substitution Grammars

Expected Usage Frequency:

$$u(t) = \sum_{d:t \in d} P(d)C(t, d)$$

Expected Occurrence Frequency:

$$\mathbf{E}[f(t)] = \sum_{p:t \in p^*} P(p)C(t, p^*),$$

(where p^* is set of all subtrees of p , and $C(t, p^*) = \#$ occurrences of t in p^*)

Data-Oriented Parsing

(Scha, 1990; Bod, 1993)

Given a corpus of phrase-tree annotated sentences

divided in a train set and a test set

all subtrees of all trees in the train set form the symbolic grammar

with which the test set sentences are parsed.

Data-Oriented Parsing

DOP1 (Bod, 1993, 1998)

$$w(t) = \frac{f(t)}{\sum_{t':r(t')=r(t)} f(t')}$$

EACL'03 (Bod, 2003)

$$w(t) = \frac{f(t)\alpha}{\sum_{t':r(t')=r(t)} f(t')\alpha}$$

(where α is a scaling factor)

Maximum Probable Parse

Excellent empirical results

E.g. on Wall Street Journal sentences of less than 100 words:

parser	LP	LR	<i>F</i>
Collins '96	.857	.853	.855
Collins '99	.883	.881	.882
Charniak '00	.895	.896	.895
L-DOP	.897	.895	.896
SL-DOP	.908	.907	.907
Charniak & Johnson '05			.910

LP=labeled precision (# correctly labeled constituents / # labeled constituents)

LR=labeled recall (# correctly labeled constituents / # target constituents)

$$F = 2 \cdot LP \cdot LR / (LP + LR) \quad \text{(harmonic mean)}$$

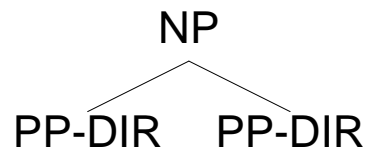
Computational problems

- Time complexity approximations of MPP
- Space complexity PCFG reduction
- Inconsistency of Estimators (EACL'06)
- Data Annotation U-DOP

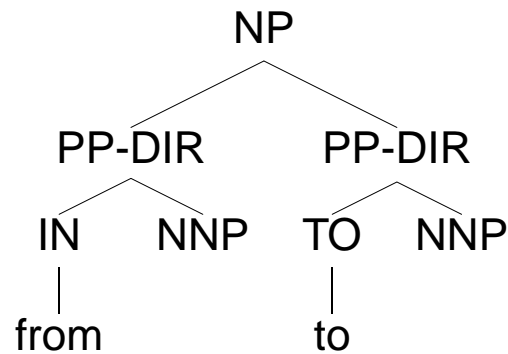
What are the relevant constructions?

TO
|
to
1.0

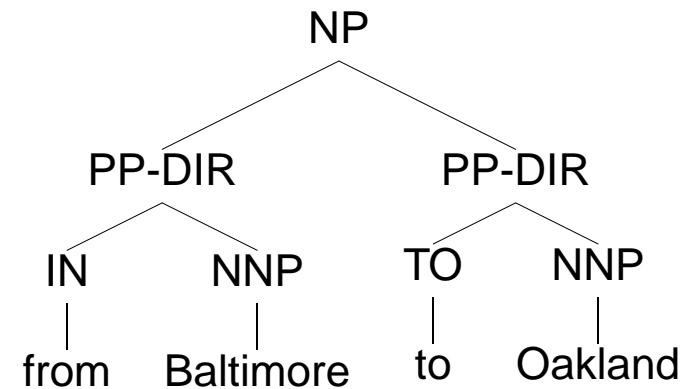
CC
|
and
46/54



$w_1 \geq$



$w_2 \geq$



w_3

Bod (2003) (“Do All Fragments Count?”, *Natural Language Engineering*, 9(4), 307-323) finds that any across-the-board restriction on the allowed fragments, decreases parse accuracy.

- DOP1 (Bod, 1993) and related estimators give observed fragments of all sizes non-zero weights, regardless of whether their frequency is already explained by smaller fragments;
- DOP* (Zollmann & Sima'an, 2005) and related estimators push all probability mass to the largest fragments, regardless of whether their frequency can be explained by smaller fragments equally well.

A new estimator: push-n-pull

- For every fragment t , calculate the difference between observed and expected frequency: $\Delta_1 = \mathbf{E}[f(t)] - f(t)$;
- If $\Delta_1 > 0$, the current grammar overestimates the frequency of t . Hence, as far as possible, *push* some of its weight to other trees.
- If $\Delta_1 < 0$, the current grammar underestimates the frequency of t . Hence, as far as possible, *pull* some weight from other trees.

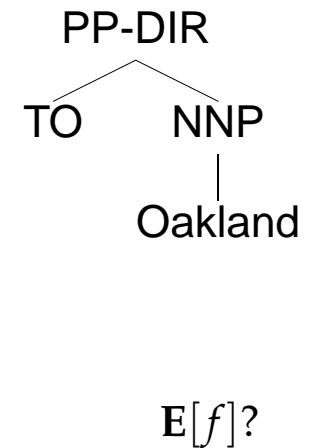
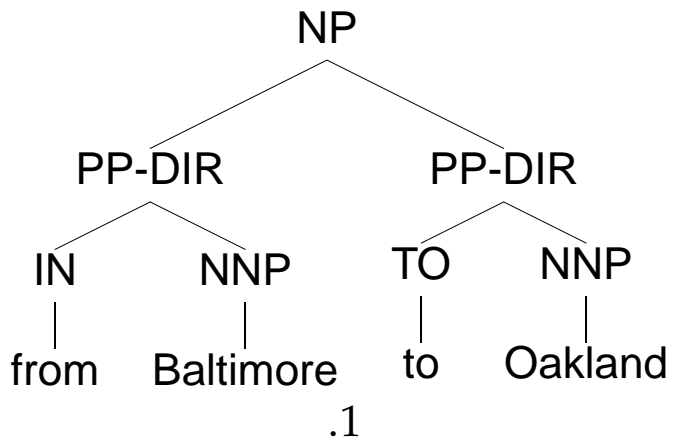
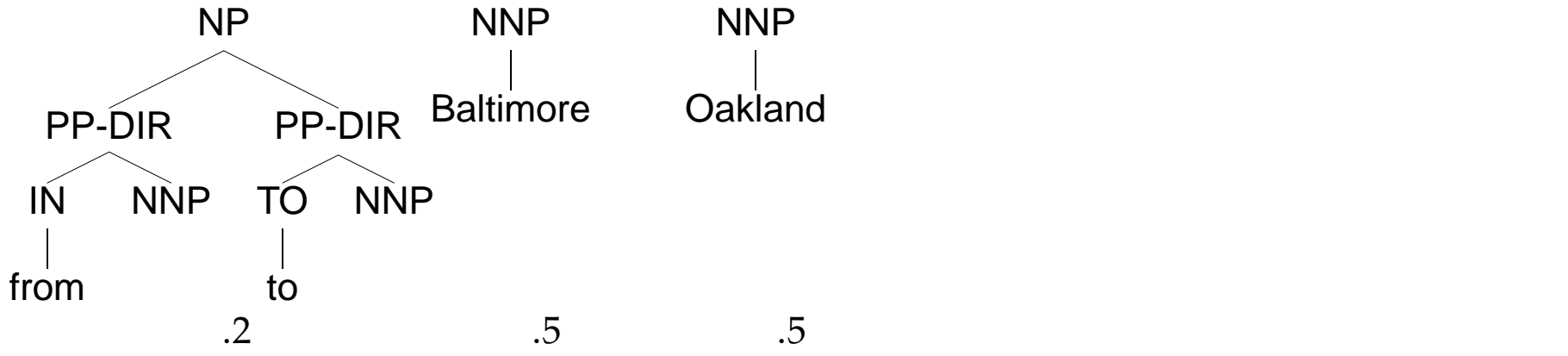
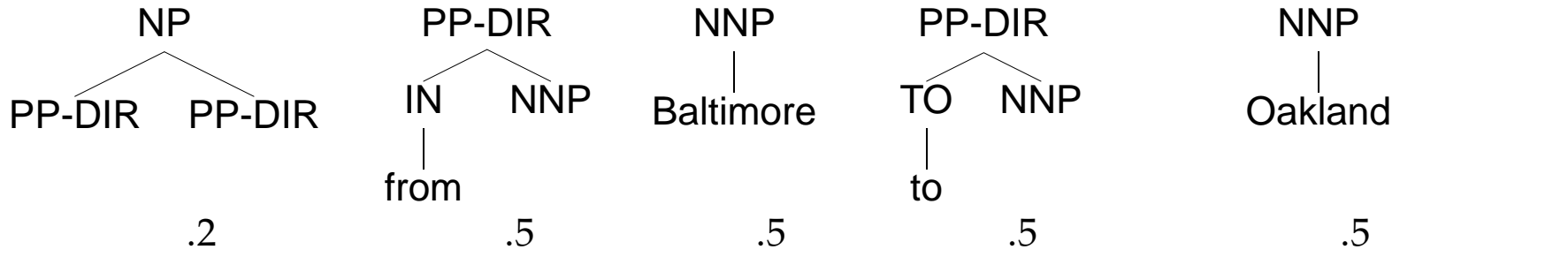
Expected Frequency

$$\mathbf{E}[f(t)] = \sum_{d \in D(t)} (\alpha(d)\beta(d))$$

$$\alpha(d) = \sum_{\tau \in \widehat{tw}(d_1)} \left(\sum_{\tau' \in \widehat{pr}_{x(t)}(\tau)} u(\tau') \right)$$

$$\beta(d) = \prod_{\substack{t' \in \\ \langle d_2, \dots, d_n \rangle}} \left(\sum_{\tau' \in \widehat{pr}_{x(t)}(t')} w(\tau') \right)$$

Expected Frequency



Push

- Pushing scores away is possible upto the current score of t . Hence, $\Delta' = \text{MINIMUM}(sc(t), \Delta)$;
- The score of t is decreased: $sc(t)_- = \Delta'$;
- The score is *pushed* towards all trees t' , which are subtrees of t involved in length-2 derivations of t . Hence: $sc(t')_+ = \Delta' / \#\text{such derivations}$

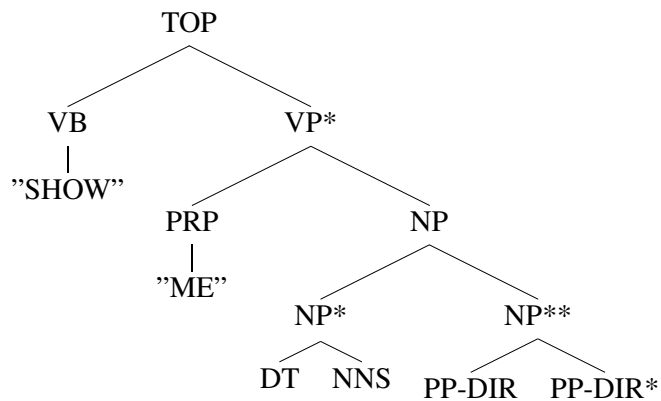
Pull

- Scores are pulled from all trees t' , which are subtrees of t involved in length-2 derivations of t .
- Pulling scores is only possible upto the point where these subtrees have score 0. If there are n length-2 derivations d^i of t , then $\delta^i = \text{MINIMUM}(sc(d_1^i), sc(d_2^i), -Delta/n)$.
- The scores of these subtrees are decreased: $sc(d_j^i)_- = \delta_i$;
- The score of t is increased: $(t)_+ = \sum_i delta_i$

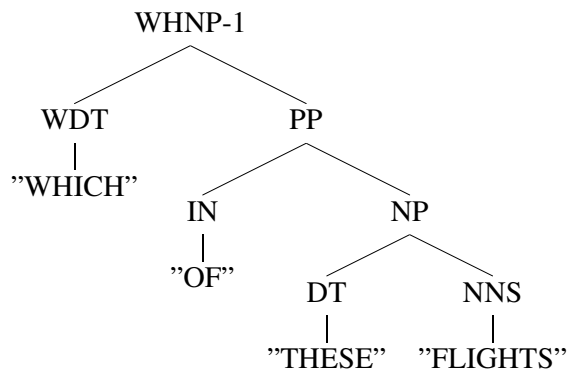
for each observed parse tree p
 for each depth-1 subtree t in p
 update-score($t, 1.0$)
 for each subtree t of p
 $\Delta = \min(sc(t), B + \gamma(\mathbf{E}[f(t)] - f(t)))$
 $\Delta' = 0$
 for each of n derivations d of t
 let $t' \dots t''$ be all elementary trees in d
 $\delta = \min(sc(t'), \dots, sc(t''), -\Delta/n)$
 $\Delta' = \delta$
 for each elementary tree t' in d
 update-score(t', δ)
 update-score (t, Δ')

Promising results on ATIS:

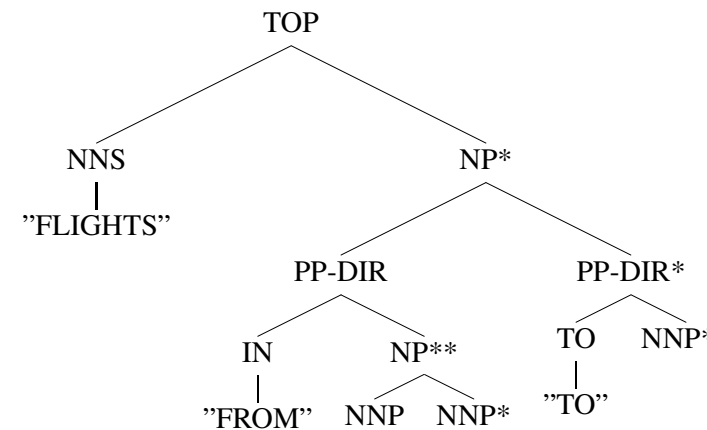
method	# rules	Cov.	LR	LP	EM
DOP1	77852	84%	95.07	95.07	83.5
p-n-p	58799	84%	95.07	95.07	83.5



(a) The “show me NP PP” frame, which occurs very frequently in the training data and is represented in several elementary trees with high weight.



(b) The complete parse tree for the sentence “Which of these flights”, which occurs 16 times in training data.



(c) The frame for “flights from NP to NP”

U-DOP

(Bod, 2006)

- All binary trees that can be assigned to sentences
- DOP as in (Bod, 2003)
- State of the art results on Unlabeled Precision and Recall
- ... but what are the relevant constructions?

Conclusions

- In the family of CG formalisms, DOP is an early and highly successful approach (in statistical parsing)
- Two crucial steps towards its use as a psycholinguistic model have been taken: (i) a new estimator, (ii) an unsupervised version
- The challenge is to combine the two approaches.

Subtree scores

$$\mathbf{E}[f(t)] = \sum_{d \in D(t)} (\alpha(d)\beta(d))$$

$$\alpha(d) = \sum_{\tau \in \widehat{tw}(d_1)} \left(\sum_{\tau' \in \widehat{pr}_{x(t)}(\tau)} u(\tau') \right) = \alpha_{d_1, x(t)}$$

$$\beta(d) = \prod_{\substack{t' \in \\ \langle d_2, \dots, d_n \rangle}} \left(\sum_{\tau' \in \widehat{pr}_{x(t)}(t')} w(\tau') \right)$$

$$= \prod_{\substack{t' \in \\ \langle d_2, \dots, d_n \rangle}} \frac{1}{\sum_{t'': r(t')=r(t'')} u(t'')} \underbrace{\left(\sum_{\tau' \in \widehat{pr}_{x(t)}(t')} u(\tau') \right)}_{\beta_{t', x(t)}^*}$$

With every change of the scores of t , keep track of the “subtree scores” of its supertrees.

update-scores(τ, δ)

$$u(\tau)_+ = \delta$$

for each prune τ' at sites x

$$\beta_{\tau',x}^* + = \delta$$

for each twig τ'' of τ

$$\alpha_{\tau'',x} + = \delta$$