

CS 294-5: Statistical Natural Language Processing



Grammar Induction
Lecture 19: 11/9/05

Idea: Lexical Affinity Models

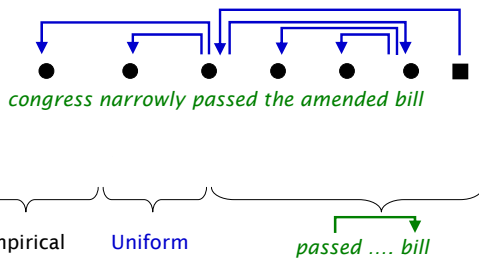
- Words select other words on syntactic grounds



- Idea: Link up pairs with high mutual information
 - [Yuret, 1998]: Greedy linkage
 - [Paskin, 2001]: Iterative re-estimation with EM
- Evaluation: compare linked pairs to a gold standard

Lexical Affinity Models

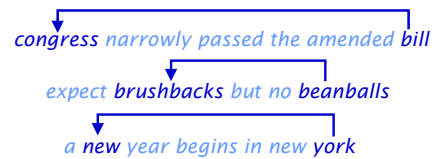
- Generative Model for [Paskin, 2001]



Problem: Non-Syntactic Affinity

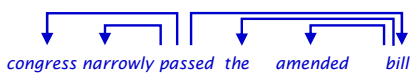
Method	Accuracy
Paskin, 2001	39.7

- Mutual information between words does not necessarily indicate syntactic selection.

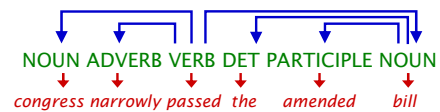


Idea: Word Classes

- Individual words like congress are entwined with semantic facts about the world.
- Syntactic classes, like NOUN and ADVERB are bleached of word-specific semantics.
- Automatic word classes more likely to look like DAYS-OF-WEEK or PERSON-NAME.
- We could build dependency models over word classes. [cf. Carroll and Charniak, 1992]





A Word-Class Model



$$P(\text{words, classes, graph}) = P(\text{length}) P(\text{graph})$$

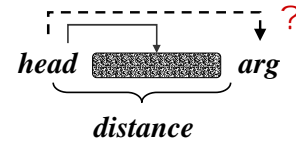
Problems: Word Class Models

Random	41.7	
Carroll and Charniak, 92	44.7	

- Issues:
 - Too simple a model – doesn't work much better supervised
 - Non-representational words can't be categorized (bill)



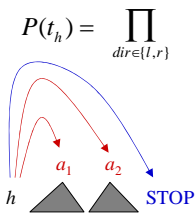
Issue: Local Representations



	Word-Classes?	Distance	Local Factor
Paskin 01	✗	✗	$P(a h)$

A Head-Outward Model (DMV)

- Supervised statistical parsers benefit from modeling tree distributions implicitly. [e.g., Collins, 99]
- A head-outward model with word classes and valence/adjacency:



Results: Dependencies

- Model is re-estimated with EM
 - Cubic dynamic program run over each sentence
 - Expected counts of each modeled configuration are aggregated
- Initialization:
 - Initial parameters from simple heuristics:

$$P(a | h, dir) \propto \sum_{dist} \frac{1}{dist} count(h, a, dist, dir)$$

Common Errors: Dependency

Overproposed Dependencies Underproposed Dependencies

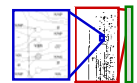
Overproposed Dependencies	Count	Underproposed Dependencies	Count
DT ← NN	3474	DET → N	3079
NNP ← NNP	2096	NNP → NNP	1898
CD → CD	760	IN ← NNS	838
IN ← DT	735	NN → VBP	714
DT ← NNS	696	DT → NNS	672
DT → IN	627	NN ← IN	669
DT → VBD	470	CD ← CD	54
DT → VBP	420	NN → VBD	54

Early Approaches: Structure Search

- Incremental grammar learning, chunking [Wolff 88, Langley 82, many others]
 - Can recover synthetic grammars
- An (extremely good) result of incremental structure search:

N-bar or zero determiner NP	Transitive VPs (complementation)	PP	Intransitive S
zNN → NN NNS	zVP → zV JJ	zPP → zN zNN	zS → PRP zV
zNN → JJ zNN	zVP → zV zNP	zPP → zN zNP	zS → zNP zV
zNN → zNN zNN	zVP → zV zNN	zPP → zN zNNP	zS → zNNP zV
NP with determiner	verb groups / intransitive VPs		
zNP → DT zNN	zV → VBZ [VBD] VBIP		
zNP → PRPS zNN	zV → MD VB		
	zV → MD RB VB		
Proper NP	zV → zV zRB		
zNNP → NNP NNPS	zV → zV zVBG		
zNNP → zNNP zNNP			
	Transitive VPs (adjunction)		
	zVP → zRB zVP		
	ZVP → zVP zPP		

- Looks good, ... but can't parse in the wild.



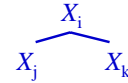
Issues with Chunk/Merge Systems

- Hard to recover from initial choices (c.f. EM, where the issue is initial state)
- Hard to make local decisions which will interact well with each other (e.g. group verb-preposition and preposition-determiner, both wrong, and not consistent)
- Good local heuristics often don't have a well-formed global objective that can be evaluated for the target grammar.

Idea: Learn PCFGs with EM

- Classic experiments on learning PCFGs with Expectation-Maximization [Lari and Young, 1990]

$\{X_1, X_2 \dots X_n\}$



- Full binary grammar over n symbols
- Parse uniformly/randomly at first
- Re-estimate rule expectations off of parses
- Repeat
- Their conclusion: it doesn't really work.

Re-estimation of PCFGs

- Basic quantity needed for re-estimation with EM:

$$P(X_c | i, j, S) = \frac{\sum_{T: (X_i, i, j) \wedge \text{yield}(T)=S} P(T)}{\sum_{T: \text{yield}(T)=S} P(T)}$$

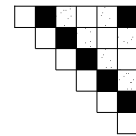
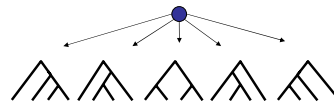
- Can calculate in cubic time with the Inside-Outside algorithm.
- Consider an initial grammar where all productions have equal weight:

$$P(X_a X_b | X_c) = 1/n^2$$

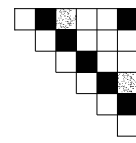
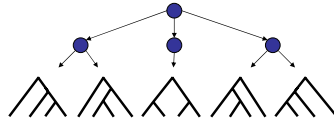
- Then all trees have equal probability initially.
- Therefore, after one round of EM, the posterior over trees will (in the absence of random perturbation) be approximately uniform over all trees, and symmetric over symbols.

Problem: "Uniform" Posteriors

Tree Uniform

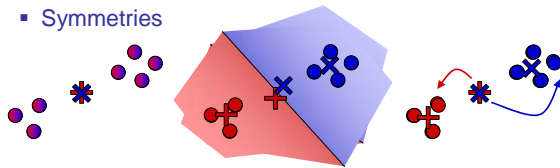


Split Uniform

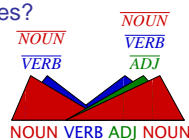
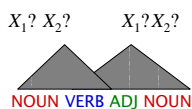


Problem: Model Symmetries

- Symmetries



- How does this relate to trees?



Other Approaches

- Evaluation: fraction of nodes in gold trees correctly posited in proposed trees (unlabeled recall)
- Some recent work in learning constituency:
 - [Adrians, 99] Language grammars aren't general PCFGs
 - [Clark, 01] Mutual-information filters detect constituents, then an MDL-guided search assembles them
 - [van Zaanen, 00] Finds low edit-distance sentence pairs and extracts their differences

Right-Branching Baseline

- English trees tend to be right-branching, not balanced



- A simple (English-specific) baseline is to choose the right chain structure for each sentence

van Zaanen, 00 | 35.6

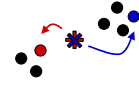
Desiderata: Practical Learnability

- To be practically learnable, models should:

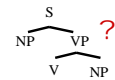
- Be as simple as possible

Congress NOUN ✓

- Make symmetries self-breaking whenever possible



- Avoid hidden structures which are not directly coupled to surface phenomena



Inspiration: Distributional Clustering

- ◆ the president said that the downturn was over ◆

president	the	of
president	the	said
governor	the	of
governor	the	appointed
said		sources
said	president	that
reported		sources

president
governor

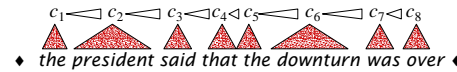
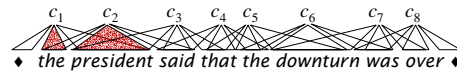
the
a

said
reported

[Finch and Chater 92, Shuetze 93, many others]

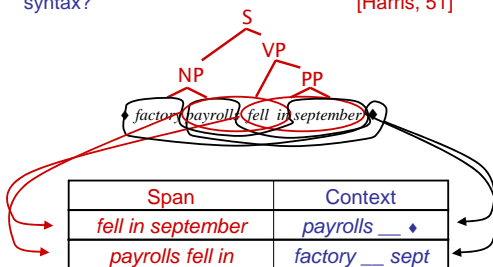
Distributional Models

$$P(S, C) = \prod_i P(c_i)P(w_i | c_i)P(w_{i-1}, w_{i+1} | c_i)$$



Idea: Distributional Syntax?

- Can we use distributional clustering for learning syntax? [Harris, 51]



Problem: Identifying Constituents

Distributional classes are easy to find...

the final vote
two decades
most people

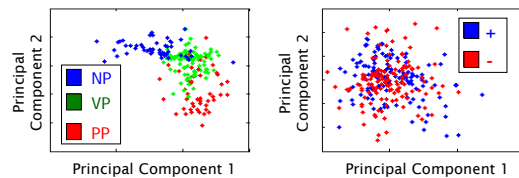
the final
the initial
two of the

of the
with a
without these

in the end
of the
for now

decided to
took most o
go with

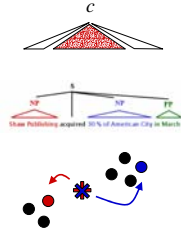
... but figuring out which are constituents is hard.



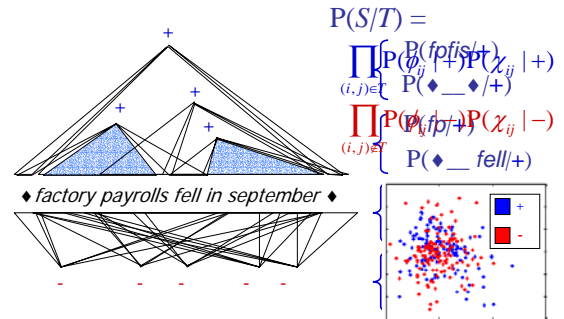
A Nested Distributional Model

- We'd like a model that:

- Ties spans to linear contexts (like distributional clustering)
- Considers only proper tree structures (like a PCFG model)
- Has no symmetries to break (like a dependency model)

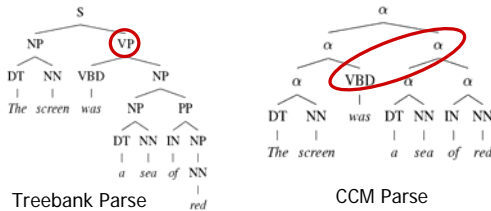


Constituent-Context Model (CCM)



Results: Constituency

Right-Branch	70.0	<div style="width: 70%; background-color: black;"></div>
--------------	------	--



Spectrum of Systematic Errors

CCM analysis better ← → Treebank analysis better

Analysis	Inside NPs	Possesives	Verb groups
CCM	<i>the [lazy cat]</i>	<i>John ['s cat]</i>	<i>[will be] there</i>
Treebank	<i>the lazy cat</i>	<i>[John 's] cat</i>	<i>will [be there]</i>
CCM Right?	Yes	Maybe	No

But the worst errors are the non-systematic ones! (~25%)

Results: Combined Models

Dependency Evaluation

Random	45.6	<div style="width: 45.6%; background-color: black;"></div>
DMV	62.7	<div style="width: 62.7%; background-color: blue;"></div>
CCM + DMV	64.7	<div style="width: 64.7%; background-color: blue;"></div>

Constituency Evaluation

Random	39.4	<div style="width: 39.4%; background-color: black;"></div>
CCM	81.0	<div style="width: 81.0%; background-color: red;"></div>
CCM + DMV	88.0	<div style="width: 88.0%; background-color: red;"></div>

- Supervised (but unannotated) PCFG constituency recall is at 92.8
- Qualitative improvements
 - Subject-verb groups gone, modifier placement improved

How General is This?

		Constituency Evaluation
English (7422 sentences)		
Random Baseline	39.4	<div style="width: 39.4%; background-color: black;"></div>
CCM+DMV	88.0	<div style="width: 88.0%; background-color: red;"></div>
German (2175 sentences)		
Random Baseline	49.6	<div style="width: 49.6%; background-color: black;"></div>
CCM+DMV	89.7	<div style="width: 89.7%; background-color: red;"></div>
Chinese (2473 sentences)		
Random Baseline	35.5	<div style="width: 35.5%; background-color: black;"></div>
CCM+DMV	46.7	<div style="width: 46.7%; background-color: red;"></div>
DMV	54.2	<div style="width: 54.2%; background-color: blue;"></div>
CCM+DMV	60.0	<div style="width: 60.0%; background-color: blue;"></div>

Dependency Evaluation

Most Common Errors: English

Overproposed Constituents

JJ NN	1022	<i>the [general partner]</i>
NNP NNP	447	<i>the [Big Board]</i>
DT NN	398	<i>[an import] order</i>
JJ NNS	294	<i>six million [common shares]</i>
NNS RB	164	<i>[seats currently] are quoted</i>

Crossing Constituents

CD CD IN CD CD	154	<i>rose to [# billion from # billion]</i>
NNS RB	133	<i>petroleum [prices also] surged</i>
NNP NNP NNP	67	<i>to [Hong Kong China] is</i>
JJ NN	66	<i>especially [strong growth]</i>

Most Common Errors: German

Overproposed Constituents

JJ NN	461	<i>der [Dalberger Hof]</i>
DT NN	430	<i>[die Moderatoren] der Zukunft</i>
DT JJ NN	94	<i>Aus [der erhofften Meisterschaft]</i>
CC NN	71	<i>Sinti [und Roma]</i>

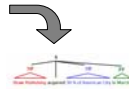
Crossing Constituents

JJ NN	30	<i>New [Yorker Aktienbörse]</i>
CD NN	18	<i>300 [000 Mark]</i>
NNP NNP	17	<i>Frankfurt A. [M. FR]</i>
IN NN	15	<i>[zwischen Schwips] und Kater</i>

Recap

Unsupervised learning:

- Constituency structure
- Dependency structure



Constituency recall:

Random Baseline	39.4	<div style="width: 39.4%; background-color: black;"></div>
CCM + DMV	88.0	<div style="width: 88.0%; background-color: red;"></div>
Supervised PCFG	92.8	<div style="width: 92.8%; background-color: black;"></div>

Why it works:

- Combination of simple models
- Representations designed for unsupervised learning