CS 294-5: Statistical Natural Language Processing

Grammar Induction
Dan Klein

Assignment 3 Honors

Idea: Lexical Affinity Models

- Words select other words on syntactic grounds

congress narrowly passed the amended bill

- 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]

congress narrowly passed the amended bill

Empirical   Uniform   *passed .... bill*

Problem: Non-Syntactic Affinity

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

congress narrowly passed the amended bill

expect brushbacks but no beanballs

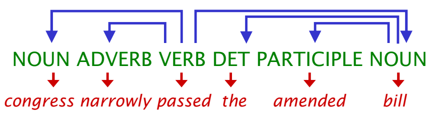
a new year begins in new york

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]

NOUN ADVERB VERB DET PARTICIPLE NOUN
congress narrowly passed the amended bill

A Word-Class Model

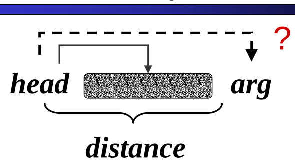


$$P(words, classes, graph) = P(length)\ \ P(graph)$$

Problems: Word Class Models

- Issues:
  - Too simple a model – doesn't work much better supervised
  - No representation of valence (number of arguments)

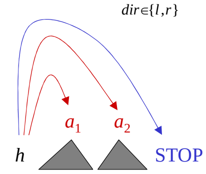
Issue: Local Representations



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:

$$P(t_h) = \prod_{dir \in \{l,r\}}$$



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:

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

Common Errors: Dependency

Overproposed Dependencies

DET ← N	3474
N-PROP ← N-PROP	2096
NUM → NUM	760
PREP ← DET	735
DET ← N-PL	696
DET → PREP	627
DET → V-PAST	470
DET → V-PRES	420

Underproposed Dependencies

DET → N	3079
N-PROP → N-PROP	1898
PREP ← N	838
N → V-PRES	714
DET → N-PL	672
N ← PREP	669
NUM ← NUM	54
N → V-PAST	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
zNN → NN | NNS
zNN → JJ zNN
zNN → zNN zNN

NP with determiner
zNP → DT zNN
zNP → PRP$ zNN

Proper NP
zNNP → NNP | NNPS
zNNP → zNNP zNNP

Transitive VPs (complementation)
zVP → zV JJ
zVP → zV zNP
zVP → zV zNN
zVP → zV zPP

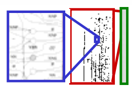
Transitive VPs (adjunction)
zVP → zRB zVP
ZVP → zVP zPP

PP
zPP → zIN zNN
zPP → zIN zNP
zPP → zIN zNNP

verb groups / intransitive VPs
zV → VBZ | VBD | VBP
zV → MD VB
zV → MD RB VB
zV → zV zRB
zV → zV zVBG

Intransitive S
zS → PRP zV
zS → zNP zV
zS → zNNP zV

Transitive S
zSt → zNNP zVP
zSt → zNN zVP
zSt → PRP zVP



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

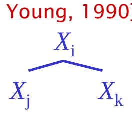
Issues with Chunk/Merge Systems

- 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]

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



  - 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:

$$P(X_c \mid i, j, S) = \frac{\sum_{T:(X_k, i, j) \wedge yield(T)=S} P(T)}{\sum_{T:yield(T)=S} P(T)}$$

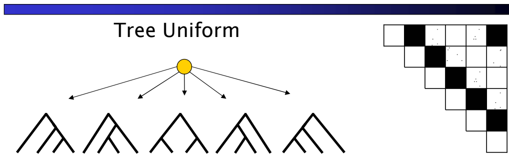
- Can calculate in cubic time with the Inside-Outside algorithm.
- Consider an initial grammar where all productions have equal weight:

$$P(X_a \; X_b \mid X_c) = 1/n^2$$

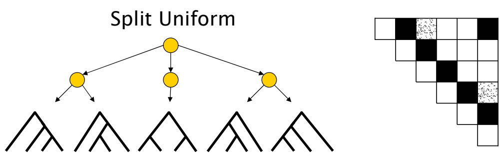
- Then all trees have equal probability initially.
- Therefore, after one round of EM, the posterior over trees will (in the absence of random perturbation) be approximately uniform over all trees, and symmetric over symbols.

Problem: "Uniform" Posteriors

Tree Uniform

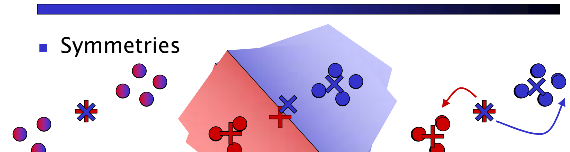


Split Uniform

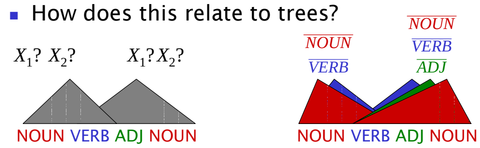


Problem: Model Symmetries

- Symmetries



- How does this relate to trees?

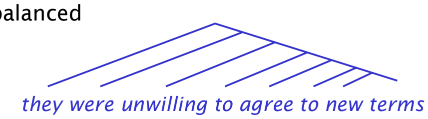


Other Approaches

- 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

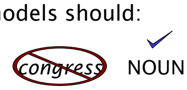
- English trees tend to be right-branching, not balanced



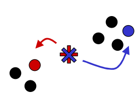
- A simple (English-specific) baseline is to choose the right chain structure for each sentence

Desiderata: Practical Learnability

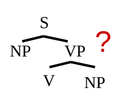
- To be practically learnable, models should:
  - Be as simple as possible



  - 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

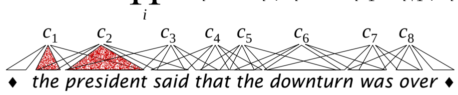
said reported

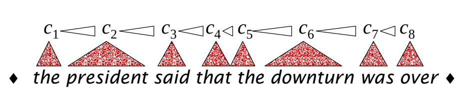
the a

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

Distributional Models

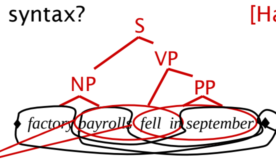
$$P(S,C) = \prod_i P(c_i)P(w_i \mid c_i)P(w_{i-1}, w_{i+1} \mid c_i)$$





Idea: Distributional Syntax?

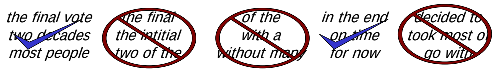
- Can we use distributional clustering for learning syntax? [Harris, 51]



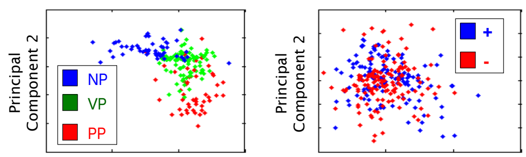
Span	Context
fell in september	*payrolls __ ♦*
payrolls fell in	*factory __ sept*

Problem: Identifying Constituents

Distributional classes are easy to find…

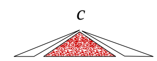


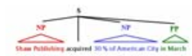
… but figuring out which are constituents is hard.

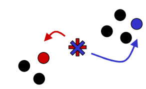


A Nested Distributional Model

- 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)



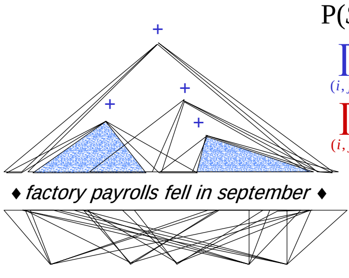


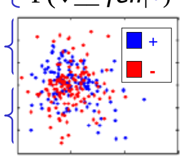


Constituent-Context Model (CCM)

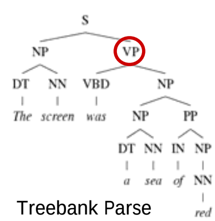
$$P(S|T) = \prod_{(i,j)\in T} P(\phi_{ij}\,|+)P(\chi_{ij}\,|+) \prod_{(i,j)\notin T} P(\phi_{ij}\,|-)P(\chi_{ij}\,|-)$$

P(fpfis|+)
P(♦__♦|+)
P(fp|−)
P(♦__ fell|+)

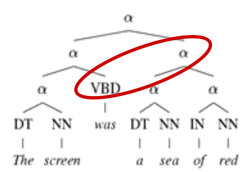




Results: Constituency

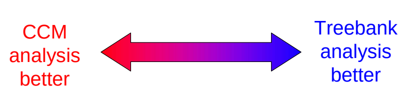


Treebank Parse



CCM Parse

Spectrum of Systematic Errors



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

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

Results: Combined Models

Dependency Evaluation

Random	45.6	
DMV	62.7	
CCM + DMV	64.7	

Constituency Evaluation

Random	39.4	
CCM	81.0	
CCM + DMV	88.0	

!

- *Supervised* 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	
CCM+DMV	88.0	
German (2175 sentences)		
Random Baseline	49.6	
CCM+DMV	89.7	
Chinese (2473 sentences)		
Random Baseline	35.5	
CCM+DMV	46.7	
DMV	54.2	
CCM+DMV	60.0	

Dependency Evaluation

Most Common Errors: English

Overproposed Constituents

ADJ N	1022	*the [general partner]*
N-PROP N-PROP	447	*the [Big Board]*
DET N	398	*[an import] order*
ADJ N-PL	294	*six million [common shares]*
N-PL ADV	164	*[seats currently] are quoted*

Crossing Constituents

NUM NUM PREP NUM NUM	154	*rose to [# billion from # billion]*
N-PL ADV	133	*petroleum [prices also] surged*
N-PROP N-PROP N-PROP	67	*to [Hong Kong China] is*
ADJ N	66	*especially [strong growth]*

Most Common Errors: German

Overproposed Constituents

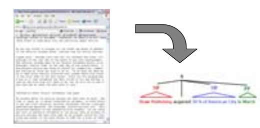
ADJ N	461	*der [Dalberger Hof]*
DET N	430	*[die Moderatoren] der Zukunft*
DET ADJ N	94	*Aus [der erhofften Meisterschaft]*
CONJ N	71	*Sinti [und Roma]*

Crossing Constituents

ADJ N	30	*New [Yorker Aktienborse]*
NUM NN	18	*300 [000 Mark]*
N-PROP N-PROP	17	*Frankfurt A. [M. FR]*
PREP N	15	*[zwischen Schwips] und Kater*

What's Been Accomplished?

- Unsupervised learning:
  - Constituency structure
  - Dependency structure



- Constituency recall:

Random Baseline	39.4	
CCM + DMV	88.0	
Supervised PCFG	92.8	

- Why it works:
  - Combination of simple models
  - Representations designed for unsupervised learning