Statistical NLP
Spring 2008

University of California
CAL NLP
Berkeley

Lecture 18: Grammar Induction

Dan Klein – UC Berkeley

Learnability

- Learnability: formal conditions under which a formal class of languages can be learned in some sense
- Setup:
  - Class of languages is $\mathscr{L}$
  - Learner is some algorithm H
  - Learner sees a sequence X of strings $x_1 \dots x_n$
  - H maps sequences X to languages L in $\mathscr{L}$
- Question: for what classes do learners exist?

Learnability: [Gold 67]

- Criterion: identification in the limit
  - A **presentation** of L is an infinite sequence of x in L in which each x occurs at least once
  - A learner H **identifies L in the limit** if for any presentation of L, from some point n onward, H always outputs L
  - A class $\mathscr{L}$ is **identifiable in the limit** if there is some H which correctly identifies in the limit any L in $\mathscr{L}$
- Theorem [Gold 67]: Any $\mathscr{L}$ which contains all finite languages and at least one infinite language (i.e. is superfinite) is unlearnable in this sense

Learnability: [Gold 67]

- Proof sketch
  - Assume $\mathscr{L}$ is superfinite
  - There exists a chain $L_1 \subset L_2 \subset \dots L_\infty$
  - Take any learner H assumed to identify $\mathscr{L}$
  - Construct the following misleading sequence
    - Present strings from L_1 until it outputs L_1
    - Present strings from L_2 until it outputs L_2
    - …
  - This is a presentation of L_∞, but H won't identify L_∞

Learnability: [Horning 69]

- Problem: IIL requires that H succeed on each presentation, even the weird ones
- Another criterion: **measure one identification**
  - Assume a distribution $P_L(x)$ for each L
  - Assume $P_L(x)$ puts non-zero mass on all and only x in L
  - Assume infinite presentation X drawn i.i.d. from $P_L(x)$
  - H measure-one identifies L if probability of drawing an X from which H identifies L is 1
- Note: there can be misleading sequences, they just have to be (infinitely) unlikely

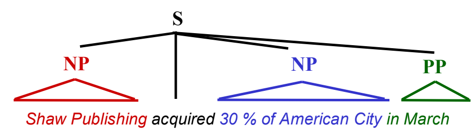
Learnability: [Horning 69]

- Proof sketch
  - Assume $\mathscr{L}$ is a recursively enumerable set of recursive languages (e.g. the set of PCFGs)
  - Assume an ordering on all strings $x_1 < x_2 < \dots$
  - Define: two sequences A and B **agree through n** if for all $x < x_n$, x in A $\Leftrightarrow$ x in B
  - Define the **error set** E(L,n,m):
    - All sequences such that the first m elements do not agree with L through n
    - These are the sequences which contain early strings outside of L (can't happen) or fail to contain all the early strings in L (happens less as m increases)
  - Claim: P(E(L,n,m)) goes to 0 as m goes to ∞
  - Let $d_L(n)$ be the smallest m such that $P(E) < 2^{-n}$
  - Let d(n) be the largest $d_L(n)$ in first n languages
  - Learner: after d(n) pick first L that agrees with evidence through n
  - Can only fail for sequence X if X keeps showing up in E(L,n,d(n)), which happens infinitely often with probability zero (we skipped some details)

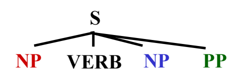
Learnability

- Gold's result says little about real learners (requirements of IIL are way too strong)
- Horning's algorithm is completely impractical (needs astronomical amounts of data)
- Even measure-one identification doesn't say anything about tree structures (or even density over strings)
  - Only talks about learning grammatical sets
  - Strong generative vs weak generative capacity

Context-Free Grammars



- Looks like a context-free grammar.
- Can model a tree as a collection of context-free rewrites (with probabilities attached).



$$P(\text{NP VERB NP PP} \mid \text{S}) = 0.1$$

Early Approaches: Structure Search

- Incremental grammar learning, chunking [Wolff 88, Langley 82, many others]
  - Can recover synthetic grammars
- An (extremely good / lucky) 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

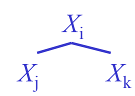
Transitive S
zSt → zNNP zVP
zSt → zNN zVP
zSt → PRP zVP

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

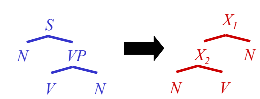
Idea: Learn PCFGs with EM

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

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

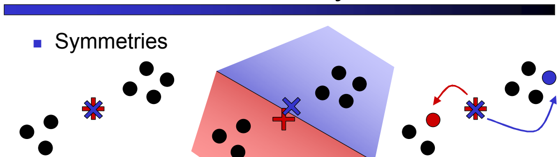


  - Full binary grammar over n symbols
  - Parse uniformly/randomly at first
  - Re-estimate rule expectations off of parses
  - Repeat
- Their conclusion: it doesn't really work.

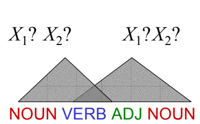


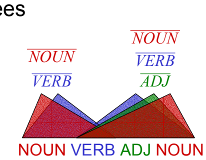
Problem: Model Symmetries

- Symmetries



- How does this relate to trees



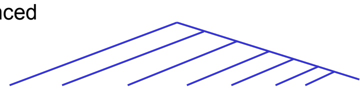


Other Approaches

- 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

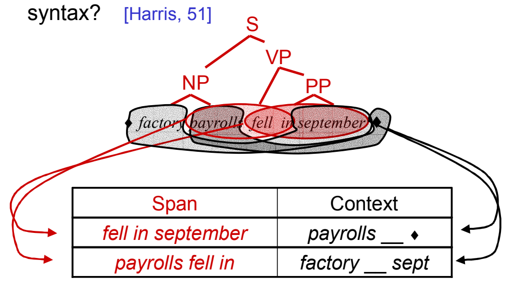


- A simple (English-specific) baseline is to choose the right chain structure for each sentence

van Zaanen, 00	35.6	

Idea: Distributional Syntax?

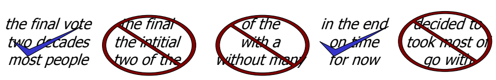
- Can we use distributional clustering for learning syntax? [Harris, 51]



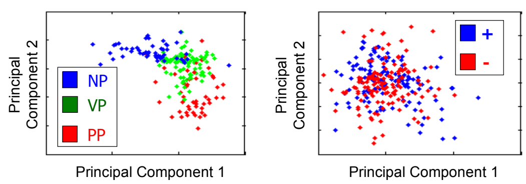
Span	Context
fell in september	*payrolls __ ♦*
payrolls fell in	*factory __ sept*

Problem: Identifying Constituents

Distributional classes are easy to find…

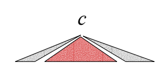


… but figuring out which are constituents is hard.

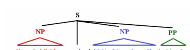


A Nested Distributional Model

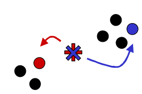
- We'd like a model that:
  - Ties spans to linear contexts (like distributional clustering)



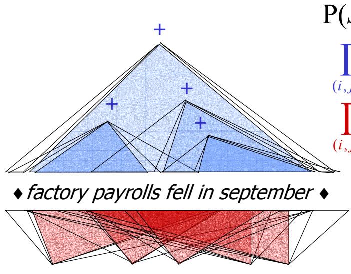
  - Considers only proper tree structures (like a PCFG model)



  - Has no symmetries to break (like a dependency model)

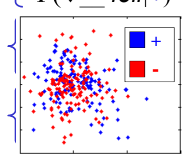


Constituent-Context Model (CCM)



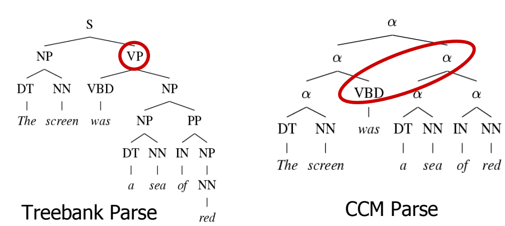
$$P(S|T) = \prod_{(i,j)\in T} P(\phi_{ij}|+)P(\chi_{ij}|+) \prod_{(i,j)\notin T} P(\phi_{ij}|-)P(\chi_{ij}|-)$$

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



Results: Constituency

Right-Branch	70.0	



Treebank Parse

CCM Parse

Spectrum of Systematic Errors

CCM analysis better ⟷ Treebank analysis better

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%)

Syntactic Parsing

- Parsing assigns structures to sentences.

Shaw Publishing acquired *30 % of American City in March*

- Dependency structure gives attachments.

WHEN

WHO WHAT

Shaw Publishing acquired *30 % of American City in March*

Idea: Lexical Affinity Models

- Words select other words on syntactic grounds

congress narrowly passed the amended bill

- 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

Method	Accuracy
Paskin, 2001	39.7

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]

congress narrowly passed the amended bill

Problems: Word Class Models

Random	41.7	
Carroll and Charniak, 92	44.7	

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

congress narrowly passed the amended bill

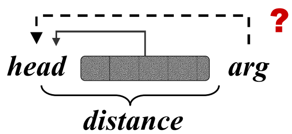
NOUN NOUN VERB

stock prices fell

NOUN NOUN VERB

stock prices fell

Local Representations

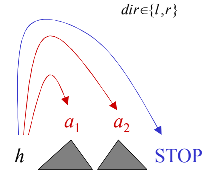


	Classes?	Distance	Local Factor
Paskin 01	✗	✗	P(a | h)

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:

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



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

Results: Dependencies

Adjacent Words	55.9	
Our Model (DMV)	62.7	

- Situation so far:
  - Task: unstructured text in, word pairs out
  - Previous results were below baseline
  - We modeled word classes [*cf.* Carroll & Charniak 92]
  - We added a model of distance [*cf.* Collins 99]
  - Resulting model is substantially over baseline
  - … but we can do much better

Results: Combined Models

Dependency Evaluation (Undir. Dep. Acc.)

Random	45.6	
DMV	62.7	
CCM + DMV	64.7	

Constituency Evaluation (Unlabeled Recall)

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

Apartment hunting

> **$1925 / 4br - Union City- 4br/2ba 2-story house w/Recent Upgrades, Gardener Included**
>
> **(fremont / union city)**
>
> Reply to: anon-87179918@craigslist.org
> Date: 2005-07-26, 10:00PM PDT
>
> Spacious, light and airy four-bedroom, two-bath house located in cul-de-sac. Perfect for entertaining guests!! Great, friendly neighbors and minutes from Union City BART, shopping centers, 880, 84, and Quarry Lakes Regional Park. Located in Union City, the lot size is approximately 5200 sq. ft with approximately 1500 sq ft of living space. Come take a look at your new home!
>
> Amenities include:
> Central Heating, no A/C
>
> Large, professionally maintained yard w/fruitbearing trees
>
> Two car garage (new garage door and opener)
>
> Living room with wood burning fireplace
>
> Dining room - great view of the backyard

- Craigslist.org classified ads
- Would like search on attributes
- Can't, because listings are largely unstructured
- Need to structure them automatically

Classified advertisements

Size Contact Terms Location Features

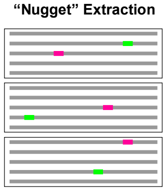
Duplex - Newly remodeled 2 Bdrms/1 Bath, spacious upper unit, located in Hilltop Mall area. Walking distance to shopping, public transportation, schools and park. Paid water and garbage, carport and plenty of street parking. Washer and dryer are provided. Private patio yard, view. Contact number (510) 691-9419, (510) 464-6581, (510) 724-6988.

Spacious 2 bd/1 ba top floor unit available now in Kentfield. Complex is located withen walking distance of many small shops and businesses. Tenants are entitled to parking, use of laundry facilites, and access to the roof top patio. This unit is available now on a 1-year lease. Monthly rent is $1147, with a security deposit of $1000.00. Cats and non-barking dogs are welcome with an additional deposit. Please call us at 456-4044.

182 Echo AVE#1, Great Campbell location, front unit 3 bedrooms, 2 full baths with new carpet and paint, patio, POOL, one car carport, laundry in the building, water and garbage incldued, available now, deposit is also $1395, contact TALI (408) 489-7149, 182 Echo Ave #1

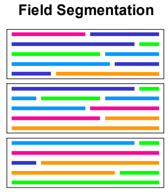
Types of IE problems

"Nugget" Extraction



- Document is mostly background text
- Information "nuggets" are defined extrinsically by the task

Field Segmentation



- Document consists entirely of a sequence of fields
- Fields are a salient and intrinsic form of structure
- Seems suitable for unsupervised learning!

Related IE Work

- Supervised field segmentation
  - McCallum et al. (1999) - HMMs for parsing citations
  - McCallum et al. (2000) - MEMMs for parsing FAQs
  - Peng and McCallum (2004) - CRFs for parsing paper headers
- Unsupervised field segmentation
  - Hearst (1997) - "TextTiling"
  - Blei and Moreno (2001) - "Aspect HMM"
  - Pasula et al. (2002) - Unsupervised citation parsing as part of a large model of "identity uncertainty"
  - Barzilay and Lee (2004) - "Content models"

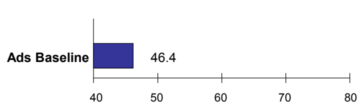
Data and Evaluation

Classified Ads

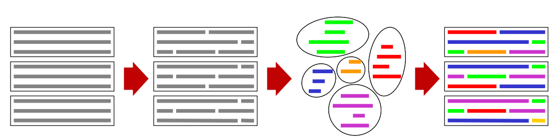
- Novel corpus
- 8767 unique rental listings collected from craigslist.org in June 2004
- 302 listings are annotated with 12 fields, including *size*, *rent*, *contact*, etc.
- Average listing has 119 tokens in 9 fields

Bibliographic Citations

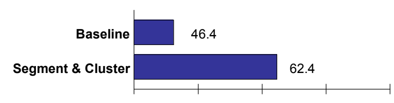
- Described in McCallum et al. (1999)
- 500 citations collected from 500 academic papers
- All are annotated with 13 fields, including *author*, *title*, *journal*, etc.
- Average citation has 35 tokens in 6 fields



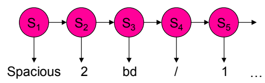
Segment and cluster



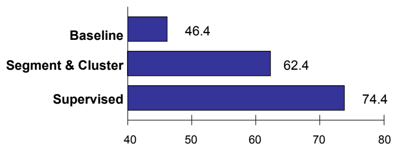
- Crude segmentation & EM clustering improve upon baseline
- We can do better: simultaneous segmentation and clustering!



Hidden Markov Models

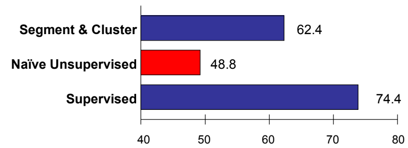


$$\mathrm{P}(\mathbf{s},\mathbf{w}) = \prod_i \mathrm{P}(s_i|s_{i-1})\mathrm{P}(w_i|s_i)$$

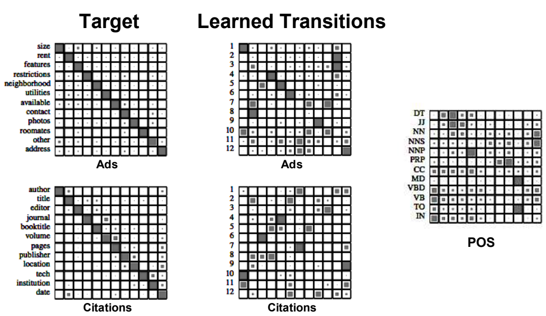


Unsupervised learning

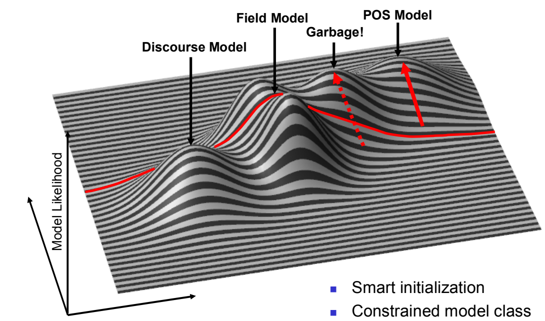
- Standard unsupervised learning in HMMs:
  - EM, with Baum-Welch for computing E-step
  - Fixed number of states (equal to number of fields)
  - Uniform initialization of transition model
  - Near-uniform initialization of emission model
- Performs terribly:



What went wrong?



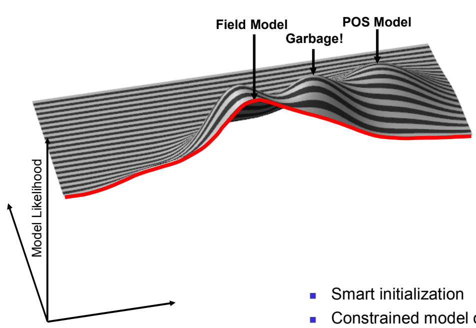
What's being learned?



HMM Parameterizations

- Smart initialization
- Constrained model class

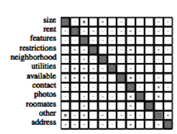
What's being learned?



HMM Parameterizations

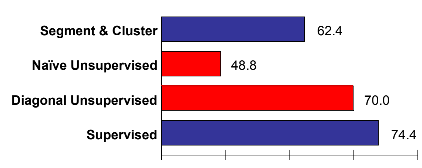
- Smart initialization
- Constrained model class

Diagonal Transition Structure



Self-loop probability

$$\mathrm{P}(s_i|s_{i-1}) = \begin{cases} \sigma + \frac{(1-\sigma)}{|S|} & \text{if } s_i = s_{i-1} \\ \frac{(1-\sigma)}{|S|} & \text{otherwise} \end{cases}$$

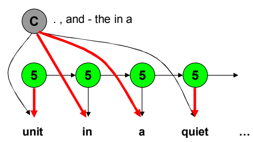


What's still wrong?

Learned Emission Model

1	. $ no ! month deposit , pets rent available
2	, . room and with in large living kitchen -
3	. a the is and for this to , in
4	[NUM2] [NUM1] , bedroom bath / - . car garage
5	, . and a in - quiet with unit building
6	- . [TIME] [PHONE] [DAY] call [NUM8] at

$$\mathrm{P}_h(w_i|s_i) = \alpha \mathrm{P}_c(w_i) + (1-\alpha)\mathrm{P}(w_i|s_i)$$

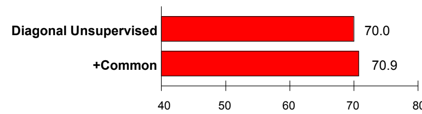


Common word model

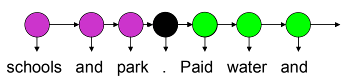
Learned Emission Model

1	. $ no ! month deposit , pets rent available
2	, . room and with in large living kitchen -
3	. a the is and for this to , in
4	[NUM2] [NUM1] , bedroom bath / - . car garage
5	, . and a in - quiet with unit building
6	- . [TIME] [PHONE] [DAY] call [NUM8] at

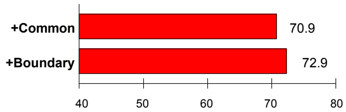
1	[NUM2] bedroom [NUM1] bath bedrooms large sq car ft garage
2	$ no month deposit pets lease rent available year security
3	kitchen room new , with living large floors hardwood fireplace
4	[PHONE] call please at or for [TIME] to [DAY] contact
5	san street at ave st # [NUM3] francisco ca [NUM4]
6	of the yard with unit private back a building floor
C	[NEWLINE] . , and - the in a / is with : of for to



Boundary model

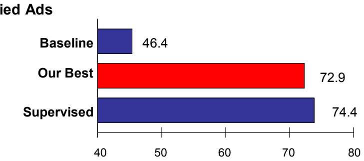


- In data, boundaries are salient, but no representation of boundaries in our model
- Add a boundary state, which emits boundary tokens
- Modify fixed transition function so that fields prefer to end with boundary state
- Boosts accuracy:



Summary of results

Classified Ads



Bibliographic Citations