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 $\mathcal{C}$ is identifiable in the limit if there is some $H$ which correctly identifies $L$ in the limit any $L$ in $\mathcal{C}$.
- Theorem [Gold 67]: Any $\mathcal{C}$ which contains all finite languages and at least one infinite language (i.e. is superfinite) is unlearnable in this sense.

Learnability: [Horning 69]

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

Learnability

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

Learnability: [Gold 67]

- Proof sketch
  - Assume $\mathcal{C}$ is superfinite.
  - There exists a chain $L_1 \subseteq L_2 \subseteq \ldots L_n$.
  - Take any learner $H$ assumed to identify $\mathcal{C}$.
  - 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 $L_n$, but $H$ won't identify $L_n$.

Learnability: [Horning 69]

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

- Gold’s result says little about real learners (requirements of IIL are too strong)
- Horn’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

Early Approaches: Structure Search

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

  
  |  
  |---
  | N-non zero determiner NP  
  | xNN -> NN | xNN  
  | xNN -> xNN xNN  
  | NP with determiner  
  | xNP -> DT xNN  
  | xNP -> PREP xNN  
  | Proper NP  
  | xNNP -> xNNP | xNNP  
  | xNNP -> xNNP xNNP  

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

Context-Free Grammars

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

Idea: Learn PCFGs with EM

- Classic experiments on learning PCFGs with Expectation-Maximization [Lan and Young, 1990]
  
  \[
  \{ X_1, X_2, \ldots, X_n \} \rightarrow X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_n
  \]

  - 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

Idea: Distributional Syntax?
- Can we use distributional clustering for learning syntax? [Harris, 51]

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)
- \[ P(S|T) = \prod_{(i,j) \in T} P(f \mid f_i \in +) P(\bullet \in +) \prod_{(i,j) \in T} P(f \mid f_i \in -) P(\bullet \in -) \]

Results: Constituency
Spectrum of Systematic Errors

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Inside NPs</th>
<th>Possessives</th>
<th>Verb groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCM</td>
<td>the [lazy cat]</td>
<td>John [s cat]</td>
<td>will be there</td>
</tr>
<tr>
<td>Treebank</td>
<td>the lazy cat</td>
<td>[John's cat]</td>
<td>will [be there]</td>
</tr>
<tr>
<td>CCM Right?</td>
<td>Yes</td>
<td>Maybe</td>
<td>No</td>
</tr>
</tbody>
</table>

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

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, 1996]: Greedy linkage
  - [Paskin, 2001]: Iterative re-estimation with EM
- Evaluation: compare linked pairs to a gold standard

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paskin, 2001</td>
<td>39.7</td>
</tr>
</tbody>
</table>

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

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]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>congress narrowly passed the amended bill</td>
<td></td>
</tr>
</tbody>
</table>

Problems: Word Class Models

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>41.7</td>
</tr>
<tr>
<td>Carroll and Charniak, 92</td>
<td>44.7</td>
</tr>
</tbody>
</table>

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

\[
P(t_b) = \prod_{d \in \{l,r\}} P(c \mid a_b)
\]

Results: Dependencies

- 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

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

How General is This?

- English (7422 sentences)
  - Random Baseline 39.4
  - CCM+DMV 88.0
- German (2175 sentences)
  - Random Baseline 49.6
  - CCM+DMV 69.7
- Chinese (2473 sentences)
  - Random Baseline 36.5
  - CCM+DMV 46.7
  - DMV 54.2
  - CCM+DMV 60.0
Apartment hunting

- 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 Bedroom/1 Bath, spacious upper unit, located in Hilltop Mall area. Walking distance to shopping, public transportation, schools and park. Paid water and garbage, carpert 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 within walking distance of many small shops and businesses. Tenants are entitled to parking, use of laundry facilities, 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 included, available now, deposit is also $1395, contact TALI (408) 490-7149, 182 Echo Ave #1

Types of IE problems

- "Nugget" Extraction
- Field Segmentation

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

- 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
- Bibliographic Citations

- 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

- 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

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?

Diagonal Transition Structure

What’s being learned?

HMMCParameterizations

Model Likelihood

Model Likelihood

HMM Parameterizations

HMM Parameterizations

HMM Parameterizations

HMM Parameterizations
What's still wrong?

**Learned Emission Model**

\[ P_h(w_i | s_i) = \alpha P_c(w_i) + (1 - \alpha) P(w_i | s_i) \]

- $\alpha$: no 1/month deposit, pets rent available
- $\alpha$: room with in large living room
- $\alpha$: the is and in the is in
- $\alpha$: (NUM, NUM1): bedroom-large sq car garage
- $\alpha$: unit in a quiet with unit building
- $\alpha$: (TIME, PHONE) [DAY] call (NUM)

**Boundary model**

- Schools and park.
- Paid water and

- 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:
  - +Common: 70.9
  - +Boundary: 72.9

**Common word model**

**Learned Emission Model**

- $\alpha$: $\text{no 1/month deposit}$, pets rent available
- $\alpha$: room with in large living room
- $\alpha$: the is and for this in
- $\alpha$: (NUM, NUM1): bedroom-large sq car garage
- $\alpha$: unit in a quiet with unit building
- $\alpha$: (TIME, PHONE) [DAY] call (NUM)

**Summary of results**

**Classified Ads**

- Baseline: 46.4
- Our Best: 72.9
- Supervised: 74.4

**Bibliographic Citations**

- Baseline: 27.9
- Our Best: 68.2
- Supervised: 72.5