Learnability: [Gold 67]

- Criterion: identification in the limit
  - A presentation of $\mathcal{L}$ is an infinite sequence of $x$ in $\mathcal{L}$ in which each $x$ occurs at least once
  - A learner $H$ identifies $\mathcal{L}$ in the limit if for any presentation of $\mathcal{L}$ from some point $n$ onward, $H$ always outputs $\mathcal{L}$
  - A class $\mathcal{L}$ is identifiable in the limit if there is some $H$ which correctly identifies in the limit any $\mathcal{L}$ in $\mathcal{L}$

- Theorem [Gold 67]: Any $\mathcal{L}$ 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 $L$
    - Assume infinite presentation $X$ drawn i.i.d. from $P_i(x)$
    - $H$-measure-one identifies $\mathcal{L}$ if probability of drawing an $X$ from which $H$ identifies $\mathcal{L}$ is 1

- Note: there can be misleading sequences, they just have to be (infinitely) unlikely

Learnability: [Horning 69]

- Proof sketch
  - Assume $\mathcal{L}$ is recursively enumerable set of recursive languages (e.g. the set of PCFGs)
  - Assume an ordering on all strings $x_1 < x_2 < ...$
  - Define: two sequences $A$ and $B$ agree through $n$ if for all $x < n$, $x \in A$ $\iff$ $x \in B$
  - Define the error set $E(L,n,m)$
    - All sequences such that the first $m$ elements do not agree with $L$ through $n$
    - These are the sequences which contain early strings outside of $L$ (can't happen)
    - 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 $\infty$
    - Let $d(n)$ be the smallest $n$ such that $P(E) < 2^{-n}$
    - Let $d(n)$ be the largest $d(n)$ 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 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).

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

**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
  - 0NN --> 02 | 0NN
  - 0NN --> 0NN 0NN
  - NP with determiner
  - 0NP --> DT 0NN
  - Proper NP
  - 0NP --> NNP \ NNP

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

**Idea: Learn PCFGs with EM**

- Classic experiments on learning PCFGs with Expectation-Maximization [Lang 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

  $$X_1 \sim X_2 \sim X_3 \sim X_4$$

  $$X_1 \sim X_6$$

- NOUN VERB ADJ NOUN
- NOUN VERB ADJ NOUN

**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 Zaanten, 00] Finds low edit-distance sentence pairs and extracts their differences
Right-Branching Baseline

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

they were unwilling to agree to new terms

- 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]

<table>
<thead>
<tr>
<th>Span</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>fell in september</td>
<td>payrolls ●</td>
</tr>
<tr>
<td>payrolls fell in</td>
<td>factory ● sept</td>
</tr>
</tbody>
</table>

Problem: Identifying Constituents

Distributional classes are easy to find...

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

Constituent-Context Model (CCM)

\[
P(S|T) = \prod_{(i,j) \in T} P(\bullet) P(\text{factory payrolls fell in september})
\]

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)

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%)*

**Syntactic Parsing**

- Parsing assigns structures to sentences.
- Dependency structure gives attachments.

**Idea: Lexical Affinity Models**

- Words select other words on syntactic grounds
- 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

<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.

**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]

**Problems: Word Class Models**

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</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)
Local Representations

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

\[
P(t_k) = \prod_{d=r}^{s} P(a | h)
\]

Common Errors: Dependency

<table>
<thead>
<tr>
<th>Overproposed Dependencies</th>
<th>Underproposed Dependencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>DET ← N</td>
<td>DET → N</td>
</tr>
<tr>
<td>N-PROP ← N-PROP</td>
<td>N-PROP → N-PROP</td>
</tr>
<tr>
<td>NUM → NUM</td>
<td>PREP ← N</td>
</tr>
<tr>
<td>PREP ← DET</td>
<td>N → V-PRES</td>
</tr>
<tr>
<td>DET ← N-PL</td>
<td>DET → N-PL</td>
</tr>
<tr>
<td>DET → PREP</td>
<td>DET → V-PAST</td>
</tr>
<tr>
<td>DET → V-PAST</td>
<td>DET → V-PAST</td>
</tr>
<tr>
<td>DET → V-PAST</td>
<td>DET → V-PAST</td>
</tr>
</tbody>
</table>

Results: Dependencies

- Situati...
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 Bdrm/1 Bath, spacious upper unit, located in Hilltop Mall area. Walking distance to shopping, public transportation, schools and park. Paid water and garbage, carpent and plenty of street parking. Washer and dryer are provided. Private patio yard, view. Contact number (510) 691-8419, (510) 444-6581, (510) 724-6988.

- Spacious 2 bd/1 ba top floor/ unit available now in Kentfield Complex. 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, one car carport, laundry in the building, water and garbage included, 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!

Baseline: 46.4
Segment & Cluster: 62.4
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?

HMM Parameterizations

Diagonal Transition Structure
What’s still wrong?

Learned Emission Model

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

Common word model

Learned Emission Model

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:

Boundary + Common

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