Lecture 1: Introduction
Dan Klein – UC Berkeley

Overview
- Learning linguistic structure
  - Probabilistic models
  - Corpus based methods
  - Models vs. procedures
- What can be learned?
- What does learning mean?
- Specific models which induce nontrivial structure
  - Not just syntax: lexicons, coreference, semantics…

Syllabus
- Introduction
- Lexicons
- Grammars I-IV
- Coreference
- Semantics

Course Details
- Requirements:
  - Come to lecture and participate
  - Some readings on web page
  - Some very short assignments TBD
  - Short final exam
- Adjacent courses:
  - Chris Manning’s parsing class will cover details of tree selection algorithms that I’ll skip
  - Roger Levy’s psycholinguistics class will cover cognitive issues that I’ll skip

Background
- I’ll assume you all know:
  - Probabilities / conditional probabilities
  - Bayes’ rule
  \[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \]
  - Conditional independence
    \[ P(X,Y) = P(X)P(Y) \text{ if } X \perp Y \]
    \[ P(X,Y|Z) = P(X|Z)P(Y|Z) \text{ if } X \perp Y|Z \]
  - Basic algorithmic methods (find best parses, maintain counts over a corpus, etc.)
  - If not, I’ll try to keep the discussion accessible

Today
- Getting warmed up:
  - Simple models of superficial structure
    - Phoneme classes
    - Word segmentation
    - Translation lexicons
    - Word classes
  - Discussion of general methodology
- What is learning and what can be learned?
Example: Phoneme Classes

- Problem: learn natural classes of phonemes
  - A start: can we distinguish vowels from consonants?
- Data: a transcribed corpus:
  
  ```plaintext
  i eat ice cream
  AY IT AY S K R I Y M
  ...```
- Probabilistic approach: describe the data using a model in which we’re looking for is an unobserved variable, fit the model

A (Too) Simple Model

- Distributional clustering: surfacy but often robust
- Treat each token as a (phoneme, context) pair

![Diagram](image)

- Assume a hidden variable which mediates as a hidden cause between phoneme and context
- Find the settings of hidden variables and model parameters given the data (objective varies)

A Simple Procedure

- Alternating optimization
  - Label each phoneme token as class 0 or 1 randomly
  - Now that the hidden variables are “observed,” update the parameter estimates
  - Relabel the phonemes using the new parameters
  - Repeat
- This is the hard expectation-maximization algorithm (hard EM)
  - Maximizes the likelihood of the data
  - Has some guarantees of convergence
  - Won’t always find a globally optimum solution
  - Almost certainly not how humans learn

Results

- So what do you get?
  - Two classes: (parameters not categorical)
    - CLASS 1 = \{AH, IH, EH, AA, AE, IY, OW, EY, ER, AO, AY, ...\}
  - Four classes:
    - CLASS 1 = \{M_N, T_N, R_N, S_N, K_N, SH_N, K_L, L_N, ...\}
    - CLASS 3 = \{AH_T, IH_AH, AH_AH, AH_Z, AH_IH, IH_T, EH_T, ...\}
    - CLASS 4 = \{N_AH, EH_AH, N_ER, EY_AH, N_IH, K_AH, K_IH, ...\}
- Not so impressive! What happened?

Model Assumptions

- What can go wrong?
  - Data: e.g., unrepresentative or insufficient data
  - Learning: e.g., bad luck in parameter search
  - Modeling: e.g., Poor model assumptions
- Here: at least a modeling problem
  - Model ignores acoustics and only captures phonotactics; also contexts and phonemes symmetric
  - Contexts:
    - CLASS 1 = \{M_N, T_N, R_N, S_N, K_N, SH_N, K_L, L_N, ...\}
    - CLASS 3 = \{AH_T, IH_AH, AH_AH, AH_Z, AH_IH, IH_T, EH_T, ...\}
    - CLASS 4 = \{N_AH, EH_AH, N_ER, EY_AH, N_IH, K_AH, K_IH, ...\}

Improving the Model

- For better results, need a model which incorporates acoustics and does not treat phonemes and contexts symmetrically
- Model from [Petrov et al. 07] (note: different data setup!)
**Phoneme Substructure**

- Mediates context and acoustics
- Learns subdivisions of phonemes, not the same as natural classes
- However, subphone transition structure reveals phone groupings

**More Phoneme Examples: /IH/**

- Several natural classes are evident, but vary by context
- These models were used to improve ASR phone classification; the degree of structure identifiability is really just a bonus...

**Learning**

- This course will focus on models and results, not learning...
- ... but
  - Always a learning phase: hidden variables and parameters are inferred from data
  - One major method for learning in probabilistic models is the expectation-maximization algorithm (EM), an alternating procedure which alternately guesses at hidden variables and updates parameters
  - Updating hidden variables often requires special dynamic programs to be efficient (and sometimes is not efficient); see Chris Manning’s parsing course for examples
  - Guesses about hidden variables and/or parameters can be hard or soft (hard or soft EM, variational Bayes), and there are other options (sampling, gradient methods)

**K-Means**

- The prototypical iterative clustering algorithm (a special case of hard EM)
  - Pick K random points as cluster centers (means)
  - Alternate:
    - Assign data instances to closest mean
    - Assign each mean to the average of its assigned points
  - Stop when no points’ assignments change
  - Why should this work?

**K-Means Example**

**Philosophy, Delayed**

- Nativism vs. Empiricism?
- A statistical learning view:
  - No tabula rasa learning
  - Models explicit about hypothesis space
  - Models explicit about biases (and priors)
  - Models tacit about where the biases come from
  - Generally, these models assume less than UG advocates, but do make assumptions!
  - Big step from computational model to any cognitive assertions
  - Even bigger step to philosophical assertions
### Word Segmentation

- **Problem:** segment a character sequence into words

\[ \text{AY IY T AY S K R IY M} \]

- **(Unigram) Model:** choose number of words, then generate each word independently [Venkataraman, 01]

\[ P(w) = P(|w|) \prod_i P(w_i) \]

\[ P(c) = \sum P(w) \]

- **Example:**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Parameters and Learning

- **The real parameters are** \( P(w) \)
  - Can learn with EM variants
  - E.g. find best segmentations, freeze segmentations, and read of \( P(w) \)
  - Length exponent can also be learned

- **What will be learned?**
  - Answer: with maximum likelihood parameters, each utterance is a single word (why?)
  - What really failed here?

### Priors over the Lexicon

- **To fix:** need a non-uniform prior on parameters \( P(w) \)
  - Common situation: need a prior, but how informative?
  - Limiting case: we use a totally uniform prior and the model has some defective behavior

- **Solution in [Goldwater et al., 2006], Dirichlet process priors**

\[ P(w') = \pi P(\theta') \prod_i P(w_i') \prod_j P(w_i') \]

- **DP priors** are a high-powered tool, but only one aspect is really important here (the base distribution)
  - The only model we’ll discuss with such priors, so I’ll skip the math

### Unigram Segmentations

<table>
<thead>
<tr>
<th><strong>Unigram DP Segmentations</strong></th>
<th><strong>Correct Segmentations</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>yu want tu si D6Bk</td>
<td>yu want tu si D6 b6k</td>
</tr>
<tr>
<td>1UK Dz 6 b7 wIt hiz hEt</td>
<td>1UK Dz 6 b7 wIt hiz hEt</td>
</tr>
<tr>
<td>&amp;nd 6dOgi</td>
<td>&amp;nd 6dOgi</td>
</tr>
<tr>
<td>yu want tu 1UK&amp;tDis</td>
<td>yu want tu 1UK &amp;t DIs</td>
</tr>
<tr>
<td>1UK&amp;tDis</td>
<td>1UK &amp;t DIs</td>
</tr>
<tr>
<td>hvv6 dRInK</td>
<td>hvv 6 dRInK</td>
</tr>
<tr>
<td>oke nQ</td>
<td>oke nQ</td>
</tr>
<tr>
<td>WAtsDIs</td>
<td>Wats DIs</td>
</tr>
<tr>
<td>WatsDkt</td>
<td>Wats Dkt</td>
</tr>
<tr>
<td>WAtiZIt</td>
<td>WAtiZ It</td>
</tr>
<tr>
<td>1UK ksexyu tek ItQt</td>
<td>1UK ksexyu tek It Qt</td>
</tr>
<tr>
<td>tek ItQt</td>
<td>tek It Qt</td>
</tr>
</tbody>
</table>

*From [Goldwater 2006]*

### Unigram vs. Bigram

- **Problem:** model undersegments

- **Cause:** only way to explain correlations between phonemes is word-externally

\[ \text{yuwant tu si D6Bk} \]

1UK ksexyu tek ItQt
tek ItQt

- **Solution:** bigram model, words chosen conditionally on previous words (the lexicon prior changes accordingly)

\[ P(w') = \pi P(\theta') \prod_i P(w_i') \prod_j P(w_i') \]

- **Now, there’s two modeled causes of correlation (inside one word or across two correlated ones)**

### Bigram Segmentations

<table>
<thead>
<tr>
<th><strong>Unigram Segmentations</strong></th>
<th><strong>Bigram Segmentations</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>yu want tu si D6Bk</td>
<td>yu want tu si D6 b6k</td>
</tr>
<tr>
<td>1UK Dz 6 b7 wIt hiz hEt</td>
<td>1UK Dz 6 b7 wIt hiz hEt</td>
</tr>
<tr>
<td>&amp;nd 6dOgi</td>
<td>&amp;nd 6dOgi</td>
</tr>
<tr>
<td>yu want tu 1UK&amp;tDis</td>
<td>yu want tu 1UK &amp;t DIs</td>
</tr>
<tr>
<td>1UK&amp;tDis</td>
<td>1UK &amp;t DIs</td>
</tr>
<tr>
<td>hvv6 dRInK</td>
<td>hvv 6 dRInK</td>
</tr>
<tr>
<td>oke nQ</td>
<td>oke nQ</td>
</tr>
<tr>
<td>WAtsDIs</td>
<td>Wats DIs</td>
</tr>
<tr>
<td>WatsDkt</td>
<td>Wats Dkt</td>
</tr>
<tr>
<td>WAtiZIt</td>
<td>WAtiZ It</td>
</tr>
<tr>
<td>1UK ksexyu tek ItQt</td>
<td>1UK ksexyu tek It Qt</td>
</tr>
<tr>
<td>tek ItQt</td>
<td>tek It Qt</td>
</tr>
</tbody>
</table>

*From [Goldwater 2006]*
Some Cognitive Evidence

- Distributional cues influence segmentation of audio streams by children [Saffran et al., 1996]
  - Their evidence is very compatible with these kinds of models, but is analyzed in a more procedural way
  - They also analyze other possible statistical cues
  - Roger Levy will be discussing their results and the relation to human child language segmentation in his class on psycholinguistics

Philosophy, Briefly

- Chomsky famously argued against:
  - Statistical models of grammaticality (on capacity grounds: "colorless green ideas..."
  - General purpose learning algorithms for language acquisition (on poverty of stimulus grounds)
- We won't recap these debates in a concentrated way
- But, he was really arguing against bad statistical models and learning algorithms (at least in the case of the green ideas, he argued against bigram models of syntax)
- Do capacity arguments or poverty of stimulus arguments apply to better models? You can decide for yourselves...
- Whatever you decide:
  - Statistical models make biases and assumptions explicit
  - They can be tested empirically (though this testing can be tricky)

Language Models

- We often want to place distributions over sentences
  - Think of these models as soft measures of fluency
  - Distinguish between the idea of a distribution over sentences and the particular ones we end up discussing
- Classic solution: n-gram models (we saw variants today)
  \[ P(w) = \prod_i P(w_i|w_{i-1} \ldots w_{i-k}) \]
- N-gram models are (weighted) regular languages
  - Natural language is not regular (of course!)
  - ... though you'd be surprised at what 5+ gram models trained on enough data can do
- This is a crude (and often useful in system building) model, but there are also language models with more linguistically plausible structure, e.g. PCFGs (we'll talk about PCFGs in detail soon)

Language Model Samples

- Unigram:
  - [fifth, an, of, futures, the, an, incorporated, a, a, the, inflation, most, quarter]
  - [that, or, limited, the]
  - [after, any, on, consistently, hospital, lake, of, of, other, and, factors, raised, analyst, too, allowed, really, never consider, full, bungled, festival, time, obtain, price, line, the, to, sale, the, free, further, band, a, details, matches, ...]
- Bigram:
  - [outside, new, car, parking, lot, of, the, agreement, reached]
  - [although, common, shares, rose, forty, six, point, four, hundred, dollars, from, thirty, seconds, at, the, greatest, play, disingenuous, to, be, reset, annually, the, buy, out, of, american, brands, vying, for, mr., womack, currently, share, data, incorporated, believe, output, process, unnecessarily, and, with, on, much, is, scheduled, to, conscientious, teaching, this, would, be, a, record, november]
- PCFG:
  - [the, year, is, surprising, independent, attack, paid, off, the, era, involving, IRS, leaders, and, transportation, price]
  - [it, could, be, announced, sometime]
  - [mr., toeland, believes, the, average, defense, economy, is, drafted, from, slightly, more, than, 12, stocks]

What's Next?

- Learning deeper linguistic structure
- Increasingly sophisticated models
- Readings on web (I don’t expect you to read them all)
  - [Pullum & Scholz, 2005] (nativism vs. empiricism, philosophical)
  - [Goldwater et al., 2006] (word segmentation)
  - [Petrov et al., 2007] (phone substructure, system building)
- Readings for future topics will be posted in advance...
- If there’s one thing you take from this class by the end:
  - There are linguistically realistic statistical models and they can be used for high-quality corpus-based learning
  - (Though, today’s models weren’t really them)