CS 679: Natural Language Processing
Lecture #3: Language Modeling Intro.

Announcements
- Reading Report #2
  - M&S 6.1-6.2
  - Due: now
- Project #1
  - Build an interpolated language model

Goals
- Introduce language modeling
- Understand language models as simple graphical models
- See examples of text generated from such models
- Motivate smoothing of language models

How’d they do that?
The retired Google Labs “Scribe” project.
Text completion.
Showing up now on your smart phone!

Probabilistic Language Models
- Goal: to build models which assign scores to sentences.
  - \(P(\text{I saw a van}) \gg P(\text{eyes awe of an})\)
  - Not really grammaticality:
    \(P(\text{artichokes intimidate zippers}) = 0\)
- One option: empirical distribution over sentences?
  - What’s that?

Empirical Distribution
- Example corpus of \(N\) sentences:
  - “the dog ran home” \(<s>\)
  - “the dog ran to work” \(<s>\)
- Probabilities:
  - \(P(\text{“the dog ran home” } <s>) = C("\text{the dog ran home” } <s>/) / N\)
  - \(P(\text{“the dog ran to work” } <s>) = C("\text{the dog ran to work” } <s>/) / N\)
- Using this model,
  - what is \(P(\text{“the dog ran to the boy” } <s>)\)?

Also called: Maximum Likelihood Estimate (MLE)
Probabilistic Language Models

- Remember the goal: to build models which assign scores to sentences.
- One option: empirical distribution over sentences?
  - Problem: doesn’t generalize (at all)
- How to generalize?
  - Decomposition: sentences generated in small steps (e.g., indiv. words); steps can be recombined in other ways
  - Smoothing: allow for the possibility of unseen events
  - Other ideas? Let’s see ...

“Gram”??

- From Ancient Greek γράμμα (grámma)
  - “letter; something written”
- Used both in classical derivations
  - anagram, epigram, program
- And in new coinages
  - Telegram
- The specific sense of “word consisting of N characters” is slightly post-Classical, first attested in monogram
  - From Latin monogrammum
  - From Byzantine Greek μονόγραμμον (monogrammon)
- Productive in English:
  - bigram, trigram, tetragram
  - pentagram, etc.
  - tetragrammation

Example: Unigram Model

- Example corpus of N sentences:
  - “the dog ran home” <

Empirical Unigram Distribution:

<table>
<thead>
<tr>
<th>w</th>
<th>C(w)</th>
<th>P(w) = C(w)/N</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>2</td>
<td>2/11</td>
</tr>
<tr>
<td>dog</td>
<td>2</td>
<td>2/11</td>
</tr>
<tr>
<td>ran</td>
<td>2</td>
<td>2/11</td>
</tr>
<tr>
<td>home</td>
<td>1</td>
<td>1/11</td>
</tr>
<tr>
<td>to</td>
<td>1</td>
<td>1/11</td>
</tr>
<tr>
<td>work</td>
<td>1</td>
<td>1/11</td>
</tr>
</tbody>
</table>
| <

Decomposition: N-Gram Language Models

- No loss of generality to break sentence probability down with the chain rule:

  \[ P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_{i-1}, \ldots, w_1) \]

- Histories are too long and probably too infrequent!

  \[ P(w_i | w_{i-1}, \ldots, w_1) \]

- N-gram solution: assume each word depends only on a short linear history

  \[ P(w_{i} | w_{i-1}, \ldots, w_{i-k}) = \prod_{j=1}^{k} P(w_j | w_{i-j}, \ldots, w_{i-j-k+1}) \]

Unigram Models

- Simplest case: unigrams

  \[ P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i) \]

- Generative process: pick a word, pick a word...

- As a graphical model:

  \[ w_1 \quad w_2 \quad \ldots \quad w_n \quad \text{STOP} \]

  - To make this a proper distribution over sentences, we have to generate a special STOP symbol last. (Why?)
- Examples:
  - [With an exclamation a the as incorporated is a the inflation rate dollars quarter in rise.] (with a right parenthesis to end a text)
  - [that or limited time]
  - [for]
  - [other as is constantly hospital duties of all other such factors raised under the shared account because the the package physician is the because the rise has been seen in other

Bigram Models

- Big problem with unigrams: P(thef the the the) >> P(like ice cream)!

- Condition on previous word taken:

  \[ P(w_{i} | w_{i-1}, \ldots, w_{i-k}) = \prod_{j=1}^{k} P(w_j | w_{i-j}, \ldots, w_{i-j-k+1}) \]

  - Any better?
  - [Texaco rose one in this issue is pursuing growth in a boiler house said mr. gurria]
  - [mexico’s motion control proposal without permission from five hundred fifty five yen]
  - [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 reject annually the buy out of american brands in ding for mr.

  - [would be a record november]
Example: Bigram Model

- Example corpus of N sentences:
  - <s>“the dog ran home”</s>
  - <s>“the dog ran to work”</s>
  - ...

| w1  | w2   | C(w2|w1) | \(\sum C(w2|w1)\) | \(P(w1)\) |
|-----|------|-------|---------------------|----------|
| <s> | the  | 200   |                     |          |
| dog | ran  | 10    |                     |          |
| ran | home | 15    |                     |          |
| home</s>| 5    |       |                     |          |
| ran | to   | 10    |                     |          |
| to  | work | 15    |                     |          |
| work</s>| 5    |       |                     |          |

Examples from a Trigram Model

\[ P(w_3 | w_2, w_1) = \prod_i P(w_i | w_{i-1}, w_{i-2}) \]

- names are resigning at an expected final maturity is the first nine months the reawakening of the sea paulo brasil and colombia two hundred ninety nine million dollars
- drink carrier completes with home depot incorporated an anthony pennsylvania law firm wither cutter 's managing director of the tax code
- people 's homes suffered physical damage and shattered glass from a gubernatorial emergency fund that uses its independence by involving an accident is to catch the thieves
- no strength lies in the bare faced mocktail the true story of women and over restoring funding to realize that only hope of keeping the listed stocks during program selling near the Howell outros as much as much as time this line method. A derivative section implies both value and expectations of prices. A figure is a few of the stock would probably have to clear up any slack
- but should still maintain earnings growth through the exclusive arrangement with shop television network and has a wary eye aimed a year earlier period according to the ideal prevented in civil texts and memorialized on the low item-vice would have to rely on the dismissal

More Examples - Trigram Model

- children from the australian company for two series of meetings to discuss internal changes
- general unsecured creditors and bidders alike concede that business has traditionally dominated the afternoon session up a nominal nine percent from fifty one point six zero percent and exports should be completed by the general trend given rising costs improve quality in every day operations
- net premiums written through september twenty sixth memo to district offices where local benefit seeking than are now being put on the roof crush resistance standard for the victims dorena bertussi denounced the ethics panel 's deliberations
- maitre'd claims in connection with the defendants are the same period last year
- both periods included non recurring credits from foreign tax rates in fairfield connecticut consumer industrial products segment where its stock bond and foreign bases to draw down its european backing and took to get the gas business were disappointing said b. alex henderson an analyst with nesbitt thomson deacon incorporated in august

Examples: 5-gram Model

- the soviet state bank announced a ninety percent devaluation of the rouble against the dollar for private transactions in an apparent attempt to maintain its bargaining position with the board
- if webster had it happened you would not be here linda bowker tells a reporter in the journal 's new york bureau
- the paris bank does n't publish that figure
- eight trading days after october twelfth the day before the quebec seagoing completed three days of emergency training and drills
- some litigators scoff at the notion that a snag in the takeover financing of united airlines instantly knocked seven percent off the value of the nation 's largest urban systems
- seventy point three million dollars or forty eight cents a share

What 's going on here?
- The model memorized the data!
- Idea #3 for generality:
  "There 's no data like more data."

More Examples: 5-gram Model

- reserves traded among commercial banks for overnight use in amounts of one million and more dollars
- in recent months mr. dorrance 's children and other family members have pushed for improved profitability and higher returns on their equity
- rick sherlund a goldman sachs analyst has raised his earnings estimates for the company twice in the past two months
- both banks have been battered as have other arizona banks by falling real estate prices

LMs are Directed Graphical Models
What’s Next

- **Project #1:**
  - Due in one week.

- **Upcoming lectures**
  - Markov Chains
  - Metrics
  - Smoothing Techniques for Language Models