A Hierarchical Bayesian Language Model based on Pitman-Yor Processes

Yee Whye Teh

Presented by Hsin-Ta Wu
Slides courtesy: Yee Whye Teh
Language Model

• Given a sentence of \( t \) words:
  \[
  \text{word}_1, \text{word}_2, \ldots, \text{word}_t
  \]

• An \( n \)-gram **LANGUAGE MODEL** defines a probability distribution over the current \( \text{word}_i \) given the prior \( n-1 \) words.
  \[
  P(\text{word}_i | \text{word}_{i-n+1}, \ldots, \text{word}_{i-1})
  \]

• This sentence then can be typically represented by the probability:
  \[
  P(\text{word}_1, \text{word}_2, \ldots, \text{word}_t) = \prod_{i=1}^{t} P(\text{word}_i | \text{word}_{i-n+1}, \ldots, \text{word}_{i-1})
  \]
Language Model (cont)

• Consider a set vocabulary $W$ with $V$ word types

• Each word $w \in W$, and a context $u$: $n$-1 prior-word
  – E.g. n=3, **bayesian nonparametric model**

• The vector of word probability estimates for n-grams:
  \[ G_u = [G_u(w)]_{w \in W} = [G_u(w_1), \ldots, G_u(w_v)] \]

• Maximum Likelihood estimation:
  \[
P(\text{word}_i = w|\text{word}_{i-n+1}, \ldots, \text{word}_{i-1} = u)\]
  
  \[
  = G_u^{ML}(w) = \frac{c_{uw}}{\sum_{w'} c_{uw'}} = \frac{c_{uw}}{c_u}.
  \]
Smoothing

• Maximum Likelihood is expected to be a very poor estimate given a realistic corpus size
  – What about a trigram $uw$ which has never occurred in the training data
    • i.e. $G_u^{ML}(w) = 0$

• *Smoothing* is used to address this problem.

\[
G_u^{ML}(w) = \frac{\delta + c_{uw}}{\delta |V| + c_u}.
\]
Back-off and Interpolated Smoothing

• Back-off approach:
  – Only use lower-order model when data for higher-order model is unavailable (i.e. count is zero).

\[
P_{\text{katz}}(w_n | w_{n-N+1}^{n-1}) = \begin{cases} 
P^*(w_n | w_{n-N+1}^{n-1}) & \text{if } C(w_{n-N+1}^{n-1}) > 1 \\ 
\alpha(w_{n-N+1}^{n-1})P_{\text{katz}}(w_n | w_{n-N+2}^{n-1}) & \text{otherwise} 
\end{cases}
\]

• Interpolated approach:
  – Linearly combine estimates of \(n\)-gram models of increasing order.

\[
\hat{P}(w_n | w_{n-2}, w_{n-1}) = \lambda_1 P(w_n | w_{n-2}, w_{n-1}) + \lambda_2 P(w_n | w_{n-1}) + \lambda_3 P(w_n)
\]

Where: \(\sum_i \lambda_i = 1\)
Bayesian Smoothing

- Estimation
  \[ P(G_u | \mathcal{D}) \propto P(\mathcal{D} | G_u)P(G_u) \]

- Predictive Inference
  \[ P(\text{word}_i = w | \text{word}_{i-n+1}, ..., \text{word}_{i-1} = u, \mathcal{D}) = \int P(w | u, G_u)P(G_u | \mathcal{D})dG_u \]

- Priors over distributions
  \[ G_u \sim \text{DP}(\theta, \mathcal{H}) \]
  \[ G_u \sim \text{PY}(d, \theta, \mathcal{H}) \]

- Inference is smoothed with respect to the distribution
Pitman-Yor Process

- **Pitman-Yor Process**
  
  \[ \mathcal{P}(d, \theta, G_0) \]
  
  - \( d \): discount parameter, \( 0 \leq d < 1 \)
  - \( \theta \): strength (concentration) parameter, \( \theta > -d \)
  - \( G_0 \): base distribution

- Generalization of the *Dirichlet process* \((d=0)\)

- Pitman-Yor processes produce distributions over words given by a power law distribution
  
  - [Goldwater et al 2006] investigated the Pitman-Yor process from this perspective.
Pitman-Yor Process for a unigram language model

• To estimate a word \( w \in W \),
  - \( P(\text{word}_i = w | \text{word}_{i-n+1}, ..., \text{word}_{i-1} = u) \)
    \[ = P(\text{word}_i = w) = G(w) \]
  - \( G = [G(w)]_{w \in W} \)

• \( G \sim PY(d, \theta, G_0) \)
  - \( d \): discount parameter, \( 0 \leq d < 1 \)
  - \( \theta \): strength parameter, \( \theta > -d \)
  - \( G_0 \): a mean vector for unigram, using uniform distribution over fixed vocabulary \( W \) of \( V \) words
Perspective by the Chinese restaurant process

- Easiest to understand them using Chinese restaurant processes.

\[ P(\text{sit at an occupied table } t_i) = \frac{c_{t_i} - d}{\theta + c}. \]
Perspective by the Chinese restaurant process

- Easiest to understand them using Chinese restaurant processes.

\[ P(\text{sit at an occupied table } t_i) = \frac{c_{t_i} - d}{\theta + c} \quad P(\text{sit at new table}) = \frac{\theta + dt}{\theta + c}. \]
Perspective by the Chinese restaurant process

• Given the seating arrangement $S$, the predictive probability of a test word $k$ is:

$$P(x_{c+1} = k | S) = \frac{c_k - dt_k}{\theta + c} + \frac{\theta + dt}{\theta + c} G_0(k)$$
What about \textit{n}-gram language model?

• Hierarchical Bayesian models
  – Capture the dependencies by statistical strength among different components of the language model
  
  – Specifically: hierarchical model based on the tree of suffixes: CONTEXT TREES
Basic assumption: words appearing later in a context are more important.
Hierarchical Bayesian Models on Context Tree

• The probability of the current word \( w \) following the context \( u \)
  \[
P(word_i = w | word_{i-n+1}, \ldots, word_{i-1} = u) = G_u(w)
  \]

• The vector of word probability estimates for \( n \)-grams
  \[
  G_u = [G_u(w)]_{w \in W} = [G_u(w_1), \ldots, G_u(w_v)]
  \]

• Tie related distribution together
  \[
  G_{u=statistical\ machine} \sim \text{DP}\left(\theta, G_{\pi(u)=machine}\right)
  \]
  \[
  G_{u=statistical\ machine} \sim \text{PY}\left(d, \theta, G_{\pi(u)=machine}\right)
  \]

\[ G_\emptyset \quad G_{\text{machine}} \quad G_{\text{statistical\ machine}} \quad G_a\ \text{machine} \quad G_{\text{Bayesian\ machine}} \]
Hierarchical Bayesian Models on Context Tree

- Tie related distribution together
  \[ G_{\text{statistical machine}} \sim \mathcal{DP}(\theta, G_{\text{machine}}) \]
  \[ G_{\text{statistical machine}} \sim \mathcal{PY}(d, \theta, G_{\text{machine}}) \]

  - Observations in one context affect inference in other context.
  - Statistical strength is shared between similar contexts
  - E.g. Observe “statistical machine learning”
Hierarchical Pitman-Yor Process for *n*-gram Language Models

- Use a Pitman-Yor process as the prior for each node $G_u = [G_u(w)]_{w \in W}$
- $G_u \sim \mathcal{PY}(d_{|u|}, \theta_{|u|}, G_{\pi(u)})$

![Diagram of hierarchical Pitman-Yor process for n-gram language models](image)
Perspective by the Chinese restaurant process

\[
P(\text{sit at an occupied table } k) = \frac{c_{uwk} - d_{|u|}}{\theta_{|u|} + c_u}\]

\[
P(\text{sit at a new table}) = \frac{\theta_{|u|} + d_{|u|}t_u}{\theta_{|u|} + c_u}\]
Perspective by the Chinese restaurant process

\[ P(\text{sit at an occupied table } i) = \frac{c_{\pi(u)}w_i - d_{|\pi(u)|}}{\theta_{|\pi(u)|} + c_{\pi(u)}} \]

\[ P(\text{sit at a new table}) = \frac{\theta_{|\pi(u)|} + d_{|\pi(u)|}t_{\pi(u)}}{\theta_{|\pi(u)|} + c_{\pi(u)}} \]
Hierarchical Pitman-Yor Process for \( n \)-gram Language Models

• Given a particular seating arrangement, \( \pi(u) \),

\[
P(w = \text{learning} | u = \text{statistical machine}) = \frac{c_{uw} - d_{|u|}t_{uw}}{\theta_{|u|} + c_{u..}} + \frac{\theta_{|u|} + d_{|u|}t_{u..}}{\theta_{|u|} + c_{u..}} P(w = \text{learning} | \pi(u) = \text{machine})
\]

\( S_u \): seating arrangement in the restaurant \( u \).
What’s next? Inference

- Based on the framework for Hierarchical Pitman-Yor Language Model, to get the probability over a word \( w \) after a context \( u \) \( P(w|u) \) given training data \( D \):

\[
p(w|u, D) = \int p(w|u, S, \Theta)p(S, \Theta|D) \, d(S, \Theta)
\]

- inference of seating arrangements \( S \) in each restaurant
- estimation of the context-specific parameters \( \Theta \)
Inference of Seating Arrangements

• Gibbs sampling is used to keep track of which table each customer sits at

• Steps:
  – Iterative over all customers present in each restaurant, resampling the table at which each customer sits
    • Randomly removing a customer from the restaurant
    • Then resampling the table at which that customer sits
Estimation of the context parameters

• For a $n$-gram language model, there are $2n$ parameters
  \[\Theta = \{d_m, \theta_m: 0 \leq m \leq n - 1\}\]

• Use the auxiliary variable sampling method, assuming
  \[\theta_m \sim Gamma(\alpha_m, \beta_m) \quad d_m \sim Beta(a_m, b_m)\]

• Further details please find the technical report [Teh, 2006]
The predictive probability:

- Approximate the integral with samples \( \{ S^{(i)}, \Theta^{(i)} \}_{i=1}^{I} \) drawn from \( p(S, \Theta | D) \):

\[
p(w | u, D) \approx \sum_{i=1}^{I} p(w | u, S^{(i)}, \Theta^{(i)}) / I
\]
Interpolated Kneser-Ney (IKN) and Modified Kneser-Ney (MKN)

\[ P^\text{ML}_u(w) = \frac{c_{uw}}{\sum_{w'} c_{uw'}} = \frac{c_{uw}}{c_u}. \]

\[ P^\text{IKN}_u(w) = \max\left(0, \frac{c_{uw} - d_{|u|}}{c_u}\right) + \frac{d_{|u|} t_u}{c_u} \cdot P^\text{IKN}_{\pi(u)}(w) \]

\[ P^\text{HPY}_u(w | \text{seating arrangement}) = \frac{c_{uw} - d_{|u|} t_{uw}}{\theta_{|u|} + c_u} + \frac{\theta_{|u|} + d_{|u|} t_u}{\theta_{|u|} + c_u} \cdot P^\text{HPY}_{\pi(u)}(w | \text{seating arrangement}) \]

- Assume that the strength parameters \( \theta_{|u|} = 0 \) for all \( u \)
- Restrict \( t_{uw} \) to be at most 1
  - all customers representing the same word token should only sit on the same table in each restaurant
- Interpret IKN as an approximate inference scheme for the HPYLM

[Chen and Goodman. 1998. An empirical study of smoothing techniques for language modeling.]
Experiments

• Test five language models on APNews corpus:
  – Interpolated Kneser-Ney (IKN)
  – Modified Kneser-Ney (MKN)
  – Hierarchical Pitman-Yor Language Model (HPYLM)
  – Optimized HPYLM (HPYCV)
  – Hierarchical Dirichlet Language Model (HDLM)

• Evaluated by Perplexities
  – Train the n-Gram model:
    \[ p(w_i|w_{i-n+1}) \]
    \[ p(T) = \prod p(t_i) \]
  – Cross-entropy:
    \[ H_p(T) = -\frac{1}{W_T} \log_2 p(T) \]
  – Perplexity:
    \[ PP_p(T) = 2^{H_p(T)} \]
Experimental Results I

<table>
<thead>
<tr>
<th>T</th>
<th>n</th>
<th>IKN</th>
<th>MKN</th>
<th>HPYLM</th>
<th>HPYCV</th>
<th>HDLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>2e6</td>
<td>3</td>
<td>148.8</td>
<td><strong>144.1</strong></td>
<td>145.7</td>
<td>144.3</td>
<td>191.2</td>
</tr>
<tr>
<td>4e6</td>
<td>3</td>
<td>137.1</td>
<td><strong>132.7</strong></td>
<td>134.3</td>
<td><strong>132.7</strong></td>
<td>172.7</td>
</tr>
<tr>
<td>6e6</td>
<td>3</td>
<td>130.6</td>
<td>126.7</td>
<td>127.9</td>
<td><strong>126.4</strong></td>
<td>162.3</td>
</tr>
<tr>
<td>8e6</td>
<td>3</td>
<td>125.9</td>
<td>122.3</td>
<td>123.2</td>
<td><strong>121.9</strong></td>
<td>154.7</td>
</tr>
<tr>
<td>10e6</td>
<td>3</td>
<td>122.0</td>
<td>118.6</td>
<td>119.4</td>
<td><strong>118.2</strong></td>
<td>148.7</td>
</tr>
<tr>
<td>12e6</td>
<td>3</td>
<td>119.0</td>
<td>115.8</td>
<td>116.5</td>
<td><strong>115.4</strong></td>
<td>144.0</td>
</tr>
<tr>
<td>14e6</td>
<td>3</td>
<td>116.7</td>
<td>113.6</td>
<td>114.3</td>
<td><strong>113.2</strong></td>
<td>140.5</td>
</tr>
<tr>
<td>14e6</td>
<td>2</td>
<td>169.9</td>
<td><strong>169.2</strong></td>
<td>169.6</td>
<td>169.3</td>
<td>180.6</td>
</tr>
<tr>
<td>14e6</td>
<td>4</td>
<td>106.1</td>
<td>102.4</td>
<td>103.8</td>
<td><strong>101.9</strong></td>
<td>136.6</td>
</tr>
</tbody>
</table>
Experimental Results II
Conclusions

• Proposed a new language model based on the hierarchical Bayesian paradigm.

• Showed that Interpolated Kneser-Ney is approximate inference in the hierarchical Pitman-Yor language model.
QUESTIONS
Power-law properties of the Pitman-Yor Process

Number of unique words

Number of words

Proportion of words appearing once

Number of words
Experimental Results II