The Infinite Markov Model

Daichi Mochihashi*† Eiichiro Sumita†

*NTT Communication Science Laboratories, Japan

†ATR Spoken Language Communication Research Laboratories, Japan

Keywords: Nonparametric Bayes, Tree prior, Pitman-Yor process, Prediction Suffix Trees

Overview

- Nonparametric Bayesian *variable-order Markov Model* that estimates latent Markov orders from which each symbol originated.
- *Tree prior* over stochastic suffix trees of diminishing branches.

Motivation and Background

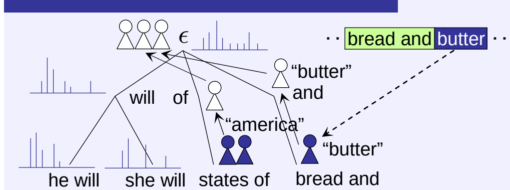
···· and she sings a song ····

- Natural language and speech processing
  → n-gram (n-1 order Markov) model is prevalent
  - Fixed (n-1) words dependency for next word
  - "less than"? "supercalifragilisticexpialidocious"?
  - Music processing, Bioinformatics, compression,..

↑

Previous works: pruning a huge model
Very interesting, but
- Often cannot build such huge models in advance (ex. Google >5 grams?)
- Difficult to integrate as other model's component

HDP and Markov Models



- Markov models can be mapped to hierarchical (Poisson-) Dirichlet process (Teh,Goldwater+ 2006)

$$p(w|h) = \frac{\#(w|h) - d\cdot t_{hw}}{\#(h)+\theta} + \frac{d\cdot t_h + \theta}{\#(h)+\theta}\cdot p(w|h')$$

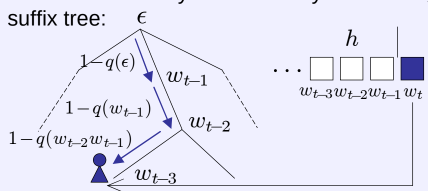
- Problem: *All real customers reside in depth (n-1)*

↓

How to deploy customers at suitable depths?

Variable-order HDP

- Add a customer by stochastically descending the suffix tree:



$$p(n|h) = q_n \prod_{i=0}^{n-1}(1-q_i) \quad q_i \sim \mathrm{Be}(\alpha,\beta).$$

➡ Posterior q_i :

$$E[q_i] = \frac{\alpha + \#(\text{stop at } i)}{\alpha+\beta+\#(\text{access of } i)}$$

- This process is still **exchangeable** over customers, so we can Gibbs sample for inference:

$$p(n_t|\mathbf{w},\mathbf{z}_{-t},\mathbf{n}_{-t}) \propto p(w_t|\mathbf{w}_{-t},\mathbf{z}_{-t},\mathbf{n}_{-t},n_t)\,p(n_t|\mathbf{w}_{-t},\mathbf{z}_{-t},\mathbf{n}_{-t}).$$

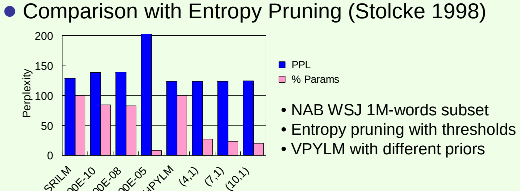
Natural Language Processing

- Bayesian ∞-gram Language Model

$$p(w|h) = \sum_{n=1}^{\infty} p(w,n|h) = \sum_{n=1}^{\infty} \underbrace{p(w|h,n)}_{\text{n-gram}}\,p(n|h)$$

- Empirical consideration

Naïve K-N 9-grams,10-grams,… might be possible (esp. with Bloom Filters), but:
  - Large n-grams are extremely noisy and bulky
  - Conveys no linguistic insights
  - Cannot generate – simply reproduces training data.
- Comparison with Entropy Pruning (Stolcke 1998)

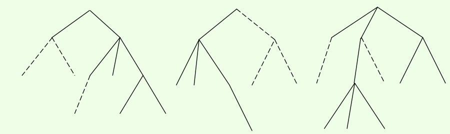


- NAB WSJ 1M-words subset
- Entropy pruning with thresholds
- VPYLM with different priors

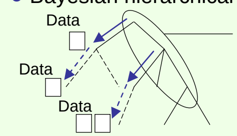
- Removes independent pruning assumption through a Gibbs
- Not depend on raw context frequency $p(\mathbf{h})$

Nonparametric Bayes Perspective

- Stochastic infinite trees

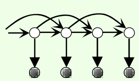


  - Top-down (<-> Coalescent tree)
  - Coalescent points known, but diminishing branches
- Bayesian hierarchical clustering



hLDA (Blei+ 2003): Multinomial for depths (maximum known)

  - "Deep semantic category" just when needed
  - Data can reside at the intermediate nodes
- Variable order HMM (Wang+ 2006)



  - Ordinary HMMs are 1-Markov
  - Estimate complex dynamics from a pure generative model

Information Theory / Compression

- Context Tree Weighting method (Willems+ 1995)
  ··· High-performance compression studied in 1990s

$$p_h(x_1^T) = \begin{cases} \gamma p_e(x_1^T) + (1-\gamma)\prod_{w\in L} p_{wh}(x_1^T) & \text{(h: non-leaf)} \\ p_e(x_1^T) & \text{(h: leaf)} \end{cases}$$

Usually 1/2!

$$p_e(x_1^T) = \int_0^1 p(x_1^T|p)\mathrm{Be}(p|\tfrac{1}{2},\tfrac{1}{2})dp \quad \text{(KT-Estimator)}$$

- γ→ Bayesian posterior, KT→Pitman-Yor ➡ Our method!
  - Infinite Markov Model = "Bayesian CTW algorithm".
  - Difference: not all histories are memoized (like CTW)
    - Memoizing only "meaningful" subsequences

Future Work

- Fast variational inference (extending VB-HDP)
- More sensitive and hierarchical prior than a single Beta
- de Finetti random measure and relationship to Tailfree processes (Fabius 1964;Ferguson 1974)