The Infinite Markov Model

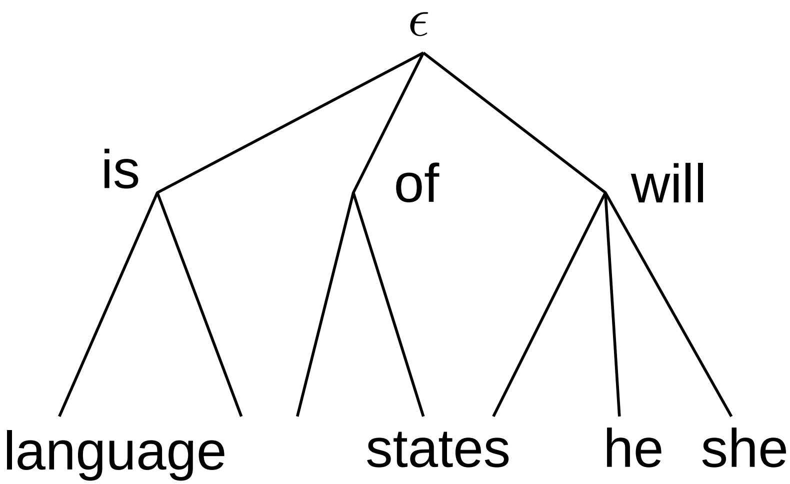
Daichi Mochihashi

NTT Communication Science Laboratories, Japan

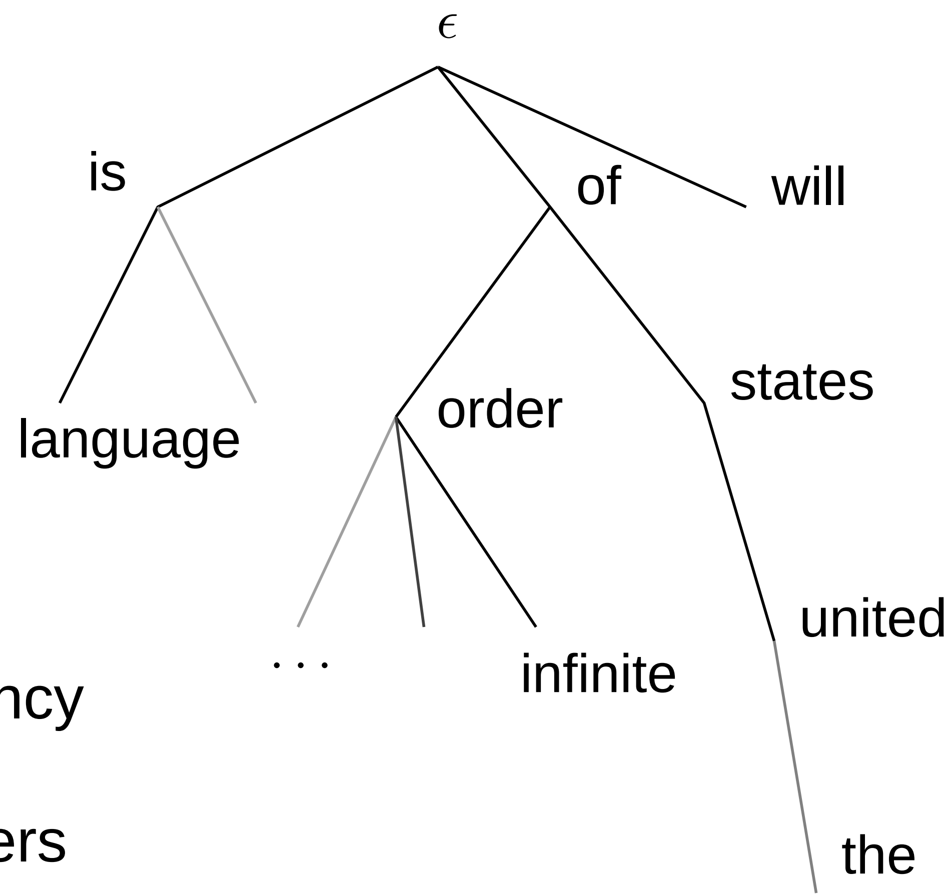
daichi@cslab.kecl.ntt.co.jp

NIPS 2007

Overview



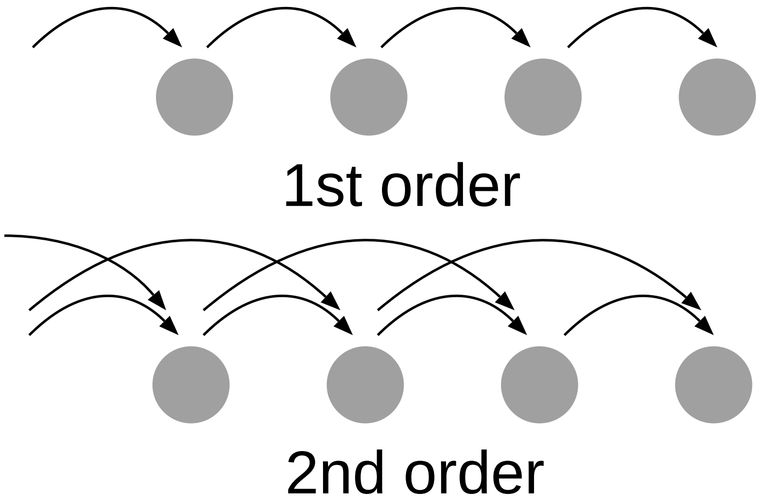
Fixed n-th order Markov model



Infinitely Variable-order Markov model

- Fixed-order Markov dependency
  $\Downarrow$
  Infinitely variable Markov orders
- Simple prior for stochastic trees (other than Coalescents)
  - How to draw an inference based on only the output sequences?

Markov Models

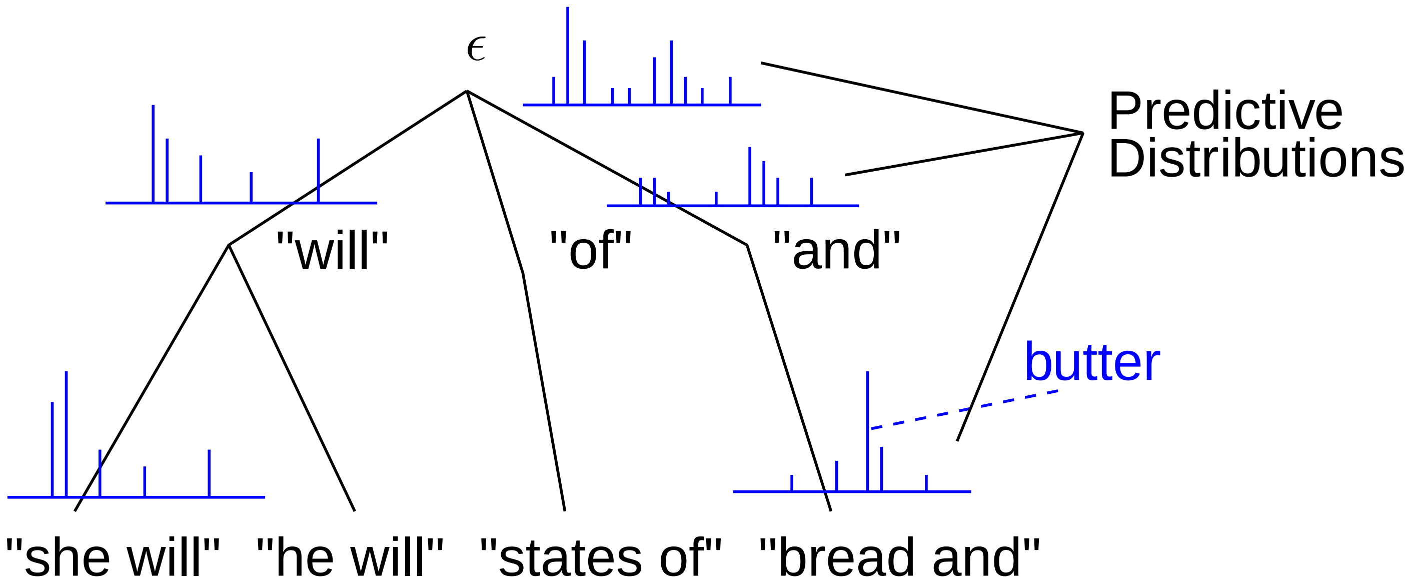


$$p(\text{"mama I want to sing"})$$
$$= p(\text{mama}) \times p(\text{I}|\text{mama})$$
$$\times\, p(\text{want}|\text{mama I})$$
$$\times\, p(\text{to}|\text{I want})$$
$$\times\, p(\text{sing}|\text{want to})$$

n-gram (3-gram)

- “n-gram” (n-1’th order Markov) model is prevalent in speech recognition and natural language processing
- Music processing, Bioinformatics, compression, $\cdots$
- Notice: HMM is a first order Markov model over hidden states
  - Emission is a unigram on the hidden state

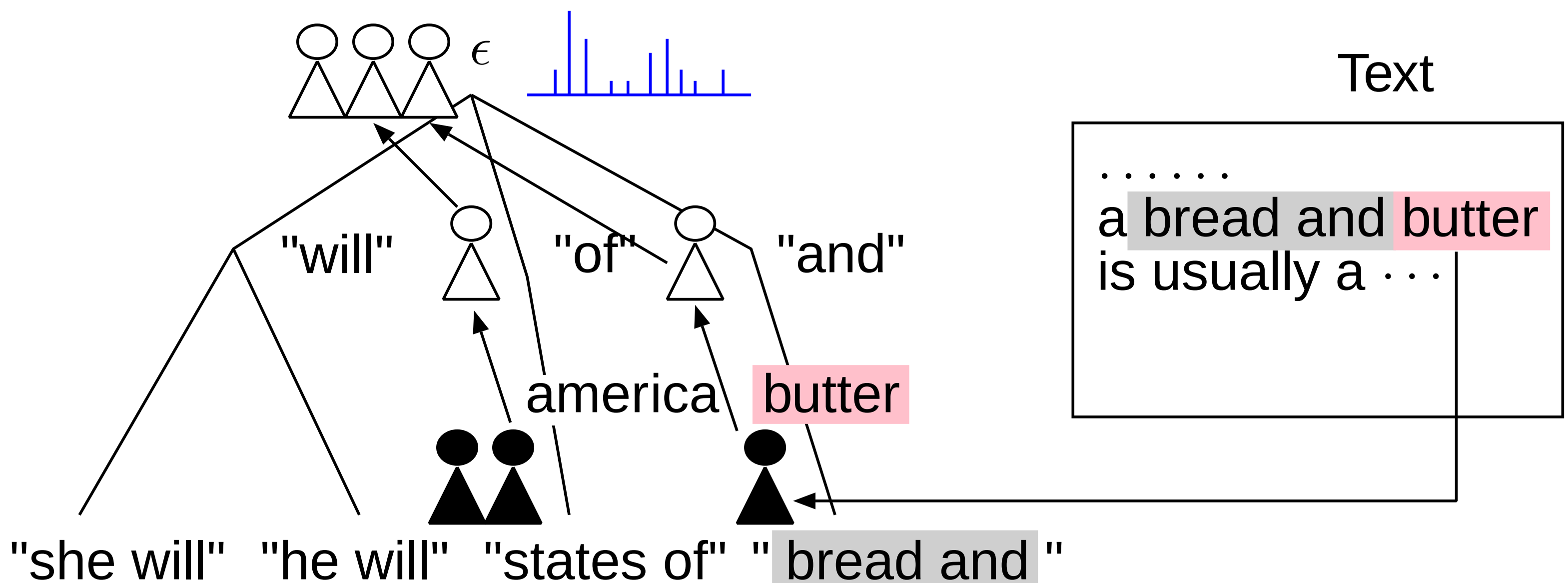
Estimating a Markov Model



- Each Markov state is a node in a Suffix Tree (Ron+ (1994), Pereira+ (1995), Buhlmann (1999))
  - Depth = Markov order
  - Each node has a predictive distribution over the next word
- Problem: # of states will explode as the order n gets larger
  - Restrict to a small Markov order ($n = 3 \sim 5$ in speech and NLP)
  - Distributions get sparser and sparser $\Rightarrow$ *using hierarchical Bayes*?

Hierarchical (Poisson-) Dirichlet Process

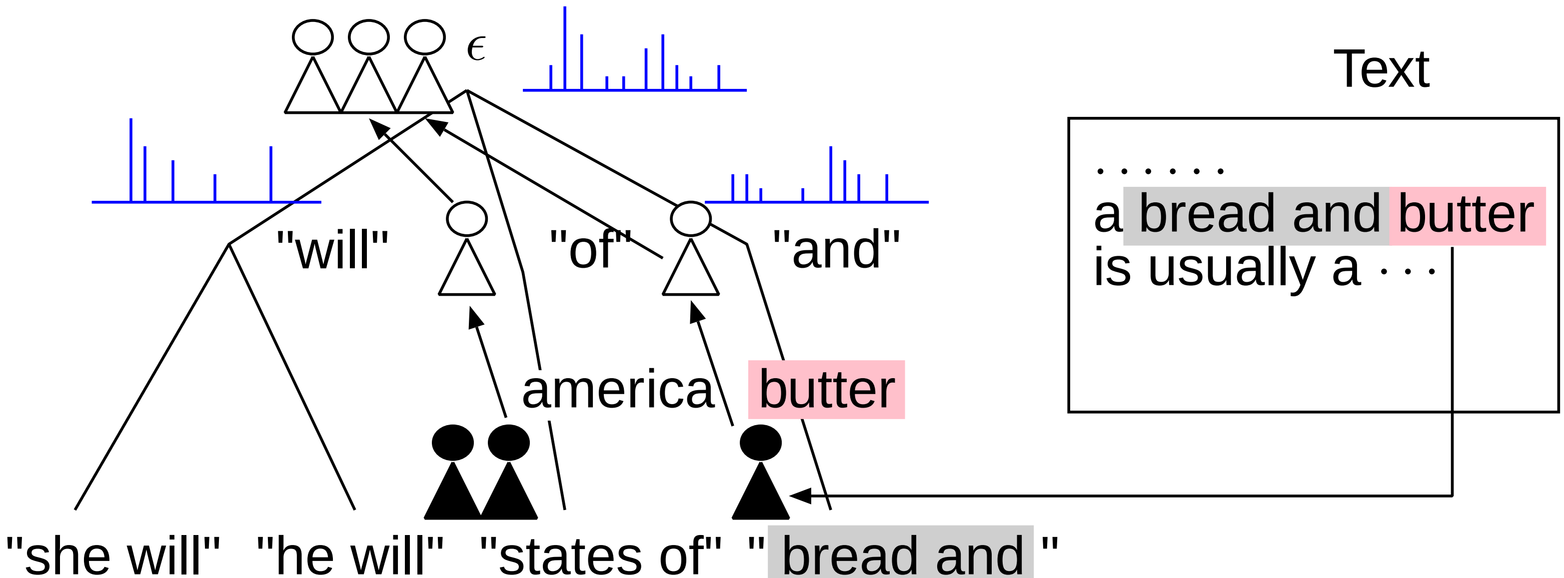
- Teh (2006), Goldwater+ (2006) mapped a hierarchical Dirichlet process to Markov Models



  - n'th order predictive distribution is a Dirichlet process draw from the $(n-1)$'th distribution
  - Chinese restaurant process representation: a customer = a count (in the training data)
  - Hierarchical Pitman-Yor Language Model (HPYLM)

Hierarchical (Poisson-) Dirichlet Process

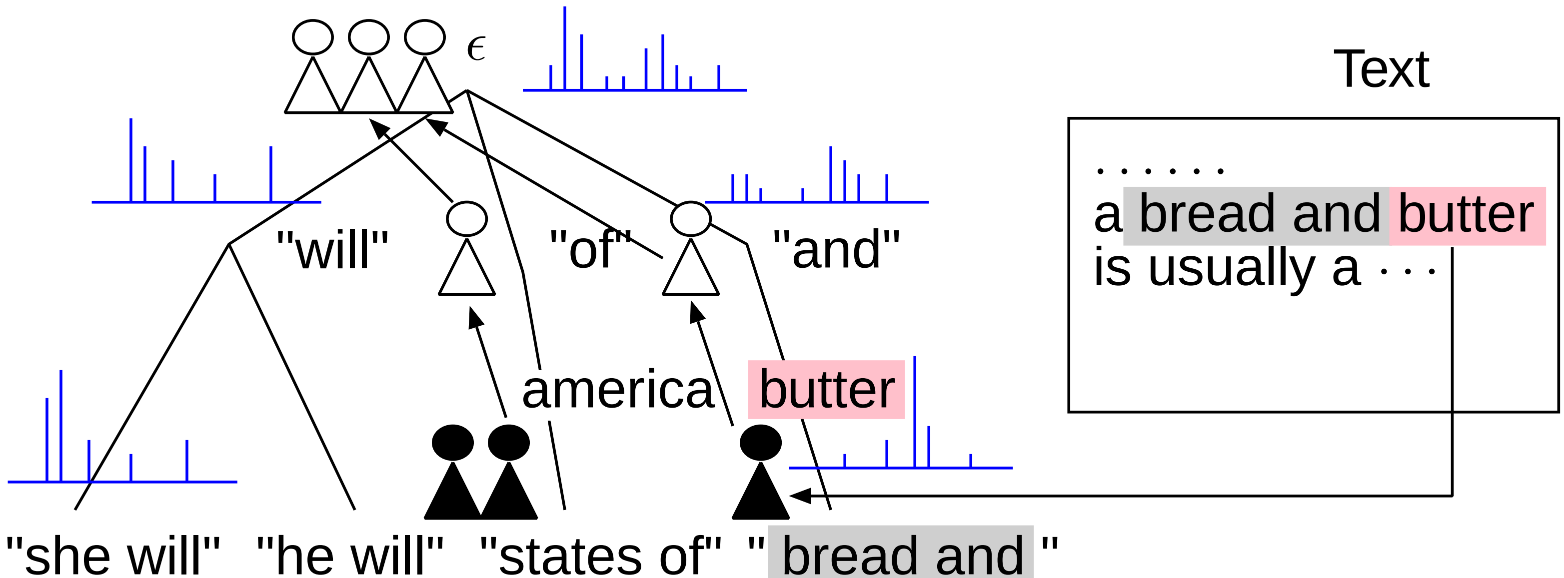
- Teh (2006), Goldwater+ (2006) mapped a hierarchical Dirichlet process to Markov Models



  - n'th order predictive distribution is a Dirichlet process draw from the $(n-1)$'th distribution
  - Chinese restaurant process representation:
    a customer = a count (in the training data)
  - Hierarchical Pitman-Yor Language Model (HPYLM)

Hierarchical (Poisson-) Dirichlet Process

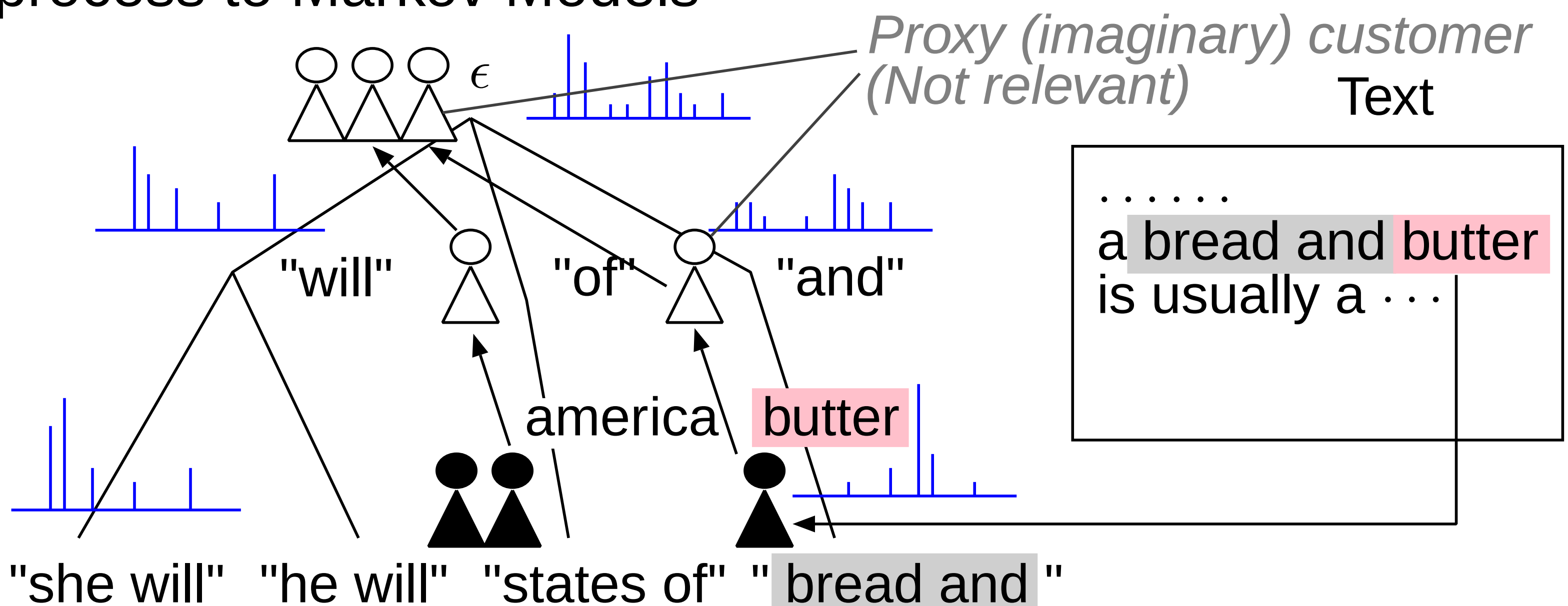
- Teh (2006), Goldwater+ (2006) mapped a hierarchical Dirichlet process to Markov Models



  - n'th order predictive distribution is a Dirichlet process draw from the $(n-1)$'th distribution
  - Chinese restaurant process representation:
    a customer = a count (in the training data)
  - Hierarchical Pitman-Yor Language Model (HPYLM)

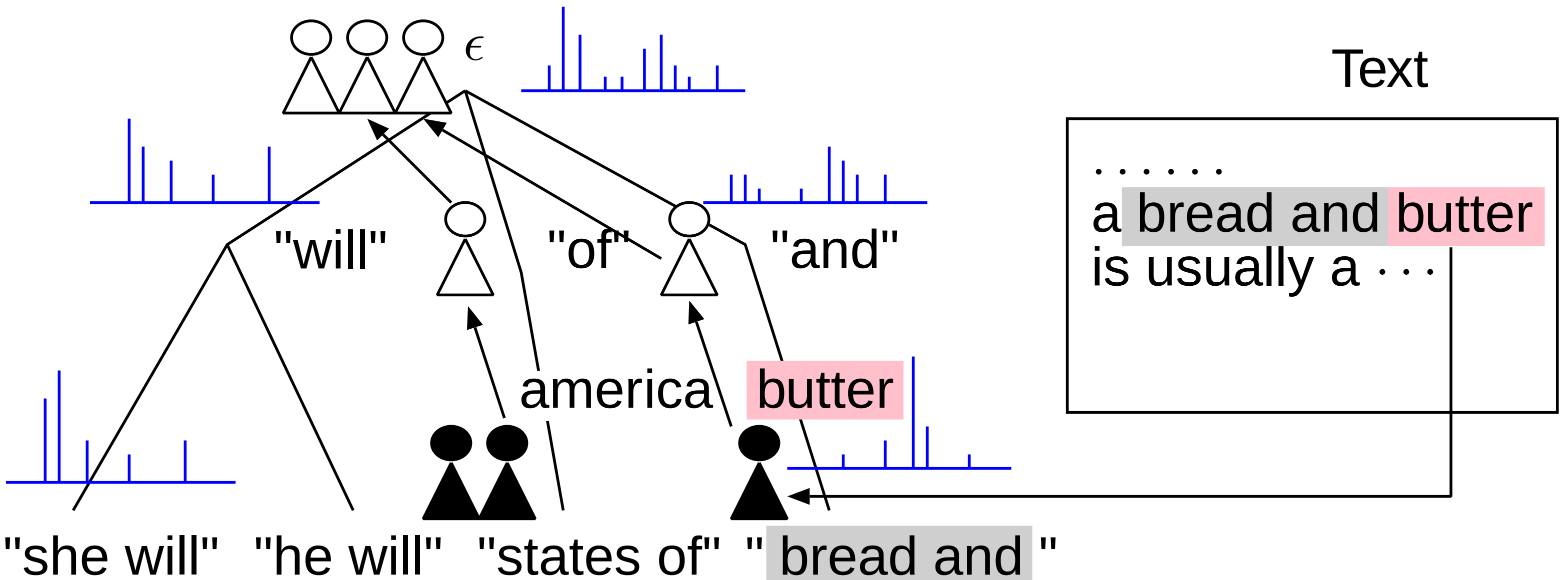
Hierarchical (Poisson-) Dirichlet Process

- Teh (2006), Goldwater+ (2006) mapped a hierarchical Dirichlet process to Markov Models



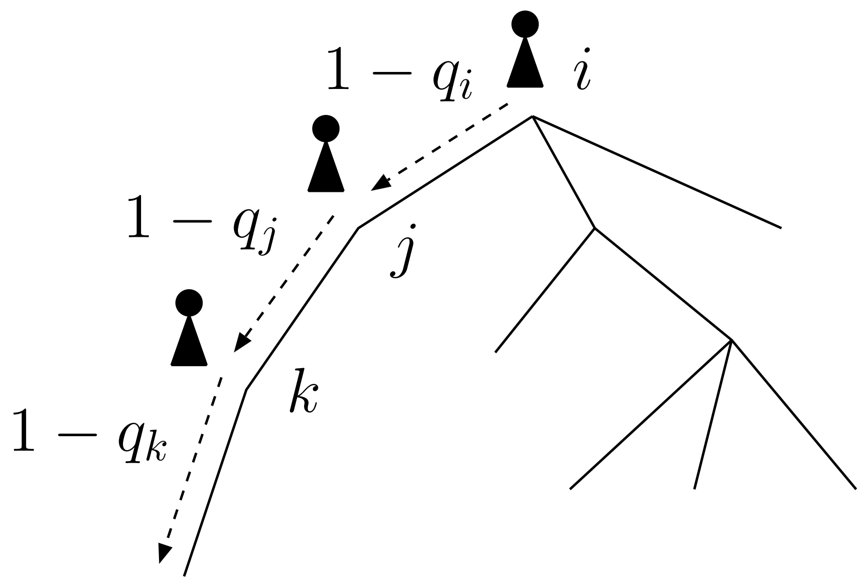
  - n'th order predictive distribution is a Dirichlet process draw from the $(n-1)$'th distribution
  - Chinese restaurant process representation: a customer = a count (in the training data)
  - Hierarchical Pitman-Yor Language Model (HPYLM)

Problem with HPYLM



- All the real customers reside in depth $(n-1)$ (say, 2) in the suffix tree
  - Corresponds to a fixed Markov order
  - “less than”; “the united states of america”
  - Character model for “supercalifragilisticexpialidocious”!
- *How can we deploy customers at suitably different depths?*

Infinite-depth Hierarchical CRP



- Add a customer by *stochastically decending* a suffix tree from its root
- Each node i has a probability to stop at that node ($1-q_i$ equals the "penetration" probability)

$$q_i \sim \mathrm{Be}(\alpha, \beta) \quad \text{i.i.d.} \tag{1}$$

- Therefore, a customer will stop at depth n by the probability

$$p(n|h) = q_n \prod_{i=0}^{n-1} (1-q_i)\,. \tag{2}$$

Variable-order Pitman-Yor language model (VPYLM)

- For the training data $\mathbf{w} = w_1 w_2 \cdots w_T$, latent Markov orders $\mathbf{n} = n_1 n_2 \cdots n_T$ exist:

$$p(\mathbf{w}) = \sum_{\mathbf{n}} \sum_{\mathbf{s}} p(\mathbf{w}, \mathbf{n}, \mathbf{s}) \tag{3}$$

  - $\mathbf{s} = s_1 s_2 \cdots s_T$: seatings of proxy customers in parent nodes

- Gibbs sample $\mathbf{n}$ for inference:

$$\begin{aligned} & p(n_t | \mathbf{w}, \mathbf{n}_{-t}, \mathbf{s}_{-t}) \\ & \qquad \propto \underbrace{p(w_t | n_t, \mathbf{w}, \mathbf{n}_{-t}, \mathbf{s}_{-t})}_{n_t\text{-gram prediction}} \cdot \underbrace{p(n_t | \mathbf{w}_{-t}, \mathbf{n}_{-t}, \mathbf{s}_{-t})}_{\text{prob to reach depth } n_t} \end{aligned} \tag{4}$$

  - Trade-off between two terms (penalty for deep n_t)
  - How to compute the second term $p(n_t | \mathbf{w}_{-t}, \mathbf{n}_{-t}, \mathbf{s}_{-t})$?

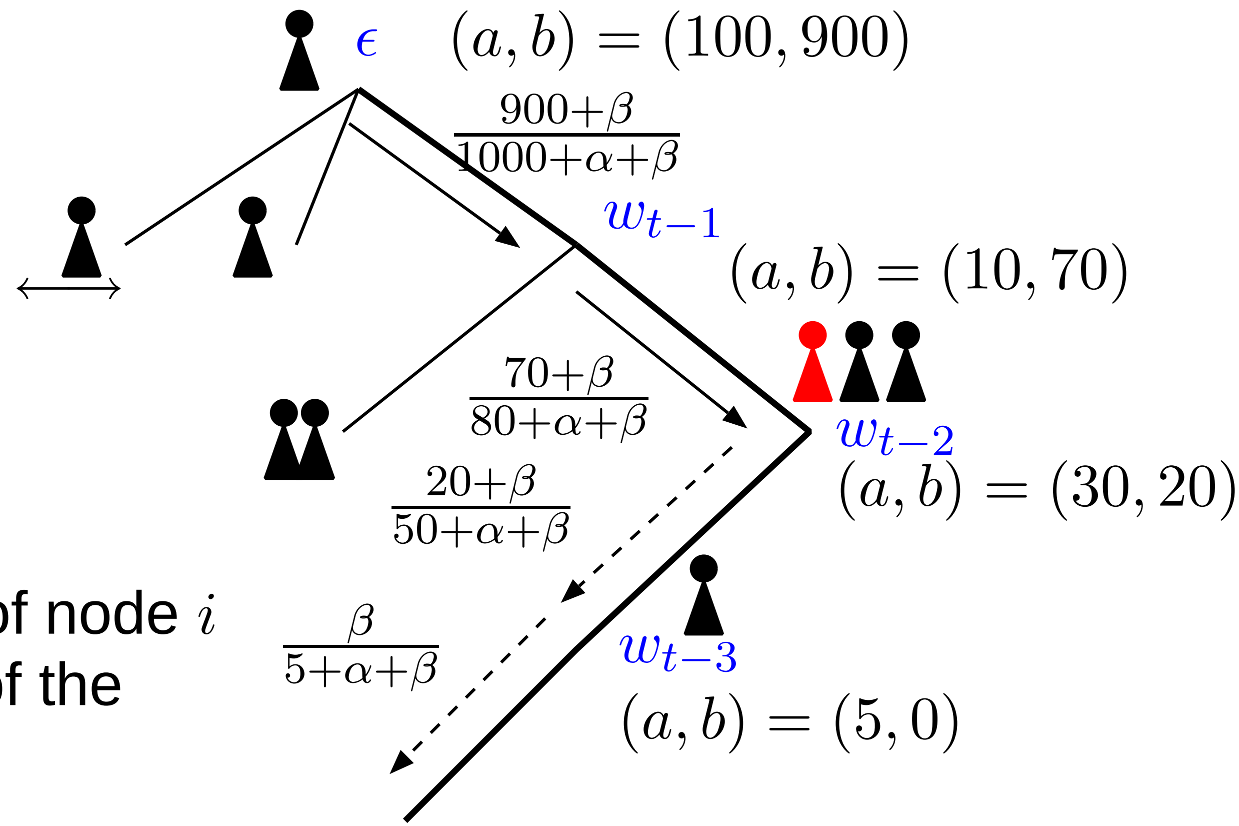
Inference of VPYLM (2)

$\mathbf{w}$

$\cdots$	w_{t-2}	w_{t-1}	w_t	w_{t+1}	$\cdots$

$\mathbf{n}$

$\cdots$	2	3	2	4	$\cdots$

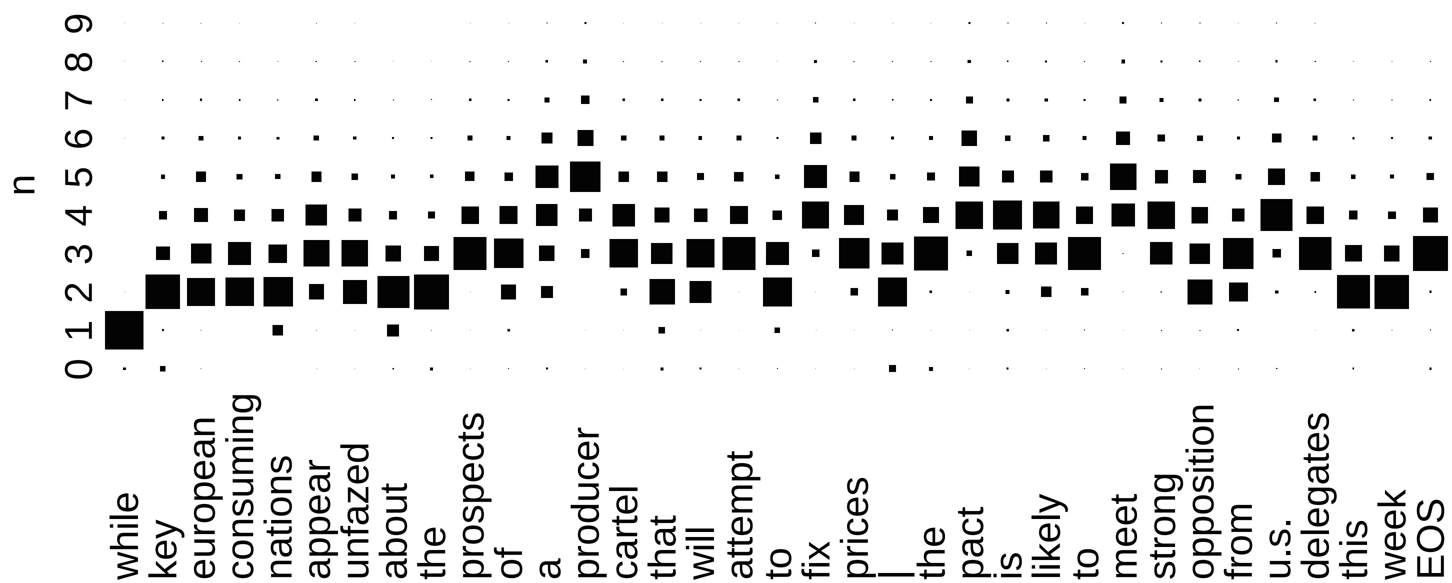


- We can estimate q_i of node i through the depths of the other customers
- Let a_i = # of times the node i was stopped at, b_i = # of times the node i was passed by:

$$p(n_t = n | \mathbf{w}_{-t}, \mathbf{n}_{-t}, \mathbf{s}_{-t}) = q_n \prod_{i=0}^{n-1} (1 - q_i) \tag{5}$$

$$= \frac{a_n + \alpha}{a_n + b_n + \alpha + \beta} \prod_{i=0}^{n-1} \frac{b_i + \beta}{a_i + b_i + \alpha + \beta} . \tag{6}$$

Estimated Markov Orders



- Hinton diagram of $p(n_t|\mathbf{w})$ used in Gibbs sampling for the training data
- Estimated Markov orders from which each word has been generated.
- NAB Wall Street Journal corpus of 10,007,108 words

Prediction

- We don't know the Markov order n beforehand $\Rightarrow$ *sum it out*

$$p(w|h) = \sum_{n=0}^{\infty} p(w,n|h) = \sum_{n=0}^{\infty} p(w|n,h)\, p(n|h). \tag{7}$$

- We can rewrite the above expression recursively:

$$\begin{aligned} p(w|h) &= p(0|h) \cdot p(w|h,0) + p(1|h) \cdot p(w|h,1) + p(2|h) \cdot p(w|h,2) + \cdots \\ &= \underline{q_0 \cdot p(w|h,0) + (1-q_0) q_1 \cdot p(w|h,1) + (1-q_0)(1-q_1) q_2 \cdot p(w|h,2) \cdots} \\ &= q_0 \cdot p(w|h,0) + (1-q_0) \left[\underline{q_1 \cdot p(w|h,1) + (1-q_1) q_2 \cdot p(w|h,2) + \cdots} \right] \end{aligned} \tag{8}$$

- Therefore,

$$p(w|h,n^+) \equiv q_n \cdot p(w|h,n) + (1-q_n) \cdot p(w|h,(n+1)^+)\,, \tag{9}$$

$$p(w|h) = p(w|h,0^+)\,. \tag{10}$$

Prediction (2)

$$p(w|h,n^+) \equiv q_n \cdot \underbrace{p(w|h,n)}_{\text{Prediction at Depth } n} + (1-q_n) \cdot \underbrace{p(w|h,(n+1)^+)}_{\text{Prediction at Depths } > n}$$

$$p(w|h) = p(w|h,0^+)\,,$$

$$q_n \sim \mathrm{Be}(\alpha,\beta)\,.$$

- Stick-breaking process on an infinite tree, where breaking proportions will differ from branch to branch.
- Bayesian sophistication of CTW (context tree weighting) algorithm (Willems+ 1995) in information theory ($\Rightarrow$ Poster)

Perplexity and Number of Nodes in the Tree

n	HPYLM	VPYLM	Nodes(H)	Nodes(V)
3	113.60	113.74	1,417K	1,344K
5	101.08	101.69	12,699K	7,466K
7	N/A	100.68	*27,193K*	10,182K
8	N/A	100.58	*34,459K*	10,434K
∞	—	100.36	—	10,629K

- Perplexity = 1/average predictive probabilities (lower is better)
- VPYLM causes no memory overflow even for large n
  - *Italic* : expected number of nodes
- Identical performance as HPYLM, but with much less number of nodes
  - ∞-gram performed the best ($\epsilon = 1e-8$)

“Stochastic phrases” from VPYLM (1/2)

- $p(w, n|h) = p(w|h, n)p(n|h)$

  $\cdots$ Probability to generate w using the last n words of h as the context

  - For example, generate “Gaussians” after “mixture of”

    $\downarrow$

    “mixture of Gaussians”: *a phrase*

- $p(w, n|h)$ = cohesion strength of the stochastic phrase
  - Will not necessarily decay with length (like an empirical probability)
  - Enumerated by traversing the suffix tree in depth-first order

"Stochastic phrases" from VPYLM (2/2)

p	Stochastic phrase in the suffix tree
0.9784	primary new issues
0.9726	ˆ at the same time
0.9556	american telephone &
0.9512	is a unit of
0.9394	to # % from # %
0.8896	in a number of
0.8831	in new york stock exchange composite trading
0.8696	a merrill lynch & co.
0.7566	mechanism of the european monetary
0.7134	increase as a result of
0.6617	tiffany & co.
:	:

- "ˆ" = beginning-of-sentence, "#" = numbers

Random Walk generation from the language model

it was a singular man , fierce and quick-tempered , very foul-mouthed when he was angry , and of her muff and began to sob in a high treble key .
“ it seems to have made you , ” said he . ’what have i to his invariable success that the very possibility of something happening on the very morning of the wedding . ”
...

- Random walk generation from the 5-gram VPYLM trained on *“The Adventures of Sherlock Holmes.”*
  - We begin with an infinite number of “beginning-of-sentence” special symbols as the context.
- If we use vanilla 5-grams, overfitting will lead to a mere reproduction of the training data.

Infinite Character Markov Model

> ‘how queershaped little children drawling-desks, which would get through that dormouse!’ said alice; ‘let us all for anything the secondly, but it to have and another question, but i shalled out, ‘you are old,’ said the you’re trying to far out to sea.

(a) Random walk generation from a character model.

Character	s a i d ␣ a l i c e ; ␣ ‘ l e t ␣ u s ␣ a l l ␣ f o r ␣ a n y t h i n g ␣ ···
Markov order	5 6 5 4 7 10 6 5 4 3 7 1 4 8 2 4 4 6 5 5 4 4 5 5 6 4 5 6 7 7 7 5 3 3 4 5 9 ···

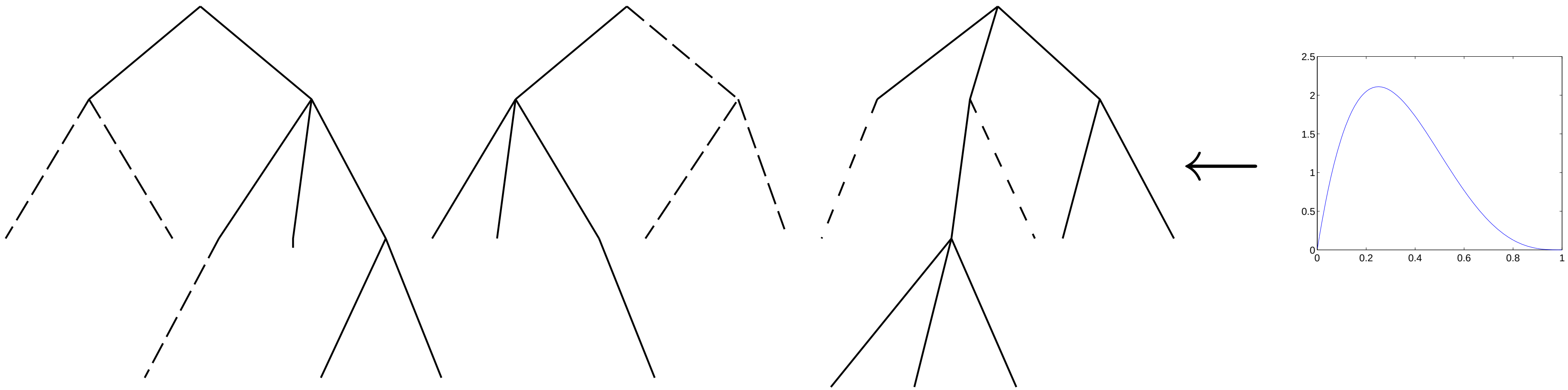
(b) Markov orders used to generate each character above.

- Character-based Markov model trained on “Alice in Wonderland”.
  - Lowercased alphabets + space
  - OCR, compression, Morphology, . . .

Final Remarks

- Hyperparameter sensitivity and empirical Bayes optimization $\Rightarrow$ Paper
- LDA extension $\Rightarrow$ Paper (but partially succeeded)
- Comparison with Entropy Pruning (Stolcke 1998) $\Rightarrow$ Poster
- Poster: W24 (near the escalator).

Summary



- We introduced the Infinite Markov model where the orders are unspecified and unbounded but can be learned from data.
- We defined a simple prior for stochastic infinite trees.
- We expect to use it for latent trees:
  - Variable resolution hierarchical clustering (cf. hLDA)
  - Deep semantic categories just when needed.
- Also for variable order HMM (pruning approach: Wang+, ICDM 2006)

Future Work

- Fast variational inference
  - Obviates Gibbs for inference and prediction
  - CVB for HDP: Teh et al. (this NIPS)
- More elaborate tree prior than a single Beta
- Relationship to Tailfree processes (Fabius 1964; Ferguson 1974)