UNIVERSITY OF
Waterloo

# A Stochastic Memoizer for Sequence Data

F. Wood, C. Archabeau, L. James, Y.W Teh (2009)

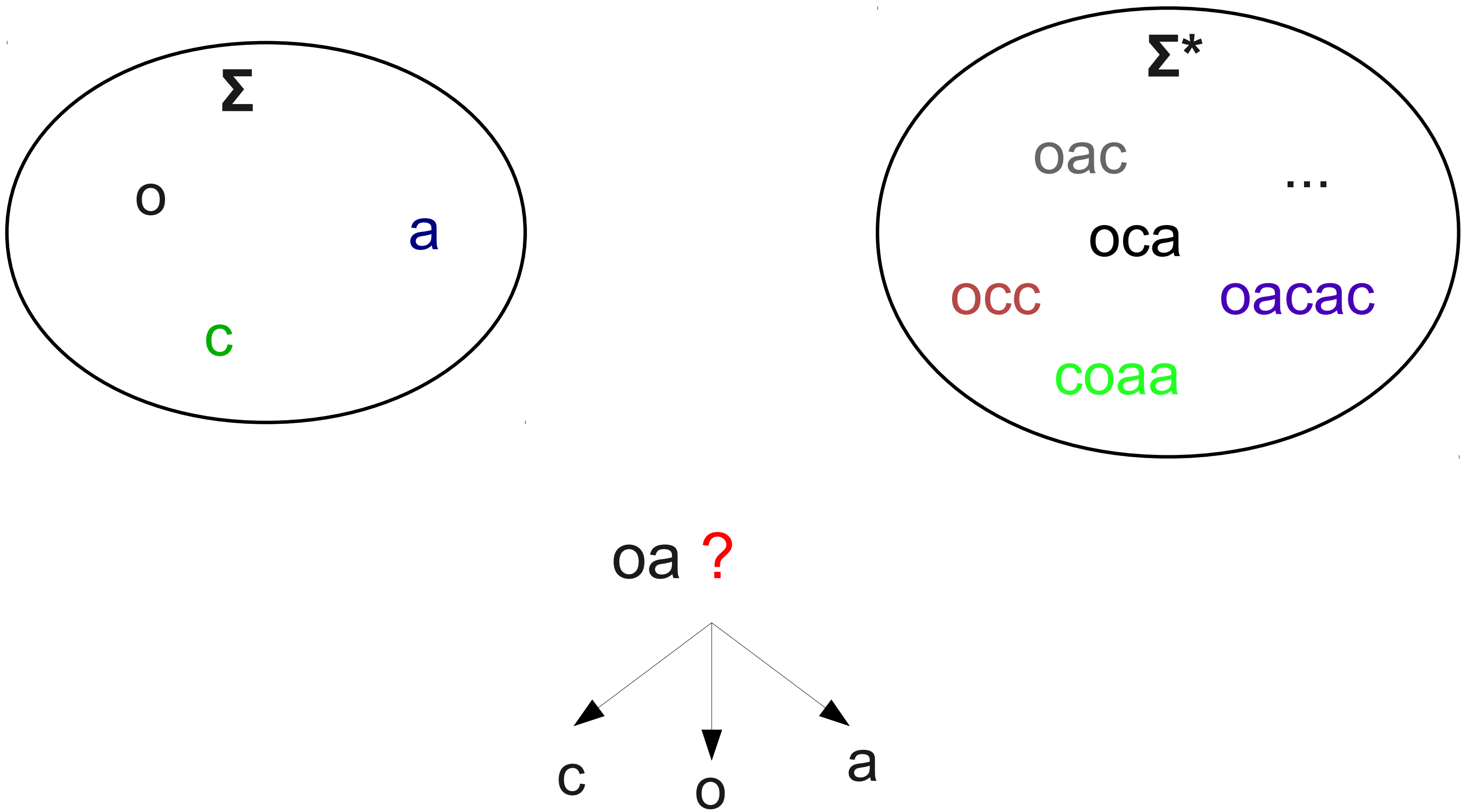
**Presented by: Sarah Nadi & Karim Ali**

**Nov. 2nd 2010 – CS 886**

# Outline

- Problem Overview
- Existing Techniques
- Proposed Solution
- Contributions
- Model Used
- Inference & Prediction
- Experiments
- Summary & Discussion

# Problem Overview



# Existing Techniques

- N-grams
  - Markov assumption applies
  - Given n-1 characters, what is the probability vector of the $n^{th}$ character (given our vocabulary)
  - Very good perplexity rates
  - Problems:
    - Determining optimal value of n
    - We need more training data

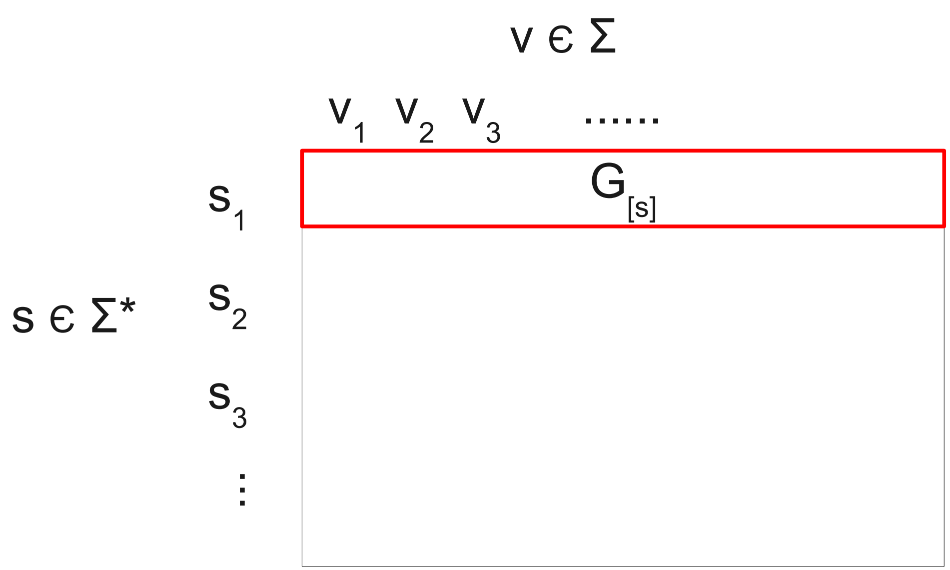
# Proposed Solution

- Use a non-Markov model
  - Next value in sequence is conditionally dependent on **all** preceding values → ∞-gram
  - This introduces a large number of latent variables
  - Use Pitman-Yor (PY) process with concentration parameter = 0
  - Use Hierarchial PY Process (HPYP) to allow comparing common suffixes

# Contribution

- Stochastic Memoizer
  - Probabilistic technique
  - Remembers previously returned values
    - Return previously returned symbols or a new symbol
  - Goals
    - Relate shorter sequences with longer sequences
    - Use a HPYP that is tractable

# Hierarchial Formulation



# Hierarchial Formulation *Cont'd*

H ~ Uniform Distribution

$G_{[\,]} \sim \mathcal{PY}(d_0, c_0, H)$

$G_{[aca]} \sim \mathcal{PY}(d_1, c_1, G_{[]})$

$G_{[oaca]} \sim \mathcal{PY}(d_{|s|}, c_{|s|}, G_{[aca]})$

H is the base measure
$d_n$ is the discount parameter
$c_n$ is the precision paramter

# Hierarchial Formulation *Cont'd*

H is the base measure
$d_n$ is the discount parameter
$c_n$ is the precision paramter

$H \sim$ Uniform Distribution

$G_{[]} \sim \mathcal{PY}(d_0, c_0, H)$

$G_{[s`]} \sim \mathcal{PY}(d_1, c_1, G_{[]})$

$G_{[s]} \sim \mathcal{PY}(d_{|\mathrm{s}|}, c_{|\mathrm{s}|}, G_{[\mathrm{s}']})$ [s] = [*v*s']

# Hierarchial Formulation *Cont'd*

$H$ ~ Uniform Distribution

$$G_{[\,]} \sim \mathcal{PY}(d_0, c_0, H)$$

$$G_{[s`\,]} \sim \mathcal{PY}(d_1, c_1, G_{[]})$$

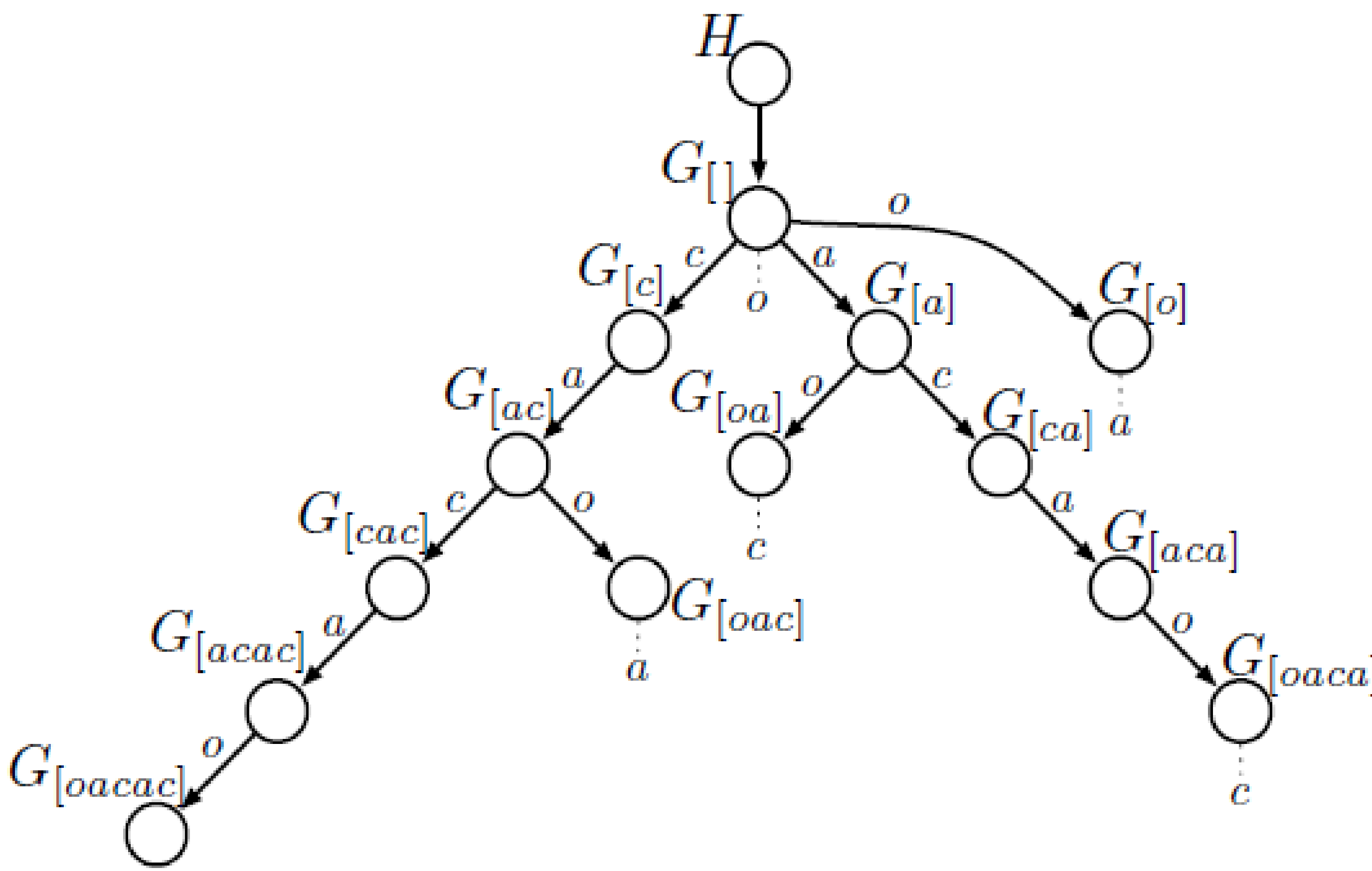
$$G_{[s]} \sim \mathcal{PY}(d_{|s|}, c_{|s|}, G_{[s']})$$

[s] = [*v*s']

...

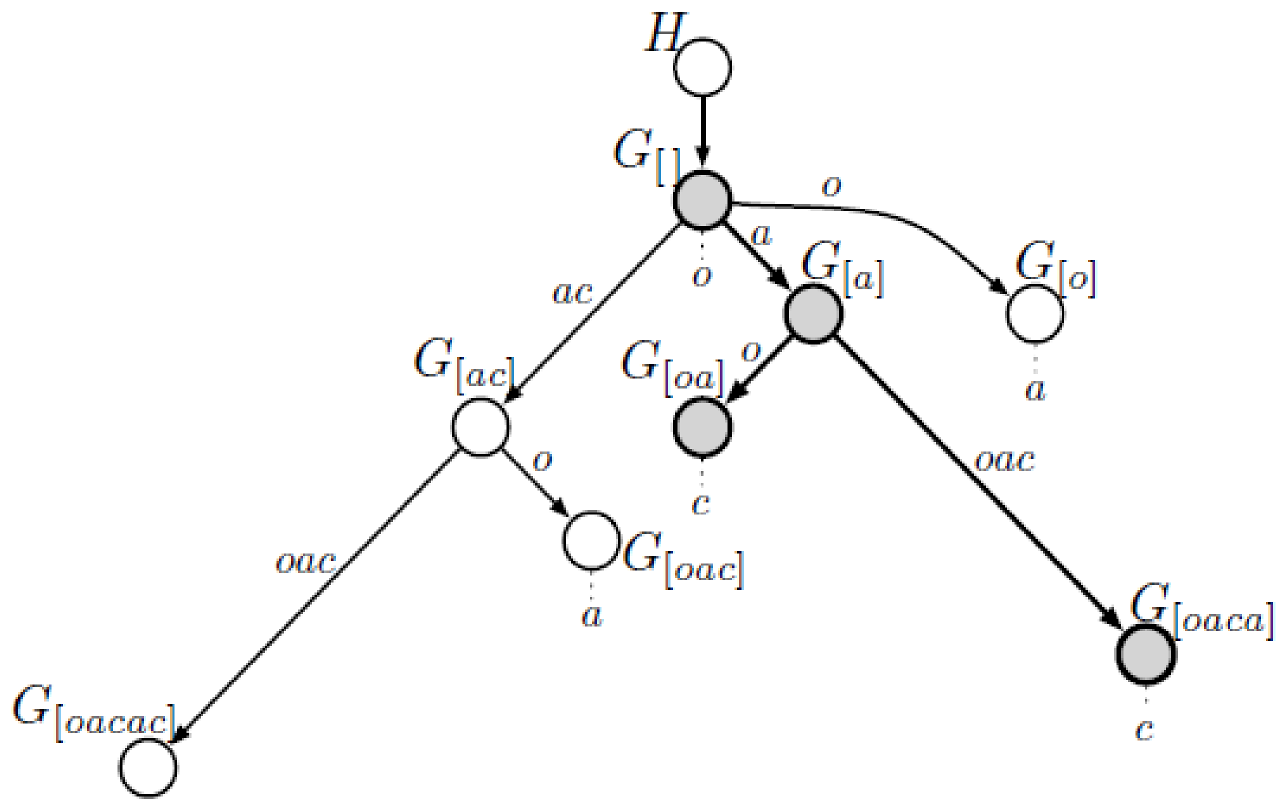
H is the base measure
$d_n$ is the discount parameter
$c_n$ is the precision paramter

# Prefix Tries



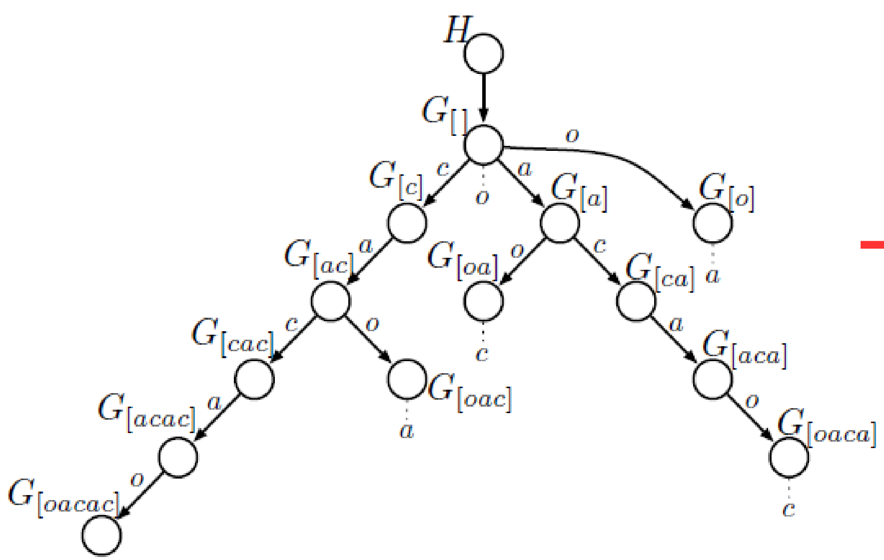

Prefix trie for *oacac*.

# Prefix Trees



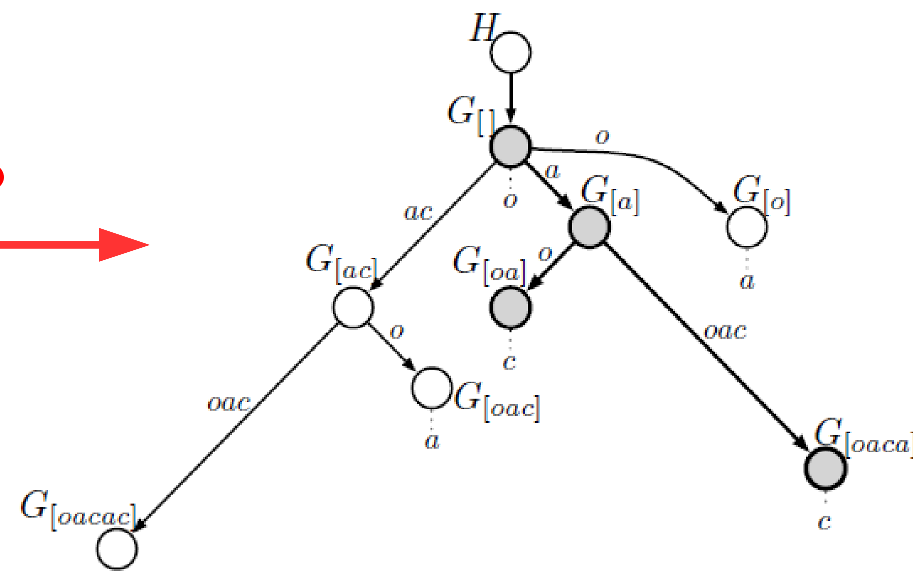

Prefix tree for *oacac*.

# Trie to Tree Conversion



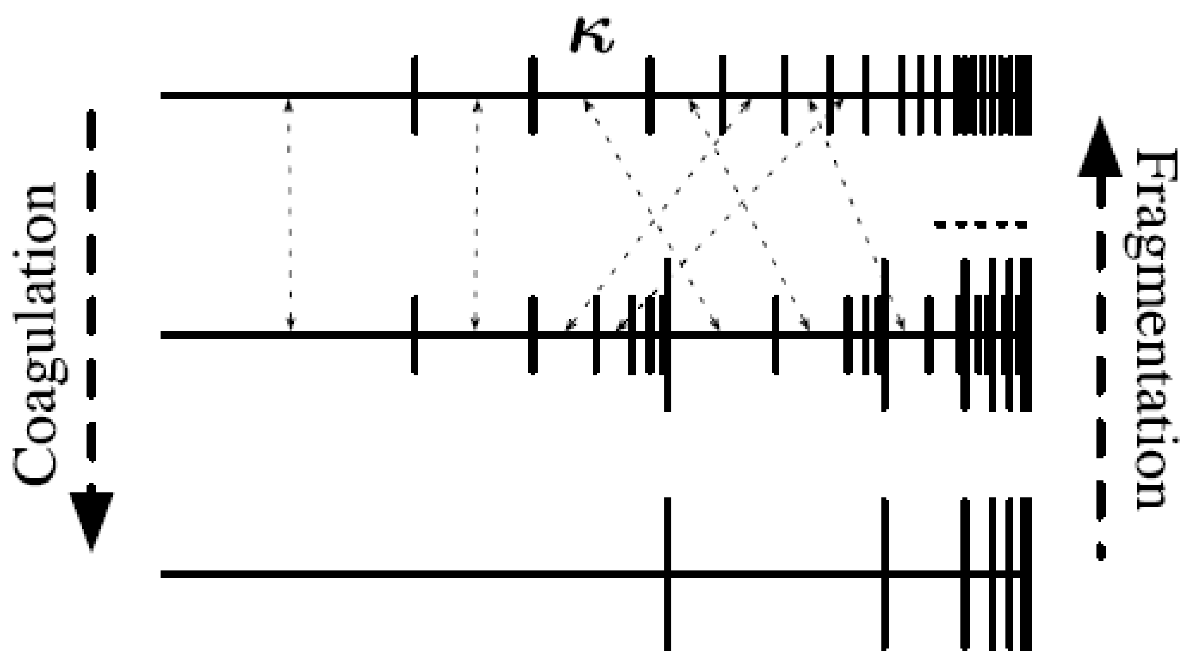

Prefix trie for *oacac*.

?



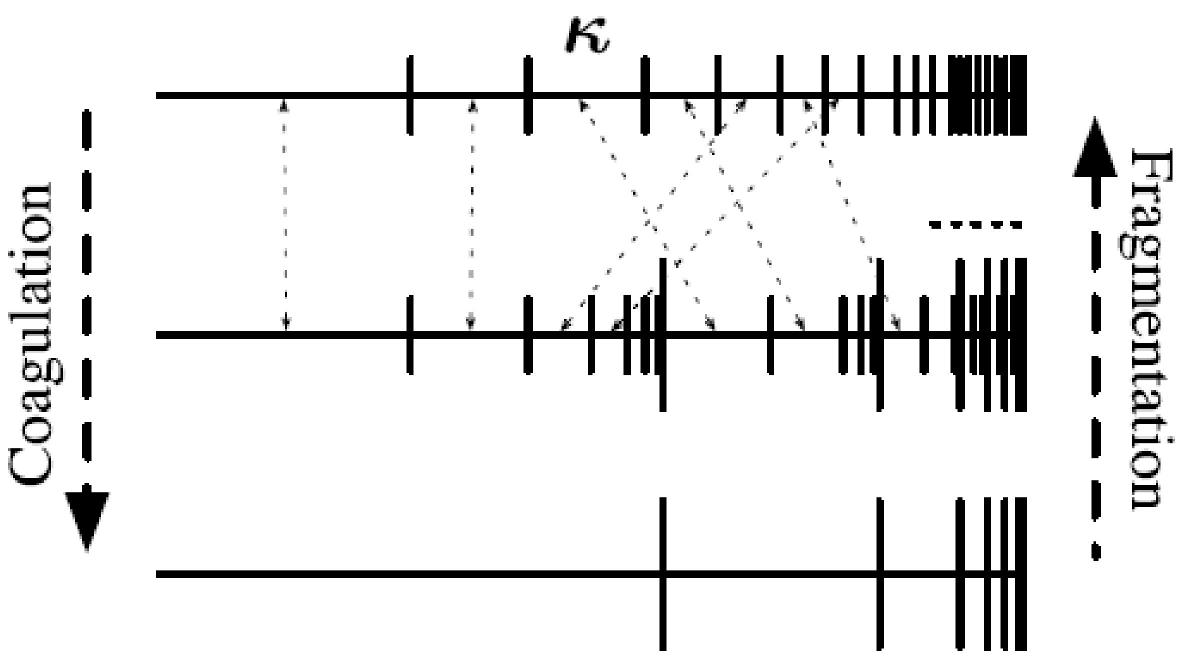

Prefix tree for *oacac*.

# Coagulation & Fragmentation



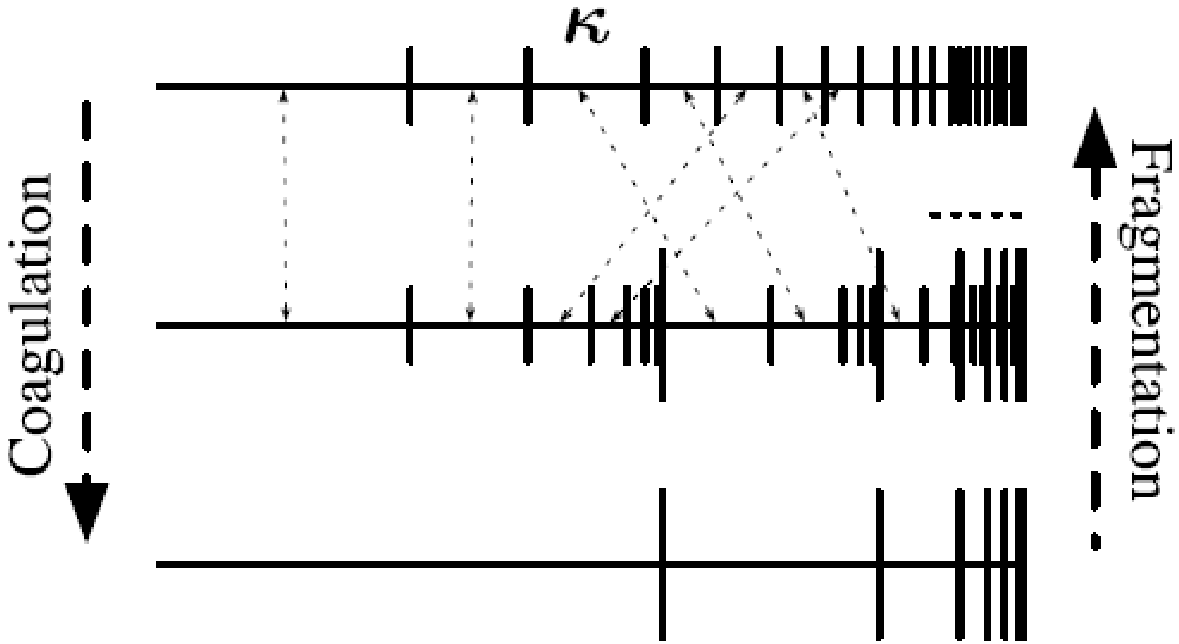
# Coagulation & Fragmentation

Construct collapsed model



# Coagulation & Fragmentation

**Construct collapsed model**




**Test sequences not in training data**

# Coagulation & Fragmentation *Cont'd*

$G_1$

$\downarrow$

$G_2 \quad \sim \mathcal{PY}(d_1, 0, G_1)$

$\downarrow$

$G_3 \quad \sim \mathcal{PY}(d_2, 0, G_2)$

# Coagulation & Fragmentation *Cont'd*

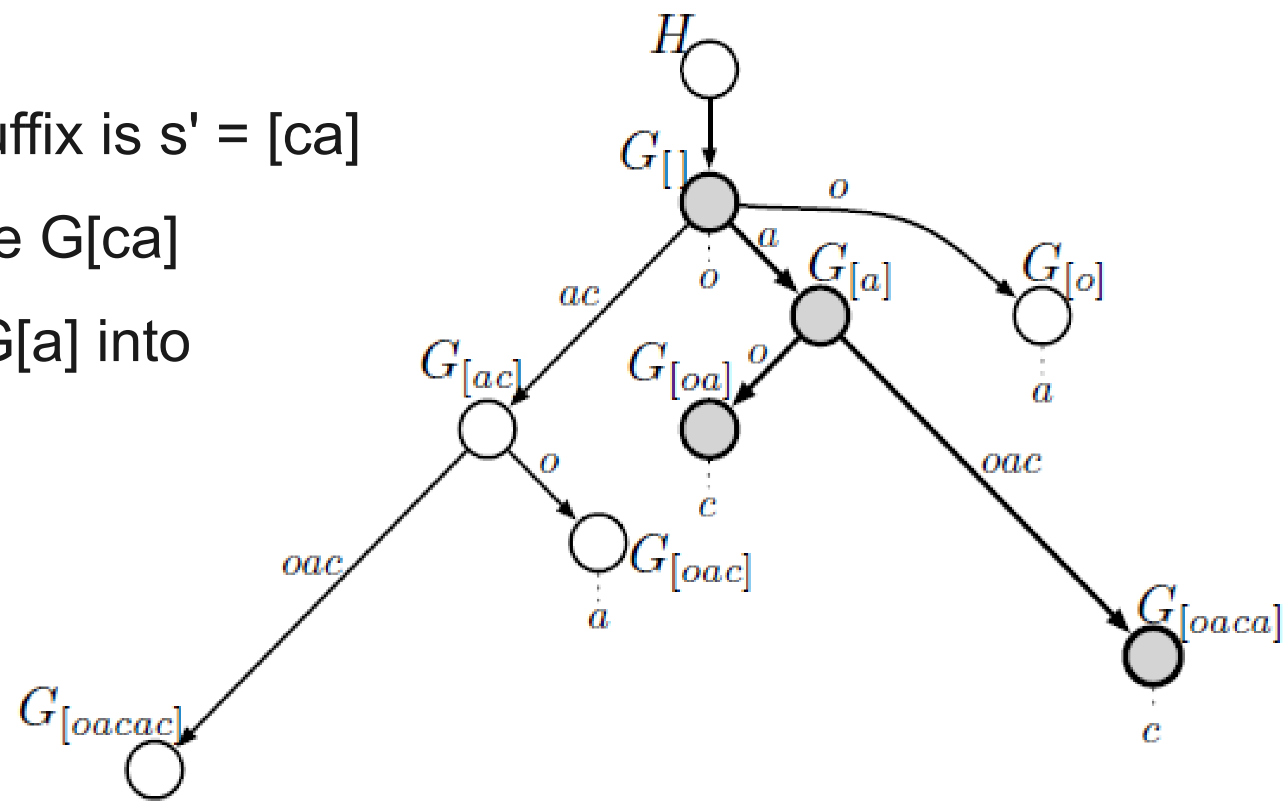
$$\begin{tikzcd}
G_1 \arrow[d] & & G_1 \arrow[dd] \\
G_2 \sim \mathcal{PY}(d_1, 0, G_1) \arrow[d] \arrow[r, "\text{Marginalize out } G_2"] & {} & \\
G_3 \sim \mathcal{PY}(d_2, 0, G_2) & & G_3 \sim \mathcal{PY}(d_1 d_2, 0, G_1)
\end{tikzcd}$$

# Inference & Prediction

- Same as HPYP
- Use Gibbs sampling in the Chinese Restaurant Franchise (CRF) representation
- $E(G_{[s]}(v)) = E(G_{[s']}(v))$ where s' is the longest common suffix of s
- Need to be able to compute the probability of a symbol ***v*** given a sequence ***s*** that is not in the training data

# Unseen Sequences

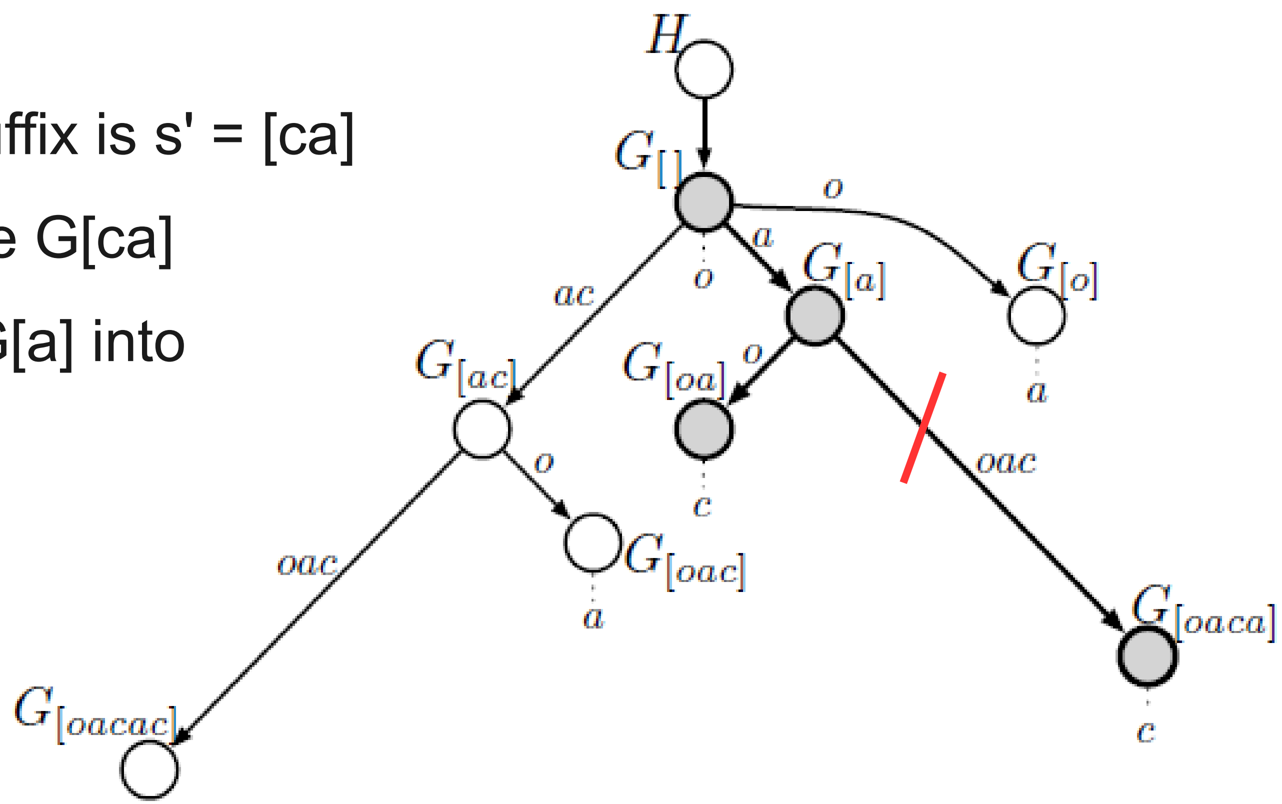
- Consider s = [oca]
- Longest common suffix is s' = [ca]
- Need to reinstantiate G[ca]
- Fragment G[oaca]|G[a] into **(G[ca] | G[a])** and **(G[oaca] | G[ca])**




Prefix tree for *oacac*.

# Unseen Sequences

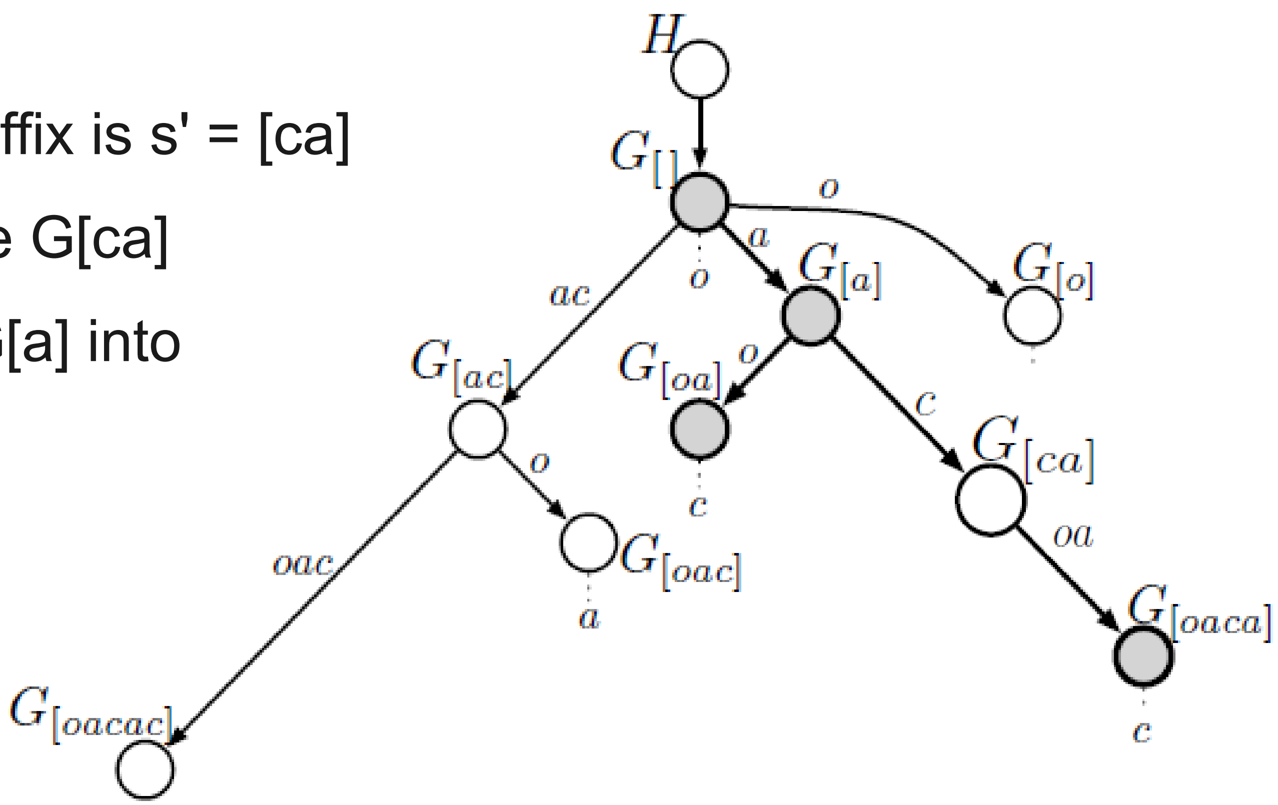
- Consider s = [oca]
- Longest common suffix is s' = [ca]
- Need to reinstantiate G[ca]
- Fragment G[oaca]|G[a] into **(G[ca] | G[a])** and **(G[oaca] | G[ca])**




Prefix tree for *oacac*.

# Unseen Sequences

- Consider s = [oca]
- Longest common suffix is s' = [ca]
- Need to reinstantiate G[ca]
- Fragment G[oaca]|G[a] into **(G[ca] | G[a])** and **(G[oaca] | G[ca])**



Prefix tree for *oacac*.

# Experiments

# Evaluation Questions

- Do prefix trees provide computational savings?
- Does the sequence memoizer's (∞-gram) performance compare to that of an n-gram model?

# Data Sets

- Associated Press (AP) corpus
  - *Vocabulary:* 1 million words
  - *Training:* 15 million words
  - *Testing:* 1 million words
  - *Preprocessing:* Replace low frequency words with a single ”unknown word” symbol
- New York Times (NYT) corpus
  - *Vocabulary:* 150,000 words
  - *Training: 13* million words
  - *Testing:* 200,000 words
  - *Preprocessing:* none

# 1) Computational Savings

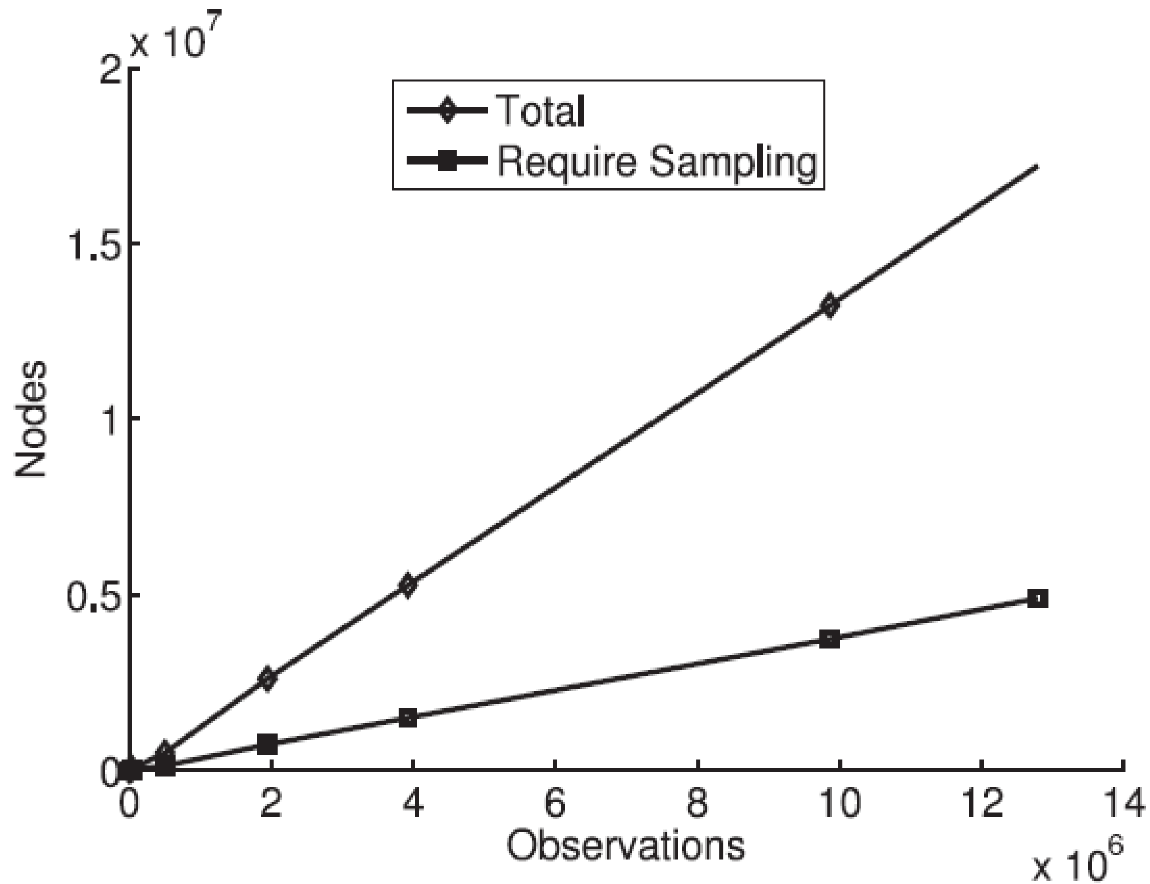
- Use the NYT data set & compare with n-gram model
- Metropolis-Hastings updates used for discount parameters
- Use distinct discount parameters for each of the first 4 levels of the trie, while levels below use a single shared discount parameter

$$d_{[0,1,2,\ldots]} = (.62, .69, .74, .80, .95, .95, \ldots)$$

# Computational Savings Results

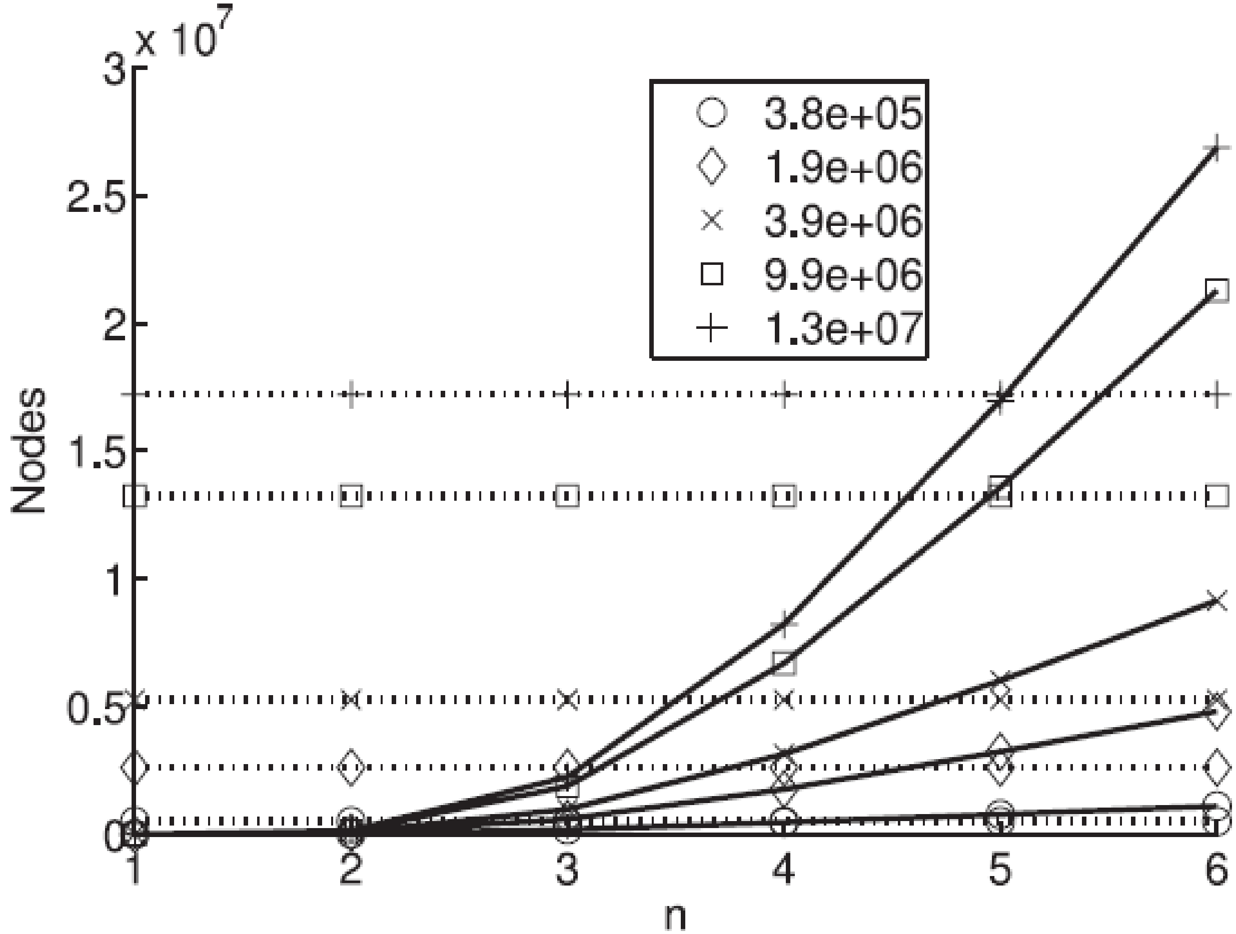
- An ∞-gram with 10 burn-in iterations & 5 samples produced same perplexity scores as a 3-gram model with 125 burn-in iterations & 175 samples

# Computational Savings Results *Cont'd*



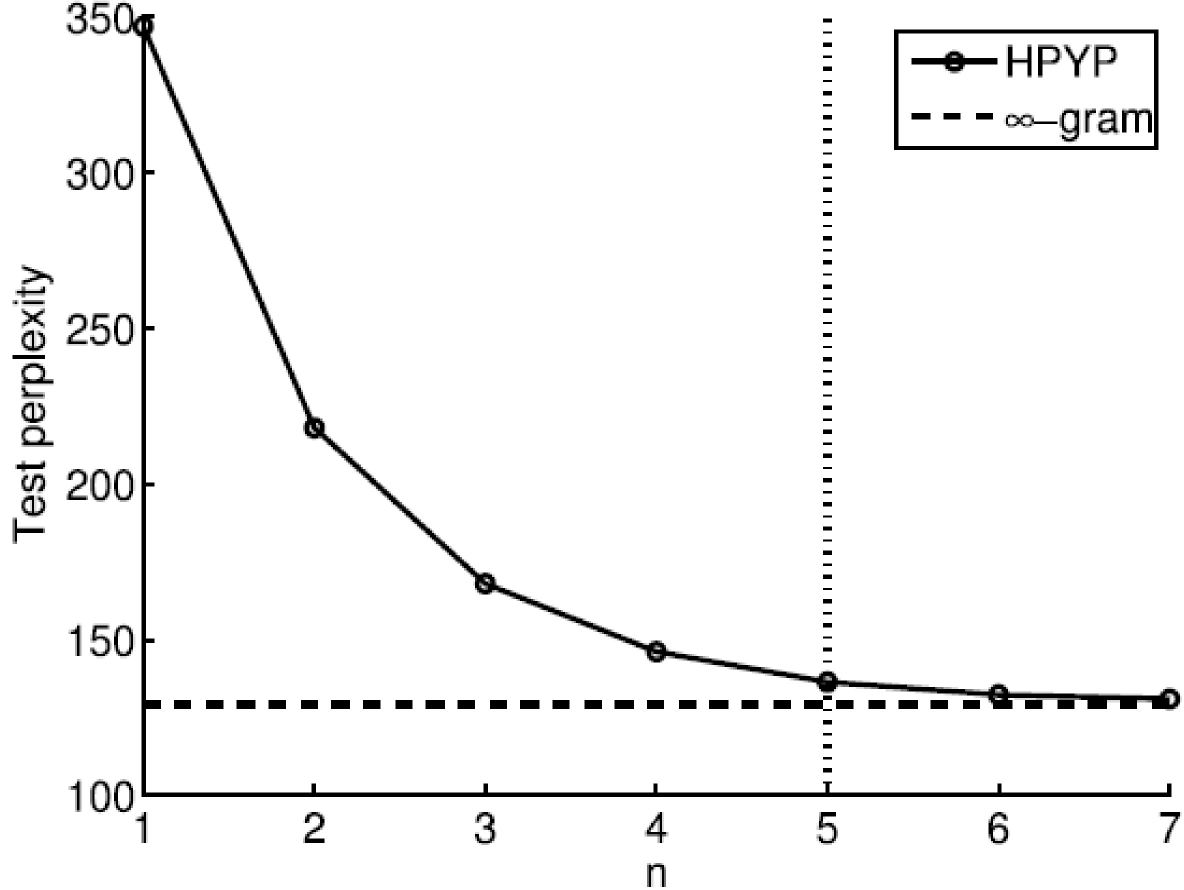

NYT data set

# Computational Savings Results *Cont'd*



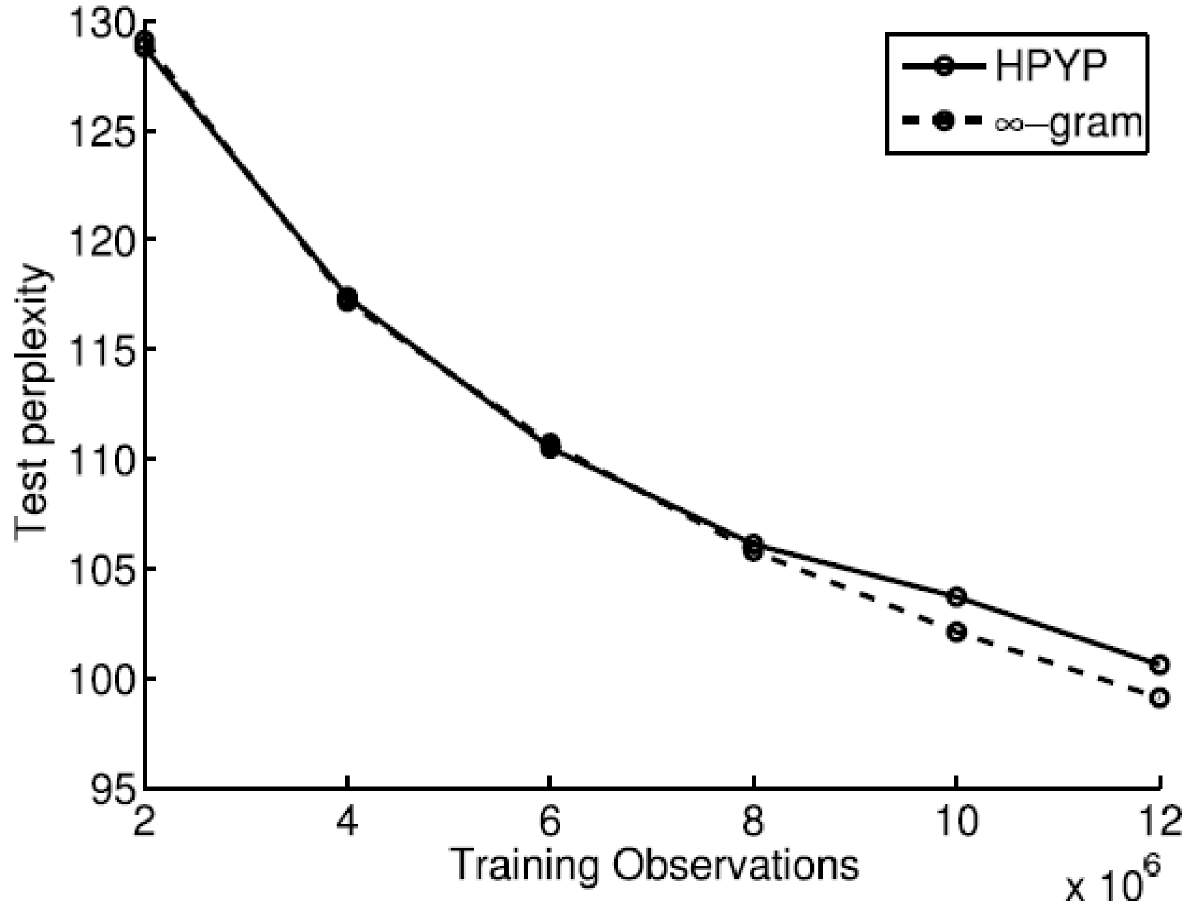

NYT data set

# 2) Performance Results




NYT data set

# Performance Results *Cont'd*




AP data set

# Take Home Message

- N-gram models perform based on the choice of n
- An ∞-gram model relieves this constraint
  - Common suffixes used to collapse prefix trie
  - Achieves **at least** same perplexity as n-gram
  - In most cases, saves computation & storage

# Discussion

- Complexity of coagulation & fragmentation processes
- Frequency of fragmentation during experiments
- Intuition of setting concentration parameters to 0
- Computational savings on AP corpus
- Did not mention which n-gram sampler had 125 burn-in & 175 samples
- Different values for discount parameters in lower levels

Thank You
??

# Reinstantiating G[ca]|G[c] "restaurant"