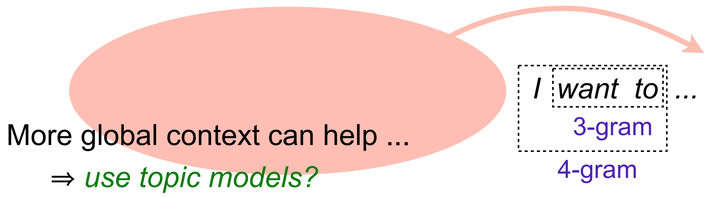
Improvements to The Bayesian Topic N-gram Models

Hiroshi Noji[1,3] Daichi Mochihashi[2,3] Yusuke Miyao[1,3]

1. National Institute of Informatics, Tokyo
2. The Institute of Statistical Mathematics, Tokyo
3. Graduate University for Advanced Studies

Overview

- The predictions of n-gram language models are very *local*

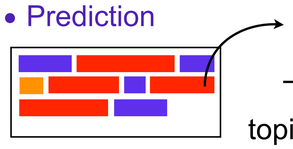


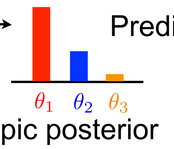
- Problems of the previous n-gram + topic models:
  - Prediction get much sparse *with higher order n-grams*
  - Inference (local Gibbs) is not very efficient
  - Resolve these with *hierarchical prior* + *blocked sampling*

Basic model (Wallach' 06)

	topic 1	topic 2	topic 3
unigram	$G^1_\phi \sim \mathrm{PYP}(\frac{1}{V})$	$G^2_\phi \sim \mathrm{PYP}(\frac{1}{V})$	$G^3_\phi \sim \mathrm{PYP}(\frac{1}{V})$
bigram	$G^1_v \sim \mathrm{PYP}(G_\phi)$	$G^2_v \sim \mathrm{PYP}(G_\phi)$	$G^3_v \sim \mathrm{PYP}(G_\phi)$
3-gram	$G^1_{wv} \sim \mathrm{PYP}(G_v)$	$G^2_{wv} \sim \mathrm{PYP}(G_v)$	$G^3_{wv} \sim \mathrm{PYP}(G_v)$

- Prediction

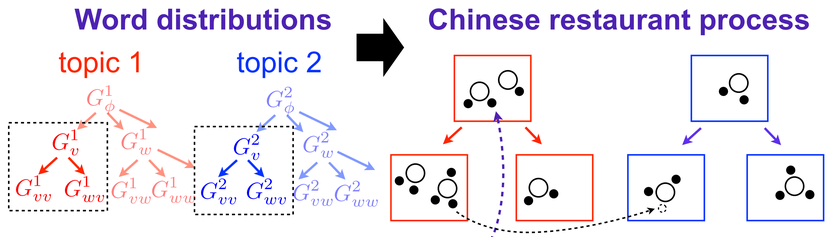




Predict by a mixture of n-grams:

$$\theta_1 G^1_{wv} + \theta_2 G^2_{wv} + \theta_3 G^3_{wv}$$

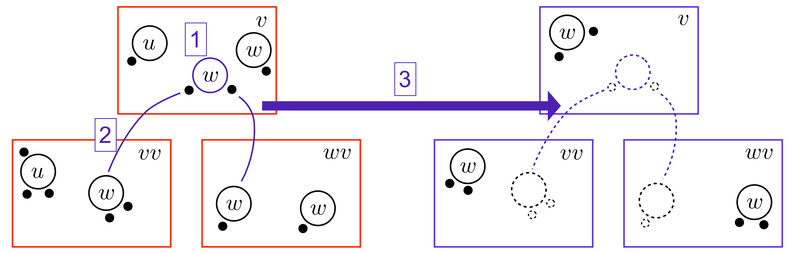
Problem of the naive Gibbs sampler



A table is connected to a customer on the parent restaurant

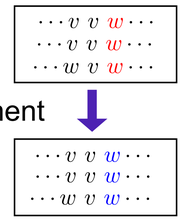
This customer *will never move* unless a table is removed in its child restaurants ⇒ slow mixing especially *with higher-order* !

Proposed blocked sampler



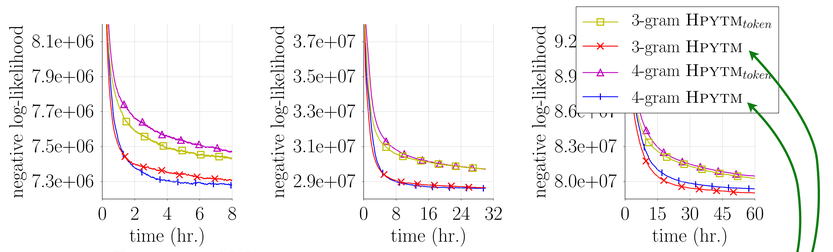
1. Select *a pivot table* on a *internal* restaurant
2. Reconstruct parent-child relations randomly
3. Move the block with holding the seating arrangement (the posterior can be calculated in a closed form)

Change many word topics across documents



Effects of the blocked sampler

- Use Wallach's model for simplicity
  - It can be applied to our extended models with a little effort



Blocked moves lead much faster convergences

Perplexity results

- BNC (10M) with 4-gram models
  - # topics=100 (not so sensitive)
- Our models are *normalization-free*:
  - much faster prediction than rescaling methods
- Conclusion to the model design:
  - *Flat structure* between the global and topics is better than *hierarchical structure* ⇒ *but why?*

HPYLM (no topic)	169.2
Wallach	140.4
Wallach **+ block**	**133.1**
Unigram rescaling	130.3
Hierarchical	129.0
Switching	**125.5**

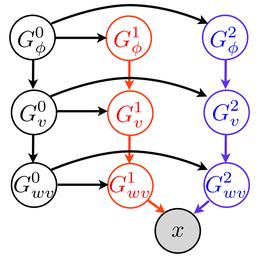
Hierarchical priors for ease of sparseness

- Motivation
  - We don't want to assign all n-grams topics equally

local context	candidates	
... in order	⇒ *to, that*	*not require* topics
... would like	⇒ *you, to*	
... state of	⇒ *Washington, the*	*require* topics
... the new	⇒ *york, algorithm*	

- Comparing two models encoding this difference
  - **Hierarchical:** the global model is used as a prior
  - **Switching:** two models are switched

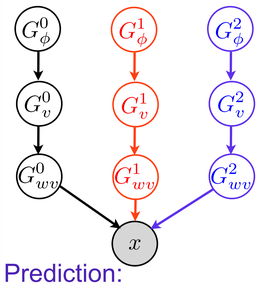
Hierarchical



Graphical-PYP (Wood/Teh '09)

$$G^1_{wv} \sim \mathrm{PYP}(\lambda G^1_v + (1-\lambda) G^0_{wv})$$

Switching



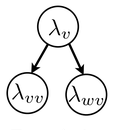
Prediction:

$$\lambda \sum_k \theta_k G^k_{wv} + (1-\lambda) G^0_{wv}$$

Topic part

lambda hierarchies

- Property
  - *the new, of new* both require topics
- Share statistical strengths with *hierarchical Betas*



$$\lambda_{\text{the new}} \sim \mathrm{Beta}(\gamma\lambda_{new}, \gamma(1-\lambda_{new}))$$
$$\lambda_{\text{of new}} \sim \mathrm{Beta}(\gamma\lambda_{new}, \gamma(1-\lambda_{new}))$$

Posterior inspections

- **Switching** assigns only some part of words topics

there has been much recent work on measuring image statistics and on learning probability distributions on images . we observe that the mapping from images to statistics is many-to-one and show it can be quantified by a phase space factor .

  - lighter words are assigned the global model (topic 0)
  - might led to more *accurate topic prediction*
- **Hierarchical** also learns differences of contexts, but all words are assigned topics

λ_h	h
0.0–0.1	in spite, were unable, a sort, on behalf, . regardless
0.5–0.6	assumed it, rand mines, plans was, other excersises
0.9–1.0	that the, the existing, the new, their own, and spatial

References

Hanna M. Wallach. 2006. Topic modeling: beyond bag- of-words. *In Proc of ICML '06*, pages 977–984.

Frank Wood and Yee Whye Teh. 2009. A hierarchical nonparametric Bayesian approach to statistical language model domain adaptation. *In Proc of AISTATS*, volume 12.