Cross Document Entity Disambiguation

August 22, 2007

Johns Hopkins Summer Workshop

Person Entities for CDC

- <u>Corpora:</u> David Day, Janet Hitzeman
- <u>Relations:</u> Su Jian, Stanley Yong, Claudio Guiliano, Gideon Mann
- <u>CDC Features:</u> Jason Duncan, Paul McNamee, Rob Hall, Mike Wick
- <u>Clustering/Machine Learning:</u> Mike Wick, Rob Hall

Problem Description

- Disambiguate entities across a corpus
  - **Document level**, **Entity level**, Mention level
- Document level disambiguation / Web People Corpora
  - SPOCK corpus (description – challenge page – discussion forum)
- High ambiguity level
  - “The Jim Smith Society” | “James Smith”

Google Search for 'James Smith'

- **James Smith Cree Nation** **James Smith** Cree Nation P.O. Box 1059 Melfort, Saskatchewan S0E 1A0. Ph: (306) 864–3636 Fx: (306) 864–3336. www.sicc.sk.ca/bands/bjames.html - 1k - Cached - Similar pages
- **James Smith (political figure) - Wikipedia, the free encyclopedia** **James Smith** (about 1719 – July 11, 1806), was a signer to the United States Declaration of Independence as a representative of Pennsylvania. **...** en.wikipedia.org/wiki/James_Smith_(political_figure) - 22k - Cached –
- **Band Details** Official Name, **James Smith**. Number, 370. Address, PO BOX 1059, MELFORT, SK. Postal Code, S0E 1A0. Phone, (306) 864-3636. Fax, (306) 864-3336 sdiprod2.inac.gc.ca/fnprofiles/FNProfiles_DETAILS.asp?BAND_NUMBER=370 - 12k
- **Comox Valley Real Estate: James Smith, your Realtor for Comox ...** **James Smith** is your realtor for the Comox Valley area, including Comox, Courtenay, Cumberland, Union Bay, Royston, and Black Creek. www.jamessmith.ca/ - 10k - Cached - Similar pages
- **Watercolor Snapshots - by James Smith** Watercolor Snapshots by **James Smith** - your portrait custom painted in watercolor, or the portrait of your relative or friend, painted from your 4 x 6 **...** 28k - Cached - Similar pages

Problem Description

- **Entity level disambiguation (ACE 2005 + CDC annotatation)**
  - PER, GPE, LOC, and ORG entities that have a NAME string on the coreference chain AND are +SPECIFIC
- Low ambiguity level (B-cubed baseline of 0.80 F versus 0.09 F for Spock corpus for “shatter all” condition, one node per cluster)

Features from previous work

- Document level bag of words/NERentities features (basically all previous systems)
- Local contextual information (bags of words/NER entities in local context.
- Syntactic Features (base NPs in document/local contexts Chen & Martin)
- Basic Relational Information (Mann & Yarowsky). DOB, POB, etc.

Three areas where workshop can make a contribution

1. **More and better features**
2. **More use of relation information** from varying sources (ground truth and system generated relations, supervised and unsupervised)
3. **Different clustering procedures** than standard greedy single-link agglomerative clustering

Experiments

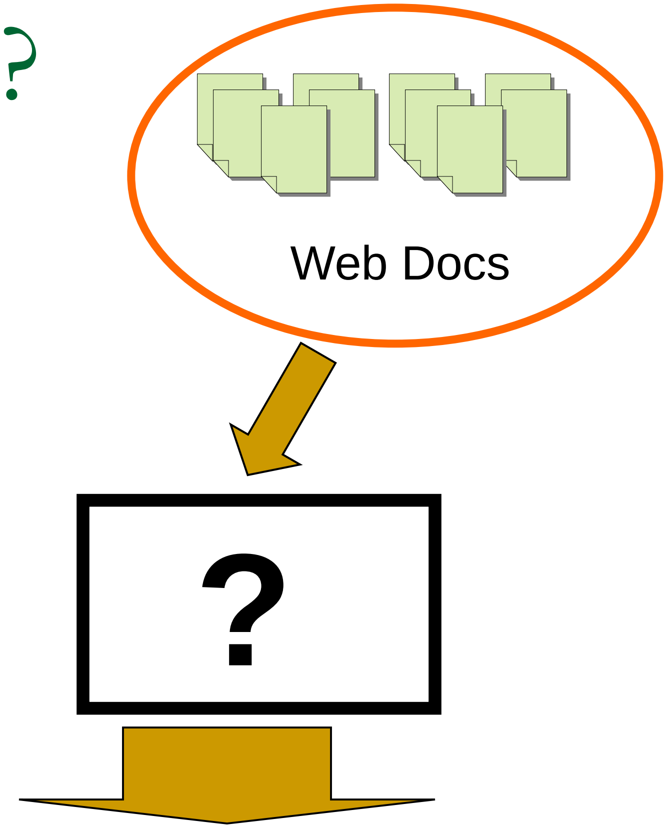
- Document Level Information (Bow, Boe)
- Mention Level Information (Bow, Boe)
- Topic Models (SPOCK)
- Relations
  - ACE (supervised): ORG-AFF 82-F, PER-SOC 91-F
  - SPOCK (unsupervised)
- Different Clustering Techniques (SPOCK)

Entity Disambiguation

Michael Wick

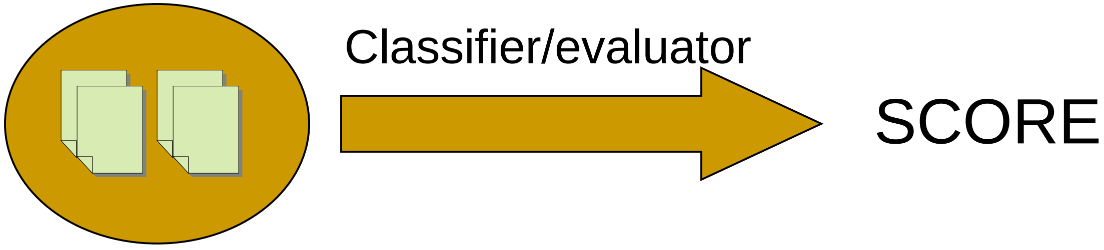
Which J. Smith?

- Jazz Musician
- Politician
- Student
- Investor
- CEO
- Historian






First Order Model



$$P(Y|X) = 1/Z_x \prod f(y_i,x_i) \prod (1-f(y_{ij},x_{ij}))$$

Find configuration to maximize this objective function

Greedy Agglomerative Approximation

? = (Traditional Features)

- Bags of Words
- Title overlaps
- Named Entities
- Chunking
- TFIDF term selection

Two articles on John Smith, the jazz musician

Which J. Smith?

- Jazz Musician
- Student
- Investor
- CEO
- Historian

…his melodic improvisation..

…and John Smith on saxophone..

1 university program learning students education
2 ashley love ash miss hey
3 tx la christi corpus san
4 company insurance financial income investor
5 contact email city state fullname
6 masters music jazz folk orchestra
7 registered status category read japan
8 museum county historical society kansas
…tography times braunfels jp rail
10 …n alex films kill night
11 senate senator section court company




Results With Topics

	Precision	**Recall**	**F1**
B-Cubed	.32	.24	.28
+topics	.23	.44	**.30**
Pairwise	.12	.19	.15
+topics	.13	.44	**.20**
MUC	.70	.65	.67
+topics	.84	.86	**.85**

Chunks+TitleOverlap + TFIDF + NER + Relations + **TOPICS!!!**

Metropolis-Hastings Sampler

$$P(Y|X) = 1/Z_x \prod f(y_i,x_i) \prod (1-f(y_{ij},x_{ij}))$$

Requires summing over all possible configurations

$$P(y'|y) = \mathbf{Max[}\ p(y')/p(y) * p(y|y')/p(y'|y), 1\mathbf{]}$$

Likelihood ratio: partition function cancels!

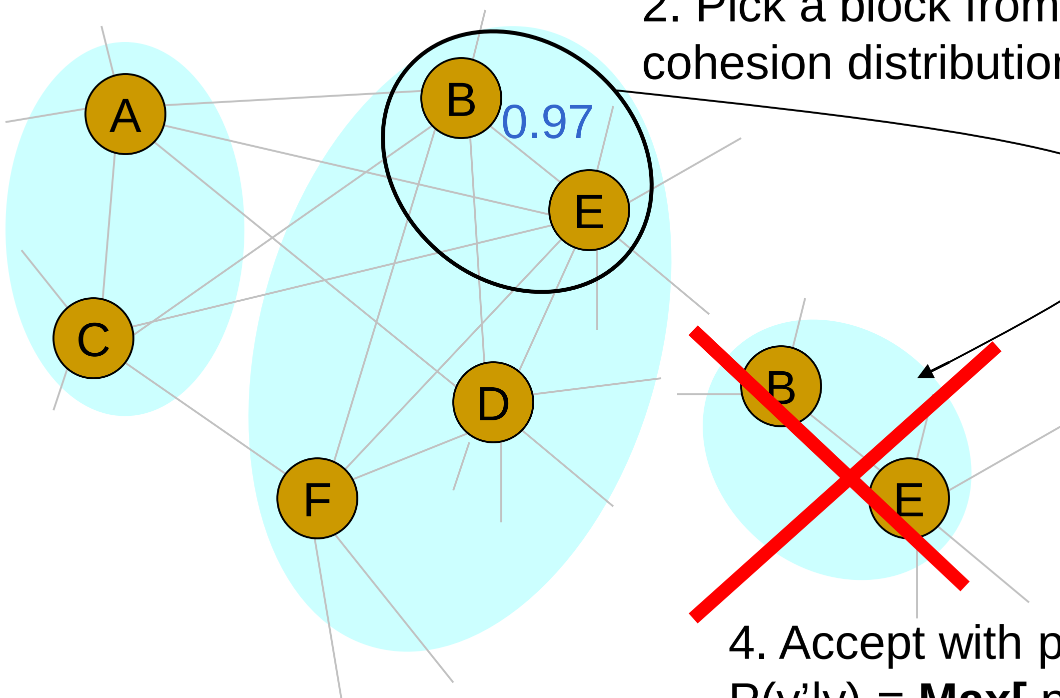
Probability of Accepting Jump

Inverse ratio of making a jump vs. reversing that jump

Metropolis-Hastings

1. Initialize with Greedy Agglomerative

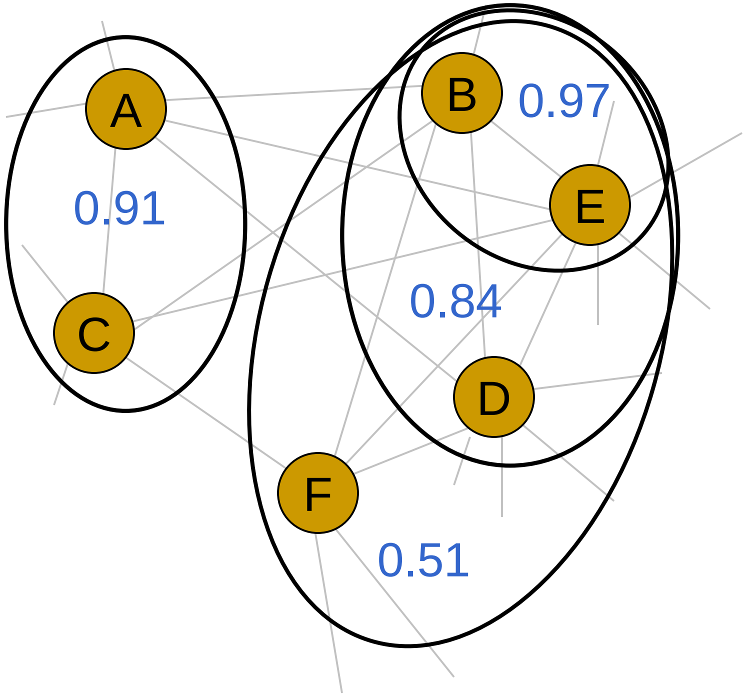
2. Pick a block from cohesion distribution



3. Pick new assignment uniformly at random

4. Accept with probability:
P(y’|y) = **Max[** p(y’)/p(y) * p(y|y’)/p(y’|y),1**]**
(in this example we reject)

Deriving Cohesion Distribution



Block Distribution

BE	.97
AC	.91
BED	.84
BEDF	.84

Results

	Precision	Recall	F1
B-Cubed	. 318	. 312	.315
w/MH	. 318	. 435	**.367**
Pairwise	.271	.243	.256
w/MH	.278	.370	**.317**
MUC	.838	.851	.844
w/MH	.863	.877	**.870**

Metropolis Hastings   Greedy Agglomerative

CDC results

	Precision	Recall	F1
B-Cubed	.96	. 88	**.92**
+DOC	.97	.90	**.93**
+MEN	.96	.90	**.93**
+CHAIN	.88	.93	**.91**
+LEX	.97	.95	**.96**
+REL	.97	.95	**.96**
Pairwise	.92	.57	**.71**
+DOC	.94	.70	**.80**
+MEN	.94	.70	**.80**
+CHAIN	.80	.87	**.84**
+LEX	.95	.88	**.91**
+REL	.95	.88	**.91**
MUC	.89	.78	**.83**
+DOC	.91	.79	**.85**
+MEN	.90	.79	**.84**
+CHAIN	.74	.83	**.79**
+LEX	.92	.87	**.89**
+REL	.92	.87	**.89**

BCubed (shattered): pr=1.0 re=.67 **F1=.80**

Conclusions

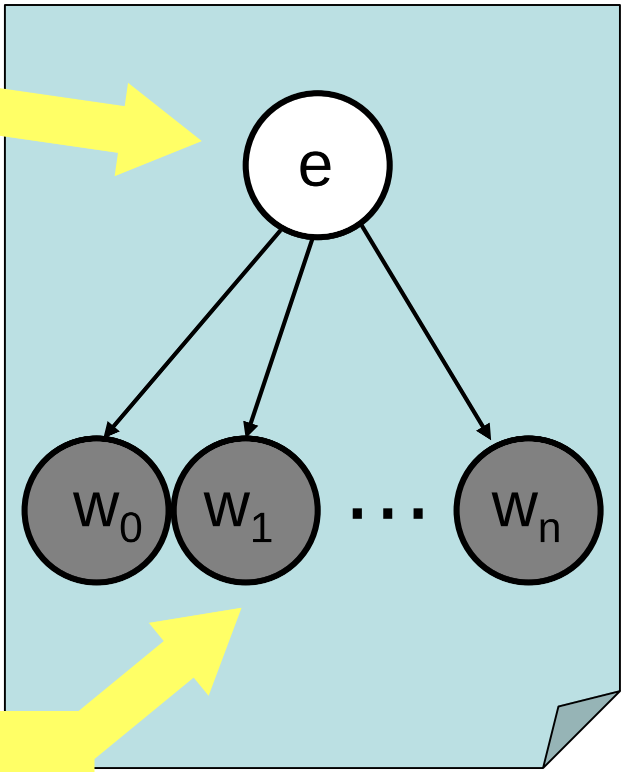
- Topics enable additional and powerful features
- Metropolis-Hastings improves upon greedy method

Generative, Unsupervised Models for Web People Disambiguation

Rob Hall (UMass Amherst)

A Simple Generative Process

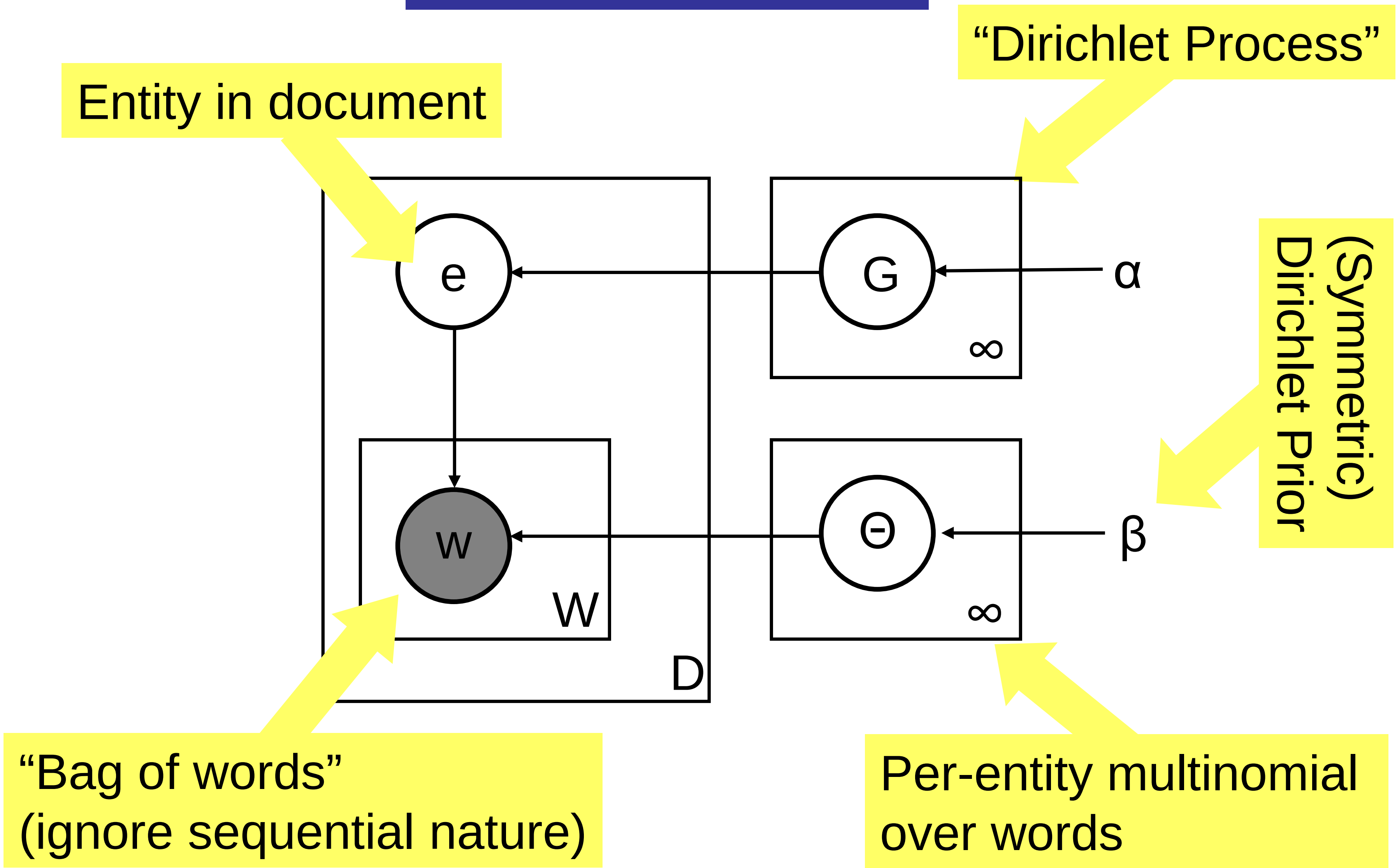
Each document "belongs" to some hidden entity.



Cluster by determining the mode of the posterior distribution of e given the words.

A sequence of observations is generated, conditioned on the entity.

Word Model



"Dirichlet Process"
Entity in document
e
G
α
∞
(Symmetric) Dirichlet Prior
w
W
Θ
β
∞
D
"Bag of words"
(ignore sequential nature)
Per-entity multinomial
over words

Approximate Inference: Collapsed Gibbs

Start with some initialization of all e.

Then resample each e^k in turn from the distribution:

$$p(e^k \mid \tilde{e}^{-k}, \tilde{w}) \propto p(\tilde{w}^k \mid e^k) \cdot p(e^k \mid \tilde{e}^{-k})$$

With $\mathbf{e}^{-k}$ sampled we can use CRP:

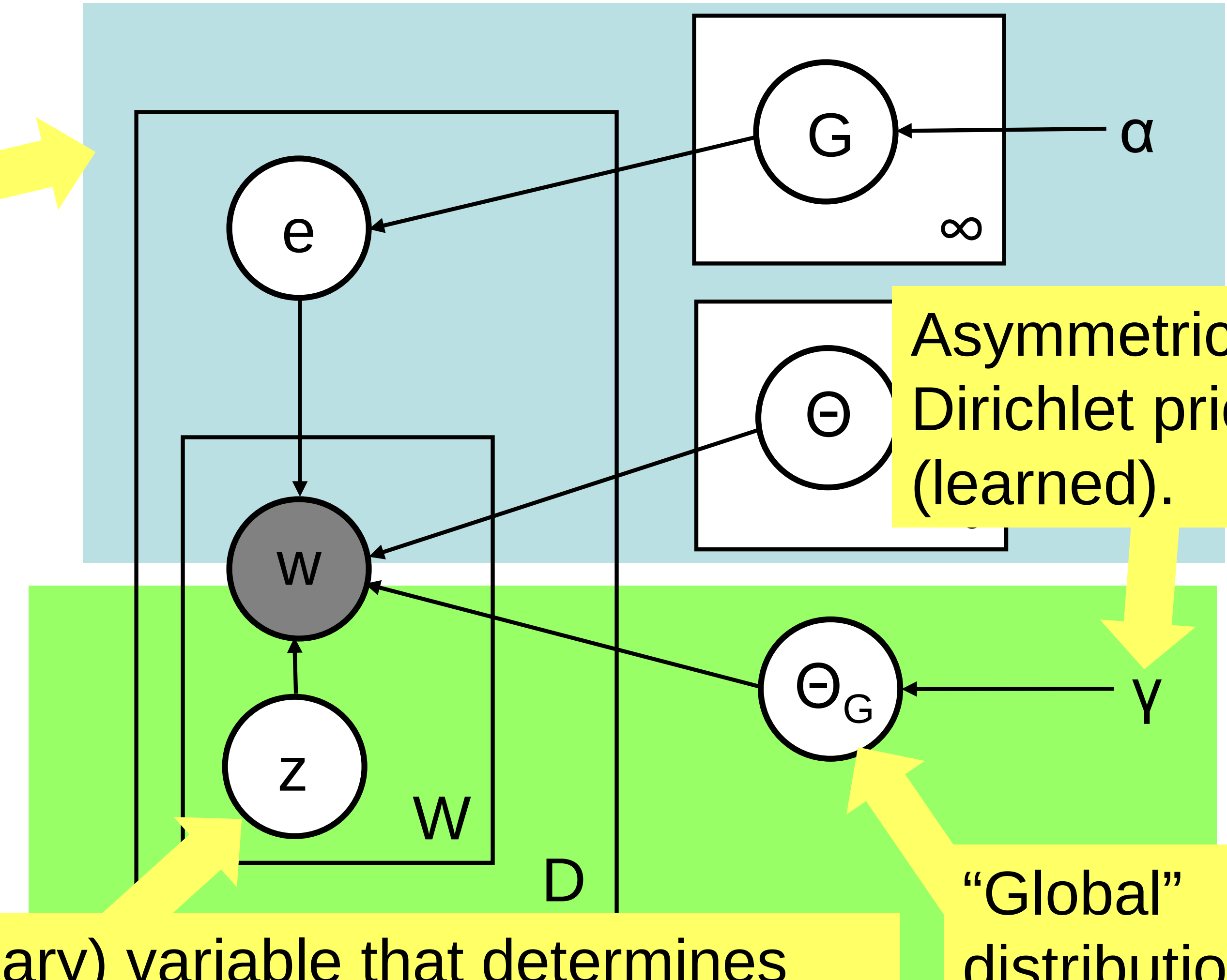
$$p(e^k = i \mid e^{-k}) \propto \frac{n_i}{n-1+\alpha} \qquad p(e^k = new \mid e^{-k}) \propto \frac{\alpha}{n-1+\alpha}$$

Then can integrate out **Θ** using Polya-urn scheme:

$$p(\tilde{w}^k = w_0 .. w_m \mid e^k) \propto \prod_i \frac{\beta + C_{e^k}^{w_i} + C_{0..i-1}^{w_i}}{\sum_{w \in V} \beta + C_{e^k}^{w} + C_{0..i-1}^{w}}$$

“Self-stopping” Word Model

Same as before



Asymmetric Dirichlet prior (learned).

“Global” distribution Over words

Bernoulli (binary) variable that determines whether w is drawn from the local or global distribution (gibbs sampled).

Approximate Inference: Collapsed Gibbs

The new probability model is:

$$p(e^k \mid \tilde{e}^{-k}, \tilde{w}) \propto p(\tilde{w}^k \mid \tilde{z}^k, e^k, \tilde{w}) \cdot p(\tilde{z}^k \mid e^k, \tilde{w}) \cdot p(e^k \mid \tilde{e}^{-k})$$

This requires sampling **z**:

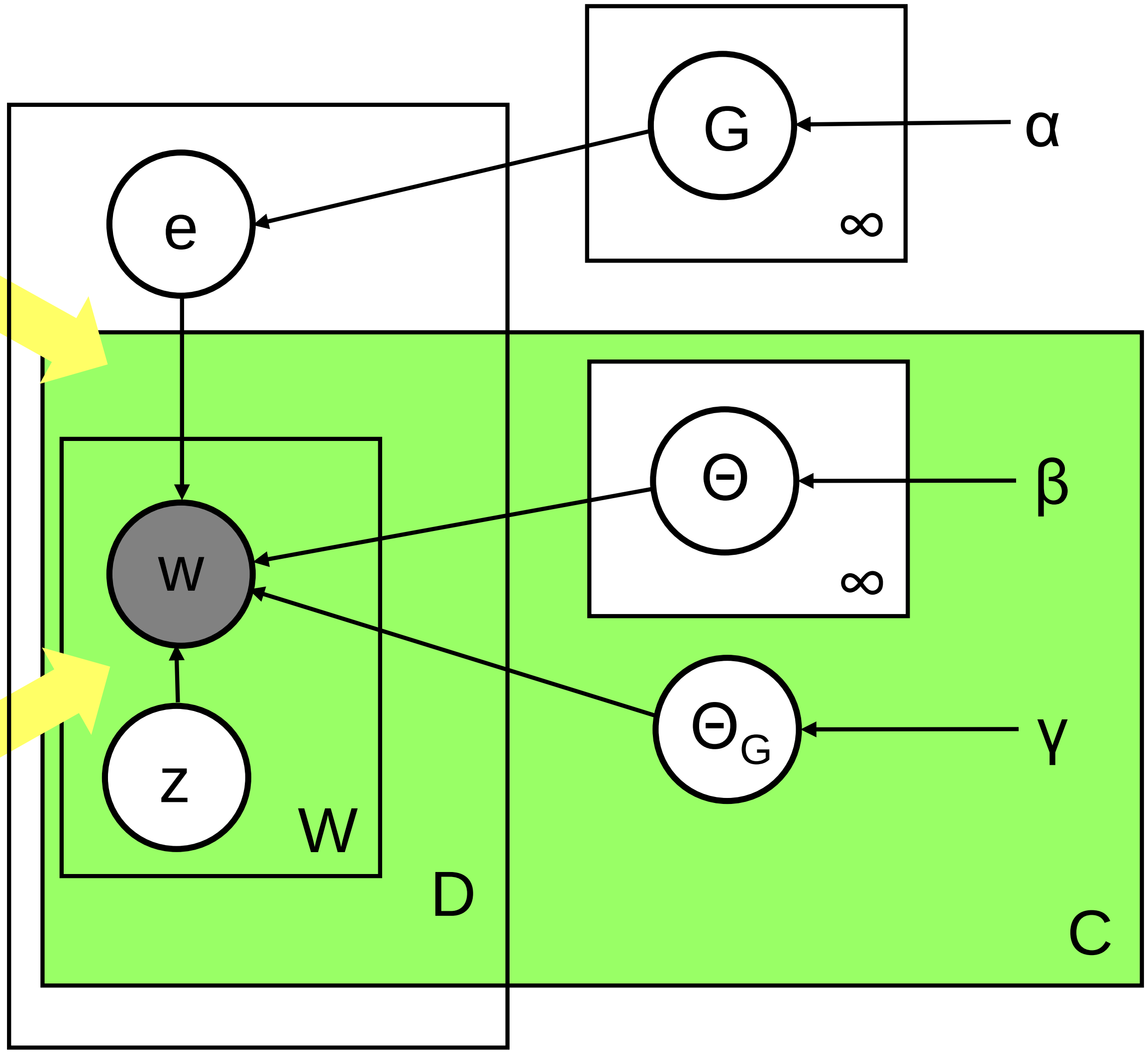
$$p(\tilde{z}_i^k = 0 \mid e^k, \tilde{w}, \tilde{z}) \propto \frac{\beta + C_{e^k}^{w_i}}{\sum_w \beta + C_{e^k}^{w}} \qquad p(\tilde{z}_i^k = 1 \mid e^k, \tilde{w}, \tilde{z}) \propto \frac{\gamma^{w_i} + C_G^{w_i}}{\sum_w \gamma^{w} + C_G^{w}}$$

Then when calculating p(e|w) *only use the w which correspond to z = 0*. (The probability of words from the global topic is absorbed into the normalizing constant).

Incorporating other Evidence

Duplicate the "bag of words" probability model for each other class of observation.

"Bag of" observations for each evidence class.



Gibbs Sampling in this Model

The new probability model is:

$$p(e^k \mid \tilde{e}^{-k}, \tilde{w}, \tilde{z}) \propto p(e^k \mid \tilde{e}^{-k}) \prod_c p(\tilde{w}_c^k \mid \tilde{z}_c^k, e^k, \tilde{w}) \cdot p(\tilde{z}_c^k \mid e^k, \tilde{z}_c^{-k} \tilde{w})$$

Start with an initialization of **e**

Iterate over each document k:

For each type of observation c:

Resample z^k_c

Resample e^k

Spock Results

Model	B-Cubed			Pairwise		
	P	R	F1	P	R	F1
Words	63.3	16.4	26.0	39.2	9.8	15.7
Words-Stop	69.7	17.3	27.7	60.8	9.6	16.5
Words-URL	51.7	18.2	26.9	12.1	11.9	12.0
Words-URL-Stop	60.9	18.6	28.5	53.2	11.1	18.4
Words-URL-WN	48.6	19.8	28.1	6.3	13.9	8.7
Words-URL-WN-Stop	63.2	19.2	**29.5**	55.5	11.6	**19.2**