



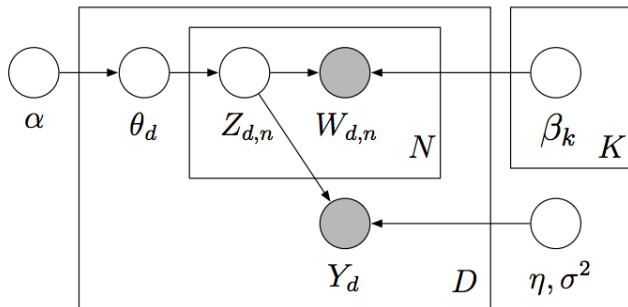
Supervised Topic Models

Advanced Machine Learning for NLP

Jordan Boyd-Graber

OVERVIEW

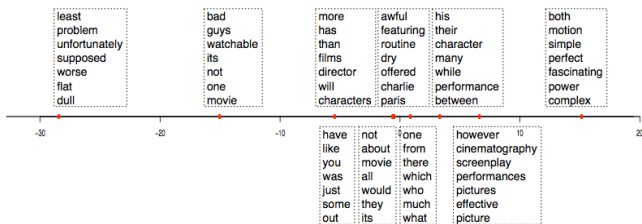
Single Language: Supervised LDA



- Normal LDA generative story
- Document also has label y_d

$$y_d \sim \mathcal{N}(y_d, \eta^\top \mathbb{E}_\theta [\bar{Z}], \sigma^2) \quad (1)$$

How does this change topics?



How does this change topics?

Recall the joint likelihood:

$$p(\mathbf{z} | \alpha, \lambda, \mathbf{w}, \boldsymbol{\eta}) \propto \quad (2)$$

$$\prod_d \prod_d \prod_d \frac{\prod_d \Gamma(n_{d,k} + \alpha_{d,k})}{\Gamma(\sum_d n_{d,k} + \alpha_{d,k})} \prod_k \frac{\prod_k \Gamma(t_{k,v} + \lambda_{k,v})}{\Gamma(\sum_k t_{k,v} + \lambda_{k,v})} \quad (3)$$

$$\prod_d \exp\{-(y_d - \boldsymbol{\eta}^\top \bar{\mathbf{z}})\} \quad (4)$$

Apply gibbs sampling equations:

$$p(z_{d,n} = k | \dots) \propto (n_{d,k}^{-d,n} + \alpha_k) \frac{t_{k,w_{d,n}}^{-d,n} + \lambda_{w_{d,n}}}{t^{-d,n} + V\lambda} \exp\{-(y_d - \boldsymbol{\eta}^\top \bar{\mathbf{z}})^2\} \quad (5)$$

How does this change topics?

Recall the joint likelihood:

$$p(\mathbf{z} | \alpha, \lambda, \mathbf{w}, \eta) \propto \quad (2)$$

$$\prod_d \prod_d \frac{\prod_k \Gamma(n_{d,k} + \alpha_{d,k})}{\Gamma(\sum_k n_{d,k} + \alpha_{d,k})} \prod_k \frac{\prod_v \Gamma(t_{k,v} + \lambda_{k,v})}{\Gamma(\sum_v t_{k,v} + \lambda_{k,v})} \quad (3)$$

$$\prod_d \exp\{-(y_d - \eta^\top \bar{\mathbf{z}})\} \quad (4)$$

Apply gibbs sampling equations:

$$p(z_{d,n} = k | \dots) \propto (n_{d,k}^{-d,n} + \alpha_k) \frac{t_{k,w_{d,n}}^{-d,n} + \lambda_{w_{d,n}}}{t^{-d,n} + V\lambda} \exp\{-(y_d - \eta^\top \bar{\mathbf{z}})^2\} \quad (5)$$

Let's expand last term

How does this change topics?

$$\exp\{-(y_d - \boldsymbol{\eta}^\top \bar{\mathbf{z}})^2\} \quad (6)$$

How does this change topics?

$$\exp\{-(y_d - \boldsymbol{\eta}^\top \bar{\mathbf{z}})^2\} = \exp\{-y_d^2 + 2\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d y_d - (\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d)^2\} \quad (6)$$

(7)

Expand product

How does this change topics?

$$\exp\{-(y_d - \boldsymbol{\eta}^\top \bar{\mathbf{z}})^2\} = \exp\{-y_d^2 + 2\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d y_d - (\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d)^2\} \quad (6)$$

$$\propto \exp\left\{2 \sum_j \eta_j z_{d,j}^{-d,n} y_d + 2 \frac{y_d}{N_d} \eta_k - (\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d)^2\right\} \quad (7)$$

$$(8)$$

Remove constant term, explicitly write dot product

How does this change topics?

$$\exp\{-(y_d - \boldsymbol{\eta}^\top \bar{\mathbf{z}})^2\} = \exp\{-y_d^2 + 2\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d y_d - (\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d)^2\} \quad (6)$$

$$\propto \exp\left\{2 \sum_j \eta_j z_{d,j}^{-d,n} y_d + 2 \frac{y_d}{N_d} \eta_k - (\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d)^2\right\} \quad (7)$$

$$\propto \exp\left\{2 \frac{y_d}{N_d} \eta_k - (\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d)^2\right\} \quad (8)$$

$$(9)$$

Break dot product into k and non- k terms

How does this change topics?

$$\exp\{-(y_d - \boldsymbol{\eta}^\top \bar{\mathbf{z}})^2\} = \exp\{-y_d^2 + 2\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d y_d - (\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d)^2\} \quad (6)$$

$$\propto \exp\left\{2 \sum_j \eta_j z_{d,j}^{-d,n} y_d + 2 \frac{y_d}{N_d} \eta_k - (\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d)^2\right\} \quad (7)$$

$$\propto \exp\left\{2 \frac{y_d}{N_d} \eta_k - (\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d)^2\right\} \quad (8)$$

$$\propto \exp\left\{2 \frac{y_d}{N_d} \eta_k - \left(\frac{\eta_k}{N_d} + \sum_j \eta_j \bar{z}_{d,j}^{-d,n}\right)^2\right\} \quad (9)$$

$$(10)$$

Expand product, drop constant terms

How does this change topics?

$$\exp\{-(y_d - \boldsymbol{\eta}^\top \bar{\mathbf{z}})^2\} \propto \exp\left\{2 \sum_j \eta_j z_{d,j}^{-d,n} y_d + 2 \frac{y_d}{N_d} \eta_k - (\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d)^2\right\} \quad (6)$$

$$\propto \exp\left\{2 \frac{y_d}{N_d} \eta_k - (\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d)^2\right\} \quad (7)$$

$$\propto \exp\left\{2 \frac{y_d}{N_d} \eta_k - \left(\frac{\eta_k}{N_d} + \sum_j \eta_j \bar{z}_{d,j}^{-d,n}\right)^2\right\} \quad (8)$$

$$\propto \exp\left\{2 \frac{y_d}{N_d} \eta_k - \left(\frac{\eta_k}{N_d}\right)^2 - \frac{2\eta_k}{N_d} \boldsymbol{\eta}^\top \bar{\mathbf{z}}^{-d,n}\right\} \quad (9)$$

$$(10)$$

Factor terms

How does this change topics?

$$\exp\{-(y_d - \boldsymbol{\eta}^\top \bar{\mathbf{z}})^2\} \propto \exp\left\{2 \frac{y_d}{N_d} \eta_k - (\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d)^2\right\} \quad (6)$$

$$\propto \exp\left\{2 \frac{y_d}{N_d} \eta_k - \left(\frac{\eta_k}{N_d} + \sum_j \eta_j \bar{z}_{d,j}^{-d,n}\right)^2\right\} \quad (7)$$

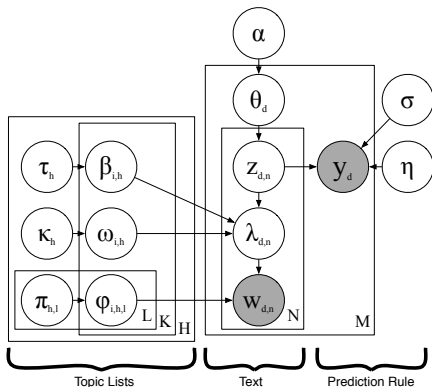
$$\propto \exp\left\{2 \frac{y_d}{N_d} \eta_k - \left(\frac{\eta_k}{N_d}\right)^2 - \frac{2\eta_k}{N_d} \boldsymbol{\eta}^\top \bar{\mathbf{z}}^{-d,n}\right\} \quad (8)$$

$$\propto \exp\left\{2 \frac{\eta_k}{N_d} (y_d - \boldsymbol{\eta}^\top \bar{\mathbf{z}}^{-d,n}) - \left(\frac{\eta_k}{N_d}\right)^2\right\} \quad (9)$$

$$(10)$$

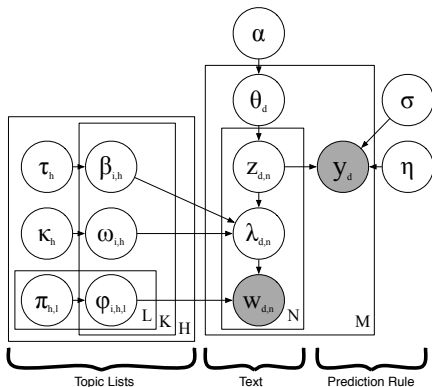
Multiple Languages

- 1 For each topic $k = 1 \dots K$, draw correlated multilingual word distribution $\{\beta_k, \omega_k, \phi_k\}$
- 2 For each document d , $\theta_d \sim \text{Dir}(\alpha)$
 - 1 $z_{d,n} \sim \text{Discrete}(\theta_d)$
 - 2 Draw path $\lambda_{d,n}$ through multilingual tree $z_{d,n}$, emit $w_{d,n}$
- 3 $y_d \sim \text{Norm}(\eta^\top \bar{z}, \sigma^2)$



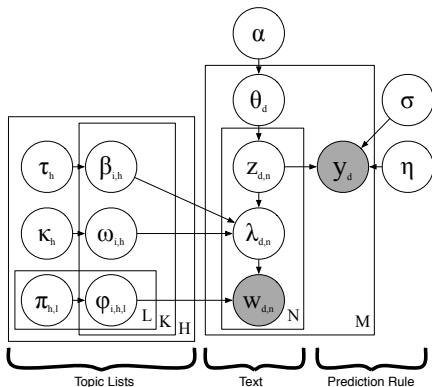
Multiple Languages

- 1 For each topic $k = 1 \dots K$, draw correlated multilingual word distribution $\{\beta_k, \omega_k, \phi_k\}$
- 2 For each document d , $\theta_d \sim \text{Dir}(\alpha)$
 - 1 $z_{d,n} \sim \text{Discrete}(\theta_d)$
 - 2 Draw path $\lambda_{d,n}$ through multilingual tree $z_{d,n}$, emit $w_{d,n}$
- 3 $y_d \sim \text{Norm}(\eta^\top \bar{z}, \sigma^2)$



Multiple Languages

- 1 For each topic $k = 1 \dots K$, draw correlated multilingual word distribution $\{\beta_k, \omega_k, \phi_k\}$
- 2 For each document d , $\theta_d \sim \text{Dir}(\alpha)$
 - 1 $z_{d,n} \sim \text{Discrete}(\theta_d)$
 - 2 Draw path $\lambda_{d,n}$ through multilingual tree $z_{d,n}$, emit $w_{d,n}$
- 3 $y_d \sim \text{Norm}(\eta^\top \bar{z}, \sigma^2)$

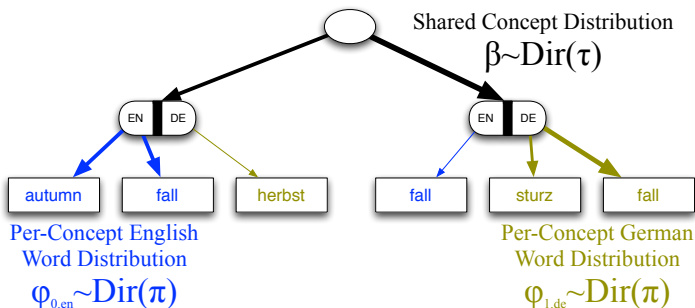


Encoding Correlations

- Statistical NLP typically uses Dirichlet distributions because of conjugacy
- Parameter of Dirichlet encode mean and variance

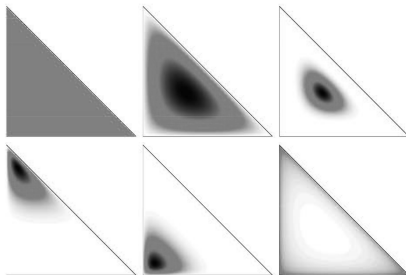
Encoding Correlations

- Statistical NLP typically uses Dirichlet distributions because of conjugacy
- Parameter of Dirichlet encode mean and variance
- But we want correlations!



Encoding Correlations

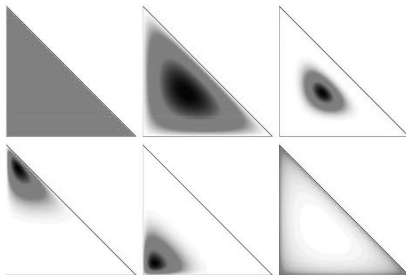
- Statistical NLP typically uses Dirichlet distributions because of conjugacy



- Parameter of Dirichlet encode mean and variance

Encoding Correlations

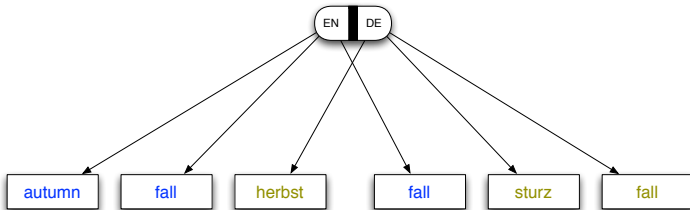
- Statistical NLP typically uses Dirichlet distributions because of conjugacy



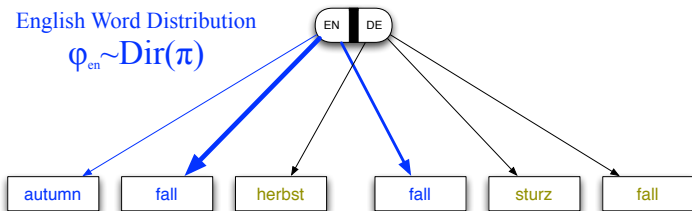
- Parameter of Dirichlet encode mean and variance
- But we want correlations!

gut hảo good

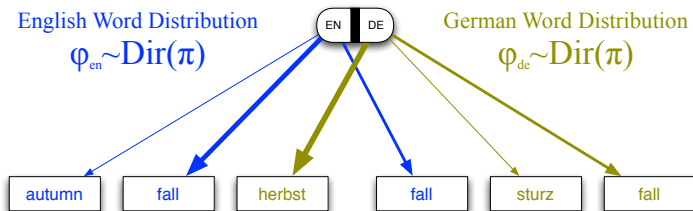
Encoding Correlations



Encoding Correlations



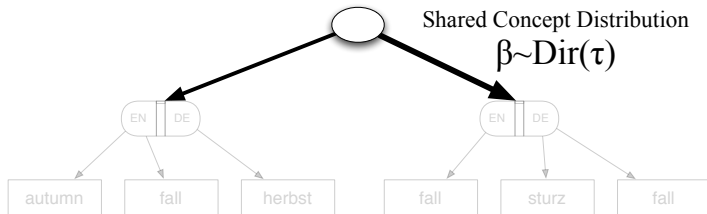
Encoding Correlations



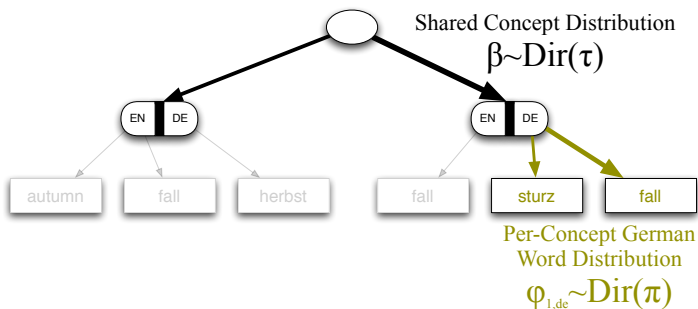
Encoding Correlations



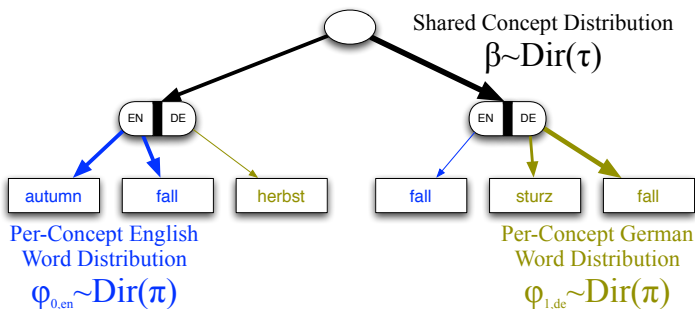
Encoding Correlations



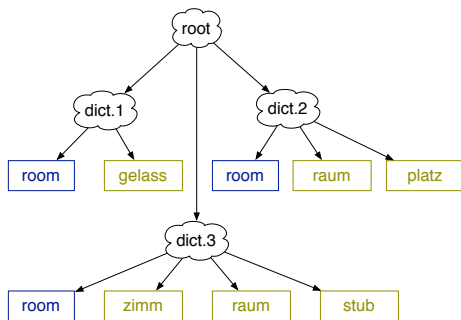
Encoding Correlations



Encoding Correlations

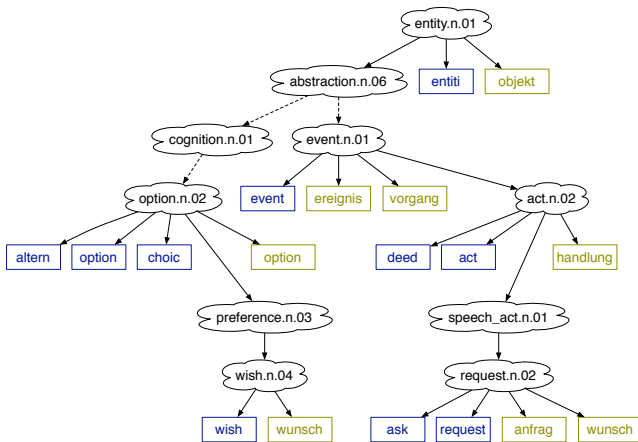


Dictionary



- CEDICT (Chinese/English) ?
- HanDeDict (Chinese/German) ?
- Ding (German/English) ?

Multilingual Ontology



GermaNet ??

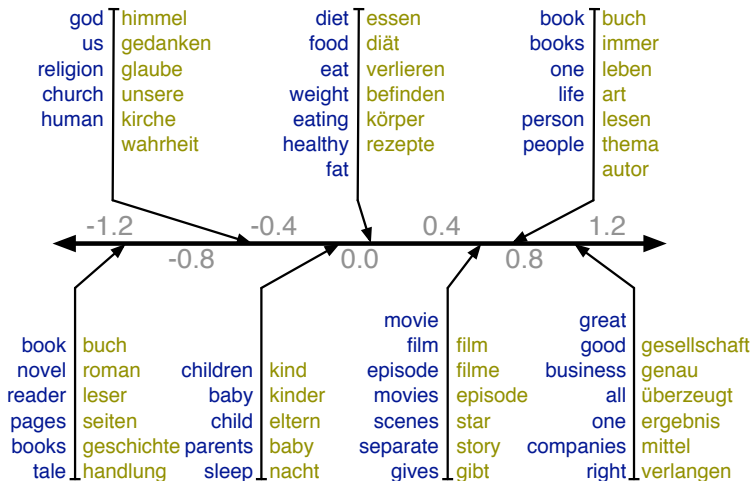
Inference

- Jointly sample z and path λ through multilingual tree

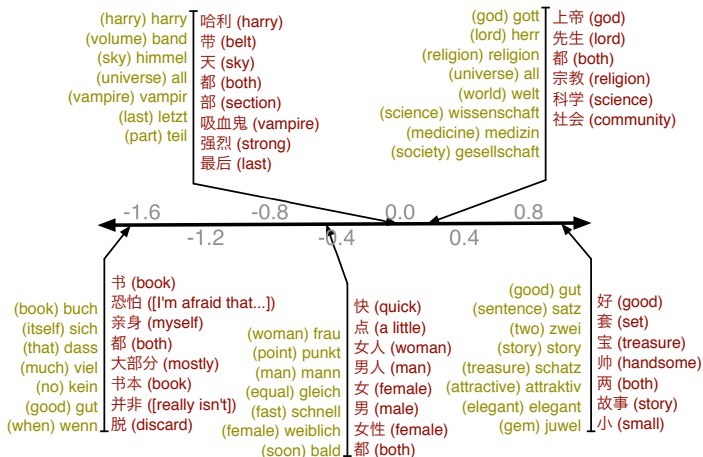
$$\begin{aligned} p(z_n = k, \lambda_n = r | \mathbf{z}_{-n}, \boldsymbol{\lambda}_{-n}, w_n, \eta, \sigma, \Theta) = \\ p(y_d | \mathbf{z}, \eta, \sigma) p(\lambda_n = r | z_n = k, \boldsymbol{\lambda}_{-n}, w_n, \boldsymbol{\tau}, \boldsymbol{\kappa}, \boldsymbol{\pi}) \\ p(z_n = k | \mathbf{z}_{-n}, \alpha). \end{aligned}$$

- Collapse out multinomial distributions in tree
- Slice sample hyperparameters
- After pass of z , update η

Multilingual Supervised LDA



Evaluation: Learned Topics (Chinese - German)



Evaluation: Prediction Accuracy

- Take large corpus (6000) of English movie reviews rated from 0-100 ?
- Combine them with smaller German corpus (300) rated using same system
- Compute mean squared error (lower is better) on held out data

Train	Test	GermaNet	Dictionary	Flat
DE	DE	73.8	24.8	92.2
EN	DE	7.44	2.68	18.3
EN + DE	DE	1.17	1.46	1.39

Moral: More data, even in another language, helps