Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

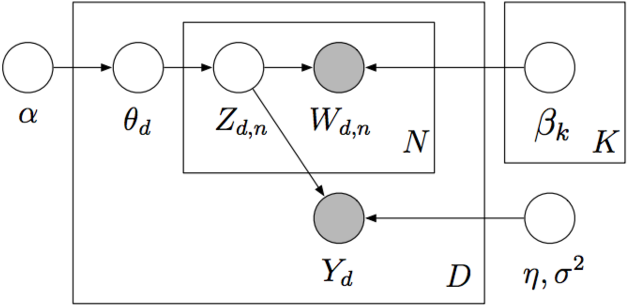
Supervised Topic Models

Advanced Machine Learning for NLP

Jordan Boyd-Graber

OVERVIEW

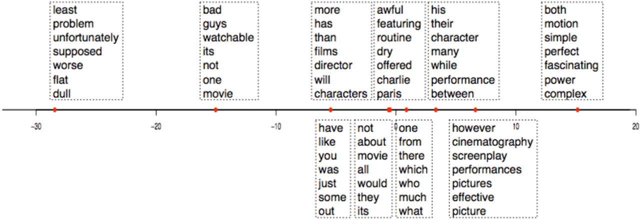
Single Language: Supervised LDA



- Normal LDA generative story
- Document also has label y_d

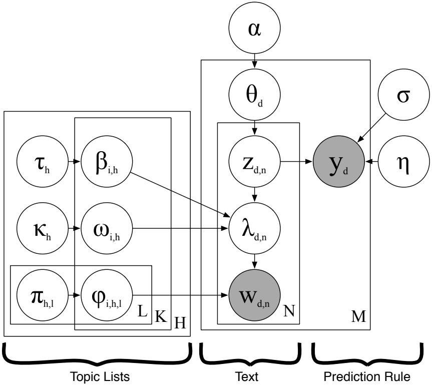
$$y_d \sim \mathcal{N}\left(y_d, \eta^\top \mathbb{E}_\theta\left[\bar{Z}\right], \sigma^2\right) \tag{1}$$

How does this change topics?

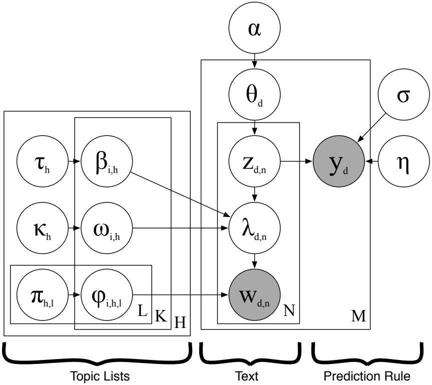


Multiple Languages

1. For each topic $k = 1 \dots K$, draw correlated multilingual word distribution $\{\boldsymbol{\beta}_k, \boldsymbol{\omega}_k, \boldsymbol{\phi}_k\}$
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. $y_d \sim \text{Norm}(\eta^\top \bar{z}, \sigma^2)$

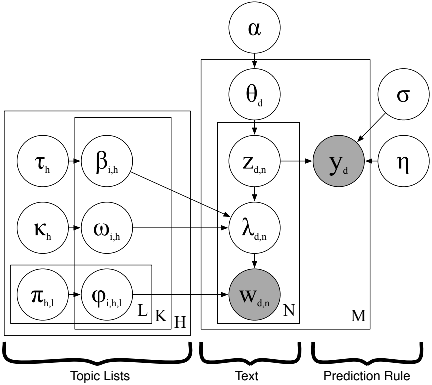


1. For each topic $k = 1 \dots K$, draw correlated multilingual word distribution $\{\boldsymbol{\beta}_k, \boldsymbol{\omega}_k, \boldsymbol{\phi}_k\}$
2. For each document d, $\theta_d \sim \text{Dir}(\alpha)$
   1. $z_{d,n} \sim \text{Discrete}(\theta_d)$
   2. Draw path $\lambda_{d,n}$ through multilingual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

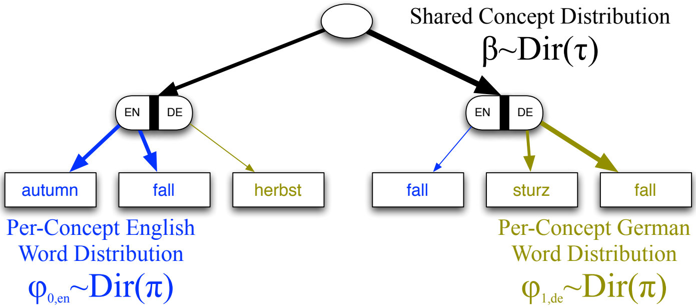


Encoding Correlations

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

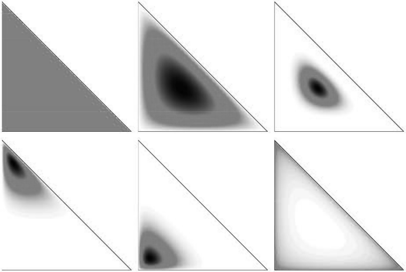
Encoding Correlations

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



Encoding Correlations

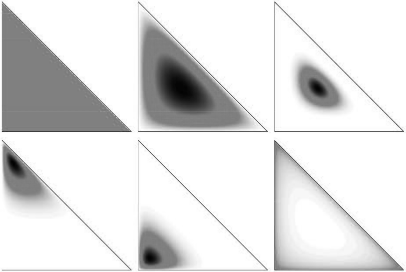
- Statistical NLP typically uses Dirichlet distributions because of conjugacy



- Parameter of Dirichlet encode mean and variance

Encoding Correlations

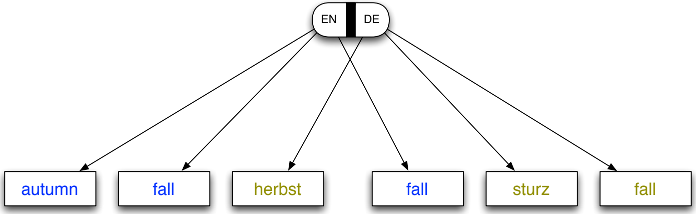
- Statistical NLP typically uses Dirichlet distributions because of conjugacy



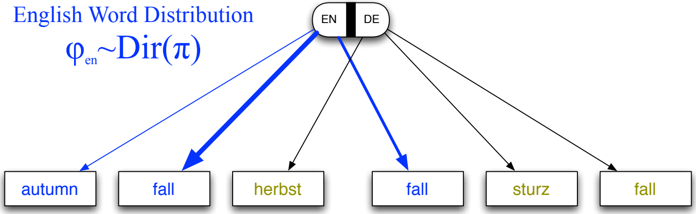
- Parameter of Dirichlet encode mean and variance
- But we want correlations!

gut   hǎo   good

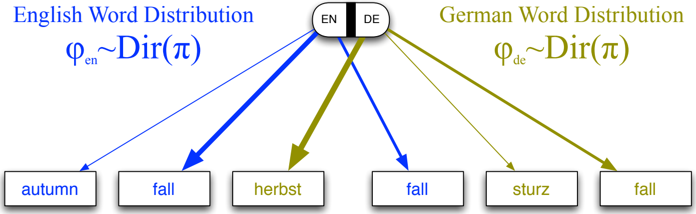
Encoding Correlations



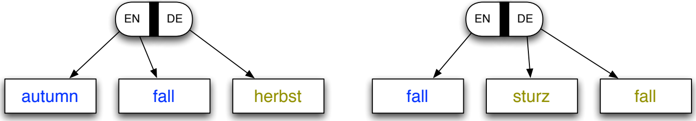
Encoding Correlations



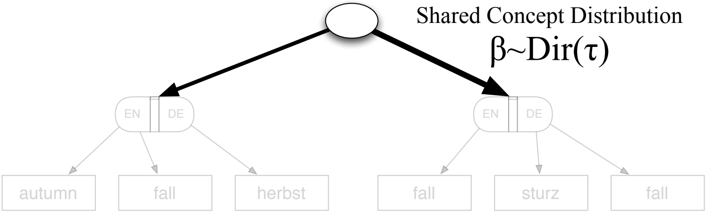
Encoding Correlations



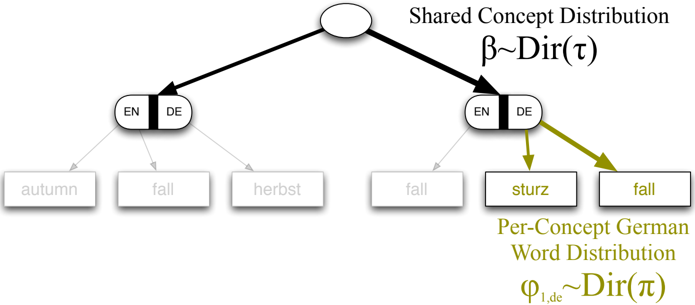
Encoding Correlations



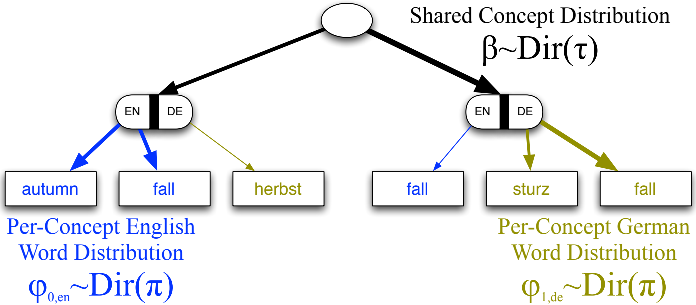
Encoding Correlations



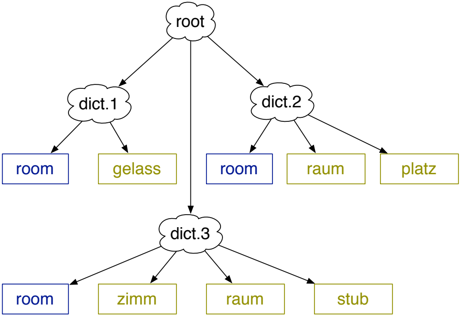
Encoding Correlations



Encoding Correlations

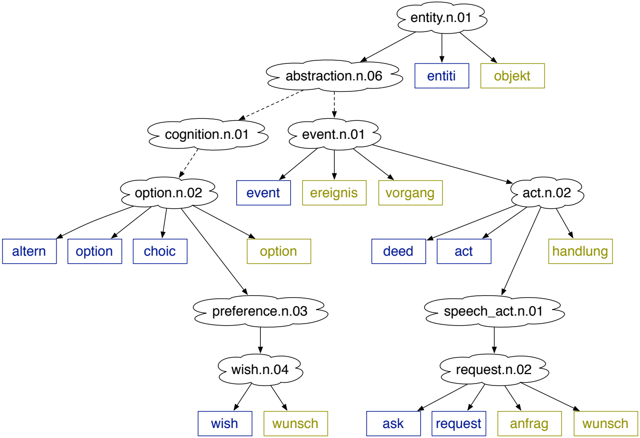


Dictionary



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

Multilingual Ontology



GermaNet **??**

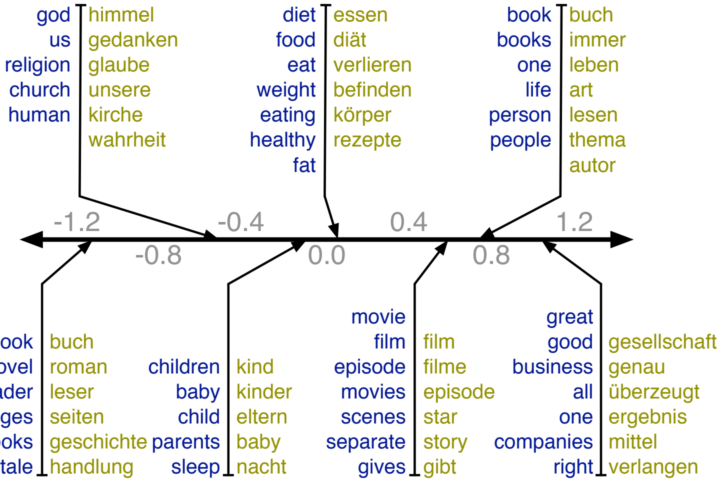
Inference

- Jointly sample z and path λ through multilingual tree

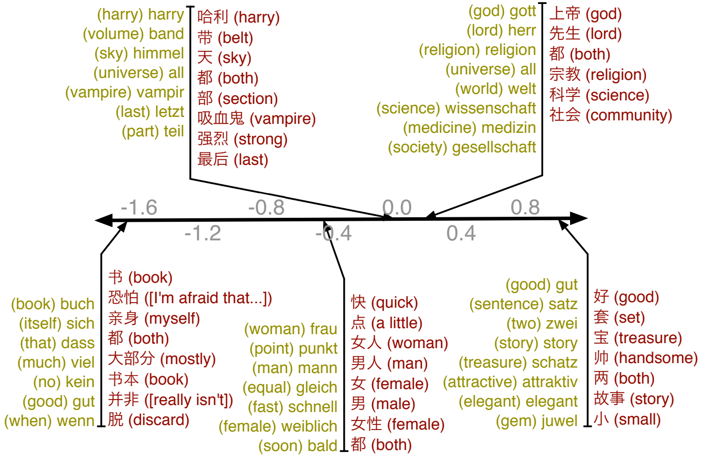
$$
\begin{aligned}
& p(z_n = k, \lambda_n = r | \boldsymbol{z}_{-n}, \boldsymbol{\lambda}_{-n}, w_n, \eta, \sigma, \Theta) = \\
& \quad p(y_d | \boldsymbol{z}, \eta, \sigma) p(\lambda_n = r | z_n = k, \boldsymbol{\lambda}_{-n}, w_n, \boldsymbol{\tau}, \boldsymbol{\kappa}, \boldsymbol{\pi}) \\
& \quad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

- 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