# Neural Models for Documents with Metadata

Dallas Card, Chenhao Tan, Noah A. Smith

July 18, 2018

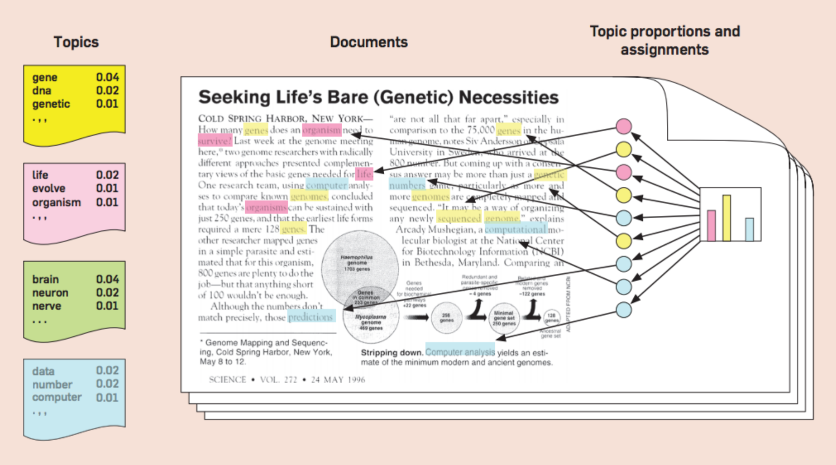
# Outline

Main points of this talk:

1. Introducing *Scholar*[1]: a neural model for documents with metadata
    - Background (LDA, SAGE, SLDA, etc.)
    - Model and related work
    - Experiments and Results
2. Power of neural variational inference for interactive modeling

[1] Sparse Contextual Hidden and Observed Language Autoencoder

# Latent Dirichlet Allocation




Blei, Ng, and Jordan. *Latent Dirichlet Allocation*. JMLR. 2003.

David Blei. *Probabilistic topic models*. Comm. ACM. 2012

# Types of metadata

- Date or time
- Author(s)
- Rating
- Sentiment
- Ideology
- etc.

# Variations and extensions

- Author topic model (Rosen-Zvi et al 2004)
- Supervised LDA (SLDA; McAuliffe and Blei, 2008)
- Dirichlet multinomial regression (Mimno and McCallum, 2008)
- Sparse additive generative models (SAGE; Eisenstein et al, 2011)
- Structural topic model (Roberts et al, 2014)
- ...

# Desired features of model

- Fast, scalable inference.
- Easy modification by end-users.

## Desired features of model

- Fast, scalable inference.
- Easy modification by end-users.
- Incorporation of metadata:
  - Covariates: features which influences text (as in SAGE).
  - Labels: features to be predicted along with text (as in SLDA).

# Desired features of model

- Fast, scalable inference.
- Easy modification by end-users.
- Incorporation of metadata:
  - Covariates: features which influences text (as in SAGE).
  - Labels: features to be predicted along with text (as in SLDA).
- Possibility of sparse topics.

# Desired features of model

- Fast, scalable inference.
- Easy modification by end-users.
- Incorporation of metadata:
  - Covariates: features which influences text (as in SAGE).
  - Labels: features to be predicted along with text (as in SLDA).
- Possibility of sparse topics.
- Incorporate additional prior knowledge.

# Desired features of model

- Fast, scalable inference.
- Easy modification by end-users.
- Incorporation of metadata:
  - Covariates: features which influences text (as in SAGE).
  - Labels: features to be predicted along with text (as in SLDA).
- Possibility of sparse topics.
- Incorporate additional prior knowledge.

$\rightarrow$ Use variational autoencoder (VAE) style of inference (Kingma and Welling, 2014)

# Desired outcome

- Coherent groupings of words (something like topics), with offsets for observed metadata

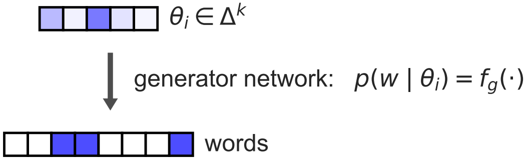
# Desired outcome

- Coherent groupings of words (something like topics), with offsets for observed metadata
- Encoder to map from documents to latent representations

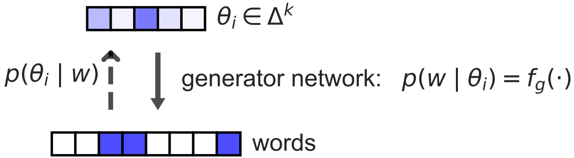
# Desired outcome

- Coherent groupings of words (something like topics), with offsets for observed metadata
- Encoder to map from documents to latent representations
- Classifier to predict labels from from latent representation

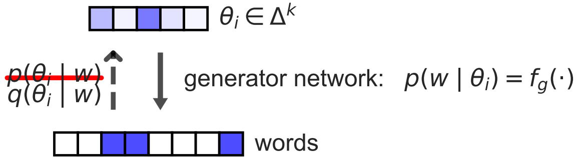
# Model


$\theta_i \in \Delta^k$
generator network: $p(w \mid \theta_i) = f_g(\cdot)$
words

# Model

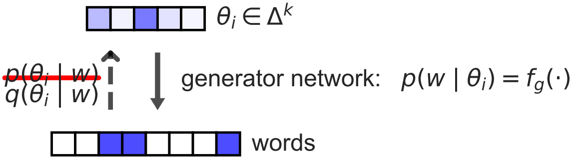

$\theta_i \in \Delta^k$
$p(\theta_i \mid w)$
generator network: $p(w \mid \theta_i) = f_g(\cdot)$
words

# Model



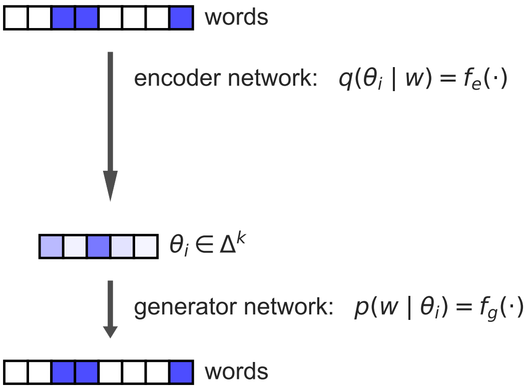
$\theta_i \in \Delta^k$
$\cancel{p(\theta_i \mid w)}$
$q(\theta_i \mid w)$
generator network: $p(w \mid \theta_i) = f_g(\cdot)$
words

# Model



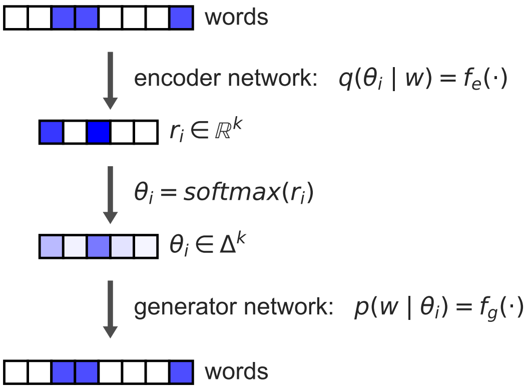

$$\mathrm{ELBO} = \mathbb{E}_q[\log p(\mathit{words} \mid \theta_i)] - D_{\mathrm{KL}}[q(\theta_i \mid \mathit{words}) \| p(\theta_i)]$$

# Model



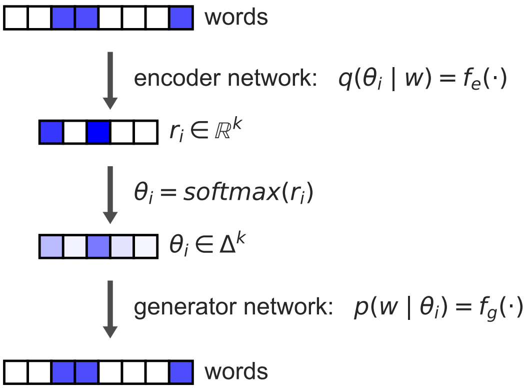

$$\mathrm{ELBO} = \mathbb{E}_q[\log p(\mathit{words} \mid \theta_i)] - D_{\mathrm{KL}}[q(\theta_i \mid \mathit{words}) \| p(\theta_i)]$$

# Model



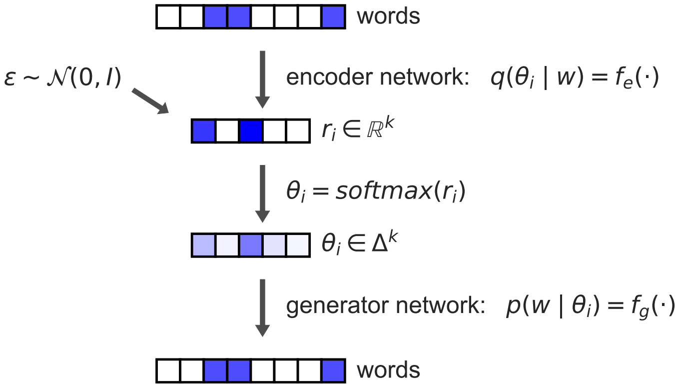

$$\mathrm{ELBO} = \mathbb{E}_q[\log p(words \mid r_i)] - D_{\mathrm{KL}}[q(r_i \mid words) \| p(r_i)]$$

# Model



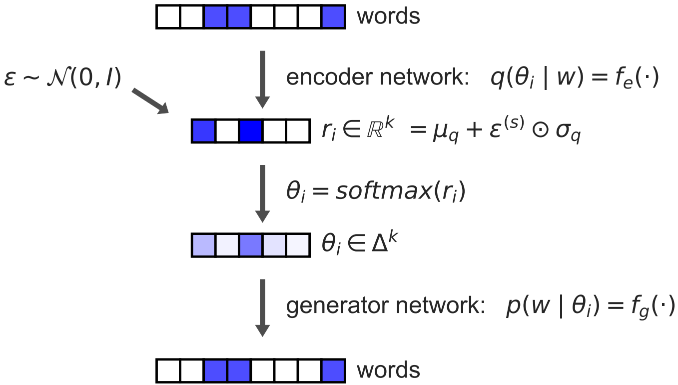

$$\mathrm{ELBO} \approx \frac{1}{S}\sum_{s=1}^{S}[\log p(words \mid r_i^{(s)})] - D_{\mathrm{KL}}[q(r_i \mid words) \| p(r_i)]$$

# Model



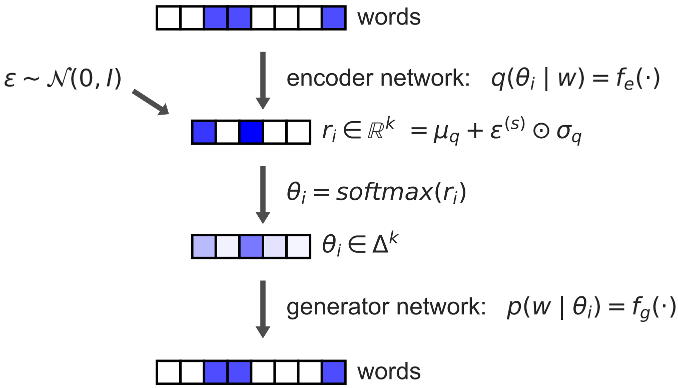

$$\mathrm{ELBO} \approx \frac{1}{S}\sum_{s=1}^{S}[\log p(words \mid r_i^{(s)})] - D_{\mathrm{KL}}[q(r_i \mid words) \| p(r_i)]$$

# Model



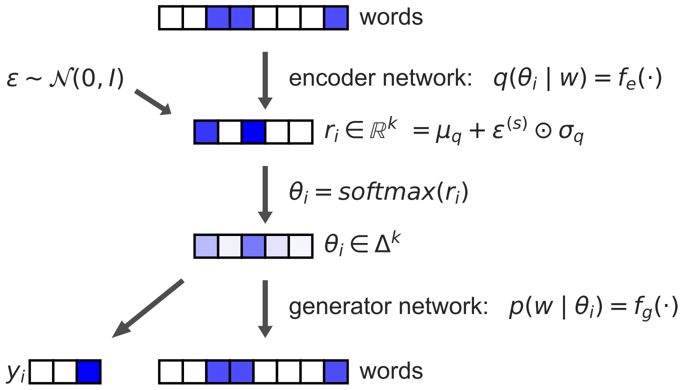

$$\mathrm{ELBO} \approx \frac{1}{S}\sum_{s=1}^{S}[\log p(words \mid r_i^{(s)})] - D_{\mathrm{KL}}[q(r_i \mid words) \| p(r_i)]$$

# Model



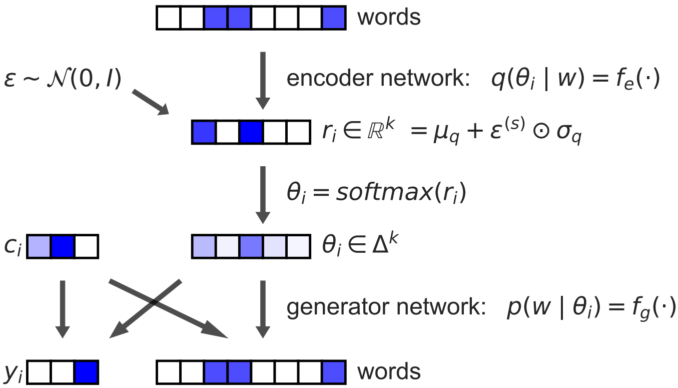

Srivastava and Sutton, 2017, Miao et al, 2016

# Model



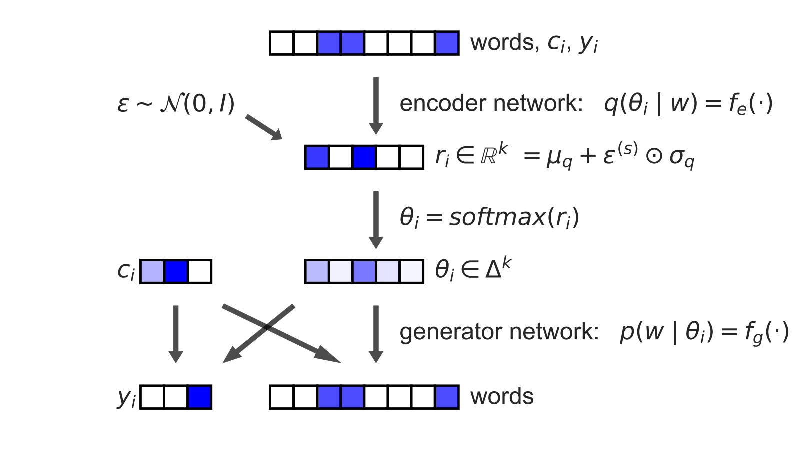
words
$\varepsilon \sim \mathcal{N}(0, I)$
encoder network: $q(\theta_i \mid w) = f_e(\cdot)$
$r_i \in \mathbb{R}^k \ = \mu_q + \varepsilon^{(s)} \odot \sigma_q$
$\theta_i = softmax(r_i)$
$\theta_i \in \Delta^k$
generator network: $p(w \mid \theta_i) = f_g(\cdot)$
$y_i$
words

# Model



words
$\varepsilon \sim \mathcal{N}(0, I)$
encoder network: $q(\theta_i \mid w) = f_e(\cdot)$
$r_i \in \mathbb{R}^k \;= \mu_q + \varepsilon^{(s)} \odot \sigma_q$
$\theta_i = softmax(r_i)$
$c_i$
$\theta_i \in \Delta^k$
generator network: $p(w \mid \theta_i) = f_g(\cdot)$
$y_i$
words

# Model



words, $c_i$, $y_i$
$\varepsilon \sim \mathcal{N}(0, I)$
encoder network: $q(\theta_i \mid w) = f_e(\cdot)$
$r_i \in \mathbb{R}^k \; = \mu_q + \varepsilon^{(s)} \odot \sigma_q$
$\theta_i = softmax(r_i)$
$c_i$
$\theta_i \in \Delta^k$
generator network: $p(w \mid \theta_i) = f_g(\cdot)$
$y_i$
words

# Scholar

Generator network:

- $p(\text{word} \mid \theta_i, c_i) = \text{softmax}(d + \theta_i^T B^{(topic)} + c_i^T B^{(cov)})$

# Scholar

Generator network:

- $p(\text{word} \mid \theta_i, c_i) = \text{softmax}(d + \theta_i^T B^{(topic)} + c_i^T B^{(cov)})$
- Optionally include interactions between topics and covariates

# Scholar

Generator network:

- $p(\text{word} \mid \theta_i, c_i) = \text{softmax}(d + \theta_i^T B^{(topic)} + c_i^T B^{(cov)})$
- Optionally include interactions between topics and covariates
- $p(y_i \mid \theta_i, c_i) = f_y(\theta_i, c_i)$

# Scholar

Generator network:

- $p(\text{word} \mid \theta_i, c_i) = \text{softmax}(d + \theta_i^T B^{(topic)} + c_i^T B^{(cov)})$
- Optionally include interactions between topics and covariates
- $p(y_i \mid \theta_i, c_i) = f_y(\theta_i, c_i)$

Encoder:

- $\mu_i = f_\mu(\text{words}, c_i, y_i)$
- $\log \sigma_i = f_\sigma(\text{words}, c_i, y_i)$
- Optional incorporation of word vectors to embed input

# Optimization

- Stochastic optimization using mini-batches of documents
- Tricks from Srivastava and Sutton, 2017:
  - Adam optimizer with high-learning rate to bypass mode collapse
  - Batch-norm layers to avoid divergence
- Annealing away from batch-norm output to keep results interpretable

# Output of Scholar

- $B^{(topic)}, B^{(cov)}$: Coherent groupings of positive and negative deviations from background ($\sim$ topics)

# Output of Scholar

- $B^{(topic)}, B^{(cov)}$: Coherent groupings of positive and negative deviations from background ($\sim$ topics)
- $f_\mu, f_\sigma$: Encoder network: mapping from words to topics: $\hat{\theta}_i = \mathrm{softmax}(f_e(\mathrm{words}, c_i, y_i, \epsilon))$

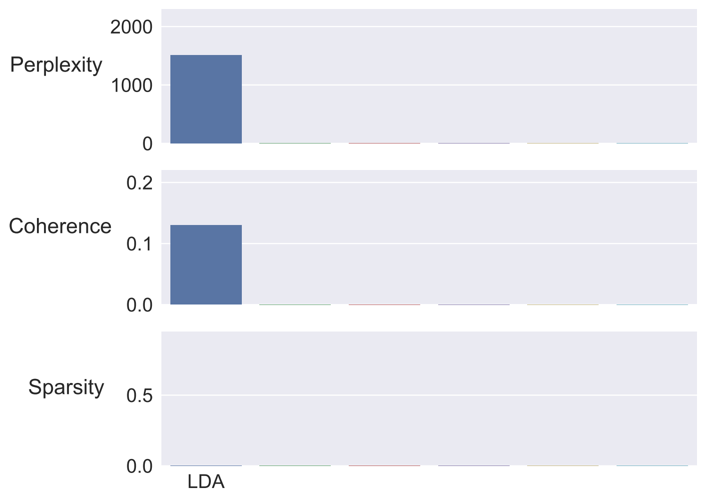
# Output of Scholar

- $B^{(topic)}, B^{(cov)}$: Coherent groupings of positive and negative deviations from background ($\sim$ topics)
- $f_\mu, f_\sigma$: Encoder network: mapping from words to topics: $\hat{\theta}_i = \text{softmax}(f_e(\text{words}, c_i, y_i, \epsilon))$
- $f_y$: Classifier mapping from $\hat{\theta}_i$ to labels: $\hat{y} = f_y(\theta_i, c_i)$

# Evaluation

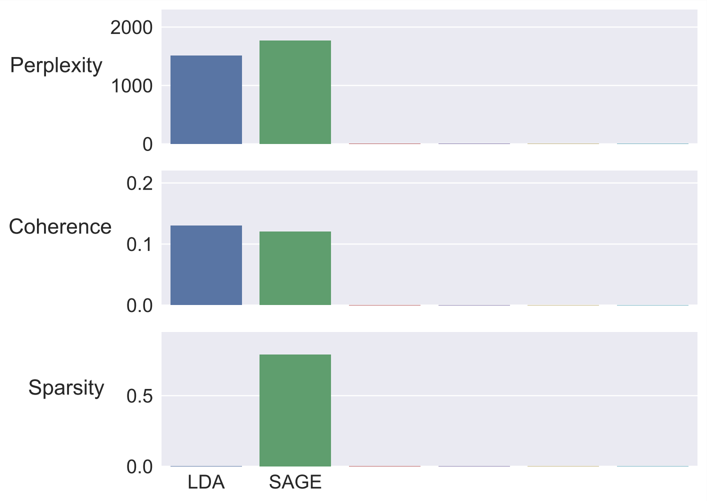
1. Performance as a topic model, without metadata (perplexity, coherence)
2. Performance as a classifier, compared to SLDA
3. Exploratory data analysis

# Quantitative results: basic model




IMDB dataset (Maas, 2011)

# Quantitative results: basic model




IMDB dataset (Maas, 2011)

# Quantitative results: basic model


IMDB dataset (Maas, 2011)

# Quantitative results: basic model


IMDB dataset (Maas, 2011)

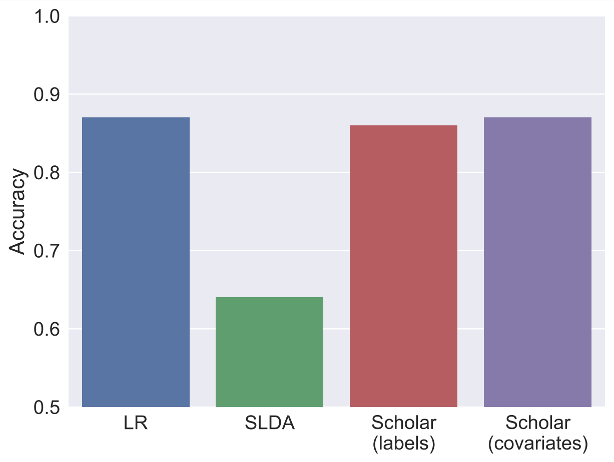
# Quantitative results: basic model


IMDB dataset (Maas, 2011)

# Quantitative results: basic model


IMDB dataset (Maas, 2011)

# Classification results



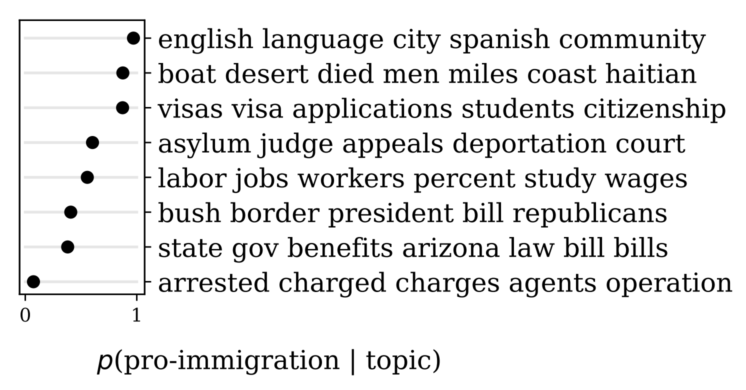

IMDB dataset (Maas, 2011)

# Exploratory Data Analysis

Data: Media Frames Corpus (Card et al, 2015)

- Collection of thousands of news articles annotated in terms of tone and framing
- Relevant metadata: year of publication, newspaper, etc.

# Tone as a label



# Tone as a covariate, with interactions

| Base topics | Anti-immigration | Pro-immigration |
|---|---|---|
| ice customs agency | **criminal** customs | **detainees** detention |
| population born percent | jobs million **illegals** | english **newcomers** |
| judge case court guilty | **guilty** charges man | **asylum** court judge |
| patrol border miles | patrol border | died authorities desert |
| licenses drivers card | foreign sept visas | green citizenship card |
| island story chinese | smuggling federal | island school ellis |
| guest worker workers | bill border house | workers tech skilled |
| benefits bill welfare | republican california | law welfare students |

# Conclusions

- Variational autoencoders (VAEs) provide a powerful framework for latent variable modeling
- We use the VAE framework to create a customizable model for documents with metadata
- We obtain comparable performance with enhanced flexibility and scalability
- Code is available: `www.github.com/dallascard/scholar`