Latent LSTM Allocation

Manzil Zaheer, Amr Ahmed and Alexander J Smola

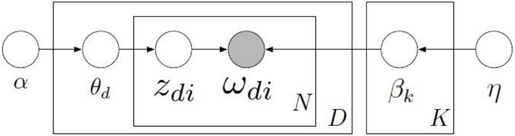
Presented by Akshay Budhkar & Krishnapriya Vishnubhotla

March 3, 2018

Outline

1. Introduction
   - Latent Dirichlet Allocation
   - LSTMs
2. Latent LSTM Allocation
   - Algorithm
   - Inference
   - Different Models
3. Results
4. Conclusion

Latent Dirichlet Allocation



1) Draw each topic $\beta_k \sim \text{Dirichlet}(\eta)$
2) For each document:
    1) Draw topic proportions $\theta_d \sim \text{Dirichlet}(\alpha)$
    2) For each word:
        1) Draw $z_{di} \sim \text{Mult}(\theta_d)$
        2) Draw $\omega_{di} \sim \text{Mult}(\boldsymbol{\beta}_{z_{di}})$

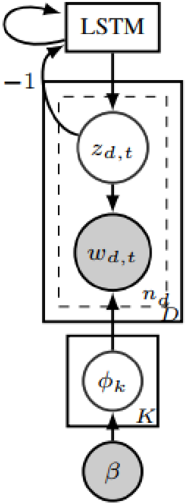
- Probabilistic graphical model
- Not sequential, but easily interpretable.

LSTMs

- Good for modeling sequential data, preserves temporal aspect
- Too many parameters
- Hard to interpret

Latent LSTM Allocation (LLA) - Algorithm

1. for $k = 1$ to K
   (a) Choose topic $\phi_k \sim \text{Dir}(\beta)$
2. for each document d in corpus $\mathcal{D}$
   (a) Initialize LSTM with $\mathbf{s}_0 = 0$
   (b) for each word index t from 1 to N_d
      i. Update $\mathbf{s}_t = \text{LSTM}(z_{d,t-1}, \mathbf{s}_{t-1})$
      ii. Get topic proportions at time t from the LSTM state, $\boldsymbol{\theta} = \text{softmax}_K(\mathbf{W}_p \mathbf{s}_t + \mathbf{b}_p)$
      iii. Choose a topic $z_{d,t} \sim \text{Categorical}(\boldsymbol{\theta})$
      iv. Choose word $w_{d,t} \sim \text{Categorical}(\phi_{z_{d,t}})$

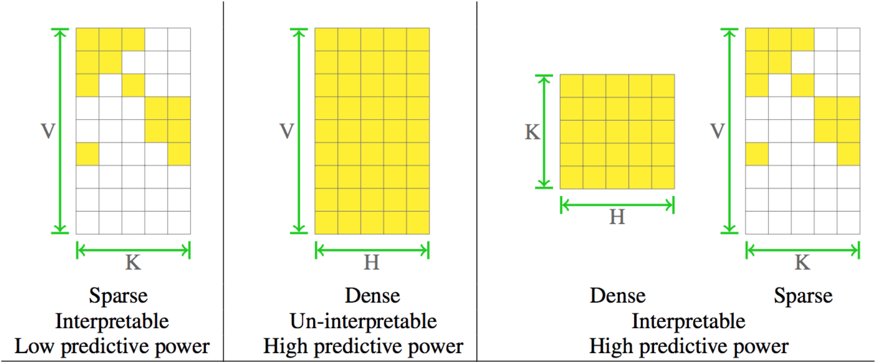


Graphical model for LLA

- Marginal probability of observing a document is

$$\begin{aligned} p(w_d|LSTM, \phi) &= \sum_{z_d} p(w_d, z_d|LSTM, \phi) \\ &= \sum_{z_d} \prod_t p(w_{d,t}|z_{d,t}; \phi) p(z_{d,t}|z_{d,1:t-1}; LSTM) \end{aligned} \tag{1}$$

- Uses a $K \times H$ dense matrix and a $V \times K$ sparse matrix.



(a) LDA | (b) LSTM | (c) LLA

Inference

- Stochastic Expectation Maximization is used to compute the posterior.
- The Evidence Lower Bound (ELBO) can be written as:

$$\begin{aligned}&\sum_d \log p(w_d|LSTM,\phi)\\ &\geq \sum_d \sum_{z_d} q(z)\log \frac{p(z_d;LSTM)\prod_t p(w_{d,t}|z_{d,t};\phi)}{q(z_d)}\end{aligned} \tag{2}$$

- Conditional probability of topic at time step t is:

$$\begin{aligned}&p(z_{d,t}=k|w_{d,t},z_{d,1:t-1}|LSTM,\phi)\\ &\propto p(z_{d,t}=k|z_{d,1:t};LSTM)p(w_{d,t}|z_{d,t}=k;\phi)\end{aligned} \tag{3}$$

- And

$$p(w_{d,t}|z_{d,t}=k;\phi)=\phi_{w,k}=\frac{n_{w,k}+\beta}{n_k+V\beta} \tag{4}$$

Algorithm 1 Stochastic EM for LLA

Input: Document corpus $\mathcal{D}$.

```
Initialize φ and LSTM with a few iterations of LDA
repeat
  SE-Step:
    for all document d ∈ D in parallel do
      for t ← 1 to N_d (possibly with padding) do
        ∀k ∈ {1, ··· , K}, i.e., for every topic index
         obtain by LSTM forward pass:
         π_k = φ_{w_{d,t} k} p(z_{d,t} = k | z_{d,1:t−1}; LSTM).
        Sample z_{d,t} ~ Categorical(π)
      end for
    end for
  M-Step:
    Collect sufficient statistics to obtain:
        φ_wk = (n_wk + β) / (n_k + Vβ),   ∀w, k
    for mini-batch of documents B ⊂ D do
      Compute the gradient by LSTM backward pass
        ∂L/∂LSTM = Σ_{d∈B} Σ_{t=1}^{N_d} ∂ log p(z_{d,t} | z_{d,1:t−1}; LSTM) / ∂LSTM
      Update LSTM parameters by stochastic gradient
      descent methods such as Adam (Kingma & Ba, 2014).
    end for
until Convergence
```

Where:

$$\pi_k = \phi_{w_{d,t}k}\, p(z_{d,t} = k \mid \boldsymbol{z}_{d,1:t-1}; \text{LSTM}).$$

$$\phi_{wk} = \frac{n_{wk} + \beta}{n_k + V\beta}, \quad \forall w, k$$

$$\frac{\partial \mathcal{L}}{\partial \text{LSTM}} = \sum_{d \in \mathcal{B}} \sum_{t=1}^{N_d} \frac{\partial \log p(z_{d,t} \mid z_{d,1:t-1}; \text{LSTM})}{\partial \text{LSTM}}$$

Mathematical Intuition

- LDA

$$\begin{aligned}\log p(w) &= \sum_t \log p(w_t|model) \\ &= \sum_t \log \sum_{z_t} p(w_t|z_t)p(z_t|doc)\end{aligned} \tag{5}$$

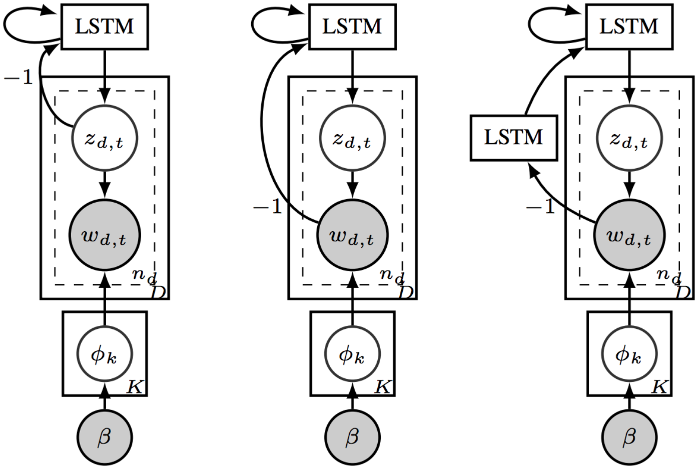
- LSTM

$$\log p(w) = \sum_t \log p(w_t|w_{t-1}, w_{t-2}, \ldots, w_1) \tag{6}$$

- LLA

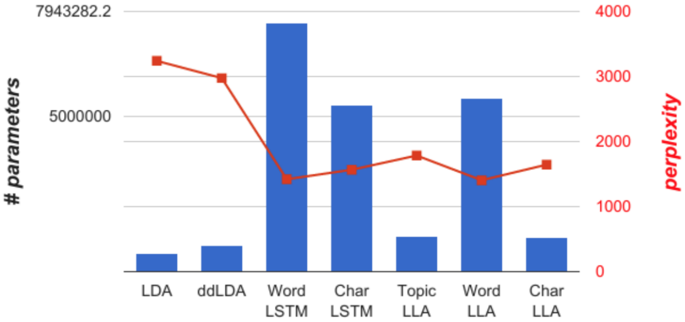
$$\log p(w) = \log \sum_{z_{1:T}} \prod_t p(w_t|z_t)p(z_t|z_{t-1}, z_{t-2}, \ldots, z_1) \tag{7}$$

Different Models

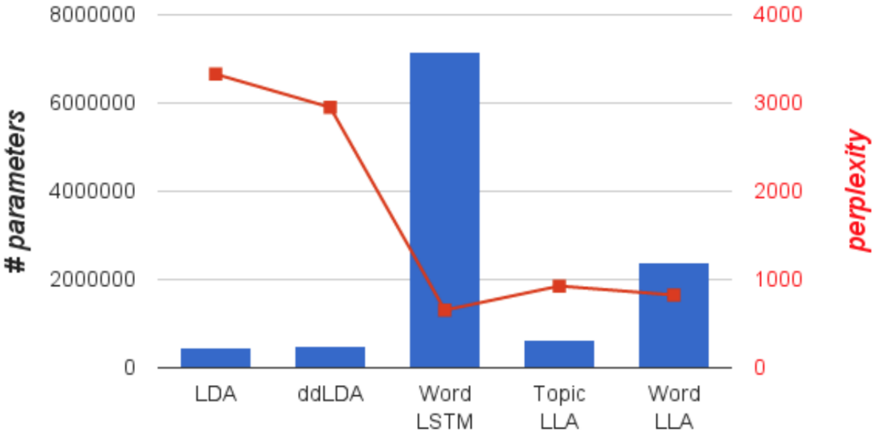


(a) Topic LLA (b) Word LLA (c) Char LLA

Perplexity vs. Number of topics (Wikipedia)

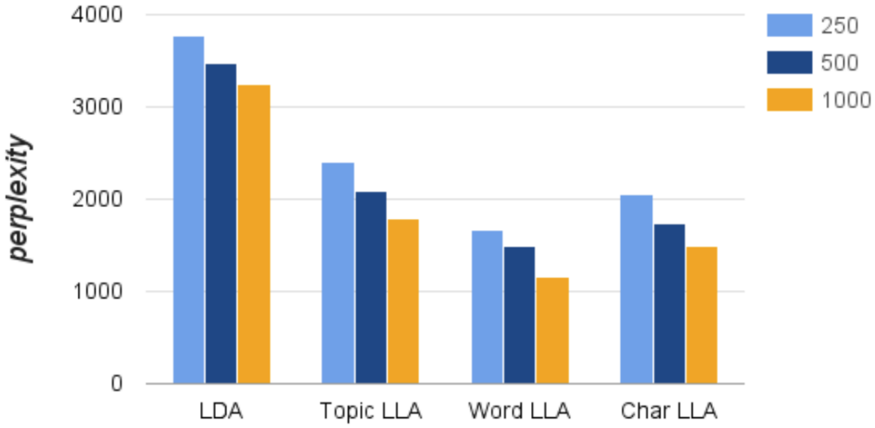


Perplexity vs. Number of topics (User Search)



- Cannot use Char LLA, since URLs lack morphological structure

LDA Ablation Study



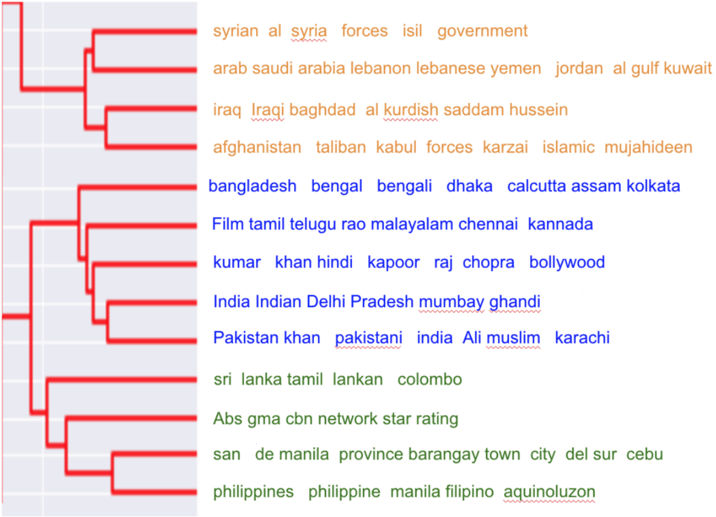
Interpreting Cleaner Topics

LDA	foundation, **iowa**, charity, fund, money, campaign, raised, donated, funds, donations, raise, support, charitable, **million**, donation
LLA	fund, foundation, money, funds, support, charity, funding, donations, campaign, raised, donated, **trust**, raising, **contributions**, **awareness**

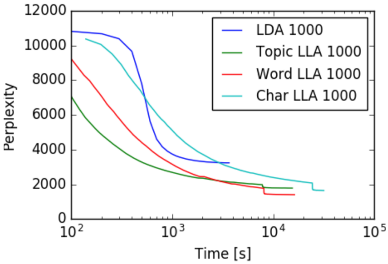
Interpreting Factored Topics

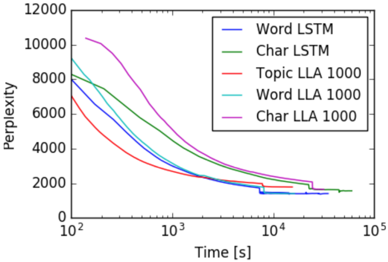
LDA	strike, strikes, striking, miners, strikers, workers workers, day, began, general, called, pinkerton, action, hour, hunger, keefe
LLA	union, unions, strike, workers, labor, federation, trade, afl, bargaining, cio, organization, relations, strikes, national, industrial
	mining, coal, mine, mines, gold, ore, miners, copper, iron, rush, silver, mineral, deposits, minerals, mined

LSTM Topic Embedding (Wikipedia)



Convergence Speed





Effect of Joint vs. Independent Training

Dataset	Independent learner	Joint learner
Wikipedia	2119	1785
User Click	1572	927

Table 4. Advantage in terms of perplexity for joint learning.

Final Thoughts

- Pros
  - Provides a knob for interpretability and accuracy
  - Less number of parameters for a reasonable perplexity
  - Cleaner factored topics
- Cons
  - Did not compare to something like hierarchical LDA
  - Can't use Char LLA for every problem
  - Perplexity is not a good measure of text generation accuracy

Bibliography

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J., and Dolan, B. (2015). deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. *arXiv preprint arXiv:1506.06863*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zaheer, M., Ahmed, A., and Smola, A. J. (2017). Latent lstm allocation: Joint clustering and non-linear dynamic modeling of sequence data. In *International Conference on Machine Learning*, pages 3967–3976.