Improving Topic Models with Latent Feature Word Representations

Dat Quoc Nguyen

Joint work with
Richard Billingsley, Lan Du and Mark Johnson

Department of Computing
Macquarie University
Sydney, Australia

September 2015

Introduction

- *Topic models* take a corpus of documents as input, and
  - learn a set of latent *topics* for the corpus
  - infer *document-to-topic* and *topic-to-word* distributions from co-occurrence of words within documents
- If the corpus is small and/or the documents are short, the topics will be noisy due to the limited information of word co-occurrence
- *Latent word representations* learnt from large external corpora capture various aspects of word meanings
  - We used the pre-trained word2vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014) word representations

High-level idea

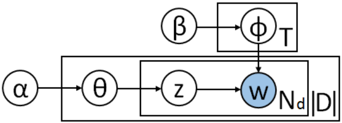
- Use the word representations learnt on a large external corpus to improve the topic-word distributions in a topic model
  - we combine Latent Dirichlet Allocation (Blei et al., 2003) and Dirichlet Multinomial Mixture (Nigam et al., 2000) with the distributed representations
  - the improvement is greatest on small corpora with short documents

LDA and DMM

- Latent Dirichlet Allocation (LDA)

$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha)$ $\qquad$ $z_{d_i} \sim \mathrm{Cat}(\boldsymbol{\theta}_d)$

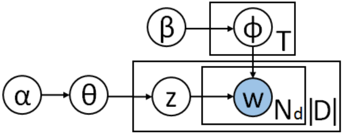
$\boldsymbol{\phi}_z \sim \mathrm{Dir}(\beta)$ $\qquad$ $w_{d_i} \sim \mathrm{Cat}(\boldsymbol{\phi}_{z_{d_i}})$



- Dirichlet Multinomial Mixture (DMM) model: one-topic-per-document

$\boldsymbol{\theta} \sim \mathrm{Dir}(\alpha)$ $\qquad$ $z_d \sim \mathrm{Cat}(\boldsymbol{\theta})$

$\boldsymbol{\phi}_z \sim \mathrm{Dir}(\beta)$ $\qquad$ $w_{d_i} \sim \mathrm{Cat}(\boldsymbol{\phi}_{z_d})$



- Inference is typically performed with a *Gibbs sampler*, integrating out $\boldsymbol{\theta}$ and ϕ (Griffiths et al., 2004; Yin and Wang, 2014)

Latent-feature topic-to-word distributions

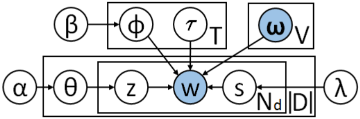
- We assume that each word w is associated with a *word vector* $\boldsymbol{\omega}_w$
- We learn a *topic vector* $\boldsymbol{\tau}_t$ for each topic t
- We use these to define a latent feature topic-to-word distribution $\mathrm{CatE}(w)$ over words:

$$\mathrm{CatE}(w \mid \boldsymbol{\tau}_t \boldsymbol{\omega}^\top) \quad \propto \quad \exp(\boldsymbol{\tau}_t \cdot \boldsymbol{\omega}_w)$$

  - $\boldsymbol{\tau}_t \boldsymbol{\omega}^\top$ is a vector of unnormalized scores, one per word
- In our topic models, we *mix the* CatE *distribution* with a multinomial distribution over words
  - combine information from a large, general corpus (via the CatE distribution) and a smaller but more specific corpus (via the multinomial distribution)
  - use a Boolean *indicator variable* that records whether a word is generated from CatE or the multinomial distribution

The Latent Feature LDA (LF-LDA) model

$$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha) \qquad z_{d_i} \sim \mathrm{Cat}(\boldsymbol{\theta}_d)$$
$$\boldsymbol{\phi}_z \sim \mathrm{Dir}(\beta) \qquad s_{d_i} \sim \mathrm{Ber}(\lambda)$$
$$w_{d_i} \sim (1 - s_{d_i})\mathrm{Cat}(\boldsymbol{\phi}_{z_{d_i}}) + s_{d_i}\mathrm{CatE}(\boldsymbol{\tau}_{z_{d_i}}\,\boldsymbol{\omega}^\top)$$



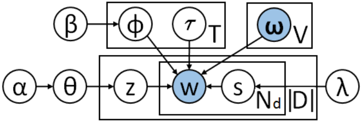
- Replace the topic-to-word Dirichlet multinomial component in LDA with a two-component mixture of a topic-to-word Dirichlet multinomial component and a latent feature topic-to-word component
- s_{d_i} is the Boolean indicator variable indicating whether word w_{d_i} is generated from the latent feature component
- λ is a user-specified hyper-parameter determining how often words are generated from the latent feature component
  - if we estimated λ from data, we expect it would never generate through the latent feature component

The Latent Feature DMM (LF-DMM) model

$\boldsymbol{\theta} \sim \text{Dir}(\alpha)$ $\qquad z_d \sim \text{Cat}(\boldsymbol{\theta})$

$\boldsymbol{\phi}_z \sim \text{Dir}(\beta)$ $\qquad s_{d_i} \sim \text{Ber}(\lambda)$

$w_{d_i} \sim (1 - s_{d_i})\text{Cat}(\boldsymbol{\phi}_{z_d}) + s_{d_i}\text{CatE}(\boldsymbol{\tau}_{z_d}\,\boldsymbol{\omega}^\top)$



- Replace the topic-to-word Dirichlet multinomial component in DMM with a two-component mixture of a topic-to-word Dirichlet multinomial component and a latent feature topic-to-word component
- s_{d_i} is the Boolean indicator variable indicating whether word w_{d_i} is generated from the latent feature component
- λ is a user-specified hyper-parameter determining how often words are generated from the latent feature component

Inference for the LF-LDA model

- We integrate out $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ as in the Griffiths et al. (2004) sampler, and *interleave MAP estimation for $\boldsymbol{\tau}$ with Gibbs sweeps for the other variables*
- Algorithm outline:
  initialize the word-topic variables z_{d_i} using the LDA sampler
  repeat:

    for each topic t:
      use LBFGS to optimize the L2-regularized log-loss
      $\boldsymbol{\tau}_t = \arg\max_{\boldsymbol{\tau}_t} \mathrm{P}(\boldsymbol{\tau}_t \mid \boldsymbol{z}, \boldsymbol{s})$
    for each document d and each word location i:
      sample z_{d_i} from $\mathrm{P}(z_{d_i} \mid \boldsymbol{z}_{\neg d_i}, \boldsymbol{s}_{\neg d_i}, \boldsymbol{\tau})$
      sample s_{d_i} from $\mathrm{P}(s_{d_i} \mid \boldsymbol{z}, \boldsymbol{s}_{\neg d_i}, \boldsymbol{\tau})$

Inference for the LF-DMM model

- We integrate out $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ as in the Yin and Wang (2014) sampler, and *interleave MAP estimation for $\boldsymbol{\tau}$ with Gibbs sweeps*
- Algorithm outline:
  initialize the word-topic variables z_{d_i} using the DMM sampler
  repeat:

  for each topic t:
  use LBFGS to optimize the L2-regularized log-loss
  $\boldsymbol{\tau}_t = \arg\max_{\boldsymbol{\tau}_t} \mathrm{P}(\boldsymbol{\tau}_t \mid \boldsymbol{z}, \boldsymbol{s})$
  for each document d:
  sample z_d and $\boldsymbol{s}_d$ from $\mathrm{P}(z_d, \boldsymbol{s}_d \mid \boldsymbol{z}_{\neg d}, \boldsymbol{s}_{\neg d}, \boldsymbol{\tau})$
- Note: $\mathrm{P}(z_d, \boldsymbol{s}_d \mid \boldsymbol{z}_{\neg d}, \boldsymbol{s}_{\neg d}, \boldsymbol{\tau})$ is *computationally expensive* to compute exactly, as it requires *summing over all possible values for $\boldsymbol{s}_d$*
- We approximate these probabilities by *assuming that the topic-word counts are "frozen"*, i.e., they don't increase within a document
  $\Rightarrow$We are able to integrate out $\boldsymbol{s}_d$

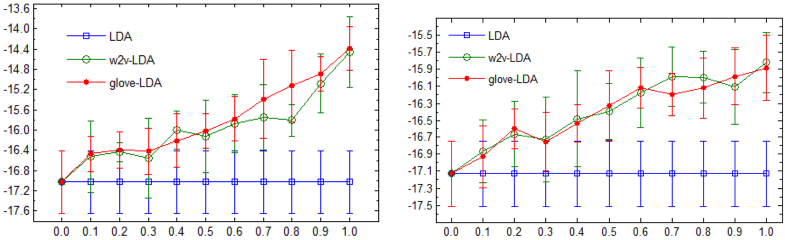
Goals of evaluation

- A topic model learns document-topic and topic-word distributions:
  - *topic coherence* evaluates the topic-word distributions
  - *document clustering* and *document classification* evaluate the document-topic distribution
- Do the word2vec and Glove word vectors behave differently in topic modelling? (w2v-LDA, glove-LDA, w2v-LDA, glove-DMM)
- We expect that the latent feature component will have *the greatest impact on small corpora*, so our evaluation focuses on them:

Dataset		# labels	# docs	words/doc	# types
N20	20 newsgroups	20	18,820	103.3	19,572
N20short	$\leq$ 20 words	20	1,794	13.6	6,377
N20small	400 docs	20	400	88.0	8,157
TMN	TagMyNews	7	32,597	18.3	13,428
TMNtitle	TagMyNews titles	7	32,503	4.9	6,347
Twitter		4	2,520	5.0	1,390

Topic coherence evaluation

- Lau et al. (2014) showed that *human scores on a word intrusion task* are highly correlated with the *normalized pointwise mutual information* (NPMI)
- We found latent feature vectors produced a *significant improvement of NPMI scores on all models and corpora*
  - greatest improvement when $\lambda = 1$ (unsurprisingly)



20 topics 40 topics

NPMI scores on the N20short dataset, varying the mixture weight λ from 0.0 to 1.0.

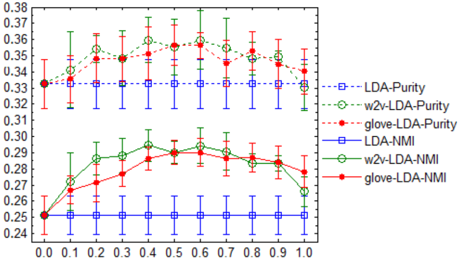
w2v-DMM on TagMyNews titles corpus

Topic 1		Topic 3		Topic 4	
DMM	w2v-DMM	DMM	w2v-DMM	DMM	w2v-DMM
japan	japan	u.s.	prices	egypt	libya
nuclear	nuclear	oil	sales	china	egypt
u.s.	u.s.	japan	oil	u.s	**iran**
crisis	plant	prices	u.s.	mubarak	**mideast**
plant	**quake**	stocks	profit	bin	**opposition**
china	radiation	sales	stocks	libya	**protests**
libya	**earthquake**	profit	japan	laden	**leader**
radiation	**tsunami**	fed	rise	france	**syria**
u.n.	**nuke**	rise	**gas**	bahrain	u.n.
vote	crisis	growth	growth	air	**tunisia**
korea	**disaster**	wall	shares	report	**chief**
europe	**power**	street	**price**	rights	**protesters**
government	**oil**	china	**profits**	court	mubarak
election	**japanese**	fall	**rises**	u.n.	**crackdown**
deal	**plants**	shares	**earnings**	war	bahrain

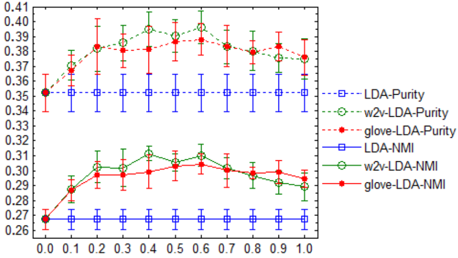
- Table shows the 15 most probable topical words found by 20-topic w2v-DMM on the TMNtitle corpus
- Words found by DMM but not by w2v-DMM are underlined
- Words found by w2v-DMM but not DMM are in bold

Document clustering evaluation (1)

- Cluster documents by assigning them to the *highest probability topic*
- Evaluate clusterings by *purity* and *normalized mutual information* (NMI) (Manning et al., 2008)



20 topics



40 topics

Purity and NMI results on the N20short dataset, varying the mixture weight λ from 0.0 to 1.0.

- In general, best results with $\lambda = 0.6$

$\Rightarrow$ Set $\lambda = 0.6$ in all further experiments

Document clustering evaluation (2)

Data	Method	Purity		NMI	
		T=4	T=20	T=4	T=20
Twitter	LDA	0.559 ± 0.020	0.614 ± 0.016	0.196 ± 0.018	0.174 ± 0.008
	w2v-LDA	**0.598** ± 0.023	**0.635** ± 0.016	**0.249** ± 0.021	**0.191** ± 0.011
	glove-LDA	0.597 ± 0.016	**0.635** ± 0.014	0.242 ± 0.013	**0.191** ± 0.007
	Improve.	**0.039**	**0.021**	**0.053**	**0.017**
Twitter	DMM	0.523 ± 0.011	0.619 ± 0.015	0.222 ± 0.013	0.213 ± 0.011
	w2v-DMM	**0.589** ± 0.017	0.655 ± 0.015	0.243 ± 0.014	0.215 ± 0.009
	glove-DMM	0.583 ± 0.023	**0.661** ± 0.019	**0.250** ± 0.020	**0.223** ± 0.014
	Improve.	**0.066**	**0.042**	**0.028**	**0.01**

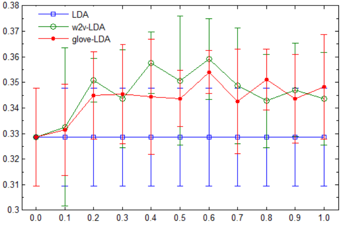
- On the short, our models obtain better clustering results than the baseline models:
  - on N20small, we get 6.0% improvement on NMI at $T = 6$
  - on TMN and TMNtitle, we obtain 6.1% and 2.5% higher Purity at $T = 80$

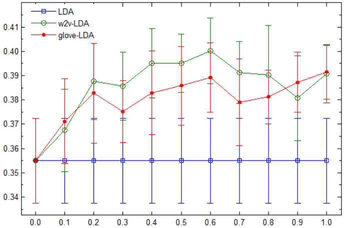
Document clustering evaluation (3)

- For small $T \leq 7$, on the large datasets of N20, TMN and TMNtitle, our models and baseline models obtain similar clustering results
- With larger T, our models perform better than baselines on the short TMN and TMNtitle datasets. On the N20 dataset, the baseline LDA model obtains slightly higher clustering results than ours
- No reliable difference between word2vec and Glove vectors

Document classification (1)

- Use SVM to predict the ground truth label from the topic-proportion vector of each document





20 topics 40 topics

F_1 scores on the N20short dataset, varying the mixture weight λ from 0.0 to 1.0.

Data	Method	$\lambda = 0.6$			
		T=6	T=20	T=40	T=80
N20small	LDA	0.204 ± 0.020	0.392 ± 0.029	0.459 ± 0.030	0.477 ± 0.025
	w2v-LDA	**0.213** ± 0.018	**0.442** ± 0.025	**0.502** ± 0.031	**0.509** ± 0.022
	glove-LDA	0.181 ± 0.011	0.420 ± 0.025	0.474 ± 0.029	0.498 ± 0.012
	Improve.	**0.009**	**0.05**	**0.043**	**0.032**

Document classification (2)

Data	Method	$\lambda = 0.6$			
		T=7	T=20	T=40	T=80
TMN	DMM	0.607 ± 0.040	0.694 ± 0.026	0.712 ± 0.014	0.721 ± 0.008
	w2v-DMM	0.607 ± 0.019	0.736 ± 0.025	**0.760** $\pm$ 0.011	0.771 ± 0.005
	glove-DMM	**0.621** $\pm$ 0.042	**0.750** $\pm$ 0.011	0.759 ± 0.006	**0.775** $\pm$ 0.006
	Improve.	**0.014**	**0.056**	**0.048**	**0.054**
TMNtitle	DMM	0.500 ± 0.021	0.600 ± 0.015	0.630 ± 0.016	0.652 ± 0.005
	w2v-DMM	0.528 ± 0.028	0.663 ± 0.008	0.682 ± 0.008	**0.681** $\pm$ 0.006
	glove-DMM	**0.565** $\pm$ 0.022	**0.680** $\pm$ 0.011	**0.684** $\pm$ 0.009	**0.681** $\pm$ 0.004
	Improve.	**0.065**	**0.08**	**0.054**	**0.029**

Data	Method	$\lambda = 0.6$			
		T=4	T=20	T=40	T=80
Twitter	LDA	0.526 ± 0.021	0.636 ± 0.011	0.650 ± 0.014	0.653 ± 0.008
	w2v-LDA	**0.578** $\pm$ 0.047	0.651 ± 0.015	0.661 ± 0.011	**0.664** $\pm$ 0.010
	glove-LDA	0.569 ± 0.037	**0.656** $\pm$ 0.011	**0.662** $\pm$ 0.008	0.662 ± 0.006
	Improve.	**0.052**	**0.02**	**0.012**	**0.011**
Twitter	DMM	0.469 ± 0.014	0.600 ± 0.021	0.645 ± 0.009	0.665 ± 0.014
	w2v-DMM	**0.539** $\pm$ 0.016	0.649 ± 0.016	0.656 ± 0.007	0.676 ± 0.012
	glove-DMM	0.536 ± 0.027	**0.654** $\pm$ 0.019	**0.657** $\pm$ 0.008	**0.680** $\pm$ 0.009
	Improve.	**0.07**	**0.054**	**0.012**	**0.015**

Conclusions and future directions

- Latent feature vectors induced from large external corpora can be used to improve topic modelling
  - latent features significantly improve topic coherence across a range of corpora with both the LDA and DMM models
  - document clustering and document classification also significantly improve, even though these depend directly only on the document-topic distribution
- The improvements were greatest for small document collections and/or for short documents
- We did not detect any reliable difference between word2vec and Glove vectors
- Retrain the word vectors to fit the topic-modeling corpus
- More sophisticated latent-feature models of topic-word distributions
- More efficient training procedures

Thank you for your attention!

- Software:
  - http://jldadmm.sourceforge.net
  - https://github.com/datquocnguyen/LFTM

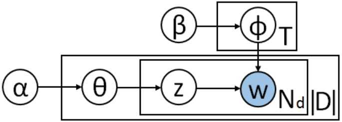
Related work

- Phan et al. (2011) assumed that the small corpus is a sample of topics from a larger corpus like Wikipedia, and use the topics discovered in the larger corpus to help shape the topic representations in the small corpus
  - if the larger corpus has many irrelevant topics, this will "use up" the topic space of the model
- Petterson et al. (2010) proposed an extension of LDA that uses external information about word similarity, such as thesauri and dictionaries, to smooth the topic-to-word distribution
- Sahami and Heilman (2006) employed web search results to improve the information in short texts
- *Neural network topic models* of a single corpus have also been proposed (Salakhutdinov and Hinton, 2009; Srivastava et al., 2013; Cao et al., 2015).

Latent Dirichlet Allocation (LDA)

$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha)$ $\quad$ $z_{d_i} \sim \mathrm{Cat}(\boldsymbol{\theta}_d)$

$\boldsymbol{\phi}_z \sim \mathrm{Dir}(\beta)$ $\quad$ $w_{d_i} \sim \mathrm{Cat}(\boldsymbol{\phi}_{z_{d_i}})$



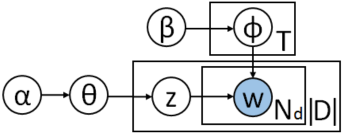
- Latent Dirichlet Allocation (LDA) is an *admixture model*, i.e., each document d is associated with a *distribution over topics* θ_d
- Inference is typically performed with a *Gibbs sampler* over the z_{d_i}, integrating out $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ (Griffiths et al., 2004)

$$\mathrm{P}(z_{d_i}{=}t \mid \boldsymbol{Z}_{\neg d_i}) \quad \propto \quad (N^t_{d_{\neg i}} + \alpha)\frac{N^{t,w_{d_i}}_{\neg d_i} + \beta}{N^t_{\neg d_i} + V\beta}$$

The Dirichlet Multinomial Mixture (DMM) model

$\boldsymbol{\theta} \sim \text{Dir}(\alpha)$ $\quad z_d \sim \text{Cat}(\boldsymbol{\theta})$

$\phi_z \sim \text{Dir}(\beta)$ $\quad w_{d_i} \sim \text{Cat}(\phi_{z_d})$



- The Dirichlet Multinomial Mixture (DMM) model is a *mixture model*, i.e., each document d is associated with a single topic z_d (Nigam et al., 2000)
- Inference can also be performed using a collapsed Gibbs sampler in which θ and ϕ_z are integrated out (Yin and Wang, 2014)

$$\begin{aligned} \text{P}(z_d = t \mid \mathbf{Z}_{\neg d}) \quad &\propto \quad (M^t_{\neg d} + \alpha) \frac{\Gamma(N^t_{\neg d} + V\beta)}{\Gamma(N^t_{\neg d} + N_d + V\beta)} \\ &\quad \prod_{w \in W} \frac{\Gamma(N^{t,w}_{\neg d} + N^w_d + \beta)}{\Gamma(N^{t,w}_{\neg d} + \beta)} \end{aligned}$$

Estimating the topic vectors $\boldsymbol{\tau}_t$

- Both the LF-LDA and LF-DMM associate each topic t with a *topic vector* $\boldsymbol{\tau}_t$, which must be learnt from the training corpus
- After each Gibbs sweep:
  - the topic variables $\boldsymbol{z}$ identify which topic each word is generated from
  - the indicator variables $\boldsymbol{s}$ identify which words are generated from the latent feature distributions CatE
- We use LBFGS to optimize the L2-regularized log-loss (MAP estimation)

$$L_t \;=\; -\sum_{w \in W} K^{t,w} \left(\boldsymbol{\tau}_t \cdot \boldsymbol{\omega}_w - \log\left(\sum_{w' \in W} \exp(\boldsymbol{\tau}_t \cdot \boldsymbol{\omega}_{w'}) \right) \right) + \mu \parallel \boldsymbol{\tau}_t \parallel_2^2$$

$$\text{NPMI-Score}(t) = \sum_{1 \leqslant i < j \leqslant N} \frac{\log \frac{\mathrm{P}(w_i, w_j)}{\mathrm{P}(w_i)\mathrm{P}(w_j)}}{-\log \mathrm{P}(w_i, w_j)}$$

- Sampling equations for inference in LF-LDA

$$\begin{aligned}
&\mathrm{P}(z_{d_i} = t \mid \boldsymbol{Z}_{\neg d_i}, \boldsymbol{\tau}, \boldsymbol{\omega}) \\
&\propto (N^t_{d_{\neg i}} + K^t_{d_{\neg i}} + \alpha) \\
&\left((1-\lambda) \frac{N^{t, w_{d_i}}_{\neg d_i} + \beta}{N^t_{\neg d_i} + V\beta} + \lambda \mathrm{CatE}(w_{d_i} \mid \boldsymbol{\tau}_t \, \boldsymbol{\omega}^\top) \right)
\end{aligned} \tag{1}$$

$$\mathrm{P}(s_{d_i}{=}s \mid z_{d_i}{=}t) \propto \begin{cases} (1-\lambda) \frac{N^{t, w_{d_i}}_{\neg d_i} + \beta}{N^t_{\neg d_i} + V\beta} \text{ for } s = 0 \\ \lambda \ \mathrm{CatE}(w_{d_i} | \boldsymbol{\tau}_t \, \boldsymbol{\omega}^\top) \text{ for } s = 1 \end{cases} \tag{2}$$

- Sampling equations for inference in LF-DMM

$$
\begin{aligned}
&\mathsf{P}(z_d = t, \boldsymbol{s}_d \mid \boldsymbol{Z}_{\neg d}, \boldsymbol{S}_{\neg d}, \boldsymbol{\tau}, \boldsymbol{\omega}) \\
&\propto \lambda^{K_d} \, (1-\lambda)^{N_d} \, (M^t_{\neg d} + \alpha) \frac{\Gamma(N^t_{\neg d} + V\beta)}{\Gamma(N^t_{\neg d} + N_d + V\beta)} \\
&\quad \prod_{w \in W} \frac{\Gamma(N^{t,w}_{\neg d} + N^w_d + \beta)}{\Gamma(N^{t,w}_{\neg d} + \beta)} \prod_{w \in W} \mathsf{CatE}(w \mid \boldsymbol{\tau}_t \, \boldsymbol{\omega}^\top)^{K^w_d}
\end{aligned} \tag{3}
$$

$$
\begin{aligned}
&Q(z_d = t, \boldsymbol{s}_d \mid \boldsymbol{Z}_{\neg d}, \boldsymbol{S}_{\neg d}, \boldsymbol{\tau}, \boldsymbol{\omega}) \\
&\propto \quad \lambda^{K_d} \, (1-\lambda)^{N_d} \, (M^t_{\neg d} + \alpha) \\
&\quad \prod_{w \in W} \left(\frac{N^{t,w}_{\neg d} + \beta}{N^t_{\neg d} + V\beta} \right)^{N^w_d} \prod_{w \in W} \mathsf{CatE}(w \mid \boldsymbol{\tau}_t \, \boldsymbol{\omega}^\top)^{K^w_d}
\end{aligned} \tag{4}
$$

$$Q(z_d = t \mid \boldsymbol{Z}_{\neg d}, \boldsymbol{\tau}, \boldsymbol{\omega}) \propto (M^t_{\neg d} + \alpha) \prod_{w \in W} \left(\begin{array}{l} (1-\lambda) \frac{N^{t,w}_{\neg d} + \beta}{N^t_{\neg d} + V\beta} \\ + \lambda \, \mathrm{CatE}(w \mid \boldsymbol{\tau}_t \, \boldsymbol{\omega}^\top) \end{array} \right)^{(N^w_d + K^w_d)} \tag{5}$$

$$Q(s_{d_i} = s \mid z_d = t) \propto \begin{cases} (1-\lambda) \frac{N^{t,w_{d_i}}_{\neg d} + \beta}{N^t_{\neg d} + V\beta} \text{ for } s = 0 \\ \lambda \, \mathrm{CatE}(w_{d_i} \mid \boldsymbol{\tau}_t \, \boldsymbol{\omega}^\top) \text{ for } s = 1 \end{cases} \tag{6}$$