Leveraging Multi-Domain Prior Knowledge in Topic Models

Zhiyuan Chen † Arjun Mukherjee † Bing Liu †
Meichun Hsu ‡ Malu Castellanos ‡ Riddhiman Ghosh ‡
† University of Illinois at Chicago, ‡ HP Labs
{czyuanacm, arjun4787}@gmail.com, liub@cs.uic.edu,
meichun.hsu, malu.castellanos, riddhiman.ghosh}@hp.com

IJCAI-13

UIC UNIVERSITY OF ILLINOIS AT CHICAGO

Introduction

❖ **Problem Definition**: Given prior knowledge from multiple domains, improve topic modeling in the **new** domain.

- Knowledge in the form of ***s-set*** containing words sharing the same semantic meaning, e.g., {Light, Heavy, Weight}.
- A novel technique to transfer knowledge to improve topic models.

❖ Existing Knowledge-based models

- DF-LDA [Andrzejewski et al., 2009], Seeded Model (e.g., [Mukherjee and Liu, 2012]).
- Two shortcomings: 1) Incapable of handling **multiple senses**, and 2) **Adverse effect** of Knowledge.

MDK-LDA

❖ **Generative Process**

1. For each topic $t \in \{1, \dots, T\}$
   i. Draw a per topic distribution over s-sets, $\varphi_t \sim Dir(\beta)$
   ii. For each s-set $s \in \{1, \dots, S\}$
      a) Draw a per topic, per s-set distribution over words, $\eta_{t,s} \sim Dir(\gamma)$
2. For each document $m \in \{1, \dots, M\}$
   i. Draw $\theta_m \sim Dir(\alpha)$
   ii. For each word $w_{m,n}$, where $n \in \{1, \dots, N_m\}$
      a) Draw a topic $z_{m,n} \sim Mult(\theta_m)$
      b) Draw an s-set $s_{m,n} \sim Mult(\varphi_{z_{m,n}})$
      c) Emit $w_{m,n} \sim Mult(\eta_{z_{m,n}, s_{m,n}})$

❖ **Plate Notation**

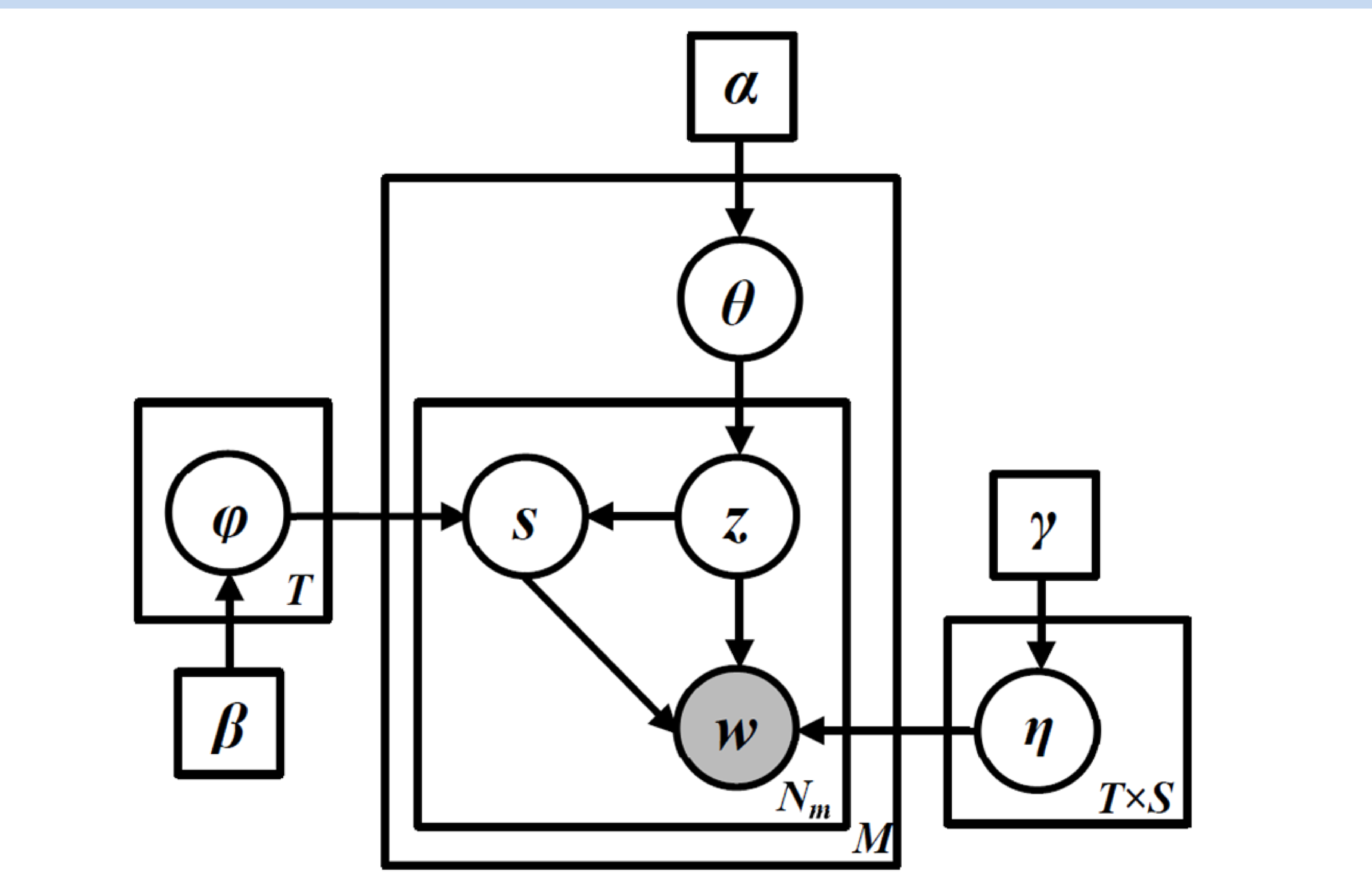

Figure 1: Plate notation of the proposed framework.

❖ **Collapsed Gibbs Sampling**

- Blocked Gibbs Sampler: Sample topic z and s-set s for word w

$$P\left(z_i = t, s_i = s \,\middle|\, \boldsymbol{z}^{-i}, \boldsymbol{s}^{-i}, \boldsymbol{w}, \alpha, \beta, \gamma\right) \propto \frac{n_{m,t}^{-i} + \alpha}{\sum_{t'=1}^{T}\left(n_{m,t'}^{-i} + \alpha\right)} \times \frac{n_{t,s}^{-i} + \beta}{\sum_{s'=1}^{S}\left(n_{t,s'}^{-i} + \beta\right)} \times \frac{n_{t,s,w_i}^{-i} + \gamma_s}{\sum_{v'=1}^{V}\left(n_{t,s,v'}^{-i} + \gamma_s\right)}$$

Generalized Pólya Urn Model

❖ Generalized Pólya urn model [Mahmoud, 2008]

- When a ball is drawn, that ball is put back along with a certain number of balls of **similar** colors.

❖ Promoting s-set as a whole

- If a ball of color w is drawn, we put back $\mathbb{A}_{s,w',w}$ balls of each color $w' \in \{1, \dots, V\}$ where w and w' share s-set s.

$$\mathbb{A}_{s,w',w} = \begin{cases} 1 & w = w' \\ \sigma & w \in s, w' \in s, w \neq w' \\ 0 & \text{otherwise} \end{cases}$$

❖ **Collapsed Gibbs Sampling**

$$P(z_i = t, s_i = s | \boldsymbol{z}^{-i}, \boldsymbol{s}^{-i}, \boldsymbol{w}, \alpha, \beta, \gamma, \mathbb{A}) \propto \frac{n_{m,t}^{-i} + \alpha}{\sum_{t'=1}^{T}\left(n_{m,t'}^{-i} + \alpha\right)} \times \frac{\sum_{w'=1}^{V}\sum_{v'=1}^{V}\mathbb{A}_{s,v',w'} \cdot n_{t,s,v'}^{-i} + \beta}{\sum_{s'=1}^{S}\left(\sum_{w'=1}^{V}\sum_{v'=1}^{V}\mathbb{A}_{s',v',w'} \cdot n_{t,s',v'}^{-i} + \beta\right)} \times \frac{n_{t,s,w_i}^{-i} + \gamma_s}{\sum_{v'=1}^{V}\left(n_{t,s,v'}^{-i} + \gamma_s\right)}$$

Experiments

❖ Datasets: reviews from six domains from Amazon.com.

❖ Baseline Models

- **LDA** [Blei et al., 2003], **LDA_GPU** [Mimno et al., 2011], and **DF-LDA** [Andrzejewski et al., 2009].

❖ Topic Discovery Results

- Evaluation measure: ***Precision @ n (p @ n)***.
- Quantitative results in Table 1, Qualitative results in Table 2.

❖ Objective Evaluation

- Topic Coherence [Mimno et al., 2011].

Domains	LDA	LDA_GPU	DF-LDA	MDK-LDA(b)	MDK-LDA
Camera	0.80	0.50	0.67	0.81	**0.93**
Computer	0.67	0.60	0.56	0.70	**0.88**
Food	0.87	0.61	0.67	0.84	**0.91**
Care	0.81	0.64	0.72	0.92	**0.91**
Average	0.79	0.59	0.66	0.82	**0.91**

Table 1 (Quantitative): Avg. precision of each model across domains.

Camera (Battery)		Computer (Price)		Food (Taste)		Care (Tooth)	
LDA	MDK	LDA	MDK	LDA	MDK	LDA	MDK
battery	extra	*acer*	cheap	taste	flavor	*price*	tooth
screen	charge	*power*	price	salt	sweet	tooth	gum
life	life	*base*	inexpensive	*almond*	sugar	*amazon*	dentist
lcd	replacement	*year*	money	*fresh*	salty	pen	dental
water	battery	*button*	expensive	*pack*	tasty	*shipping*	whitening
usb	charger	*amazon*	cost	tasty	tasting	gum	pen
cable	aa	*control*	dollar	*oil*	delicious	dentist	refill
case	power	price	buck	*roasted*	taste	whitening	*year*
charger	rechargeable	*color*	worth	pepper	salt	refill	*date*
hour	time	purchase	low	*easy*	spice	*worth*	*product*

Table 2 (Qualitative): Example topics (MDK is short for MDK-LDA); *errors* are marked in red/italic.

Poster Unique Code: 3e2315