Birds of a Feather Linked Together: A Discriminative Topic Model using Link-based Priors

Weiwei Yang[1], Jordan Boyd-Graber[2], and Philip Resnik[1,3,4]

[1]Computer Science, [3]Linguistics, [4]UMIACS, University of Maryland, College Park, MD

[2]Computer Science, University of Colorado, Boulder, CO

Abstract

- A topic model for link prediction using:
  (1) Cluster priors.
  (2) Seeding based on distributed representations.
  (3) Lexical term weights.
  (4) Max-margin learning criterion.

(1) Cluster Priors

- Clusters are identified from links, using **strongly connected component**.
- Each cluster l has its **own Dirichlet prior** π_l over its topic distribution.

(2) Seeding

- Selected from **high frequency words**, using **word2vec representations**.
- Cluster the words into K word-clusters using k-means.
- Within each topic k, compute each word's skip-gram transition probability sum to the other words.
- Select top three words as the seed words for topic k.

(3) Lexical Term Weights

- The regression value of document d and d' is

$$R_{d,d'} = \eta^T(\overline{z}_d \circ \overline{z}_{d'}) + \tau^T(\overline{w}_d \circ \overline{w}_{d'}).$$

- $\overline{z}_{d,k} = \frac{1}{N_d}\sum_{n=1}^{N_d}\mathbb{I}\left[z_{d,n} = k\right]$.
- $\overline{w}_{d,v} = \frac{1}{N_d}\sum_{n=1}^{N_d}\mathbb{I}\left[w_{d,n} = v\right]$.
- $\circ$ denotes the Hadamard product.

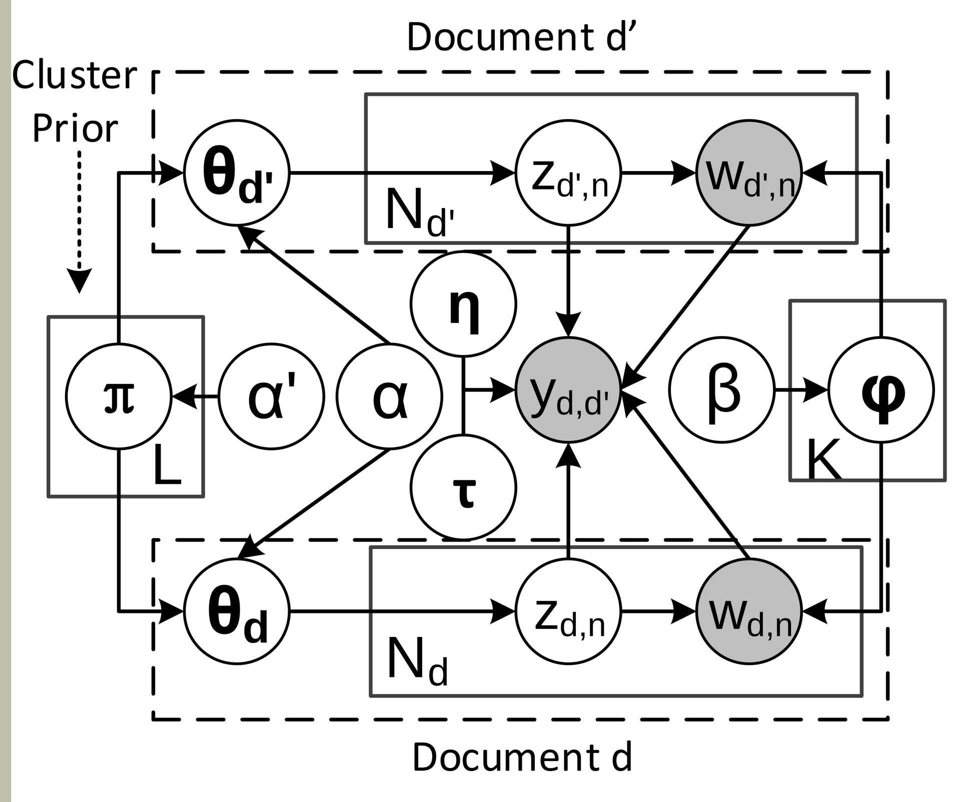
(4) Max-margin Learning

- We use **hinge loss** as the link prediction function Ψ

$$p(y_{d,d'} = 1) = \exp(-2c\max(0, 1 - R_{d,d'})).$$

- c is the regularization parameter.

Relational Topic Model with Cluster Priors and Lexical Term Weights

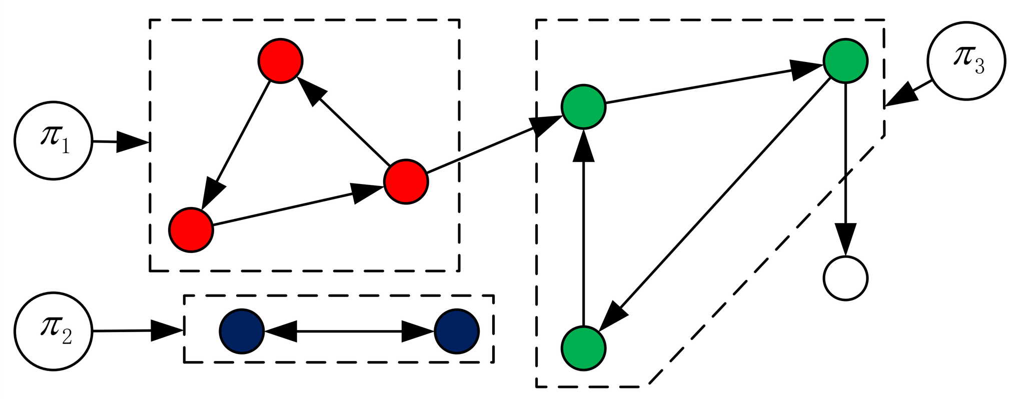


A two-document segment of our model

1. For each document cluster $l \in \{1, \ldots, L\}$
   (a) Draw $\pi_l \sim \mathrm{Dir}(\alpha')$
2. For each topic $k \in \{1, \ldots, K\}$
   (a) Draw word distribution $\phi_k \sim \mathrm{Dir}(\beta)$
   (b) Draw topic regression parameter $\eta_k \sim \mathcal{N}(0, \nu^2)$
3. For each word $v \in \{1, \ldots, V\}$
   (a) Draw lexical regression parameter $\tau_v \sim \mathcal{N}(0, \nu^2)$
4. For each document $d \in \{1, \ldots, D\}$
   (a) Draw topic proportions $\theta_d \sim \mathrm{Dir}(\alpha\pi_{l_d})$
   (b) For each word $t_{d,n}$ in document d
      i. Draw a topic assignment $z_{d,n} \sim \mathrm{Mult}(\theta_d)$
      ii. Draw a word $t_{d,n} \sim \mathrm{Mult}(\phi_{z_{d,n}})$
5. For each linked pair of documents d and d'
   (a) Draw link indicator $y_{d,d'} \sim \Psi(\cdot \mid z_d, z_{d'}, w_d, w_{d'}, \eta, \tau)$

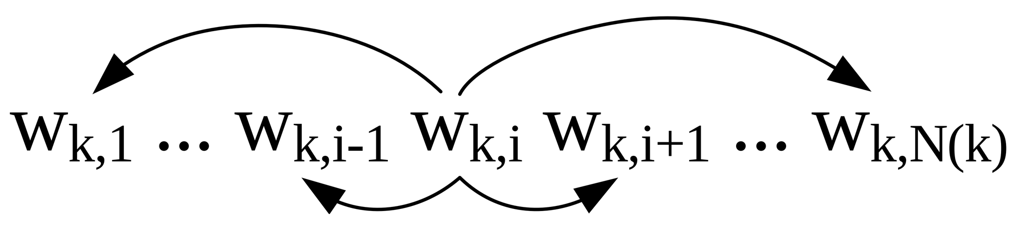
Examples

Strongly Connected Component Example



Each color denotes a component

Seed Word Transition Prob. Sum Example



An example for the i-th word in topic k

Qualitative Example

@A: Just finished a TV play. Heading for Beijing now. Exhausted. @B

Mentioning link example

A → Topic 16 ← B

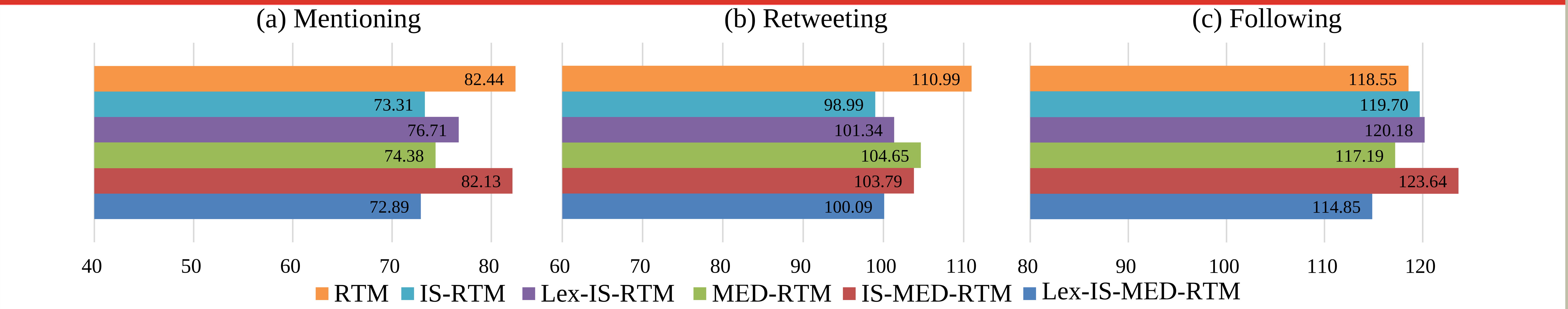
TV play, movie, song, music, program, director, drama

A and B share a common interest in entertainment

Model	PLR	%T16	
		A	B
RTM	52	.141	.108
IS-RTM	40	.129	.112
Lex-IS-RTM	**24**	**.157**	.119
MED-RTM	40	.100	.094
IS-MED-RTM	26	.082	.099
Lex-IS-MED-RTM	26	.137	**.139**

- A is an actor and also interested in food: meat, soup, popcorn, roasted duck, snack, etc.
- B is a host and also interested in sports: Olympic, badminton, gold medal, champion, referee, etc.

Link Prediction: Predictive Link Rank



- **Dataset**: Tweets from 2,000 Weibo users, with mentioning, retweeting and following links.
- **Task**: Predicting links between held-out documents.
- **Baseline**: Relational Topic Model (RTM).
- **Evaluation**: Predictive link rank (lower is better).

- **Prefixes**:
  - **IS-**: The model incorporates user interactions and seed words.
  - **Lex-**: Lexical terms were included in the link probability function.
  - **MED-**: Max-margin learning is applied.

Document Modeling

- **Dataset**: Same as link prediction.
- **Task**: Predicting held-out words in documents using various links.
- **Baseline**: LDA and Markov Random Topic Fields (MRTF).
- **Split**: Each document's 80% tokens for training. The rest for test.
- **Evaluation**: Perplexity (lower is better).

Model	LDA	MRTF	I-LDA
Mentioning		2582.08	**2522.58**
Retweeting	2605.06	2588.30	**2519.27**
Following		2587.26	**2530.67**

- I-LDA incorporates user interactions, but doesn't predict links.

Link Prediction: Quantitative Analysis

Model		RTM	IS-RTM	Lex-IS-RTM	MED-RTM	IS-MED-RTM	Lex-IS-MED-RTM
Topic PMI ↑		1.186	1.224	1.216	1.214	**1.294**	1.229
Avg Reg Values	**Linked/All ↑**	3.621	4.777	**5.026**	2.909	3.097	3.158
	SD/Avg ↓	0.9415	1.2081	1.2671	**0.6364**	0.7254	0.7353

- **Topic PMI**: Each topic's top 20 words' PMI value. Higher is better ↑.
- **Linked/All**: Ratio of linked pairs' average regression values to all pairs' values. Higher is better ↑.
- **SD/All**: Ratio of standard deviation to all pairs' average regression values. Lower is better ↓.

Future Directions

- Introduce hierarchical topic models.
- Use other clustering methods to obtain clusters.
- Explore the predicted links for downstream tasks.
  - Friend recommendation.
  - Inference of user attributes.

Paper

Supplemental

wwyang@cs.umd.edu, Jordan.Boyd.Graber@colorado.edu, resnik@umd.edu