PAPER Special Section on Data Engineering and Information Management

Relation Prediction in Multilingual Data Based on Multimodal Relational Topic Models

Yosuke SAKATA[†*], ***Nonmember and*** **Koji EGUCHI**[†a)], ***Member***

SUMMARY There are increasing demands for improved analysis of multimodal data that consist of multiple representations, such as multilingual documents and text-annotated images. One promising approach for analyzing such multimodal data is latent topic models. In this paper, we propose conditionally independent generalized relational topic models (CI-gRTM) for predicting unknown relations across different multiple representations of multimodal data. We developed CI-gRTM as a multimodal extension of discriminative relational topic models called generalized relational topic models (gRTM). We demonstrated through experiments with multilingual documents that CI-gRTM can more effectively predict both multilingual representations and relations between two different language representations compared with several state-of-the-art baseline models that enable to predict either multilingual representations or unimodal relations.
key words: *latent topic models, relational topic models, multimodal data, margin maximization*

1. Introduction

The use of multimodal data such as multilingual parallel/comparable documents and text-annotated images has increased explosively with the growth of social media services. Therefore, research on search and analysis of such multimodal data is becoming more important than ever. One promising approach for analyzing multimodal data is latent topic models [1], [2]. Latent Dirichlet allocation (LDA) [2] is one such model that is widely used. For handling multimodal data, conditionally independent LDA (CI-LDA) [3]–[5] has often been used as an extension of LDA. CI-LDA can model shared latent topics across different modes or modalities (e.g., languages) for multimodal data. However, CI-LDA cannot directly predict the relation between two different modes in multimodal data.

For predicting the relation or link between two documents, relational topic models (RTM) [6] and their generalized versions [7] have been developed in previous studies. For instance, these models consider a citation in a research paper as a link between the two papers and predict unknown links using latent topics. The links are assumed to be generated in accordance with a function with latent topics that outputs binary representations with '1' and '0' indicating the presence and absence of a link between two documents, respectively. These relational topic models aim to predict links between unimodal data, so they cannot be directly applied to multimodal data.

In this paper, we aim to predict relations across different modes (e.g., languages) in multimodal data and also predict the multimodal data themselves (e.g., words in each language mode). As illustrated in Fig. 1, we can view mode representations that are linked to each other in each multimodal data, for instance, an English *automobile* article and a Spanish *automóvil* article are connected via an inter-language link, as seen in Wikipedia. For these objectives, we propose conditionally independent gRTM (CI-gRTM) that can predict both multimodal data themselves and the relations between two different modes differently from previous models such as CI-LDA and RTM. We evaluated our CI-gRTM and several state-of-the-art baseline models through experiments with multilingual parallel and comparable documents and demonstrated that our model effectively predicts both multimodal data themselves and inter-mode relations.

2. Related Work

In this section, we briefly review some previous topic models: LDA, CI-LDA, RTM, and gRTM. CI-LDA can represent shared latent topics among multiple modes. RTM and gRTM can predict relations between *unimodal* documents. There are different types of multilingual topics models, such as in [8]–[10]. However, those models are based on the premise of using some additional knowleage or resources, such as multilingual dictionaries, while our assumption in

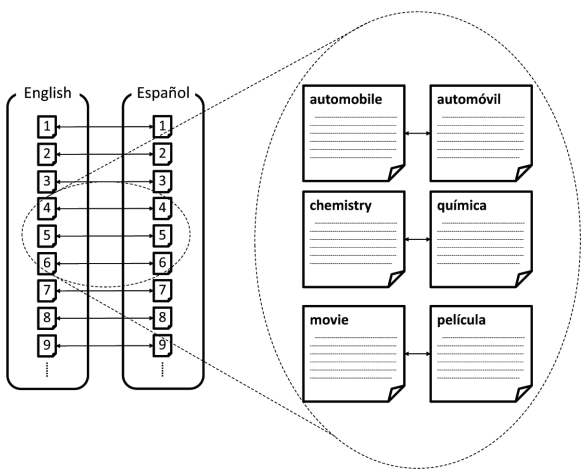


Fig. 1 Illustration of cross-modal links.

Manuscript received June 22, 2016.
Manuscript revised November 8, 2016.
Manuscript publicized January 17, 2017.
[†]The authors are with Kobe University, Kobe-shi, 657–8501 Japan.
[*]Presently, with Sumitomo Life Information Systems Co., Ltd.
a) E-mail: eguchi@port.kobe-u.ac.jp
DOI: 10.1587/transinf.2016DAP0021

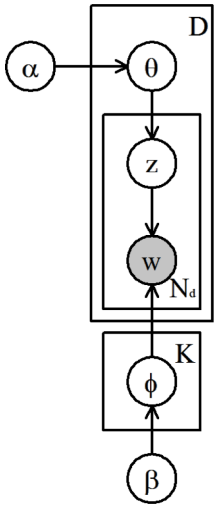


Fig. 2 Graphical model of LDA.

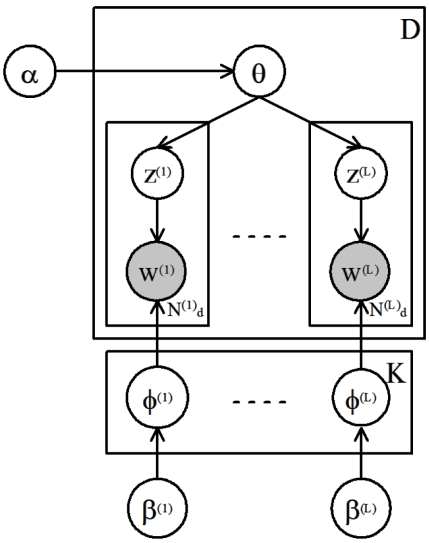


Fig. 3 Graphical model of CI-LDA.

this paper is that only multilingual comparable documents are available, as in the work on CI-LDA.

2.1 LDA

Latent Dirichlet allocation (LDA) assumes that each document is represented as a mixture of topics, where each topic is represented as a multinomial distribution over words. Figure 2 shows a graphical model representation of LDA, where D, N_d, and K indicate the number of documents, the number of words in document d, and the number of latent topics, respectively. In this figure, the shaded circle represents an observed variable. $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ indicate multinomial parameters over topics with respect to document d and multinomial parameters over words with respect to topic k, respectively. α and β are hyperparameters of Dirichlet priors for each of the multinomial distributions. The generating process can be described as:

1. Draw per-document multinomial $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha)$ (where $d \in \{1, \ldots, D\}$).
2. Draw per-topic multinomial $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\beta)$ (where $k \in \{1, \ldots, K\}$).
3. For each word in document d:
   a. Draw topic assignment $z_{di} \sim \mathrm{Mult}(\theta_d)$.
   b. Draw word $w_{di} \sim \mathrm{Mult}(\phi_k)$.

Here, Mult$(\cdot)$ and Dir$(\cdot)$ indicate a multinomial distribution and a Dirichlet distribution, respectively.

2.2 CI-LDA

CI-LDA [3]–[5] is an extension of LDA that can handle multimodal data, such as multilingual parallel documents. When the target is multilingual parallel documents, each mode corresponds to a language. We show a graphical model representation of CI-LDA in Fig. 3, assuming that the number of modes is L, where each superscript variable indicates a mode.

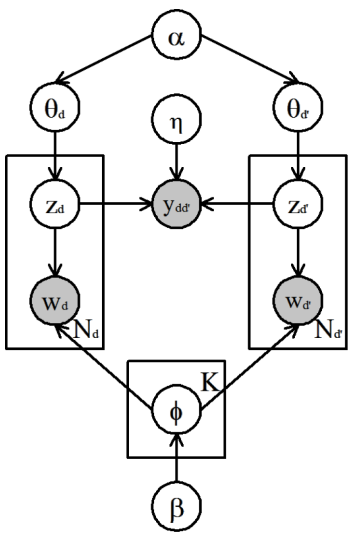


Fig. 4 Graphical model of RTM.

The generative process of CI-LDA can be described as:

1. Draw per-document multinomial $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha)$ (where $d \in \{1, \ldots, D\}$).
2. Draw per-topic, per-language multinomial $\boldsymbol{\phi}_k^{(\ell)} \sim \mathrm{Dir}(\beta)$ (where $k \in \{1, \ldots, K\}$ and $\ell \in \{1, \ldots, L\}$).
3. For each word of document d and language ℓ:
   a. Draw topic assignment $z_{di}^{(\ell)} \sim \mathrm{Mult}(\theta_d)$.
   b. Draw word $w_{di}^{(\ell)} \sim \mathrm{Mult}(\boldsymbol{\phi}_k^{(\ell)})$.

2.3 RTM and gRTM

Relational topic models (RTM) [6] are an extension of LDA that considers both text content and network structure. Figure 4 shows a graphical model representation of RTM, where the document plate is omitted for convenience. In

this figure, $\boldsymbol{\eta} = (\eta_k)$ is the weight vector for link generation, indicating that the larger the weight is, the more the corresponding topic contributes to link generation. The generating process can be described as:

1. Draw per-document multinomial $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha)$ (where $d \in \{1, \ldots, D\}$).
2. Draw per-topic multinomial $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\beta)$ (where $k \in \{1, \ldots, K\}$).
3. For each word of document d:
   a. Draw topic assignment $z_{di} \sim \mathrm{Mult}(\theta_d)$.
   b. Draw word $w_{di} \sim \mathrm{Mult}(\phi_k)$.
4. Draw link $y_{dd'} \sim \psi$.

Here, ψ indicates the link probability function that defines the probability of generating a link between two documents, depending on the latent topics in these documents, as below:

$$\psi(y_{dd'} = 1 | \mathbf{z}_d, \mathbf{z}_{d'}, \boldsymbol{\eta}) = \sigma(\boldsymbol{\eta}^{\mathrm{T}}(\bar{\mathbf{z}}_d \circ \bar{\mathbf{z}}_{d'})) \tag{1}$$

where $\mathbf{z}_d = \{z_{di}\}$ while $\bar{\mathbf{z}}_d = (\bar{z}_{d,k})$ and $\bar{z}_{d,k} = \frac{1}{N_d} C_d^k$. Here, C_d^k is the number of times topic k is assigned to any words in document d. Operator '$\circ$' denotes the elementwise product, and σ denotes the sigmoid function. In the work by Chang et al. [6], some other choices of σ were also used, for example, using the exponential function and the cumulative distribution function of a normal distribution. In this paper, we use the commonly used logistic likelihood model [11] with the sigmoid function.

RTM's link probability function is based on the elementwise product so that it only considers interactions between the same topics. Therefore, some of the weight η_k values are positive while others may be negative. The negative interactions are undesirable when trying to understand the process of generating links in document networks. To address this issue, generalized RTM (gRTM) [7] extends the link probability function of RTM, as below:

$$\psi(y_{dd'} = 1 | \mathbf{z}_d, \mathbf{z}_{d'}, U) = \{\sigma(\bar{\mathbf{z}}_d^{\mathrm{T}} U \bar{\mathbf{z}}_{d'})\}^c \tag{2}$$

where U is the full $K \times K$ weight matrix, and K denotes the number of topics. Regularization parameter c controls the likelihood of generating links. The graphical model representation of gRTM is essentially the same as that of RTM and is derived by replacing $\boldsymbol{\eta}$ with U as shown in Fig. 4.

3. Multimodal Relational Topic Models

3.1 CI-gRTM

As mentioned in Sect. 2.2, CI-LDA represents topics shared across multiple modes in multimodal data. However, CI-LDA cannot directly predict relations or links across the modes. To address this problem, we propose conditionally independent generalized RTM (CI-gRTM). Figure 5 shows a graphical model representation of CI-gRTM, where each superscript variable indicates a mode or a pair of modes. This model uses the link probability function, which is similar to

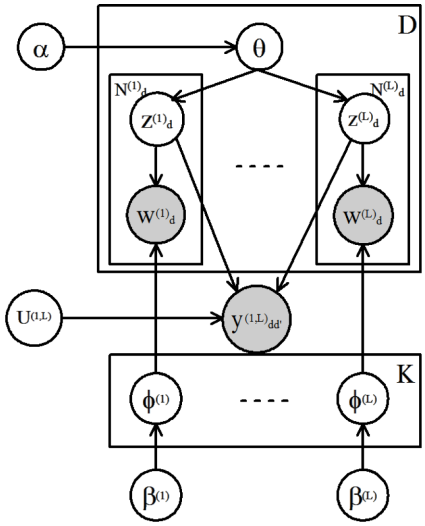


Fig. 5 Graphical model of CI-gRTM.

that of gRTM in Sect. 2.3. $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_d\}$ are the per-document multinomial parameters that are common in all modes in each document. The generating process can be described as:

1. Draw per-document multinomial $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha)$ (where $d \in \{1, \ldots, D\}$).
2. Draw per-topic multinomial $\boldsymbol{\phi}_k^{(\ell)} \sim \mathrm{Dir}(\beta)$ (where $k \in \{1, \ldots, K\}$ and $\ell \in \{1, \ldots, L\}$).
3. For each word of document d:
   a. Draw topic assignment $z_{di}^{(\ell)} \sim \mathrm{Mult}(\theta_d)$.
   b. Draw word $w_{di}^{(\ell)} \sim \mathrm{Mult}(\phi_k^{(\ell)})$.
4. Draw link $y_{dd'}^{(\ell_1,\ell_2)} \sim \psi$ (where $\ell_1, \ell_2 \in \{1, \ldots, L\}$).

The link probability function of CI-gRTM is defined as follows:

$$\psi(y_{dd'}^{(\ell_1,\ell_2)} = 1 | \mathbf{z}_d^{(\ell_1)}, \mathbf{z}_{d'}^{(\ell_2)}, U^{(\ell_1,\ell_2)}) = \{\sigma(\bar{\mathbf{z}}_d^{(\ell_1)\mathrm{T}} U^{(\ell_1,\ell_2)} \bar{\mathbf{z}}_{d'}^{(\ell_2)})\}^c \tag{3}$$

where $\bar{\mathbf{z}}_d^{(\ell_1)}$ and $\bar{\mathbf{z}}_{d'}^{(\ell_2)}$ mean the expectations of topic assignments in mode ℓ_1 of document d and mode ℓ_2 of document d', respectively. Here, document d' can be the same as document d, for instance, modes ℓ_1 and ℓ_2 of document d are assumed to be linked to each other.

3.2 Inference with Collapsed Gibbs Sampling

MedLDA [12] is a model that can infer unknown parameters and latent variables as optimizations. It requires assuming hard constraints in the process of inference of the model. In this section, we discuss an inference with collapsed Gibbs sampling [13] that is simple and efficient. The algorithm of the collapsed Gibbs sampling is based on data augmentation [14]. We discuss below the inference of CI-gRTM in accordance with the previous study [7]. First, we show the posterior distribution for all unknown parameters and latent

variables.

$$q(\mathbf{U},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}) = \frac{p_0(\mathbf{U},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})p(\mathbf{W}|\mathbf{Z},\boldsymbol{\Phi})\psi(\mathbf{y}|\mathbf{Z},\mathbf{U})}{\varphi(\mathbf{y},\mathbf{W})} \tag{4}$$

where $\mathbf{W} = \{w_{di}^{(\ell)}\}$, $\mathbf{Z} = \{z_{di}^{(\ell)}\}$, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_d\}$, $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_k^{(\ell)}\}$, $\mathbf{U} = \{U^{(\ell_1,\ell_2)}\}$, and $\mathbf{y} = \{y_{dd'}^{(\ell_1,\ell_2)}\}$. $p_0(\mathbf{U},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})$ is the prior, $p(\mathbf{W}|\mathbf{Z},\boldsymbol{\Phi})$ is the likelihood of generating words, $\psi(\mathbf{y}|\mathbf{Z},\mathbf{U})$ is the pseudo-likelihood of generating links, and $\varphi(\mathbf{y},\mathbf{W})$ is the normalizing constant. Using data augmentation [7], [14], ψ is formulated as follows:

$$\psi(y_{dd'}^{(\ell_1,\ell_2)}|\mathbf{z}_d^{(\ell_1)},\mathbf{z}_{d'}^{(\ell_2)},U^{(\ell_1,\ell_2)}) = \frac{\exp(\kappa_{dd'}^{(\ell_1,\ell_2)}\omega_{dd'}^{(\ell_1,\ell_2)})}{2^c} \int_0^\infty \exp\left(-\frac{\lambda_{dd'}^{(\ell_1,\ell_2)}\omega_{dd'}^{(\ell_1,\ell_2)2}}{2}\right) p(\lambda_{dd'}^{(\ell_1,\ell_2)}|c,0)d\lambda_{dd'}^{(\ell_1,\ell_2)} \tag{5}$$

where $\kappa_{dd'}^{(\ell_1,\ell_2)} = c(y_{dd'}^{(\ell_1,\ell_2)} - 1/2)$, and $\omega_{dd'}^{(\ell_1,\ell_2)} = \bar{\mathbf{z}}_d^{(\ell_1)\mathrm{T}}U^{(\ell_1,\ell_2)}\bar{\mathbf{z}}_{d'}^{(\ell_2)}$. $\boldsymbol{\lambda} = \{\lambda_{dd'}^{(\ell_1,\ell_2)}\}$ is a variable for data augmentation. The greater the regularization parameter c is, the more misclassification can be allowed. $p(\lambda_{dd'}^{(\ell_1,\ell_2)}|c,0)$ follows Polya-Gamma distribution [14], as below:

$$p(\lambda_{dd'}^{(\ell_1,\ell_2)}|a,b) = \frac{1}{2\pi^2}\sum_{i=1}^{\infty}\frac{g_i}{(i-1/2)^2 + b^2/(4\pi^2)} \tag{6}$$

where g_i follows Gamma distribution $\mathcal{G}(a,1)$. Therefore, the posterior distribution with $\boldsymbol{\lambda}$ is:

$$q(\mathbf{U},\boldsymbol{\lambda},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}) = \frac{p_0(\mathbf{U},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})p(\mathbf{W}|\mathbf{Z},\boldsymbol{\Phi})\psi(\mathbf{y},\boldsymbol{\lambda}|\mathbf{Z},\mathbf{U})}{\varphi(\mathbf{y},\mathbf{W})} \tag{7}$$

where the pseudo-joint distribution of $\mathbf{y}$ and $\boldsymbol{\lambda}$ is:

$$\psi(\mathbf{y},\boldsymbol{\lambda}|\mathbf{Z},\mathbf{U}) = \prod_{dd'}\exp\left(\kappa_{dd'}^{(\ell_1,\ell_2)}\omega_{dd'}^{(\ell_1,\ell_2)} - \frac{\lambda_{dd'}^{(\ell_1,\ell_2)}\omega_{dd'}^{(\ell_1,\ell_2)2}}{2}\right) p(\lambda_{dd'}^{(\ell_1,\ell_2)}|c,0) \tag{8}$$

We then marginalize out $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$, and the resulting collapsed posterior is:

$$q(\mathbf{U},\boldsymbol{\lambda},\mathbf{Z}) \propto p_0(\mathbf{U})\prod_{k=1}^{K}\frac{\delta(\mathbf{C}_k^{(\ell)} + \beta^{(\ell)})}{\delta(\beta^{(\ell)})}\prod_{d=1}^{D}\frac{\delta(\mathbf{C}_d + \alpha)}{\delta(\alpha)} \prod_{dd'}\exp\left(\kappa_{dd'}^{(\ell_1,\ell_2)}\omega_{dd'}^{(\ell_1,\ell_2)} - \frac{\lambda_{dd'}^{(\ell_1,\ell_2)}\omega_{dd'}^{(\ell_1,\ell_2)2}}{2}\right) p(\lambda_{dd'}^{(\ell_1,\ell_2)}|c,0) \tag{9}$$

where $\delta(\mathbf{x}) = \frac{\prod_{i=1}^{\dim(\mathbf{x})}\Gamma(x_i)}{\Gamma(\prod_{i=1}^{\dim(\mathbf{x})}x_i)}$. $\mathbf{C}_k^{(\ell)} = \{C_k^{v^{(\ell)}}\}_{v=1}^{V^{(\ell)}}$ when $C_k^{v^{(\ell)}}$ is the number of counts when topic k is assigned to word type v in language ℓ. Here, $V^{(\ell)}$ indicates the number of word types in language ℓ. Similarly, $\mathbf{C}_d = \{C_d^k\}_{k=1}^K$ when C_d^k is the number of counts when topic k is assigned to any words in document d.

Below we show the full conditional probability of each parameter, assuming the use of collapsed Gibbs sampling.

3.2.1 Inference of U

We assume $\bar{\mathbf{z}}_{dd'}^{(\ell_1,\ell_2)} = \mathrm{vec}(\bar{\mathbf{z}}_d^{(\ell_1)}\bar{\mathbf{z}}_{d'}^{(\ell_2)\mathrm{T}})$ and $\boldsymbol{\eta}^{(\ell_1,\ell_2)} = \mathrm{vec}(U^{(\ell_1,\ell_2)})$. Here, $\mathrm{vec}(A)$ defines the vector concatenating the row vectors of A. We then have $\omega_{dd'}^{(\ell_1,\ell_2)} = \boldsymbol{\eta}^{(\ell_1,\ell_2)\mathrm{T}}\bar{\mathbf{z}}_{dd'}^{(\ell_1,\ell_2)}$. When a Gaussian prior of $U^{(\ell_1,\ell_2)}$ is assumed to be $p_0(U^{(\ell_1,\ell_2)}) = \prod_{kk'}\mathcal{N}(U_{kk'}^{(\ell_1,\ell_2)};0,\nu^2)$, we obtain:

$$\begin{aligned} q(\boldsymbol{\eta}^{(\ell_1,\ell_2)}|\mathbf{Z},\boldsymbol{\lambda}) &\propto p_0(\boldsymbol{\eta}^{(\ell_1,\ell_2)})\prod_{dd'}\exp\Big(\kappa_{dd'}^{(\ell_1,\ell_2)}\boldsymbol{\eta}^{(\ell_1,\ell_2)\mathrm{T}}\bar{\mathbf{z}}_{dd'}^{(\ell_1,\ell_2)} \\ &\quad - \frac{\lambda_{dd'}^{(\ell_1,\ell_2)}(\boldsymbol{\eta}^{(\ell_1,\ell_2)\mathrm{T}}\bar{\mathbf{z}}_{dd'}^{(\ell_1,\ell_2)})^2}{2}\Big) \\ &= \mathcal{N}(\boldsymbol{\eta}^{(\ell_1,\ell_2)};\boldsymbol{\mu},\boldsymbol{\Sigma}) \end{aligned} \tag{10}$$

where the posterior mean is $\boldsymbol{\mu} = \boldsymbol{\Sigma}(\sum_{dd'}\kappa_{dd'}^{(\ell_1,\ell_2)}\bar{\mathbf{z}}_{dd'}^{(\ell_1,\ell_2)})$ and the covariance is $\boldsymbol{\Sigma} = (\frac{1}{\nu^2}I + \sum_{dd'}\lambda_{dd'}^{(\ell_1,\ell_2)}\bar{\mathbf{z}}_{dd'}^{(\ell_1,\ell_2)}\bar{\mathbf{z}}_{dd'}^{(\ell_1,\ell_2)\mathrm{T}})^{-1}$. We can easily draw a sample from this K^2-dimensional Gaussian distribution. This enables the inference of $\mathbf{U}$ by the procedure above for all pairs of languages.

3.2.2 Inference of Z

The full conditional probability of $\mathbf{Z}$ is:

$$\begin{aligned} q(z_{di}^{(\ell)} = k|\mathbf{Z}_{-di},\mathbf{U},\boldsymbol{\lambda},\mathbf{W}) &\propto \frac{(C_{d,-i}^k + \alpha)(C_{k,-i}^{v^{(\ell)}} + \beta^{(\ell)})}{\sum_{v^{(\ell)}}C_{k,-i}^{v^{(\ell)}} + V^{(\ell)}\beta^{(\ell)}} \\ &\prod_{\ell'\in\mathcal{L}_{-\ell}}\prod_{d'\in\mathcal{N}_d}\psi(y_{dd'}^{(\ell,\ell')}|\boldsymbol{\lambda},\mathbf{Z}_{-di},z_{di}^{(\ell)} = k) \\ &\prod_{\ell'\in\mathcal{L}_{-\ell}}\prod_{d'\in\mathcal{N}_d}\psi(y_{d'd}^{(\ell',\ell)}|\boldsymbol{\lambda},\mathbf{Z}_{-di},z_{di}^{(\ell)} = k) \end{aligned} \tag{11}$$

where $\psi(y_{dd'}^{(\ell,\ell')}|\boldsymbol{\lambda},\mathbf{Z}) = \exp(\kappa_{dd'}^{(\ell,\ell')}\omega_{dd'}^{(\ell,\ell')} - \frac{\lambda_{dd'}^{(\ell,\ell')}\omega_{dd'}^{(\ell,\ell')2}}{2})$. $\mathcal{N}_d = \{d' : (d,d') \in \mathcal{I}\}$ denotes a set of document d' that are observed to be linked or unlinked to document d in the training set. Note that document d' can be the same as document d, as mentioned at the end of Sect. 3.1. Subscript '$-i$' means that word i is removed. $\mathcal{L}_{-\ell}$ indicates that language ℓ is removed from a set of language modes $\mathcal{L}$. We can see that the first term in the right-hand side corresponds to the word counts in LDA, and the second and third terms are derived from a set of links $\mathbf{y}$.

3.2.3 Inference of λ

Finally, the full conditional probability of the data-augmentation variable $\boldsymbol{\lambda}$ is:

$$\begin{aligned} q(\lambda_{dd'}^{(\ell_1,\ell_2)}|\mathbf{Z},\mathbf{U}) &\propto \exp\left(-\frac{\lambda_{dd'}^{(\ell_1,\ell_2)}\omega_{dd'}^{(\ell_1,\ell_2)2}}{2}\right)p(\lambda_{dd'}^{(\ell_1,\ell_2)}|c,0) \\ &= \mathcal{PG}(\lambda_{dd'}^{(\ell_1,\ell_2)};c,\omega_{dd'}^{(\ell_1,\ell_2)}) \end{aligned}$$

Table 1 Dataset-A after pre-processing.

	Dataset-A		Dataset-B		
	Japanese	English	Japanese	English	Spanish
No. of documents	14111		5818		
No. of word tokens	2983135	4338115	1827463	5062266	2507643
No. of word types	23979	34398	21618	47650	33721

Table 2 Examples of estimated multimodal topics in English and Spanish.

Topic: **football**

English		Español (*English translation*)	
cup	0.0147	copa (*cup*)	0.0124
league	0.0126	club (*club*)	0.0121
football	0.0110	fútbol (*football*)	0.0117
team	0.0110	equipo (*equipment*)	0.0098
season	0.0096	temporada (*season*)	0.0079
club	0.0092	liga (*league*)	0.0078
game	0.0062	final (*final*)	0.0075
match	0.0059	mundial (*world*)	0.0064
player	0.0058	selección (*selection*)	0.0062
players	0.0057	partido (*match*)	0.0059

Topic: **plane**

English		Español (*English translation*)	
air	0.0358	aeropuerto (*airport*)	0.0098
aircraft	0.0202	air (*air*)	0.0092
airlines	0.0187	avión (*plane*)	0.0086
airport	0.0101	vuelo (*flight*)	0.0085
international	0.0096	guerra (*war*)	0.0083
flight	0.0095	aviones (*planes*)	0.0074
service	0.0067	servicio (*service*)	0.0069
war	0.0066	internacional (*international*)	0.0059
force	0.0059	aérea (*area*)	0.0056
airways	0.0058	fueron (*they*)	0.0055

As can be seen above, $\boldsymbol{\lambda}$ follows Polya-Gamma distribution. Note that $\boldsymbol{\lambda}$ takes a different value for each language pair. For instance, when J and E indicate Japanese and English, respectively, $\lambda_{dd'}^{(J,E)}$ and $\lambda_{dd'}^{(E,J)}$ are different. This enables the inference of $\boldsymbol{\lambda}$ by the procedure above for all languages pairs.

With the conditional distributions above, we can estimate the model by iteratively drawing samples of $\mathbf{U}$, $\mathbf{Z}$ and $\boldsymbol{\lambda}$.

4. Experiments

In this section, we optimize regularization parameter c in CI-gRTM, since it is sensitive to predict links. Through the experiments with two datasets, we then evaluated our CI-gRTM compared with CI-RTM and gRTM in a link prediction task and compared it with CI-LDA, CI-RTM, gRTM, and LDA in a word prediction task. For the details of LDA, CI-LDA, gRTM, and CI-gRTM, see Sects. 2.1, 2.2, 2.3, and 3.1, respectively. CI-RTM is a model that only uses diagonals of CI-gRTM's weight matrix U.

4.1 Datasets

The first dataset we used for experiments is Japanese-English bilingual documents on Kyoto, a historical town in Japan[†]. This dataset, which we refer to as Dataset-A, consists of 14,111 Japanese articles on Kyoto's people and buildings that were extracted from Wikipedia and their English translations. We removed low frequency words that appear in less than five articles [13]. For Japanese articles, we also removed symbols and function words, such as conjunctions and particles, using part-of-speech tags annotated by MeCab[††]. For English articles, we removed 418 types of standard stopwords [15]. The statistics of Dataset-A after preprocessing are shown in Table 1.

Dataset-A can be said to be a bilingual *parallel* corpus, since it consists of a pair of translations. The second dataset (which we refer to as Dataset-B) is a trilingual *comparable* corpus in English, Spanish, and Japanese, all of which were extracted from Wikipedia, where each set of Wikipedia articles are connected via inter-language links. Here, each article is not a translation of the other article that is connected via an inter-language link; however, the main subjects of the two articles are the same. We extracted text content from the original Wikipedia articles, removing link information and revision history information. We used WP2TXT[†††] for this purpose. For simplicity, we only used the articles whose titles begin with 'A' from the collection of English articles. We applied the same pre-processing that was used for Dataset-A. As for Spanish articles, we removed 351 types of standard stopwords[††††]. The statistics of Dataset-B after preprocessing are shown in Table 1. In Datasets A and B, presence of relations is assumed for the Wikipedia article pairs that are explicitly associated each other by document-level alignment for Dataset A or inter-language links for Dataset B (referred to as positive pairs), while absence of relations is assumed for all the other article pairs (referred to as negative pairs). For the model estimation, we randomly selected the double number of negative pairs compared with that of positive pairs to balance the data. Following the terminology of topic modeling, we refer to each unit of Datasets-A and -B as a *document* that consists of multiple language parts. Table 2 shows two examples of estimated multimodal topics in English and Spanish using CI-gRTM with Dataset-B, omitting the Japanese part for simplicity. Under the same setting, Fig. 6 gives an example of the estimated weight matrix and the corresponding topics. In this figure, the weight matrix's diagonal elements are positive (as colored in red) but non-diagonal elements are negative

[†]http://alaginrc.nict.go.jp/WikiCorpus/
[††]http://mecab.googlecode.com/svn/trunk/mecab/doc/
[†††]http://wp2txt.rubyforge.org/
[††††]http://members.unine.ch/jacques.savoy/clef/spanishSmart.txt

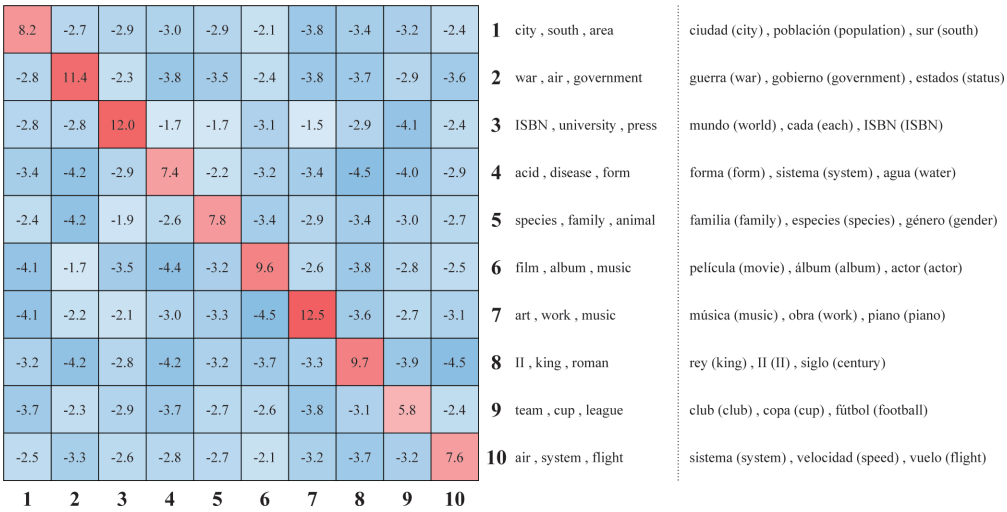


Fig. 6 Example of estimated weight matrix and corresponding topics. The elements with positive weights are colored in red, while the elements with negative weights are colored in blue.

(as colored in blue). It indicates that any pair of the same topics contributes to the link generatation, while any pair of different topics does not.

4.2 Evaluation of Regularization Parameter c

In this section, we determine the optimal regularization parameter c for CI-gRTM, CI-RTM, and gRTM using Dataset-A. We randomly sampled 20% of the documents from the dataset as the *test set*, and the remaining documents were used for the cross-validation. We applied 4-fold cross-validation where, for each experiment, a fourth part was used for validation and the other parts were used for training. For training, we estimated the unknown parameters and latent variables by collapsed Gibbs sampling. Both Datasets-A and -B contain mostly negative links rather than positive links. Following the previous study [7], we randomly selected negative links for training so that the number of negative links was double that of positive links. For validation, we carried out the experiments in the task of link prediction and obtained the F-measure. We also obtained the perplexity of validation documents to evaluate prediction performance of unseen words. The perplexity is defined as the reciprocal of the per-word geometric mean of the likelihood as shown below:

$$p(D_{test}) = \prod_{\ell=1}^{L} \prod_{d=1}^{D_{test}} \prod_{i=1}^{N_d^{(\ell)}} \sum_{k=1}^{K} \frac{C_d^k + \alpha}{\sum_{k'} C_d^{k'} + K\alpha} \frac{C_k^{w_i^{(\ell)}} + \beta^{(\ell)}}{\sum_{w_i'^{(\ell)}} C_k^{w_i'^{(\ell)}} + V^{(\ell)}\beta^{(\ell)}} \tag{12}$$

The smaller the perplexity is, the more effectively the model works.

We evaluated the F-measure and perplexity while varying the regularization parameter c. We set c to be one for negative links while varying it to be $c \in \{1, 2, 4, 8, 16\}$ for positive links. We determined the optimal c for positive links via the 4-fold cross-validation. The hyperparameters were set to be $\alpha = 0.1$ and $\beta^{(J)} = \beta^{(E)} = 0.01$. Here, $\beta^{(J)}$ and $\beta^{(E)}$ are used for Japanese and English articles, respectively. gRTM cannot handle multiple modes, so we mixed Japanese and English words to make up a single-mode representation, assuming $\beta = 0.01$. We set λ in Eq. (5) to be one in accordance with the previous study [7]. We also assumed the number of topics $K \in \{5, 10, 15\}$.

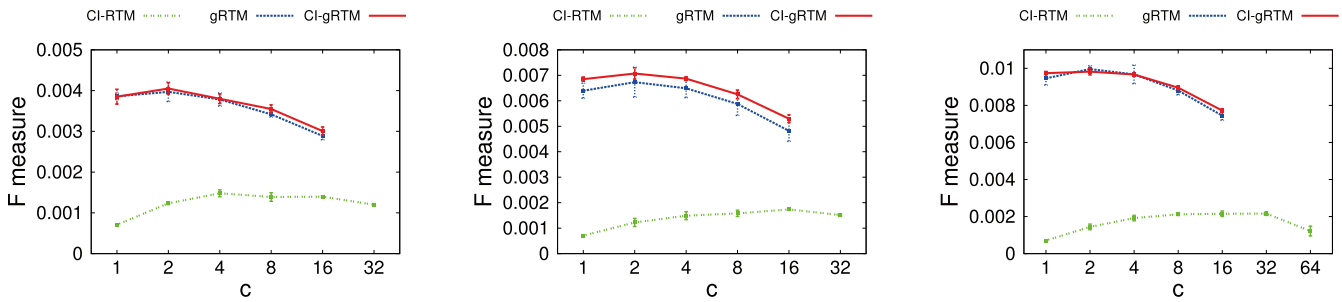
In Fig. 7, we show the results of F-measure with varying regularization parameter c for three models: CI-gRTM, CI-RTM, and gRTM. In these figures, we can see that our CI-gRTM performs better than CI-RTM in terms of F-measure. Therefore, it is important to consider all pairwise topic interactions for modeling the multimodal data. We can also see that the F-measure of CI-gRTM and gRTM are comparable for any number of topics. CI-gRTM and gRTM have the best F-measure when $c = 2$, while CI-RTM works best when $c = 16$. This is probably because CI-gRTM and gRTM compute the link likelihood using all pairwise topics, bringing a larger number of parameters, so a larger c results in a smaller F-measure because of overfitting.

In Fig. 8, we show the perplexity with varying c for three models: CI-gRTM, CI-RTM, and gRTM. As can be seen in this figure, gRTM performed the worst in every condition. Unlike the other models, gRTM cannot distinguish multiple languages, so it has to handle a larger number of word types together.

From all the evaluation results above, we determined regularization parameter c to be $c = 2$ for CI-gRTM, $c = 16$ for CI-RTM, and $c = 2$ for gRTM.

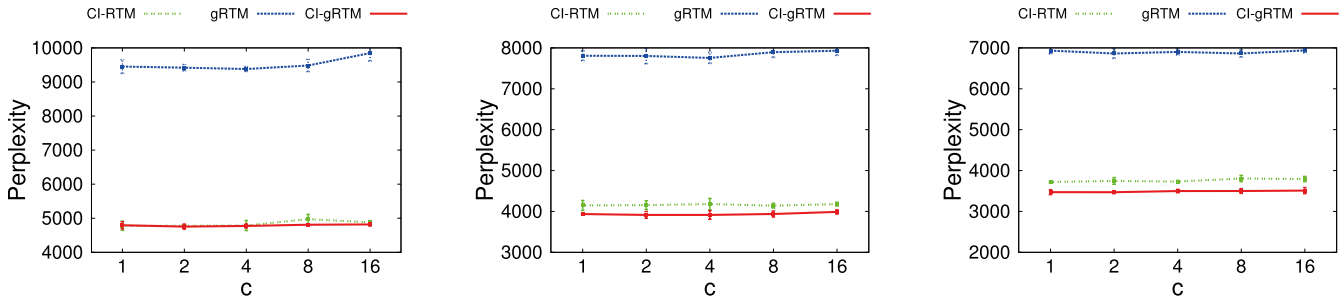
4.3 Testing

Using regularization parameter c that was determined previously, we evaluated the models with unseen documents in terms of F-measure and perplexity. For the testing, we used



(a) $K = 5$ (b) $K = 10$ (c) $K = 15$

Fig. 7 F-measure with varying regularization parameter c when number of topics K is 5, 10, and 15. Error bars represent one sample standard deviation.



(a) $K = 5$ (b) $K = 10$ (c) $K = 15$

Fig. 8 Perplexity with varying regularization parameter c when number of topics K is 5, 10, and 15. Error bars represent one sample standard deviation.

Table 3 F-measure and perplexity with Dataset-A.

	F-measure				Perplexity			
	$K = 10$	$K = 15$	$K = 20$	$K = 25$	$K = 10$	$K = 15$	$K = 20$	$K = 25$
CI-LDA	—	—	—	—	3695.20	3412.06	3056.85	2877.88
CI-RTM	0.00198	0.00227	0.00327	0.00446	3671.30	3568.07	3212.77	2931.03
gRTM	0.00627	0.00999	0.01202	0.01431	7815.39	6532.47	6092.29	5683.37
LDA	—	—	—	—	7250.17	6516.22	5981.70	5647.45
CI-gRTM	0.00708	0.00978	0.01215	0.01528	3661.51	3449.69	3094.65	2895.75

some baselines that are appropriate for target tasks: CI-RTM and gRTM for the task of link prediction and CI-LDA, CI-RTM, and gRTM for the task of word prediction. To avoid loss of generality, we used both Datasets-A and -B. We first estimated unknown parameters and latent variables for each model using 80% of the documents of each dataset, as in Sect. 4.2. We then evaluated the models using the remaining documents as test sets. The number of topics was set to $\{10, 15, 20, 25\}$ for Dataset-A and $\{5, 10, 15, 20\}$ for Dataset-B. The other settings were the same as in Sect. 4.2. We obtained the F-measure for the link prediction task and test-set perplexity for the word prediction task.

Table 3 shows the results of the F-measure and test-set perplexity for each model with Dataset-A. First, let us take a look at CI-gRTM and gRTM in these tables. As you can see in the left side of Table 3, CI-gRTM's and gRTM's link prediction performances are higher than those of the others, while CI-RTM's performance is significantly lower. This is probably because CI-gRTM and gRTM can consider all pairwise topic interactions, while CI-RTM cannot. As for perplexity, CI-gRTM works significantly better than gRTM and LDA, as shown in the right side of Table 3. Here, note that the smaller the perplexity is, the more effectively the model works. This is probably because CI-gRTM can handle multiple modes, Japanese and English, while gRTM (and LDA) cannot. Second, let us compare CI-gRTM with CI-LDA. CI-gRTM's perplexity is comparable with that of CI-LDA, as can be seen in the right side of Table 3. This is probably because usually links are partially observed. However, CI-LDA cannot predict links; therefore, there is no F-measure result for this model, while CI-gRTM achieved a high performance in link prediction, as you can see in the left side of Table 3.

We also performed evaluation with Dataset-B, as shown in Table 4. As seen in Table 3, the results of CI-RTM and LDA are clearly worse than those of the others, so we omitted them for this evaluation. We used the same regularization parameter c as was used with Dataset-A for simplicity. In Table 4, you can see tendencies that are similar to those in Table 3. In particular, CI-gRTM significantly

Table 4 F-measure and perplexity with Dataset-B.

	F-measure				Perplexity			
	$K = 5$	$K = 10$	$K = 15$	$K = 20$	$K = 5$	$K = 10$	$K = 15$	$K = 20$
CI-LDA	—	—	—	—	7059.01	6124.24	5453.43	5112.33
gRTM	0.00694	0.01050	0.02046	0.02054	22077.2	14586.8	11970.9	10546.7
CI-gRTM	0.01061	0.01836	0.02440	0.02959	7452.76	6005.85	5532.09	5113.61

outperforms the link prediction performance compared with gRTM, as shown in the left side of Table 4. This is probably because CI-gRTM with the three languages makes use of richer data to estimate the model parameters than the models with two languages.

We further performed Wilcoxon's signed rank testing for the F-measures for the best number of topics for each model: $K = 25$ in Table 3 (left) and $K = 20$ in Table 4 (left), resulting in that our CI-gRM is significantly more effective than all the other baselines at the 0.05 significance level.

From the overall results above, CI-gRTM achieves high performance in both link prediction and word prediction, while the other baseline models only achieve high performance in either link prediction or word prediction.

5. Conclusions

We proposed conditionally independent generalized relational topic models (CI-gRTM) that can predict the relation or link between multiple modes in multimodal data. For instance, the model predicts links across different languages in multilingual parallel/comparable data and also predicts unseen words in each language mode. Our CI-gRTM has advantages of both multimodal topic models [3]–[5] and relational topic models [6], [7]. Our experimental results with two multilingual datasets show that our CI-gRTM has both link prediction ability and word prediction ability with multimodal data, which has not been achieved by a single model in previous studies. For real applicatoins, for instance, our model can discover unknown relationships across multiple languages, such as in a situation, given a known article in a source language, to find unknown but related articles in different languages. This paper covers a start-up study on CI-gRTM, and therefore, comparing with non-generative methods or evaluation under more practical situations are positioned as beyond the scope of this paper and left for our future work. We are also planning to apply our model to other kinds of multimodal data, such as text-annotated image data and video data.

Acknowledgments

This work was partially supported by the Grant-in-Aid for Scientific Research (#23300039 and #15H02703) from JSPS, Japan.

References

[1] T. Hofmann, "Probabilistic latent semantic indexing," Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.50–57, 1999.

[2] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation," J. Machine Learning Research, vol.3, pp.993–1022, 2003.

[3] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed-membership models of scientific publications," Proc. National Academy of Sciences of the United States of America, vol.101, no.Suppl 1, pp.5220–5227, 2004.

[4] D. Newman, C. Chemudugunta, and P. Smyth, "Statistical entity-topic models," Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.680–686, ACM, 2006.

[5] D. Mimno, H.M. Wallach, J. Naradowsky, D.A. Smith, and A. McCallum, "Polylingual topic models," Proc. 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, pp.880–889, ACL, 2009.

[6] J. Chang and D.M. Blei, "Relational topic models for document networks," International Conference on Artificial Intelligence and Statistics, pp.81–88, 2009.

[7] N. Chen, J. Zhu, F. Xia, and B. Zhang, "Generalized relational topic models with data augmentation," Proc. Twenty-Third International Joint Conference on Artificial Intelligence, pp.1273–1279, AAAI Press, 2013.

[8] J. Boyd-Graber and D.M. Blei, "Multilingual topic models for unaligned text," Proc. 25th Conference on Uncertainty in Artificial Intelligence, pp.75–82, 2009.

[9] J. Jagarlamudi and H. Daumé III, "Extracting multilingual topics from unaligned comparable corpora," Advances in Information Retrieval, Lecture Notes in Computer Science, vol.5993, pp.444–456, Springer, 2010.

[10] D. Zhang, Q. Mei, and C. Zhai, "Cross-lingual latent topic extraction," Proc. 48th Annual Meeting of the Association for Computational Linguistics, pp.1128–1137, 2010.

[11] K. Miller, M.I. Jordan, and T.L. Griffiths, "Nonparametric latent feature models for link prediction," Advances in Neural Information Processing Systems, pp.1276–1284, 2009.

[12] J. Zhu, A. Ahmed, and E.P. Xing, "MedLDA: Maximum margin supervised topic models for regression and classification," Proc. 26th Annual International Conference on Machine Learning, pp.1257–1264, ACM, 2009.

[13] T.L. Griffiths and M. Steyvers, "Finding scientific topics," Proc. National Academy of Sciences of the United States of America, vol.101, no.Suppl 1, pp.5228–5235, 2004.

[14] N.G. Polson, J.G. Scott, and J. Windle, "Bayesian inference for logistic models using Pólya–Gamma latent variables," Journal of the American Statistical Association, vol.108, no.504, pp.1339–1349, 2013.

[15] J.P. Callan, W.B. Croft, and S.M. Harding, "The inquery retrieval system," Database and Expert Systems Applications, pp.78–83, Springer, 1992.

Yosuke Sakata received the B.E. and M.E. degrees in computer science from Kobe University, Japan in 2014 and 2016, respectively. He is currently with Sumitomo Life Information Systems Co., Ltd.

Koji Eguchi is an Associate Professor at the Graduate School of System Informatics, Kobe University, Japan. His research interests include information retrieval, statistical machine learning, and data mining.