マージン最大化マルチモーダル関係トピックモデルと多言語間関係予測による評価

Maximum-margin multimodal relational topic models for multilingual relation prediction

坂田洋介† 江口浩二†
Yosuke SAKATA Koji EGUCHI

1. はじめに

複数の表現によるマルチモーダルデータとして，多言語対訳文書データ [1] やテキストアノテーション付き画像データ [2]，テキストデータと引用論文リストで表現された学術文献データ [3] などが挙げられるが，その有望な分析手段の一つに潜在トピックモデルが挙げられる．潜在トピックモデルとは，単語の分布として表現されるトピックの混合分布によって文書を確率的に表現するモデルである．潜在ディリクレ配分法 (Latent Dirichlet allocation: LDA) [4] は広く受け入れられたトピックモデルの 1 つである．

前述のようなマルチモーダルデータを扱える潜在トピックモデルとして Conditionally independent LDA(CI-LDA) [1], [3], [5] が提案されている．CI-LDA は LDA の拡張モデルであり，多言語対訳文書データのモデル化にも使われている．しかし，CI-LDA はマルチモーダルデータに対してモード (例えば言語) を横断して共有される潜在トピックをモデル化することは可能であるが，異なるモード間の関係を直接予測することはできない．

一方で，文書間の未知の関係を予測する研究の一つとして，LDA を拡張した関係トピックモデル (Relational topic models: RTM) [6] がある．文書間の関係として，学術文献の参照・被参照の関係や，ウェブのハイパーリンクが想定されている．文書をノードとし，文書間の関係をリンクとみなせばネットワーク (グラフ) で表現できる．文書間の関係 (リンク) は両文書の潜在トピックの関数 (リンク評価関数) として表される．また，従来の RTM はリンクの有無に関する観測数が不均衡である問題に対処していなかったのに対して，Generalized RTM(gRTM) [7] は正則化パラメータを導入することによってこの問題を解決している．これらのモデルはリンクの予測は可能であるが，マルチモーダルデータの想定を行っていない．そこで本稿では，複数の異なる表現によるマルチモーダルデータに対してモード (例えば言語) 間の関係及びデータ (例えば複数の言語の単語) を同時に予測するために，マルチモーダル関係トピックモデル (Conditionally independent generalized RTM: CI-gRTM) を提案する．各モード間の関係を gRTM のリンク評価関数で表現することによってこのモデルを実現する．CI-gRTM は，CI-LDA や gRTM それぞれ単独では実現できなかった異なるモード間の関係及び各モードに対するデータの予測を両方同時に実現できる．

本稿では，CI-gRTM の評価のため，二言語の対訳文書集合 (平行コーパス) 及び三言語の準対訳文書集合 (比較可能コーパス) を用いて，与えられた日本語表現と英語表現が同一の内容を表したものであるか否かを予測する問題を考える．その際のベースラインとして，gRTM，及び RTM に基づく CI-RTM と比較する．同様に，各言語に対する単語の予測性能について CI-gRTM を評価するために，ベースラインとして CI-LDA，CI-RTM，gRTM と比較する．これらの実験によって，CI-gRTM が他のモデルと比較して精度を落とすことなく関係及び単語の予測を同時に行えることを示す．

† 神戸大学, Kobe University

2. 関連研究

ここでは提案手法に関連した研究として，LDA 及び多モードの潜在トピックを推定するモデルである CI-LDA，文書間の関係を予測するモデルである RTM と gRTM について説明する．

2.1 LDA

潜在的ディリクレ配分法 (Latent Dirichlet allocation: LDA) [4] は代表的なトピックモデルの一つであり，ディリクレ分布を導入することによって文書を潜在トピックの多項分布として表現する．同様に潜在トピックを単語の多項分布として表現する．LDA のグラフィカルモデルを Fig.1 に示す．図中の D, N_d, K がそれぞれ文書数, 文書 d の単語数，トピック数を表している．θ_d, ϕ_k はそれぞれ文書 d に関するトピックの多項分布パラメータ，トピック k に関する単語の多項分布パラメータである．α, β はそれぞれ θ, ϕ に対応するディリクレ事前分布のハイパーパラメータである．また，図中の網掛け部分は観測変数を表している．LDA における文書の生成過程を以下に示す．

（1）D 個の文書に対して多項分布パラメータ $\theta_d \sim \mathrm{Dir}(\alpha)$ を選択する．($d \in \{1, ..., D\}$)
（2）K 個のトピックに対して多項分布パラメータ $\phi_k \sim \mathrm{Dir}(\beta)$ を選択する．($k \in \{1, ..., K\}$)
（3）文書 d の N_d 個の単語 w_i に対し:
　（a）トピック $z_i \sim \mathrm{Mult}(\theta_d)$ を選択する．
　（b）単語 $w_i \sim \mathrm{Mult}(\phi_k)$ を選択する．

ここで Dir はディリクレ分布，Mult は多項分布を表している．周辺化ギブスサンプリング [8] を用いて LDA の推定を行う場合に用いる完全条件付き確率は以下の式の通りである．

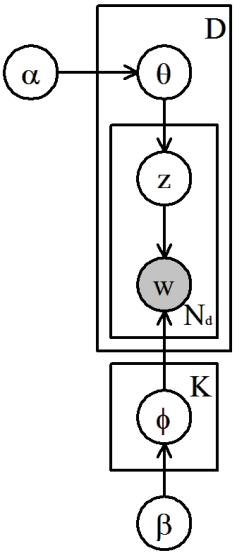

Fig. 1　LDA のグラフィカルモデル

$$q(z_i = k|w_i = v, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \alpha, \beta) \propto \frac{(C^k_{d,-i} + \alpha)(C^v_{k,-i} + \beta)}{\sum_{v'} C^{v'}_{k,-i} + V\beta}$$

ここで，$\mathbf{w} = \{w_i\}$ であり，$\mathbf{w}_{-i}$ は $\mathbf{w}$ から w_i を除いた集合である．同様に，$\mathbf{z} = \{z_i\}$ であり，$\mathbf{z}_{-i}$ は $\mathbf{z}$ から z_i を除いた集合である．$C^k_{d,-i}, C^v_{k,-i}$ はそれぞれ i 番目の単語の割り当てを除いたトピック k が d 番目の文書に割り当てられた回数，トピック k が v 番目の語彙に割り当てられた回数である．V は語彙数を表している．

2.2　CI-LDA

CI-LDA [1], [3], [5] は多言語対訳文書データやテキストアノテーション付き画像データなど，多モードを扱える LDA の拡張モデルである．マルチモーダルデータとして多言語対訳文書データを扱う場合，モードは言語に対応する．L 個の言語を想定した時の CI-LDA のグラフィカルモデルを Fig.2 に示す．上付き文字は言語を示している．CI-LDA における多言語対訳文書の生成過程を以下に示す．以降，対訳関係があるものの中で全ての言語の記述を集めたものを 1 つの文書と定義する．但し，LDA や RTM のような言語を区別しないモデルに関しては，各言語それぞれの記述を 1 つの文書とする．

（1）D 個の文書に対して $\theta_d \sim \mathrm{Dir}(\alpha)$ を選択する．($d \in \{1, ..., D\}$)

（2）L 個の言語及び K 個のトピックに対して $\phi^{(\ell)}_k \sim \mathrm{Dir}(\beta)$ を選択する．($\ell \in \{1, ..., L\}, k \in \{1, ..., K\}$)

（3）文書 d の言語 ℓ の $N^{(\ell)}_d$ 個の単語 $w^{(\ell)}_i$ に対し:

（a）トピック $z^{(\ell)}_i \sim \mathrm{Mult}(\theta_d)$ を選択する．

（b）単語 $w^{(\ell)}_i \sim \mathrm{Mult}(\phi^{(\ell)}_k)$ を選択する．

LDA との違いは，言語の数だけ異なる ϕ を仮定するという点である．但し，異なる言語でも記事が同じものならば θ は共通のパラメータとなる．CI-LDA の周辺化ギブスサンプリングのための完全条件付き確率は以下の式の通りである．

$$q(z^{(\ell)}_i = k|w^{(\ell)}_i = v^{(\ell)}, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \alpha, \beta^{(\ell)}) \propto \frac{(C^k_{d,-i} + \alpha)(C^{v^{(\ell)}}_{k,-i} + \beta^{(\ell)})}{\sum_{v'^{(\ell)}} C^{v'^{(\ell)}}_{k,-i} + V^{(\ell)}\beta^{(\ell)}}$$

ここで，上付き文字は ℓ 番目の言語であることを示している．

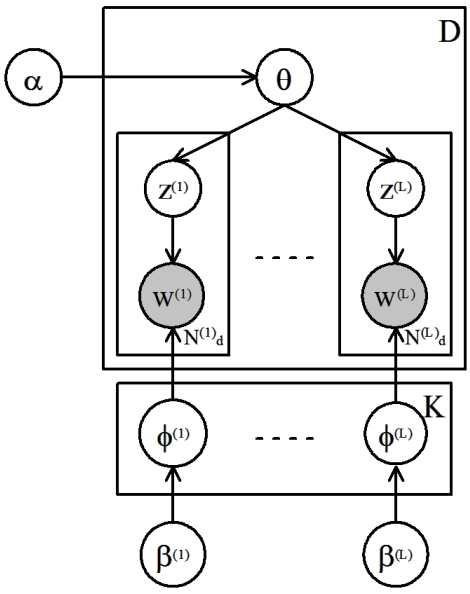

Fig. 2　CI-LDA のグラフィカルモデル

2.3　RTM

関係トピックモデル (Relational topic models: RTM) [6] はテキストコンテンツ及びネットワーク構造 (リンク) の両方を考慮する LDA の拡張モデルである．RTM のグラフィカルモデルを Fig.3 に示す．図中の η はリンク評価時の各トピックに対する係数であり，値が大きい程対応したトピックが重視されることを表している．RTM における文書の生成過程を以下に示す．

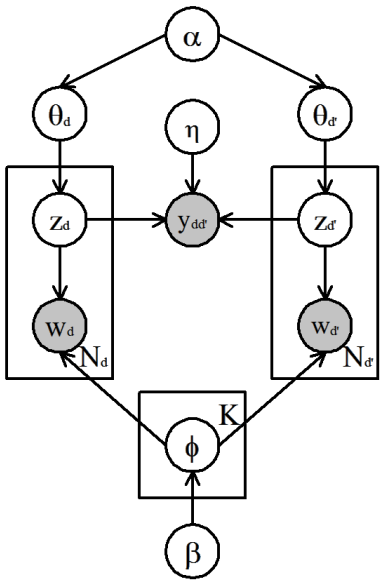

Fig. 3　RTM のグラフィカルモデル

（1）D 個の文書に対して $\theta_d \sim \mathrm{Dir}(\alpha)$ を選択する．($d \in \{1, ..., D\}$)

（2）K 個のトピックに対して $\phi_k \sim \mathrm{Dir}(\beta)$ を選択する．$(k \in \{1, ..., K\})$

（3）文書 d の N_d 個の単語 w_i に対し:

（a）トピック $z_i \sim \mathrm{Mult}(\theta_d)$ を選択する．

（b）単語 $w_i \sim \mathrm{Mult}(\phi_k)$ を選択する．

（4）文書ペア d, d' に対してリンク $y_{dd'} \sim \psi$ を選択する．

関数 ψ はリンクの評価関数で，2 つの文書間のリンクの確率分布を定義している．この関数は各文書のトピックの分布に依存している．ψ の中身は以下のように示される．

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

ここで，$\bar{\mathbf{z}}_d = \frac{1}{N_d} \sum_i z_{d,i}$ であり，$\circ$ はアダマール積，すなわち要素毎の積を表し，σ はシグモイド関数である．[6] では，指数関数や正規分布の累積分布など，σ に代わる他の分布も利用されている．ここでは，一般的に用いられているシグモイド関数を用いたロジスティック尤度モデルを用いる [9]．これは，2 つの文書のトピック分布の類似性を考慮した関数となっている．

2.4 gRTM

Generalized RTM(gRTM) [7] は RTM のリンクの評価関数を拡張したモデルである．RTM では，リンクの評価関数はアダマール積を含んでいるため，同じトピック同士でしか相互作用を与えない．その結果，η の要素が正の値であるものと負の値であるものに分かれてしまう．負の値に対応するトピックはリンクの予測に対して直感的な理解に混乱を招くものとなる．これを解決するために，gRTM では評価関数を以下のように定義している．

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

ここで，U は $K \times K$ の重み行列であり，K はトピック数を表している．c は，リンクがあるペアと無いペアに偏りがある時，それを解決するための正則化パラメータである．[7] では，gRTM が全てのペアのトピックを考慮したモデルとなっているため，結果的に U の全ての対角要素が正の値となることが示されている．

3. マルチモーダルデータに対応した関係トピックモデル

3.1 CI-gRTM

2.2 節でも述べたように，マルチモーダルデータに対してモードを横断するトピックを表現するには CI-LDA を利用すればよい．しかし，CI-LDA は多モード間の関係を直接予測することはできない．この問題を解決するために，我々はマルチモーダル関係トピックモデル (Conditionally independent generalized RTM: CI-gRTM) を提案する．CI-gRTM のグラフィカルモデルを Fig.4 に示す．このモデルでは 2.4 節で述べた gRTM のリンクの評価関数を利用してモードを横断した関係の有無に関する評価を行う．この時，$\bar{\mathbf{z}}_d$ 及び $\bar{\mathbf{z}}_{d'}$ は互いに異なるモードに関するトピック割り当てとなる．また，各モードのトピックを同じトピック番号に対応させるため，CI-LDA と同様に θ は全モードで共通とする．マルチモーダルデータとして多言語対訳文書データを扱う場合，言語数を L 個と仮定した時の CI-gRTM における多言語対訳文書の生成過程を以下に示す．

（1）D 個の文書に対して $\theta_d \sim \mathrm{Dir}(\alpha)$ を選択する．$(d \in \{1, ..., D\})$

（2）L 個の言語のそれぞれ ℓ 及び K 個のトピックに対して $\phi_k^{(\ell)} \sim \mathrm{Dir}(\beta)$ を選択する．$(\ell \in \{1, ..., L\}, k \in \{1, ..., K\})$

（3）文書 d の言語 ℓ の $N_d^{(\ell)}$ 個の単語 $w_i^{(\ell)}$ に対し:

（a）トピック $z_i^{(\ell)} \sim \mathrm{Mult}(\theta_d)$ を選択する．

（b）単語 $w_i^{(\ell)} \sim \mathrm{Mult}(\phi_k^{(\ell)})$ を選択する．

（4）文書 d の言語 ℓ_1 と文書 d' の言語 ℓ_2 のペアに対してリンク $y_{dd'}^{(\ell_1,\ell_2)} \sim \psi$ を選択する．$(\ell_1, \ell_2 \in \{1, ..., L\})$

CI-gRTM では，CI-LDA における多言語対訳文書の生成過程に加えて，リンクの評価関数である ψ によるリンクの予測が追加されている．ψ は以下の式で与えられる．

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

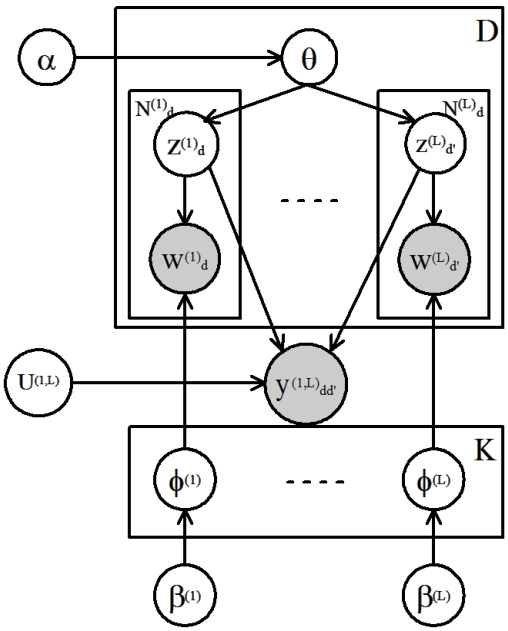

Fig. 4　CI-gRTM のグラフィカルモデル

3.2 周辺化ギブスサンプリングを用いた推定

MedLDA [10] は，変分ベイズ法に加えて最適化手法を導入して潜在変数と未知パラメータの推定を行うモデルである．これは推定時に強い制約を仮定する必要があり，計算が複雑になる問題が発生する．この節では，制約仮定を置かない単純で効率的な周辺化ギブスサンプリング [8] による推定手法について述べる．周辺化ギブスサンプリングのアルゴリズムはデータ拡張 [11] に基づいている．以下では，gRTM の周辺化ギブスサンプリングに関する論文 [7] を参考にして，CI-gRTM の推定手法について述べる．まず，全ての潜在変数と未知パラメータに関する事後分布を以下に示す．

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

ここで，$q(\mathbf{U}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$ は事後分布，$p_0(\mathbf{U}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$ はモデル

によって定義された事前分布，$p(\mathbf{W}|\mathbf{Z},\mathbf{\Phi})$ は単語に関する尤度，$\psi(\mathbf{y}|\mathbf{Z},\mathbf{U})$ はリンクの疑尤度，$\phi(\mathbf{y},\mathbf{W})$ は $q(\mathbf{U},\mathbf{\Theta},\mathbf{Z},\mathbf{\Phi})$ を正規分布にするための正規化定数である．ここからデータ拡張により，リンクの疑尤度 ψ を以下の式のように変形する [7], [11].

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

$\kappa_{dd'}^{(\ell_1,\ell_2)} = c(y_{dd'}^{(\ell_1,\ell_2)} - 1/2)$，$\omega_{dd'}^{(\ell_1,\ell_2)} = \bar{\mathbf{z}}_d^{(\ell_1)\mathrm{T}} U^{(\ell_1,\ell_2)} \bar{\mathbf{z}}_{d'}^{(\ell_2)}$ であり，上式の λ が新たに拡張されたデータ拡張変数となる．c が大きいほど誤分類を許容した推定となる．また，$p(\lambda_{dd'}^{(\ell_1,\ell_2)}|c,0)$ は Polya-Gamma 分布 [11] に従い，以下の式を取る．

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

ここで，g_i はガンマ分布 $\mathcal{G}(a,1)$ に従う．これにより，λ を含んだ周辺事後分布は以下のようになる．

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

ここで $\psi(\mathbf{y},\mathbf{\lambda}|\mathbf{Z},\mathbf{U}) = \prod_{dd'}\exp(\kappa_{dd'}^{(\ell_1,\ell_2)}\omega_{dd'}^{(\ell_1,\ell_2)} - \frac{\lambda_{dd'}^{(\ell_1,\ell_2)}\omega_{dd'}^{(\ell_1,\ell_2)2}}{2})$ $\times p(\lambda_{dd'}^{(\ell_1,\ell_2)}|c,0)$ であり，この式は $\mathbf{y}$ 及び $\mathbf{\lambda}$ の疑同時確率分布である．

更に，ここから $(\mathbf{\Theta},\mathbf{\Phi})$ を周辺化し，マルコフ連鎖を構築する必要がある．CI-gRTM では，周辺事後分布は以下の式で表される．

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

ここで，$\delta(\mathbf{x}) = \frac{\prod_{i=1}^{\dim(\mathbf{x})}\Gamma(x_i)}{\Gamma(\prod_{i=1}^{\dim(\mathbf{x})}x_i)}$ であり，$\mathbf{C}_k = \{C_k^v\}_{v=1}^V$ であり，C_k^v は v 番目の語彙にトピック k が割り当てられた回数を表している．同様に，$\mathbf{C}_d = \{C_d^k\}_{k=1}^K$ であり，C_d^k は d 番目の文書内でトピック k が割り当てられた回数を表している．

以下では，周辺化ギブスサンプリングで用いる各パラメータの完全条件付き確率を示す．

重み行列 $\mathbf{U}$ について:

$\bar{\mathbf{z}}_{dd'}^{(\ell_1,\ell_2)} = \mathrm{vec}(\bar{\mathbf{z}}_d^{(\ell_1)}\bar{\mathbf{z}}_{d'}^{(\ell_2)\mathrm{T}})$ とし，$\boldsymbol{\eta}^{(\ell_1,\ell_2)} = \mathrm{vec}(U^{(\ell_1,\ell_2)})$ とする．ここで，$\mathrm{vec}(A)$ とは A の列ベクトルを連接したものである．この時，$\omega_{dd'}^{(\ell_1,\ell_2)} = \boldsymbol{\eta}^{(\ell_1,\ell_2)\mathrm{T}}\bar{\mathbf{z}}_{dd'}^{(\ell_1,\ell_2)}$ となる．また，$U^{(\ell_1,\ell_2)}$ のガウス事前分布を $p_0(U^{(\ell_1,\ell_2)}) = \prod_{kk'}\mathcal{N}(U_{kk';}^{(\ell_1,\ell_2)}0,\nu^2)$ と仮定する時，以下の式が成立する．

$$q(\boldsymbol{\eta}^{(\ell_1,\ell_2)}|\mathbf{Z},\mathbf{\lambda}) \propto p_0(\boldsymbol{\eta}^{(\ell_1,\ell_2)})\times \prod_{dd'}\exp\left(\kappa_{dd'}^{(\ell_1,\ell_2)}\boldsymbol{\eta}^{(\ell_1,\ell_2)\mathrm{T}}\bar{\mathbf{z}}_{dd'}^{(\ell_1,\ell_2)} - \frac{\lambda_{dd'}^{(\ell_1,\ell_2)}(\boldsymbol{\eta}^{(\ell_1,\ell_2)\mathrm{T}}\bar{\mathbf{z}}_{dd'}^{(\ell_1,\ell_2)})^2}{2}\right) = \mathcal{N}(\boldsymbol{\eta}^{(\ell_1,\ell_2)};\boldsymbol{\mu},\mathbf{\Sigma})$$

$\boldsymbol{\mu} = \mathbf{\Sigma}\big(\sum_{dd'}\kappa_{dd'}^{(1,2)}\bar{\mathbf{z}}_{dd'}^{(1,2)}\big)$ は事後平均であり，$\mathbf{\Sigma} = \big(\frac{1}{\nu^2}I + \sum_{dd'}\lambda_{dd'}^{(1,2)}\bar{\mathbf{z}}_{dd'}^{(1,2)}\bar{\mathbf{z}}_{dd'}^{(1,2)\mathrm{T}}\big)^{-1}$ である．この K^2 次元の多変量正規分布から $U^{(\ell_1,\ell_2)}$ の各要素をサンプリングすることが可能となる．全ての言語ペアに対してこの操作を行うことにより，重み行列 $\mathbf{U}$ の推定が可能となる．

トピック割り当て $\mathbf{Z}$ について:

$\mathbf{Z}$ の完全条件付き確率は以下の通りである．

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

ここで，$\mathcal{N}_d = \{d' : (d,d') \in \mathcal{I}\}$ は文書 d に関するリンク先の集合，$\mathcal{I}$ はリンクの集合を表しており，$\psi(y_{dd'}^{(\ell,\ell')}|\mathbf{\lambda},\mathbf{Z}) = \exp(\kappa_{dd'}^{(\ell,\ell')}\omega_{dd'}^{(\ell,\ell')} - \frac{\lambda_{dd'}^{(\ell,\ell')}\omega_{dd'}^{\ell,\ell')2}}{2})$ である．また，$L_{-\ell}$ は言語 $1,...,L$ の内，言語 ℓ を除いた集合である．上式より，初項が LDA モデルの単語に関する完全条件付き確率に比例し，第 2 項がリンク構造 $\mathbf{y}$ を表していることが分かる．

データ拡張変数 $\boldsymbol{\lambda}$ について:

最後に，データ拡張変数 $\boldsymbol{\lambda}$ の完全条件付き確率は以下の通りである．

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

上式から分かるように，$\boldsymbol{\lambda}$ の完全条件付き確率は Polya-Gamma 分布に従う．ここで注意しなければならないのは，データ拡張変数 $\boldsymbol{\lambda}$ は言語ごとに異なる値となることである．例えば，J を日本語，E を英語と仮定した時の $\lambda_{dd'}^{(J,E)}$ と $\lambda_{dd'}^{(E,J)}$ は異なるものとなる．全ての言語ペアに対してこの操作を行うことにより，データ拡張変数 $\boldsymbol{\lambda}$ の推定が可能である．

上記の 3 つの完全条件付き確率を反復させながら各パラメータを推定することによって，マルコフ連鎖を構築することができる．周辺化ギブスサンプリングではこれらのパラメータの推定が収束するまで繰り返される．

4. 実　　験

この章では，実験によって，提案手法である CI-gRTM で用いられる正則化パラメータ c の最適な値を導出する．そして，2 つのデータセットに対して，リンク予測性能に関して CI-RTM 及び gRTM と比較し，更に，単語予測性能に関して CI-LDA, CI-RTM, gRTM, LDA と比較し，それぞれの結果について考察する．

4.1 データセット

実験において，2 つのデータセットを用いた．1 つ目のデータセットとして日英京都関連文書対訳コーパス (注1) を使用した (以下データセット A とする)．これは，京都の人物や建造物な

(注1)：http://alaginrc.nict.go.jp/WikiCorpus/

どを対象とした合計 14111 もの Wikipedia の記事データであり，日本語の Wikipedia 記事とその英語翻訳からなる二言語の記事が用意されている．前処理として，日本語英語共に全記事中で 5 記事以下しか出現しない低頻度語 [8] を除外した．また，日本語記事では MeCab(注2) を用いて形態素解析を行い，記号・接続詞などの機能語を除外し，英語記事ではストップワード [12] を除外した．前処理後のデータセット A の情報を Table 1 に示す．本実験では，日本語-英語間の対訳記事をリンクと仮定した．

Table 1　前処理後のデータセット A

	日本語	英語
文書数	14111	
単語数	2983135	4338115
語彙数	23979	34398

2 つ目のデータセットとして過去の Wikipedia の記事データ (注3)(日本語，英語，スペイン語の三言語に言語間リンクによる対応関係のある記事) を使用した (以下データセット B とする)．データセット A が翻訳関係のある (二言語) 平行コーパスであるのに対して，データセット B は厳密な翻訳ではないものの主題レベルで対応関係があるような (三言語) 比較可能コーパスであると言える．

全ての記事データを扱おうとすると記事数が異常に膨大なものとなったため，本実験では，対訳関係のある記事のうち，英語記事のタイトルが”A”で始まる記事のみを使用した．データセット A と条件を合わせるため，上記と同じ前処理を行った．更に，スペイン語記事ではストップワード (注4) を除外した．前処理後のデータセット B の情報を Table 2 に示す．本実験では日本語-英語間，英語-スペイン語間，日本語-スペイン語間の対応記事をリンクと仮定した．

Table 2　前処理後のデータセット B

	日本語	英語	スペイン語
文書数	5818		
単語数	1827463	5062266	2507643
語彙数	21618	47650	33721

4.2　正則化パラメータ c の検証実験

ここでは，提案手法である CI-gRTM，推定及び予測時に重み行列 U の対角成分のみを使用する CI-RTM，CI-gRTM のベースとなっている gRTM の 3 つのモデルで使用する正則化パラメータ c の最適な値を求めるための検証実験を行った．検証実験ではデータセット A のみを用いた．データセット A に関して，14111 文書を文書単位でランダムに 5 分割した．その内の 1 セットは予め 4.4 節のテスト実験のためのテスト文書として確保し，残りの 4 セットで 4 分割交差検定を行った．交差検定時の検証に用いる文書セットを検証文書，訓練時に用いる文書セットを訓練文書と呼ぶこととする．訓練文書に対して，ギブスサンプリングを用いて未知パラメータの推定を行った．この時，データセットの仕様として，リンクがある (正例) ペアに対してリンクが無い (負例) ペアの方が圧倒的に多い. [7] では，このような大きな偏りがあるデータでそのまま推定を行うと，正例の影響が負例に埋もれてしまうため，文書数の割合が正例文書数 1 に対し負例文書数 2〜10 となるように実験を行っている．本稿では，訓練時は正例と負例の割合が 1:2 になるように負例文書をランダムに抽出した．訓練時に推定した各言語に対するトピック-単語分布を用いて，検証文書に関する文書-トピック分布を再推定した．

そして，検証文書に対してリンクの予測性能を評価するために，リンクの評価関数を用いて検証文書の全ペアについてリンクの有無を判別し，リンクに対する F 値を求めた．F 値とは，予測結果の評価指標の 1 つであり，再現率と適合率の調和平均を取ったものである．また同様に，検証文書に対して単語の予測性能を評価するために，検証文書のパープレキシティを求めた．パープレキシティは尤度の幾何平均の逆数であり，尤度は以下の式から導出される．尤度の幾何平均は 1 単語ごとの平均を取る．ここでの尤度はモデルからテスト文書 (又は検証文書) 中の単語が生成される確率を指す．尤度は 0〜1 を取り，高いほど汎化能力すなわち新たなデータに対する予測能力が高いことを表している．そのため，パープレキシティは 1 以上の値を取り，1 に近いほど，すなわち値が小さいほど予測能力が高いことを表している．

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

F 値及びパープレキシティの 2 つの評価指標について，正則化パラメータ c を変化させた時のそれぞれの結果を測定した．c は負例に対しては 1 で固定し，正例に対しては {1,2,4,8,16} の 5 通りで変化させて交差検定で決定した．対称ディリクレハイパーパラメータについては，$\alpha = 0.1, \beta^{(J)} = \beta^{(E)} = 0.01$ に設定した．なお，$\beta^{(J)}, \beta^{(E)}$ はそれぞれ日本語，英語に対する β である．多言語文書データに対応していない gRTM は β を言語ごとに区別することができないため，両言語に対して $\beta = 0.01$ とした．トピック数は {5,10,15} の 3 通りでそれぞれ変化させて実験を行った．初期設定として，トピックの割り当ては全てランダムに選択し，また λ の各要素は全て 1 とした．ギブスサンプリングにおける収束条件は，テストセット対数尤度を 10 回ごとに測定し，その変化率が 0.1%以下に収まった時とした．

4.3　検証実験の実験結果及び考察

CI-gRTM，CI-RTM，gRTM の 3 つのモデルを用いて，正則化パラメータ c をそれぞれ変化させた時の F 値の測定結果を Fig.5，Fig.6，Fig.7 に示す．Fig.5，Fig.6，Fig.7 でそれぞれトピック数は 5，10，15 となっている．横軸は，正則化パラメータ c である．

全てのトピック数において，CI-RTM よりも CI-gRTM の方が F 値が高いことが分かる．これは，多言語対訳文書のようなマルチモーダルデータに対しても，全トピックペアを考慮する

(注2) : http://mecab.googlecode.com/svn/trunk/mecab/doc/
(注3) : http://dumps.wikimedia.org/
(注4) : http://members.unine.ch/jacques.savoy/clef/spanishSmart.txt

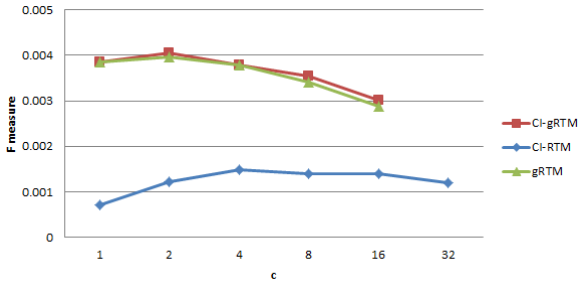

Fig. 5　トピック数 5 における F 値

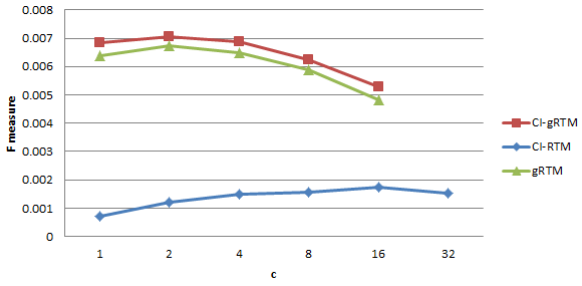

Fig. 6　トピック数 10 における F 値

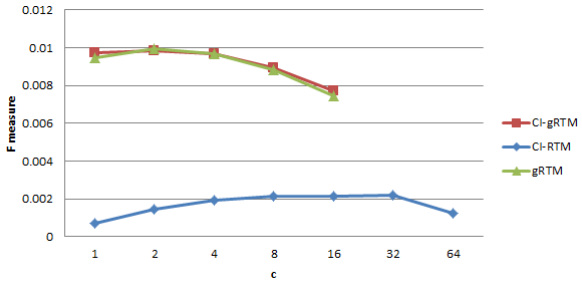

Fig. 7　トピック数 15 における F 値

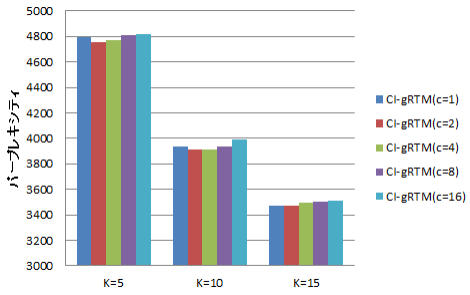

Fig. 8　CI-gRTM におけるパープレキシティ

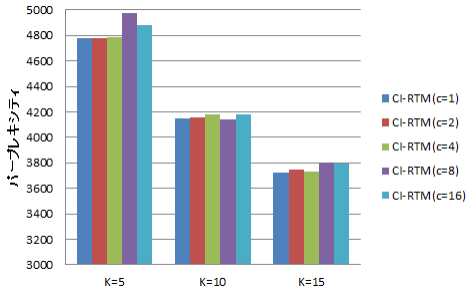

Fig. 9　CI-RTM におけるパープレキシティ

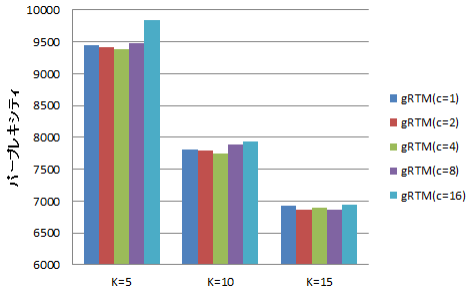

Fig. 10　gRTM におけるパープレキシティ

gRTM のリンクの予測関数の方がより良い予測が可能であるからと考えられる．また，gRTM と CI-gRTM の F 値はどのトピック数においてもほぼ同じ値となっていることが分かる．これは，CI-gRTM が gRTM のリンクの評価関数を利用したモデルになっているために，リンク予測において gRTM と同等な結果になったと考えられる．また，gRTM と CI-gRTM に関しては，全てのトピック数において $c = 2$ で最も F 値が高くなっている一方で，CI-RTM に関しては，いずれのトピック数でも $c = 16$ で安定して F 値が高くなっている．これらの結果の違いは，gRTM 及び CI-gRTM は全トピックペアをリンクの評価の計算に用いているため，トピック間の相互作用が大きくなり，大きな c では過学習を引き起こすため，最適な c は比較的小さな値となっていると考えられる．また，全体的に絶対的な F 値が低くなってしまっている．これは，今回の実験ではデータセットとして対訳文書を用いていることにより，正例は 1 文書につき 1 つしか存在しない．そのため，類似した内容であったとしても正解とせずに評価しているからと考えられる．

次に，3 つのモデルに対するパープレキシティの測定結果を Fig.8，Fig.9，Fig.10 に示す．Fig.8 が CI-gRTM，Fig.9 が CI-RTM，Fig.10 が gRTM のパープレキシティとなっている．横軸はトピック数を表しており，図の左側から順にトピック数は 5，10，15 となっている．縦軸はパープレキシティを表しており，値が小さい程良い性能であることを示している．

Fig.8，Fig.9，Fig.10 より，全てのトピック及び c において gRTM のパープレキシティの方が最も大きな値となっていることが分かる．これは，gRTM は言語を区別できないモデルであるため，1 つの記事には 1 つの言語の語彙しか現れないにもかかわらず，両方の言語の語彙を考慮する必要があり，考慮する語彙数が他のモデルに比べて多いからと考えられる．また，Fig.8 より，CI-gRTM では全てのトピックにおいて $c = 2$ の時に最もパープレキシティが小さくなっていることが分かる．Fig.9 より，多少ばらつきがあるものの，CI-RTM では $c = 1$ の時に最もパープレキシティが小さくなっていることが分かる．Fig.10 より，gRTM では $c = 2, 4$ の時にほぼ同じ値でパープレキシティが小さくなっていることが分かる．

リンク及び単語の 2 種類の評価より，CI-gRTM では $c = 2$，CI-RTM ではリンク予測においては $c = 16$，単語予測においては $c = 1$，gRTM では $c = 2$ が最適値であると言える．次節

のテスト実験では上記の c の条件で実験を行う．

4.4 テスト実験による既存手法との比較

ここでは，検証実験で導出した c を利用して，CI-gRTM がリンク及び各言語の単語を如何なる精度で予測できるかに関して既存手法とのテスト実験を行った．リンク予測の比較モデルは CI-RTM 及び gRTM，単語予測の比較モデルは CI-LDA，CI-RTM，gRTM を用いた．テスト実験では，より一般性を失わないようにするため，2 つのデータセットをそれぞれ用いた．データセット A については，4.2 節の交差検定時に用いた 4 セットを用いて各モデルの潜在変数と未知パラメータを推定し，テスト文書を用いて予測を行った．データセット B については，データセット A と同様に，全体の文書の 80%を用いて各モデルの潜在変数と未知パラメータを推定し，残りの 20%を用いて予測を行った．推定法，正例文書に対する負例文書の割合，再推定，評価指標，ハイパーパラメータ，初期設定，収束条件は，いずれのデータセットに対しても 4.2 節と全く同じ過程で実験を行った．トピック数は，データセット A に関しては {10,15,20,25}，データセット B に関しては {5,10,15,20} でそれぞれ変化させて実験を行った．比較モデルに関しても CI-gRTM と同様の手順で実験を行った．そして，この実験では CI-gRTM と既存モデルとの比較が目的であるため，2 つの評価指標に関して，既存モデルから見た CI-gRTM の改善率を測定した．

4.5 テスト実験の実験結果及び考察

4.3 節で導出した各モデルに対する正則化パラメータ c を用いてリンク予測における各モデルの性能比較を行った．データセット A に対する，CI-gRTM，CI-RTM，gRTM の 3 つのモデルの F 値及び CI-gRTM の改善率を Table 3 に示す．K はトピック数，パーセンテージは改善率を表している．ここでの CI-RTM の c の値は 4.3 節の 4 分割交差検定におけるリンク予測性能に基づいて 16 に設定した．Table 3 より，全てのモデルでトピック数が上昇するに連れて F 値も上昇しているが，その中でも CI-RTM は他のモデルに比べて低い値となっていることが分かる．これは，4.3 節でも述べたように，全トピックペアを考慮する gRTM がより正確なリンク予測を可能にしているからと考えられる．

次に，単語予測における比較を行った．CI-gRTM，CI-LDA，CI-RTM，gRTM，LDA の 5 つのモデルに対するパープレキシティ及び CI-gRTM の改善率を Table 4 に示す．K はトピック数，パーセンテージは CI-gRTM の改善率を表している．パープレキシティは小さい方が良いため，改善率は負である方がより改善されたということを表している．ここでの CI-RTM の c の値は 4.3 節の 4 分割交差検定における単語予測性能に基づいて 1 に設定した．Table 4 より，gRTM と LDA が他に比べて明らかに大きな値となることが分かる．4.3 節の考察同様，gRTM と LDA は日本語と英語を区分することができないモデルであるため，考慮する語彙数が他のモデルよりも多くなった分パープレキシティも大きくなったと考えられる．

Table 3 及び Table 4 より，まず CI-LDA と比較すると，パープレキシティは同等か僅かに (+1%程度) 性能が落ちたが，CI-LDA はリンク予測ができないモデルであるため，CI-gRTM と異なり F 値は測定できない．一方，言語を区別しないモデル同士である LDA と gRTM を比較すると，Table 5 より，パープレキシティは gRTM の方が若干大きくなることが分かる．これらの結果より，gRTM は言語を区別するかどうか (gRTM か CI-gRTM かどうか) にかかわらず，リンク予測を行う代わりに単語予測性能を条件次第では若干犠牲にするモデルであると言える．そのため，CI-LDA と CI-gRTM の単語予測性能の差は gRTM の性質から生じているものと考えられる．次に CI-RTM と比較すると，パープレキシティは同等かやや改善され (-1%程度)，多少のばらつきはあるが F 値は約+250%の改善が見られた．gRTM と比較すると，F 値はトピック数 15 以外で改善が見られ，パープレキシティも約-50%の改善が見られた．

次に，データセット B に対してテスト実験を行った．Table 3 及び Table 4 より，明らかに CI-gRTM よりも性能が劣化したモデルを省略し，比較モデルを CI-LDA 及び gRTM の 2 種類に限定した．まず，CI-gRTM，gRTM の 2 つのモデルの F 値及び CI-gRTM の改善率を Table 6 に示す．ここでの c の値はデータセット A での実験と同じ値を使用している．Table 6 より，いずれのトピックに関しても CI-gRTM が大きく改善されており，Table 3 と比較しても改善率が更に上昇していることが分かる．これは，言語数を二言語から三言語に拡張したことにより，各言語を区別できない gRTM が二言語の時よりも

Table 3　リンク予測における CI-gRTM の改善率

	K=10	K=15	K=20	K=25
CI-LDA	—	—	—	—
CI-gRTM	0.007076	0.009784	0.012152	0.015282
CI-RTM	0.001980	0.002268	0.003271	0.004464
	⇓ **+257.4%**	⇓ **+331.4%**	⇓ **+271.4%**	⇓ **+242.4%**
CI-gRTM	0.007076	0.009784	0.012152	0.015282
gRTM	0.006267	0.009990	0.012020	0.014311
	⇓ **+12.90%**	⇓ **-2.060%**	⇓ **+1.094%**	⇓ **+6.785%**
CI-gRTM	0.007076	0.009784	0.012152	0.015282
LDA	—	—	—	—
CI-gRTM	0.007076	0.009784	0.012152	0.015282

Table 4　単語予測における CI-gRTM の改善率

	K=10	K=15	K=20	K=25
CI-LDA	3695.20	3412.06	3056.85	2877.88
	⇓ **-0.912%**	⇓ **+1.103%**	⇓ **+1.237%**	⇓ **+0.621%**
CI-gRTM	3661.51	3449.69	3094.65	2895.75
CI-RTM	3671.30	3568.07	3212.77	2931.03
	⇓ **-0.267%**	⇓ **-3.318%**	⇓ **-3.677%**	⇓ **-1.204%**
CI-gRTM	3661.51	3449.69	3094.65	2895.75
gRTM	7815.39	6532.47	6092.29	5683.37
	⇓ **-53.150%**	⇓ **-47.191%**	⇓ **-49.204%**	⇓ **-49.049%**
CI-gRTM	3661.51	3449.69	3094.65	2895.75
LDA	7250.17	6516.22	5981.70	5647.45
	⇓ **-49.498%**	⇓ **-47.060%**	⇓ **-48.265%**	⇓ **-48.725%**
CI-gRTM	3661.51	3449.69	3094.65	2895.75

Table 5　単語予測による gRTM の改善率

	K=10	K=15	K=20	K=25
LDA	7250.17	6516.22	5981.70	5647.45
	⇓ **+7.796%**	⇓ **+0.249%**	⇓ **+1.849%**	⇓ **+0.636%**
gRTM	7815.39	6532.47	6092.29	5683.37

Table 6　リンク予測における CI-gRTM の改善率

	K=5	K=10	K=15	K=20
CI-LDA	—	—	—	—
CI-gRTM	0.01061	0.018357	0.024395	0.029585
gRTM	0.006944	0.010497	0.020461	0.020542
	⇓ **+52.80%**	⇓ **+74.89%**	⇓ **+19.23%**	⇓ **+44.03%**
CI-gRTM	0.01061	0.018357	0.024395	0.029585

Table 7　単語予測における CI-gRTM の改善率

	K=5	K=10	K=15	K=20
CI-LDA	7059.01	6124.24	5453.43	5112.33
	⇓ **+5.578%**	⇓ **-1.933%**	⇓ **+1.442%**	⇓ **+0.025%**
CI-gRTM	7452.76	6005.85	5532.09	5113.61
gRTM	22077.2	14586.8	11970.9	10546.7
	⇓ **-66.242%**	⇓ **-58.827%**	⇓ **-53.787%**	⇓ **-51.515%**
CI-gRTM	7452.76	6005.85	5532.09	5113.61

悪い結果を生成していると考えられる．

次に，CI-gRTM，gRTM，CI-LDA の 3 つのモデルのパープレキシティ及び CI-gRTM の改善率を Table 7 に示す．ここでの c の値はデータセット A での実験と同じ値を使用している．Table 7 より，gRTM が他のモデルに比べて明らかに大きな値となっていることが分かる．これは，データセット A の時と同様に，gRTM は考慮する語彙数が多くなった分パープレキシティも大きくなったと考えられる．

以上より，CI-gRTM はリンク予測，単語予測それぞれに特化した各既存モデルに対してその性能を概ね維持しつつ，既存モデルではできなかったリンク及び単語の予測を同時に行えるモデルであると言える．

5. おわりに

本稿では，マルチモーダルデータに対するモード間の関係の予測を行うモデルとして CI-gRTM を提案し，既存のモデルではできなかった多モードのデータ予測とモードを横断した関係予測を同時に実現した．更に，CI-gRTM が既存のモデルである gRTM に匹敵するリンク予測性能を持ち，同じく既存のモデルである CI-LDA に匹敵する単語予測性能を持つことを示した．また，マルチモーダルデータである多言語対訳文書データに対して，CI-gRTM が CI-RTM よりもリンク及び単語の予測において有効なモデルであることを示した．また，gRTM と LDA による比較では，gRTM の方が単語予測性能が僅かに劣ることが分かり，言語を区別するかどうかにかかわらず，gRTM(CI-gRTM) は単語予測性能を僅かに犠牲にするもののリンク予測を可能とするモデルであることが分かった．

今後の課題として，他のマルチモーダルデータとして，テキストアノテーション付き画像データなどへの適用が考えられる．これは，bag-of-visual-words を用いて画像データを局所特徴量で表現し，この特徴量をテキストデータの単語とみなしてモード間の関係を定義することによって実現可能となる．

謝　辞

本研究の一部は，科学研究費補助金基盤研究 (B)（23300039）の援助による．

文　献

[1] David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pp. 880–889. Association for Computational Linguistics, 2009.

[2] David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 127–134. ACM, 2003.

[3] Elena Erosheva, Stephen Fienberg, and John Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, No. Suppl 1, pp. 5220–5227, 2004.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, Vol. 3, pp. 993–1022, 2003.

[5] David Newman, Chaitanya Chemudugunta, and Padhraic Smyth. Statistical entity-topic models. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 680–686. ACM, 2006.

[6] Jonathan Chang and David M Blei. Relational topic models for document networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 81–88, 2009.

[7] Ning Chen, Jun Zhu, Fei Xia, and Bo Zhang. Generalized relational topic models with data augmentation. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pp. 1273–1279. AAAI Press, 2013.

[8] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, Vol. 101, No. Suppl 1, pp. 5228–5235, 2004.

[9] Kurt Miller, Michael I Jordan, and Thomas L Griffiths. Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems*, pp. 1276–1284, 2009.

[10] Jun Zhu, Amr Ahmed, and Eric P Xing. Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1257–1264. ACM, 2009.

[11] Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using polya-gamma latent variables. *Journal of the American Statistical Association*, No. just-accepted, 2013.

[12] James P Callan, W Bruce Croft, and Stephen M Harding. The inquery retrieval system. In *Database and expert systems applications*, pp. 78–83. Springer, 1992.