Vol.2011-NL-201 No.4
Vol.2011-SLP-86 No.4
2011/5/16

多言語トピックモデルによる言語横断リンク検出

福 増 康 佑[†1] 松 浦 愛 美[†1,*1] 江 口 浩 二[†1]

トピックモデルは大規模なテキストデータコレクションの解析に広く使用されているアプローチである．最近，Wikipedia を典型とする並列または比較可能な多言語データにおいて潜在トピックを発見する多言語トピックモデルが研究されている．また，元々は内部構造を持つ文書を対象として開発されたトピックモデルのうち，多言語の文書にも適用可能なものがある．しかしながら，現在まで多言語トピックモデルの比較評価を行った報告は我々の知る限りない．我々は多言語文書データに適用可能ないくつかのトピックモデルの性能を，テストセット対数尤度，トピック割り当てのヒストグラム，そして言語横断ストーリーリンク検出タスクに着目して比較評価した．実験により，これまで多言語に関連した研究に用いられてこなかったトピックモデルのいくつかが，従来研究で用いられた多言語トピックモデルより優れていることを示した．

Cross-Lingual Link Detection using Multilingual Topic Models

Kosuke Fukumasu,[†1] Manami Matsuura[†1,*1]
and Koji Eguchi[†1]

Topic modeling is a widely-used approach to analyze large text collections. Recently a few number of multilingual topic models have been explored to discover latent topics among parallel or comparable documents, such as Wikipedia. Moreover, there are some other topic models that were originally proposed for documents with structure and are also applicable for multilingual documents. However, no comparative studies have been reported for the purpose of multilingual topic modeling, to our knowledge. We compared the performance of various topic models that can be applied to multi-language documents in terms of test-set log-likelihood, histograms of topic assignments, and also in the task of cross-lingual story link detection. We demonstrated through the experiments that several topic models that have not ever used for multilingual context work better than the other multilingual topic models that were used in prior work.

1. はじめに

トピックモデルは大規模なテキストデータコレクションの有用な解析手法として広く知られている．トピックモデルでは，各文書はトピックの混合分布として表現され，各トピックは単語分布で表現される．トピックモデルとしてよく知られているのは，確率的潜在意味解析法 (Probabilistic Latent Semantic Indexing: PLSI)[1] と潜在的ディリクレ配分法 (Latent Dirichlet Allocation: LDA)[2] であり，目的や対象データの特徴に応じてこれらを拡張したトピックモデルが提案されている．それらのトピックモデルのほとんどは，テキストが単一言語であることを想定しているが，一部のトピックモデルではテキスト表現が複数種のクラスで構造化されていることを想定し，クラス間の統計的依存性を捉えることができる．これらのトピックモデルは，並列または比較可能文書に対して使用可能なトピックモデルであると言える．その典型的な例の一つに CI-LDA[3] を挙げることができ，最近これと等価な多言語トピックモデルに関して報告されている[4),5)]．

ところで，インターネット百科事典の Wikipedia は多言語の比較可能文書の大規模コレクションである．Wikipedia では，特定の項目に関する各記事が 250 以上の言語で執筆されている．そのような記事は複数言語間で同じトピックを持つと考えることができる．

本論文では，多言語テキストデータに適用可能なトピックモデルの性能を比較評価する．我々は，CI-LDA[3)–6)]，SwitchLDA[3)]，CorrLDA[7)] と LDA について，Wikipedia の日本語と英語の比較可能記事を用いて実験し，内的評価と外的評価の観点からそれらのモデルを評価する．とくに，与えられた記事に対応する他の言語で記述された記事を，多言語トピックモデルを用いて発見するタスクに着目する．これはストーリーリンク検出[8)] を言語を横断して実現するタスクといえる．実験により，多言語テキストデータの研究にこれまで用いられてこなかった CorrLDA が，従来研究で多言語トピックモデルとして用いられた CI-LDA より優れていることを示す．

2. モデル

この節では，複数のクラスを扱うことができ，多言語テキストに適用できる LDA 型のト

†1 神戸大学
Kobe University
*1 現在，東京大学
Presently with the University of Tokyo

Vol.2011-NL-201 No.4
Vol.2011-SLP-86 No.4
2011/5/16

ピックモデルを紹介する．

Mimno ら[4] と Ni ら[5] はそれぞれ，CI-LDA[3),6)] と等価なモデルを多言語テキストデータから複数言語にまたがって共通のトピックを発見する目的で使用し，様々な方法で評価した．Cohn と Hofmann[6] は PLSI の拡張として学術文献の各々に出現する単語と引用を同時にモデル化するトピックモデルを提案した．また，LDA 型のトピックモデルである CI-LDA[3] も同様の目的で提案された．SwitchLDA[3] は元々は固有名詞のアノテーションが付与された文書を対象とした LDA 型のトピックモデルとして提案された．さらに，CorrLDA[7] は元々はテキストアノテーションが付与された画像データを対象とし，単語と画像特徴量を同時にモデル化するトピックモデルとして提案された．これらのモデルは多言語の比較可能文書にも適用可能であるが，複数の多言語トピックモデルに関する比較評価を行った報告は我々の知る限りない．本論文では，多言語の比較可能文書に対する CI-LDA，SwitchLDA，CorrLDA の性能をテストセット対数尤度，トピック割り当てのヒストグラム，そして言語横断ストーリーリンク検出タスクの観点から比較評価する．

2.1 LDA

複数のクラスを扱うことができるトピックモデルを紹介する前に，これらのモデルのベースになっている LDA について簡単に述べる．

潜在的ディリクレ配分法 (Latent Dirichlet Allocation: LDA)[2] は，各文書が潜在トピックの混合分布から生成されているとするモデルである．図 1 は LDA のグラフィカルモデルを示している．グラフィカルモデルでは，確率変数およびパラメータは頂点，それらの依存関係は有向辺で表現される．網掛けの頂点は顕在変数，他の頂点は潜在変数または未知パラメータを示している．矩形は角に記された数だけ矩形内の変数の生成が繰り返されることを示している．D は文書数，T はトピック数，N_d は文書 d 内の総延べ語数 (トークン数) を示している．θ と ϕ はそれぞれ文書・トピック多項分布パラメータとトピック・単語多項分布パラメータである．α と β はそれぞれ θ と ϕ のディリクレハイパーパラメータである．以下にグラフィカルモデルに従った文書生成過程を示す．

(1) 全て文書 d に対して，$\theta_d \sim Dir(\alpha)$ を選択する．
(2) 全てトピック t に対して，$\phi_t \sim Dir(\beta)$ を選択する．
(3) 文書 d の N_d 個の各単語 w_i に対して:
- トピック $z_i \sim Mult(\theta_d)$ を選択する．
- 単語 $w_i \sim Mult(\phi_{z_i})$ を選択する．

本論文では，トピックモデルの推定にギブスサンプリング[9] を用いる．LDA のギブスサンプリングで用いる完全条件付確率（すなわち文書 d の i 番目の単語に割り当てられるトピックを除いてすべてが観測されているという仮定の上で，そのトピックが t である確率）は，以下の式で与えられる．

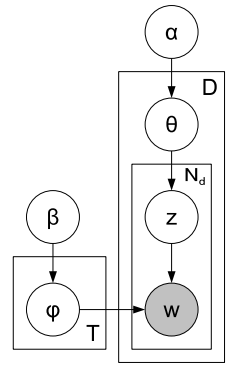


図 1 LDA のグラフィカルモデル

$$p(\mathbf{z}_i = t|\mathbf{w}_i = w, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \alpha, \beta) \propto \frac{C^{TD}_{td,-i} + \alpha}{\sum_{t'} C^{TD}_{t'd,-i} + T\alpha} \frac{C^{WT}_{wt,-i} + \beta}{\sum_{w'} C^{WT}_{w't,-i} + W\beta} \quad (1)$$

ここで，$\mathbf{w} = \{w_i\}$，$\mathbf{z} = \{z_i\}$ である．$\mathbf{w}_{-i}$ は $\mathbf{w}$ から w_i を除外した集合，$\mathbf{z}_{-i}$ は $\mathbf{z}$ から z_i を除外した集合である．$C^{TD}_{td,-i}$ はトピック t が文書 d に割り当てられた回数，$C^{WT}_{wt,-i}$ は単語 w にトピック t が割り当てられた回数であり，共に i 番目の単語を除外している．

2.2 CI-LDA

CI-LDA[3] は複数のクラスを扱うために LDA を拡張したモデルである．Mimno ら[4] や Ni ら[5] は多言語にまたがって共通のトピックを持つ文書に対してこのモデルで実験を行った．簡単のため，CI-LDA が扱うクラスは 2 つであるとする．図 2 は CI-LDA のグラフィカルモデルを示している．このとき，CI-LDA の生成過程は以下の通りである:

(1) 全ての文書 d に対して，$\theta_d \sim Dir(\alpha)$ を選択する．
(2) 全てのトピック t に対して，$\phi_t \sim Dir(\beta)$ と $\tilde{\phi}_t \sim Dir(\tilde{\beta})$ を選択する．
(3) 文書 d の N_d 個の各単語 w_i に対して:
- トピック $z_i \sim Mult(\theta_d)$ を選択する．

Vol.2011-NL-201 No.4
Vol.2011-SLP-86 No.4
2011/5/16

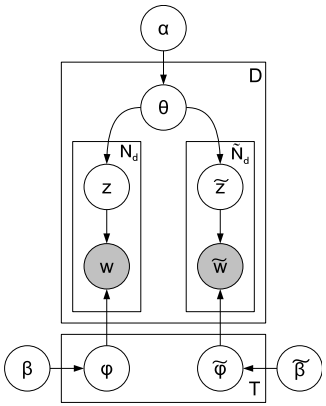


図 2 CI-LDA のグラフィカルモデル

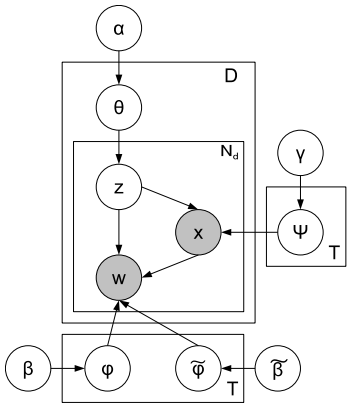


図 3 SwitchLDA のグラフィカルモデル

- 単語 $w_i \sim Mult(\phi_{z_i})$ を選択する．

(4) 文書 d の $N_{\tilde{w}_d}$ 個の各単語 $\tilde{w}_i$ に対して:

- トピック $\tilde{z}_i \sim Mult(\theta_d)$ を選択する．
- 単語 $\tilde{w}_i \sim Mult(\tilde{\phi}_{\tilde{z}_i})$ を選択する．

CI-LDA のギブスサンプリングにおける完全条件付き確率は以下の式で与えられる:

$$p(\mathbf{z}_i = t|\mathbf{w}_i = w, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \alpha, \beta) \propto \frac{C^{TD}_{td,-i} + \alpha}{\sum_{t'} C^{TD}_{t'd,-i} + T\alpha} \frac{C^{WT}_{wt,-i} + \beta}{\sum_{w'} C^{WT}_{w't,-i} + W\beta} \tag{2}$$

$$p(\mathbf{z}_i = t|\mathbf{w}_i = w, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \alpha, \beta) \propto \frac{C^{TD}_{td,-i} + \alpha}{\sum_{t'} C^{TD}_{t'd,-i} + T\alpha} \frac{C^{\tilde{W}T}_{\tilde{w}t,-i} + \tilde{\beta}}{\sum_{\tilde{w}'} C^{\tilde{W}T}_{\tilde{w}'t,-i} + \tilde{W}\tilde{\beta}} \tag{3}$$

2 つのクラスを英語と日本語とする時，w と $\tilde{w}$ はそれぞれ英語の語彙と日本語の語彙になる．W と $\tilde{W}$ はそれぞれ英語の語彙数と日本語の語彙数である．

2.3 SwitchLDA

SwitchLDA[3] は CI-LDA と同様に，複数のクラスを扱うために LDA を拡張したモデルである．しかし，CI-LDA と異なり，SwitchLDA は二項分布に従って 2 つの異なるクラスの各トピックにおける割合を調節できる．図 3 は SwitchLDA のグラフィカルモデルを示している．なお，簡単のためにクラス数は 2 を想定している．SwitchLDA の生成過程を以下に記す．

(1) 全ての文書 d に対して，$\theta_d \sim Dir(\alpha)$ を選択する．

(2) 全てのトピック t に対して，$\phi_t \sim Dir(\beta)$，$\tilde{\phi}_t \sim Dir(\tilde{\beta})$，$\psi_t \sim Beta(\gamma)$ を選択する．

(3) 文書 d の N_d 個の各単語 w_i に対して:

- トピック $z_i \sim Mult(\theta_d)$ を選択する．
- フラグ $x_i \sim Binomial(\psi_{z_i})$ を選択する．
- (x_i=0) の場合，単語 $w_i \sim Mult(\phi_{z_i})$ を選択する．
- (x_i=1) の場合，単語 $w_i \sim Mult(\tilde{\phi}_{z_i})$ を選択する．

ψ_{z_i} は各トピックの 2 つのクラスの割合を調節するための二項分布パラメータを示している．SwitchLDA のギブスサンプリングにおける完全条件付き確率は以下の式で与えられる:

$$p(\mathbf{z}_i = t|\mathbf{w}_i = w, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \alpha, \beta) \propto \frac{C^{TD}_{td,-i} + \alpha}{\sum_{t'} C^{TD}_{t'd,-i} + T\alpha} \frac{n_{t,-i} + \gamma}{n_{t,-i} + \tilde{n}_t + 2\gamma} \frac{C^{WT}_{wt,-i} + \beta}{\sum_{w'} C^{WT}_{w't,-i} + W\beta} \tag{4}$$

Vol.2011-NL-201 No.4
Vol.2011-SLP-86 No.4
2011/5/16

$$p(\mathbf{z}_i = t|\mathbf{w}_i = w, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \alpha, \beta) \propto \frac{C^{TD}_{td,-i} + \alpha}{\sum_{t'} C^{TD}_{t'd,-i} + T\alpha} \frac{\tilde{n}_{t,-i} + \gamma}{n_t + \tilde{n}_{t,-i} + 2\gamma} \frac{C^{\tilde{W}T}_{\tilde{w}t,-i} + \tilde{\beta}}{\sum_{\tilde{w}'} C^{\tilde{W}T}_{\tilde{w}'t,-i} + \tilde{W}\tilde{\beta}} \quad (5)$$

2 つのクラスを英語と日本語とする時，n_t はトピック t が英語部分に割り当てられた回数，$\tilde{n}_t$ はトピック t が日本語部分に割り当てられた回数である．

ベータ分布のハイパーパラメータ γ が十分に大きい場合，式 (4) および式 (5) の右辺の第二項は定数になり，従ってこれらの式はそれぞれ CI-LDA の式 (2) と (3) に等価になる．

2.4 CorrLDA

多言語トピックモデルとして CI-LDA や SwitchLDA を用いる時，同一文書の各言語の部分において支配的な潜在トピックが乖離することがある．CorrLDA[7] は，各文書の異なる言語の部分における潜在トピックの間の依存性の制約をより強く課すことが特徴である．このモデルは最初に文書の一方の言語の部分からトピックを生成する．本論文では，この言語を基軸言語と呼ぶ．そしてもう一方の言語に対しては，基軸言語で生成されたトピックのみを用いる．図 4 は CorrLDA のグラフィカルモデルを示している．CorrLDA の生成過程は以下の通りである．

(1) 全ての文書 d に対して，$\theta_d \sim Dir(\alpha)$ を選択する．
(2) 全てのトピック t に対して，$\phi_t \sim Dir(\beta)$ と $\tilde{\phi}_t \sim Dir(\tilde{\beta})$ を選択する．
(3) 文書 d の N_d 個の各単語 w_i に対して:
- トピック $z_i \sim Mult(\theta_d)$ を選択する．
- 単語 $w_i \sim Mult(\phi_{z_i})$ を選択する．

(4) 文書 d の $N_{\tilde{w}_d}$ 個の各単語 $\tilde{w}_i$ に対して:
- トピック $\tilde{z}_i \sim Unif(z_{w_1}, \cdots, z_{w_{N_{w_d}}})$ を選択する．
- 単語 $\tilde{w}_i \sim Mult(\tilde{\phi}_{\tilde{z}_i})$ を選択する．

このモデルのギブスサンプリングにおける完全条件付確率は以下の式で与えられる:

$$p(\mathbf{z}_i = t|\mathbf{w}_i = w, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \alpha, \beta) \propto \frac{C^{TD}_{td,-i} + \alpha}{\sum_{t'} C^{TD}_{t'd,-i} + T\alpha} \frac{C^{WT}_{wt,-i} + \beta}{\sum_{w'} C^{WT}_{w't,-i} + W\beta} \quad (6)$$

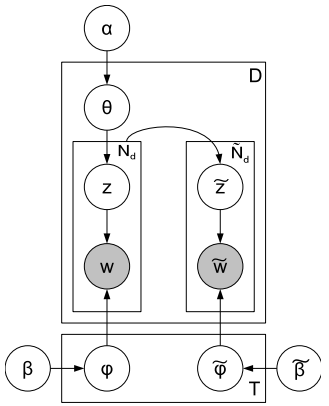


図 4 CorrLDA のグラフィカルモデル

$$p(\mathbf{z}_i = \tilde{t}|\mathbf{w}_i = w, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \alpha, \beta) \propto \frac{C^{TD}_{td,-i}}{N_{w_d}} \frac{C^{\tilde{W}T}_{\tilde{w}t,-i} + \tilde{\beta}}{\sum_{\tilde{w}'} C^{\tilde{W}T}_{\tilde{w}'t,-i} + \tilde{W}\tilde{\beta}} \quad (7)$$

N_{w_d} は文書 d の基軸言語の総延べ語数である．

3. 比較評価

この節では，Wikipedia データを用いた様々な評価手法を用いて多言語トピックモデルを比較する．本論文ではテストセット対数尤度，トピック割り当てのヒストグラム，そして日本語の記事と同じトピックを持つ英語の記事を発見するタスクを用いて評価する．

3.1 データセット

この論文で用いる Wikipedia データは，2009 年 11 月 2 日時点での，言語間で相互リンクがある英語と日本語の Wikipedia の記事で構成されている．我々は Wikipedia の各記事から本文を抽出し，リンク情報と編集履歴情報などを除去した．この処理に WP2TXT[⋆1] を用いた．

⋆1 http://wp2txt.rubyforge.org/

Vol.2011-NL-201 No.4
Vol.2011-SLP-86 No.4
2011/5/16

日本語の記事に対して，MeCab[⋆1]で自動付与した品詞タグを利用して記号，接続詞，助詞などの機能語を除去した．さらに，英語の記事に対してはストップワード 418 語を除去した[10]．前処理後の Wikipedia データの概要を表 1 に示す．

表 1 前処理後の Wikipedia データの概要

	Japanese	English
No. of documents	229855	
No. of word types (vocab)	124,046	173,157
No. of word tokens	61,187,469	80,096,333

3.2 実験設定

本論文では，英語と日本語で相互にリンクで繋がっている Wikipedia 記事を 2 つの言語部分を持つ単一の文書と見なした．評価のため，Wikipedia 文書コレクションを文書レベルでランダムに分割し，80% を訓練用文書， 20% をヘルドアウト文書と呼ぶ．さらに，テストセット対数尤度を計算するために，訓練用文書の各々を単語レベルでランダムに分割し，80% を訓練セット，20% をテストセットとして用いる．

まず，訓練用文書を用いてギブスサンプリングにより CI-LDA，SwitchLDA，CorrLDA，そしてベースラインとして LDA を推定した．このとき，$\alpha = 50/T$，$\beta = 0.01$ の対称ディリクレ分布を仮定した．ギブスサンプリングの収束条件はテストセット対数尤度の変化率が 0.1% 未満であることとした．SwitchLDA の場合，ハイパーパラメータ $\gamma = 1$ の対称ベータ分布を仮定した．LDA の場合，言語を区別しないので，我々は言語間に相互リンクがある英語/日本語の記事を，言語の区別なく混合して単一の文書と見なした．

$T = 1000$ の CI-LDA，CorrLDA で推定したトピックの例を表 2 と表 3 に示す．

3.3 テストセット対数尤度

テストセット対数尤度を測定することで，各トピックモデルの精度を評価できる．テストセット対数尤度が高いほど，モデルの予測能力は高いといえる．本論文では，訓練用文書の訓練セットを用いて多言語トピックモデルを推定し，テストセットの対数尤度を測定した．

表 4 は $T = 500, 1000$ それぞれで推定した各多言語トピックモデルの単語毎テストセット対数尤度を示している．これ以後，*CorrLDA1* を日本語を基軸言語として推定した CorrLDA とする．2.4 節で言及したように，CorrLDA は最初に文書の基軸言語の部分からトピックを生成する．そして，文書の他の言語の部分に対しては，モデルは基軸言語で生成されたトピックのみを用いる．逆に，*CorrLDA2* は英語を基軸言語として推定した CorrLDA とする．表を見てわかる通り，各モデルのテストセット対数尤度は $T = 500$ より $T = 1000$ のほうが高い．よって，各モデルは $T = 500$ より $T = 1000$ のほうが精度が高いといえる．ま

⋆1 http://mecab.sourceforge.net/

表 2 CI-LDA で推定されたトピックの例

English		Japanese	
word	frequency	word	frequency
awarded	1857	選挙	9912
confidence	1271	候補	4964
twothirds	1255	投票	4281
choosing	1066	者	2277
republican	1031	票	2236
ran	696	大統領	1842
partylist	569	議員	1551
sluggish	358	党	1545
chairman	303	当選	1404
hung	288	支持	1253

表 3 CorrLDA で推定されたトピックの例

English		Japanese	
word	frequency	word	frequency
overseeing	2090	選挙	3533
awarded	1863	大統領	3452
civilzation	1590	候補	2750
catholic	1560	民主党	1792
chairman	1383	共和党	1766
ancestry	1329	議員	1627
wood	1283	州	1576
atlases	1208	支持	1442
house	1170	上院	1400
republican	982	知事	1234

表 4 単語毎テストセット対数尤度

	Japanese		English	
	T=500	T=1000	T=500	T=1000
SwitchLDA	-8.1393	-8.01203	-8.6409	-8.5494
CI-LDA	-8.1359	-8.0081	-8.6436	-8.5485
CorrLDA1	-7.4630	-7.3449	-8.4026	-8.3459
CorrLDA2	-7.7768	-7.6629	-8.1969	-8.1092
LDA	-8.1273	-7.9922	-8.6329	-8.5295

Vol.2011-NL-201 No.4
Vol.2011-SLP-86 No.4
2011/5/16

た，CorrLDA1 と CorrLDA2 のテストセット対数尤度は他のモデルより遥かに高い．特に，CorrLDA1 は日本語のテストセット対数尤度が最も高く，CorrLDA2 は英語のテストセット対数尤度が最も高い．これは CorrLDA が基軸言語部分 (CorrLDA1 の場合は日本語) からは特別な制約なしにトピックを推定できるが，他の言語に対しては強い制約 (基軸言語で選択されたトピックからトピックが選択される) が課されるためと考えられる．

3.4 トピック割り当てヒストグラム

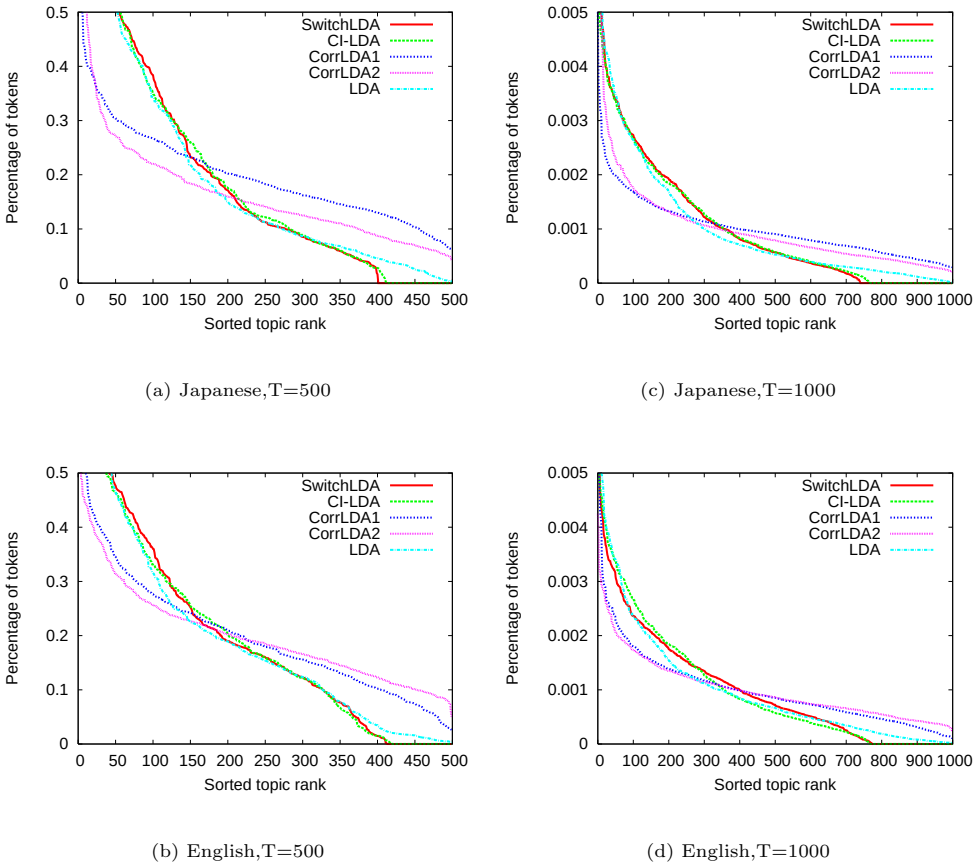
図 5 は各トピックに割り当てられた全てのトークンの割合を降順で並べたものを示している．図を見てわかる通り，CI-LDA，SwitchLDA，LDA の場合は使用されるトピックの少なさが目立つ．一方，CorrLDA(CorrLDA1，CorrLDA2) の場合，より多くのトピックが割り当てられている．これは CorrLDA のトピックが他のモデルと比べてより特定的で，粒度が細かいことを示している．トピックモデルにおいて，トピックの粒度の精細化はユーザの直観的な有用性と強く結びつくと言える[4]．

3.5 対 訳 発 見

与えられた記事に対応する他の言語で記述された記事を，多言語トピックモデルを用いて発見するタスクに着目する．これはストーリーリンク検出[8] を言語を横断して実現するタスクといえる．このタスクを評価するために，訓練用文書から各々の多言語トピックモデルによって推定したトピック・単語分布を用いて，言語毎にヘルドアウト文書の文書・トピック分布を推定した．そして，日本語の記事をクエリとして用いて，英語の対訳記事を発見する性能を評価した．

ヘルドアウト文書の文書・トピック分布を推定するため，それぞれの多言語トピックモデルで前もって推定したトピック単語分布を用いて，LDA のリサンプリングを行った[11]．そして，各ヘルドアウト文書に対して，日本語の文書・トピック分布と英語の文書・トピック分布間の Jensen-Shannon ダイバージェンス (JS ダイバージェンス)[12] を計算した．言語間が相互リンクで繋がっている英語と日本語のヘルドアウト記事のペアは同じトピックを持つと考えられ，従って，潜在トピックが正確ならば，JS ダイバージェンスは小さくなると予測される．私たちは日本語の各ヘルドアウト記事をクエリとし，英語の全てのヘルドアウト記事に対する JS ダイバージェンスを求め，その昇順で並べたランキングを得る．クエリに対応する英語の記事を適合記事と見なして評価する．

図 6 はランキングにおいて上位 K 件までに適合記事が見つかったクエリ記事の割合を示している．図に示されているように，LDA はこのタスクではあまり性能が良くない．これは LDA が言語を全く区別できないためである．一方，CorrLDA はこのタスクにおいて他

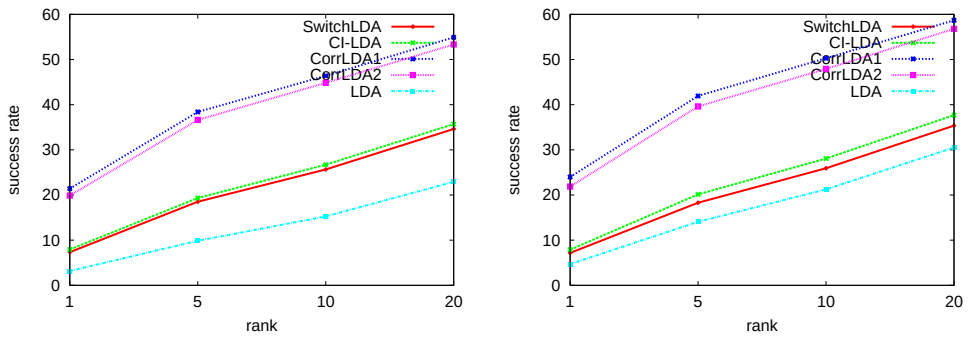


(a) Japanese,T=500　(c) Japanese,T=1000

(b) English,T=500　(d) English,T=1000

図 5 割り当てられた語の数によってトピックをソートした結果

Vol.2011-NL-201 No.4
Vol.2011-SLP-86 No.4
2011/5/16



(a) T=500　　(b) T=1000

図 6　上位 K 件までにクエリ記事に対する対訳記事が見つかった成功率

のモデルより遥かに性能が良い．さらに，CorrLDA1 は CorrLDA2 より僅かに性能が良い．CorrLDA1 の基軸言語は日本語であり，従ってクエリ記事で使用された言語はターゲットとなる言語より重要と考えられる．

4. おわりに

本論文では，多言語の比較可能文書データに適用可能な様々なトピックモデルの性能をテストセット対数尤度，トピック割り当てのヒストグラム，そして言語横断ストーリーリンク検出のタスクによって比較した．実験によって，これまで多言語に関連した研究で使用されてこなかった CorrLDA が，多言語トピックモデルの従来研究で使用された CI-LDA より遥かに性能が良いことを示した．より多くの言語を用いた評価は今後の課題とする．

謝　辞

本研究の一部は，科学研究費補助金基盤研究 (B)（20300038，23300039）の援助による．

参 考 文 献

1) Hofmann, T.: Probabilistic Latent Semantic Indexing, *Proceedings of the 22nd Anuual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, California, USA, pp.50–57 (1999).
2) Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
3) Newman, D., Chemudugunta, C., Smyth, P. and Steyvers, M.: Statistical Entity-Topic Models, *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, Pennsylvania, USA, pp. 680–686 (2006).
4) Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A. and McCallum, A.: Polylingual Topic Models, *Proceeding EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, pp.880–889 (2009).
5) Ni, X., Sun, J.-T., Hu, J. and Chen, Z.: Mining Multilingual Topics from Wikipedia, pp.1155–1156 (2009).
6) Cohn, D. and Hofmann, T.: The missing link - a probabilistic model of document content and hypertext connectivity, *Advances in Neural Information Processing Systems 13*, Berkeley, California, USA, pp.430–436 (2001).
7) Blei, D.M. and Jordan, M.I.: Modeling annotated data, *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, New York, NY, USA, pp.127–134 (2003).
8) Allan, J.: *Topic Detection and Tracking: Event-based Information Organization*, chapter1: Introduction to Topic Detection and Tracking, Kluwer Academic Publishers (2002).
9) Griffiths, T.L. and Steyvers, M.: Finding Scientific Topics, *Proceedings of the National Academy of Sciences of the United States of America*, Vol.101, pp.5228–5235 (2004).
10) Callan, J.P., Croft, W.B. and Harding, S.M.: The INQUERY Retrieval System, *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, Valencia, Spain, pp.78–83 (1992).
11) Wallach, H.M., Murray, I., Salakhutdinov, R. and Mimno, D.: Evaluation Methods for Topic Models, *Proceedings of the 26th International Conference on Machine Learning*, New York, NY, USA, pp.1105–1112 (2009).
12) Lin, J.: Divergence Measures based on the Shannon Entropy, *IEEE Transactions on Information Theory*, Vol.37, No.1, pp.145–151 (1991).