情報処理学会論文誌 Vol.58 No.12 2080–2092 (Dec. 2017)

セグメント構造を持つバイリンガルトピックモデル

田村 晃裕[1,†1,a)] 隅田 英一郎[1,b)]

受付日 2016年8月19日, 採録日 2017年9月5日

概要：本稿では，各文書を「文書–セグメント（たとえば，段落やセクション）–単語」の階層構造でモデル化する新たな多言語トピックモデル「Bilingual Segmented Topic Model（BiSTM）」を提案する．Bilingual Latent Dirichlet Allocation（BiLDA）などの従来の多言語トピックモデルは，対応関係がある文書のトピック分布を共有させることで，異言語の文書間の対応関係を反映したモデル化を行う．一方で，BiSTMは，文書間の対応関係に加えて，対応関係のあるセグメントのトピック分布も共有させることにより，異言語のセグメント間の対応関係も反映したモデル化を行う．また，本稿では，セグメントが与えられていない場合にも提案モデルを適用できるようにするため，Du ら (2013) の教師なしトピック分割手法を BiSTMに導入し，潜在トピックとセグメント境界を同時に推定するモデルも提案する．日英および仏英の多言語コーパスを使った評価実験を通じて，提案モデルは BiLDA よりパープレキシティの観点で優れたモデルであることを示し，対訳対抽出の性能も改善できることを示す．

キーワード：多言語トピックモデル，階層モデル，対訳対抽出

Bilingual Segmented Topic Model

Akihiro Tamura[1,†1,a)] Eiichiro Sumita[1,b)]

Received: August 19, 2016, Accepted: September 5, 2017

Abstract: This paper proposes the bilingual segmented topic model (BiSTM), which hierarchically models documents by treating each document as a set of segments, e.g., sections. While previous bilingual topic models, such as bilingual latent Dirichlet allocation (BiLDA), consider only cross-lingual alignments between entire documents, the proposed model considers cross-lingual alignments between segments in addition to document-level alignments and assigns the same topic distribution to aligned segments. This paper also presents a method for simultaneously inferring latent topics and segmentation boundaries, incorporating unsupervised topic segmentation into BiSTM. Experiments using a Japanese–English and French–English Wikipedia corpus show that the proposed model significantly outperforms BiLDA in terms of perplexity and demonstrates improved performance in translation pair extraction.

Keywords: bilingual topic model, hierarchical model, translation pair extraction

1. はじめに

これまで，probabilistic latent semantic analysis（PLSA）[13] や latent Dirichlet allocation（LDA）[2] など，文書に隠れた潜在トピックを教師なしで解析するトピックモデルが数多く提案されている．トピックモデルは，当初，単言語文書集合を対象としていたが，近年では，多言語文書集合に対して言語共通のトピックを解析する，多言語トピックモデルが提案され，言語横断文書分類や対訳対抽出など数多くの多言語処理タスクに活用されている（詳細はサーベイ論文 [33] 参照）．

Bilingual LDA（BiLDA）[18], [22] を筆頭に，多言語トピックモデルの多くは，ウィキペディアの記事集合など，直接の対訳関係はないが話題や分野を文書単位で共有する多言語文書集合（以降，「コンパラブルコーパス」と呼ぶ）

[1] 情報通信研究機構
National Institute of Information and Communications Technology, Soraku, Kyoto 619–0289, Japan
[†1] 現在，愛媛大学
Presently with Ehime University
[a)] tamura@cs.ehime-u.ac.jp
[b)] eiichiro.sumita@nict.go.jp

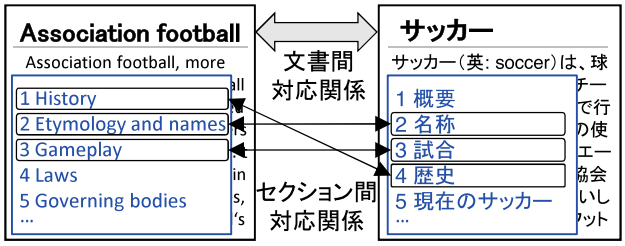


図 1 ウィキペディア記事の例
Fig. 1 Wikipedia article example.

をモデル化する[*1]．具体的には，コンパラブルコーパスは文書間で対応付いているという特徴を利用し，対応関係がある文書のトピック分布を共有させることで，文書間の対応関係を反映したモデル化を行う．

従来の多言語トピックモデルは，文書単位の対応関係しか考慮しない．しかしながら，多くの文書は，「文書–セグメント（たとえば，段落やセクション）–単語」のように階層的に構成されており，文書より小さいセグメント単位で対応付く場合が多い．**図 1** にウィキペディアの記事の例を示す．図 1 では，各記事（文書）は複数のセクションで構成されており，英語記事のセクション 1，2，3 は，それぞれ，日本語記事のセクション 4，2，3 に対応付いている．従来の多言語トピックモデルでは，このようなセグメント間の対応関係は考慮されていなかった．しかし我々は，対応関係のある文書同様，対応関係のあるセグメントも共通のトピック分布を持つべきであると考えた．

また，Du ら [8] は，セグメント単位のトピックとそれらの関係性をとらえることにより，単言語の文書集合のモデル化性能を改善できることを示している．この研究からも，我々は，セグメント単位のトピックが多言語文書集合のモデル化に役立つのではないかと考えた．

そこで，本稿では，BiLDA を拡張し，セグメント間の対応関係をとらえる新たな多言語トピックモデル「Bilingual Segmented Topic Model（BiSTM）」を提案する[*2]．BiSTM では，各文書をセグメント集合と見なし，「文書–セグメント–単語」の階層構造で多言語文書集合をモデル化する．各セグメントのトピック分布は，属する文書のトピック分布を基底測度とした Pitman–Yor 過程（PYP）[25] により生成し，各単語のトピックは，属するセグメントのトピック分布に基づき生成する．また，BiSTM では，異言語のセグメント間が対応関係にあるかを示す二値変数を導入する．そして，その変数に基づき，対応関係にあるセグメントのトピック分布は共通化し，対応関係にないセグメントのトピック分布は独立に生成する．

BiSTM は，各文書がセグメントに分割されていることを前提とする．しかしながら，文書は必ずしもセグメントに分割されているとは限らない．また，分割されているセグメントが，必ずしも文書をモデル化するにあたって最適な分割であるとは限らない．そこで，本稿では，教師なしトピック分割を BiSTM に組み込んだモデルも提案する．以降，このトピック分割（Topic Segmentation，略して「TS」）を組み込んだ提案モデルを「BiSTM+TS」と呼ぶ．BiSTM+TS では，Du ら [10] により提案されたトピック境界のサンプリングを BiSTM に導入することで，セグメントと潜在トピックを同時に推定する．

まとめると，本稿では，以下の 3 つのいずれの場合であっても，セグメント間の対応関係を反映できる多言語トピックモデルを提案する．(i) 文書のセグメントおよびセグメント間の対応関係が与えられている場合（セグメント間の対応関係を表す変数を推定しない BiSTM），(ii) 文書のセグメントは与えられているが，セグメント間の対応関係が与えられていない場合（セグメント間の対応関係を表す変数を推定する BiSTM），(iii) 文書のセグメントおよびセグメント間の対応関係が与えられていない場合（BiSTM+TS）．

日本語と英語（日英）およびフランス語と英語（仏英）のウィキペディア記事からなるコンパラブルコーパスを使った評価実験を通じて，提案モデル（BiSTM，BiSTM+TS）は BiLDA よりパープレキシティの観点で優れたモデルであることを示し，対訳対抽出の性能も改善できることを示す．また，セグメントを自動推定する BiSTM+TS は，人手で設定されたセグメント（具体的には，ウィキペディア記事中のセクション）を用いた BiSTM に近い性能を達成できることを示す．

以降，2 章で提案モデルの基となる従来の多言語トピックモデル BiLDA を説明し，3 章で提案モデル BiSTM を提案する．4 章では，BiSTM に教師なしトピック分割を統合した BiSTM+TS を提案し，5 章では，日英および仏英のコンパラブルコーパスを用いた評価実験を通じて提案モデルの有効性を示す．6 章で提案モデルの効果や性質についての考察を行い，7 章で本研究の関連研究について述べる．最後に，8 章で本稿のまとめを行う．

2. 従来モデル：BiLDA

本章では，提案モデルのベースラインとなる BiLDA モデル [18], [22] を説明する．BiLDA は，単言語の LDA [2] をコンパラブルコーパス用に多言語に拡張したモデルである．単言語 LDA では，各文書は独自のトピック分布を持つのに対して，BiLDA では，対応関係がある文書のトピック分布を共有させることにより，言語横断の潜在トピックを解析する．

アルゴリズム 1 と**図 2** に，それぞれ，BiLDA により，言

[*1] 本研究では，多言語トピックモデルの中でも，文書単位で対応付いているコンパラブルコーパスを解析するモデルに着目する．その他の多言語トピックモデルに関しては，その欠点とともに 7 章で記述する．
[*2] 本稿では，言語数が 2 つであるバイリンガルな設定を扱うが，提案モデルは，3 言語以上の設定に対しても単純に拡張が可能である．

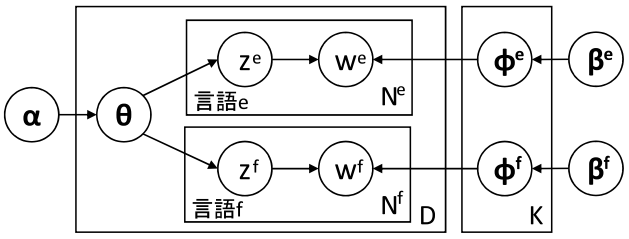


図 2 BiLDA のグラフィカルモデル
Fig. 2 Graphical model of BiLDA.

アルゴリズム 1 BiLDA の生成過程

for each topic $k \in \{1,\dots,K\}$ **do**
  for each language $\ell \in \{e,f\}$ **do**
    choose $\boldsymbol{\phi}^{\ell}_{k} \sim \text{Dirichlet}(\boldsymbol{\beta}^{\ell})$
  end for
end for
for each document pair d_i $(i \in \{1,\dots,D\})$ **do**
  choose $\boldsymbol{\theta}_i \sim \text{Dirichlet}(\boldsymbol{\alpha})$
  for each language $\ell \in \{e,f\}$ **do**
    for each word w^{ℓ}_{im} $(m \in \{1,\dots,N^{\ell}_i\})$ **do**
      choose $z^{\ell}_{im} \sim \text{Multinomial}(\boldsymbol{\theta}_i)$
      choose $w^{\ell}_{im} \sim \text{Multinomial}(\boldsymbol{\phi}^{\ell}_{z^{\ell}_{im}})$
    end for
  end for
end for

語 e と f で記述された D 個の文書対からなるコンパラブルコーパスを生成する生成過程とグラフィカルモデルを示す．以降，各文書対 d_i $(i \in \{1,\dots,D\})$ における言語 e の文書を d^e_i，言語 f の文書を d^f_i と表記する．つまり，d_i=(d^e_i, d^f_i) である．BiLDA では，各トピック $k \in \{1,\dots,K\}$ は，言語ごとの単語分布（言語 e の単語分布 $\boldsymbol{\phi}^{\boldsymbol{e}}_{\boldsymbol{k}}$ と言語 f の単語分布 $\boldsymbol{\phi}^{\boldsymbol{f}}_{\boldsymbol{k}}$）を持つ．各単語分布 $\boldsymbol{\phi}^{\boldsymbol{\ell}}_{\boldsymbol{k}}$ $(\ell \in \{e,f\})$ は，$\boldsymbol{\beta}^{\boldsymbol{\ell}}$ をパラメータとするディリクレ分布から生成される（ステップ 1–5）．文書対 d_i の生成過程では，まず，d_i に対するトピック分布 $\boldsymbol{\theta}_{\boldsymbol{i}}$ が，$\boldsymbol{\alpha}$ をパラメータとするディリクレ分布から生成される（ステップ 7）．これにより，対応関係にある d^e_i と d^f_i は共通のトピック分布 $\boldsymbol{\theta}_{\boldsymbol{i}}$ を持つことになる．その後，言語 ℓ の文書 d^{ℓ}_i 中の各単語位置 $m \in \{1,\dots,N^{\ell}_i\}$ に対して，潜在トピック z^{ℓ}_{im} が，$\boldsymbol{\theta}_{\boldsymbol{i}}$ をパラメータとする多項分布から生成される（ステップ 10）．そして，単語 w^{ℓ}_{im} が，ステップ 10 で具体化された潜在トピック z^{ℓ}_{im} に関する単語分布 $\boldsymbol{\phi}^{\boldsymbol{\ell}}_{\boldsymbol{z^{\ell}_{im}}}$ をパラメータとする多項分布から生成される（ステップ 11）．

3. 提案モデル：BiSTM

本章では，BiLDA を拡張し，セグメント間の対応関係を考慮する多言語トピックモデル BiSTM を提案する．アルゴリズム 2 と図 **3** に，それぞれ，BiSTM の生成過程とグラフィカルモデルを示す．ここで，各文書 d^{ℓ}_i は，S^{ℓ}_i 個のセグメントで構成されているものとする．つまり，$d^{\ell}_i = \bigcup_{j=1}^{S^{\ell}_i} s^{\ell}_{ij}$ である．図 3 に示されているとおり，BiSTM では，各言語（言語 e，f）において，文書に関する層と単語に関する層の間にセグメントに関する層を設け，セグメントのトピック分布（$\boldsymbol{\nu}^{\boldsymbol{e}}$，$\boldsymbol{\nu}^{\boldsymbol{f}}$）を，文書のトピック分布（$\boldsymbol{\theta}$）と単語のトピック（$\boldsymbol{z}^{\boldsymbol{e}}$，$\boldsymbol{z}^{\boldsymbol{f}}$）の間に導入することで，文書を階層的にモデル化する．また，異言語のセグメント間に対応関係があるかを示す二値変数 y を導入することで，セグメント間の対応関係を反映したモデル化を行う．ここで，$y_{ijj'} = 1$ は 2 つのセグメント s^e_{ij} と $s^f_{ij'}$ に対応関係があることを示し，$y_{ijj'} = 0$ は対応関係がないことを示す．また，各文書対 d_i に対して，セグメント間の対応関係全体を行列 Y_i で表現する．Y_i は，$S^e_i \times S^f_i$ 行列であり，(j,j') 成分が $y_{ijj'}$ である．

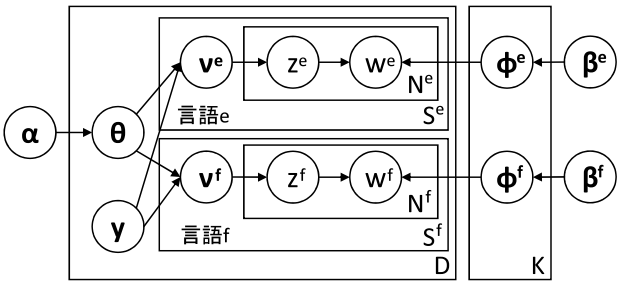


図 3 BiSTM のグラフィカルモデル
Fig. 3 Graphical model of BiSTM.

アルゴリズム 2 BiSTM の生成過程

for each topic $k \in \{1,\dots,K\}$ **do**
  for each language $\ell \in \{e,f\}$ **do**
    choose $\boldsymbol{\phi}^{\ell}_{k} \sim \text{Dirichlet}(\boldsymbol{\beta}^{\ell})$
  end for
end for
for each document pair d_i $(i \in \{1,\dots,D\})$ **do**
  choose $\boldsymbol{\theta}_i \sim \text{Dirichlet}(\boldsymbol{\alpha})$
  if Y_i is not given **then**
    choose $\gamma_i \sim \text{Beta}(\eta_0,\eta_1)$
    choose $Y_i \sim \text{Bernoulli}(\gamma_i)$
  end if
  generate aligned segment sets $\boldsymbol{A}_i = \text{gen\_A}(Y_i)$
  for each set $\boldsymbol{A}_{ig}$ $(g \in \{1,\dots,|\boldsymbol{A}_i|\})$ **do**
    choose $\boldsymbol{\nu}_{ig} \sim \text{PYP}(a,b,\boldsymbol{\theta}_i)$
  end for
  for each language $\ell \in \{e,f\}$ **do**
    for each segment s^{ℓ}_{ij} $(j \in \{1,\dots,S^{\ell}_i\})$ **do**
      get index of s^{ℓ}_{ij} in $\boldsymbol{A}_i$: g =get_idx($\boldsymbol{A}_i$,s^{ℓ}_{ij})
      for each word w^{ℓ}_{ijm} $(m \in \{1,\dots,N^{\ell}_{ij}\})$ **do**
        choose $z^{\ell}_{ijm} \sim \text{Multinomial}(\boldsymbol{\nu}_{ig})$
        choose $w^{\ell}_{ijm} \sim \text{Multinomial}(\boldsymbol{\phi}^{\ell}_{z^{\ell}_{ijm}})$
      end for
    end for
  end for
end for

3.1 BiSTM の生成過程

本節では，提案モデル BiSTM の生成過程をアルゴリズム 2 と図 3 を用いて説明する．まず，BiLDA 同様，各トピックに対して言語固有の単語分布 $\boldsymbol{\phi}^{\boldsymbol{\ell}}_{\boldsymbol{k}}$ がディリクレ分布により生成される（ステップ 1–5）．そして，文書対 d_i の

生成過程では，最初に，d_i に対するトピック分布 $\boldsymbol{\theta_i}$ が生成される（ステップ 7）．したがって，BiSTM においても，各文書対は共通のトピック分布を持つ．

その後，セグメント間の対応関係が与えられていない場合，Y_i が生成される（ステップ 8–11）．Y_i の生成過程では，各文書対 d_i に対して，d_i^e 中のセグメントと d_i^f 中のセグメントは確率 γ_i で対応付くことを仮定する．この仮定に基づき，まず，γ_i が η_0 と η_1 をパラメータとするベータ分布から生成され（ステップ 9），その後，Y_i の各成分 $y_{ijj'}$ が，互いに独立に，γ_i をパラメータとするベルヌーイ分布から生成される（ステップ 10）．セグメント間の対応関係が与えられている場合は，ステップ 8 から 11 は実行されずに所与の Y_i を用いることに注意されたい．

その後，具体化されている Y_i に基づき，対応関係のあるセグメント集合（以降，「対応セグメント集合」と呼ぶ）の集合 $\boldsymbol{A_i}$ が生成される（ステップ 12）．アルゴリズム 2 中の gen_A() が，Y_i を受け取り，$y = 1$ の関係となっているセグメントを 1 つの対応セグメント集合にまとめることにより対応セグメント集合の集合 $\boldsymbol{A_i}$ を生成する関数である．たとえば，$d_i^e = \{s_{i1}^e, s_{i2}^e\}$，$d_i^f = \{s_{i1}^f, s_{i2}^f, s_{i3}^f\}$，$y_{i11}$ と y_{i12} が 1，その他の y が 0 のとき，ステップ 12 では，gen_A() により $\boldsymbol{A_i} = \{\boldsymbol{A_{i1}} = \{s_{i1}^e, s_{i1}^f, s_{i2}^f\}, \boldsymbol{A_{i2}} = \{s_{i2}^e\}, \boldsymbol{A_{i3}} = \{s_{i3}^f\}\}$ が生成される．続いて，$\boldsymbol{A_i}$ 中の各対応セグメント集合 $\boldsymbol{A_{ig}}$ ($g \in \{1, \ldots, |\boldsymbol{A_i}|\}$) に対して，トピック分布 $\boldsymbol{\nu_{ig}}$ が，基底測度 $\boldsymbol{\theta_i}$，集中度パラメータ a，ディスカウントパラメータ b の Pitman–Yor 過程から生成される（ステップ 14）．ステップ 12 から 15 を通じて，Y_i で示唆される対応関係のあるセグメントは，共通のトピック分布を持つ．たとえば，上記の例では，s_{i1}^e，s_{i1}^f および s_{i2}^f は共通のトピック分布 $\boldsymbol{\nu_{i1}} \sim \mathrm{PYP}(a, b, \boldsymbol{\theta_i})$ を持つ．

最後に，セグメント s_{ij}^ℓ の各単語位置 $m \in \{1, \ldots, N_{ij}^\ell\}$ に対して，潜在トピック z_{ijm}^ℓ が，$\boldsymbol{\nu_{ig}}$ をパラメータとする多項分布から生成され（ステップ 20），単語 w_{ijm}^ℓ が，具体化された z_{ijm}^ℓ に関する単語分布 $\boldsymbol{\phi}_{\boldsymbol{z}_{ijm}^\ell}^{\boldsymbol{\ell}}$ をパラメータとする多項分布から生成される（ステップ 21）．ここで，g は，セグメント s_{ij}^ℓ を含む対応セグメント集合のインデックスであり，ステップ 18 で get_idx() により具体化されている．get_idx() は，セグメントと対応セグメント集合の集合を引数として受付け，入力のセグメントを含む対応セグメント集合のインデックスを返す関数である．たとえば，上記の例では，s_{i1}^e や s_{i2}^f に対する g は 1 (get_idx($\boldsymbol{A_i}$, s_{i1}^e)=get_idx($\boldsymbol{A_i}$, s_{i2}^f)=1) であり，s_{i2}^e に対する g は 2（get_idx($\boldsymbol{A_i}$, s_{i2}^e)=2）である．

3.1.1 中華料理店過程による表現

本項以前で BiSTM の生成過程の説明をひととおり行ったが，本項では，BiSTM の推定を簡単にするために，BiSTM の階層性，つまり，$\boldsymbol{\nu}$ と $\boldsymbol{z}$ の生成過程を中華料理店過程で表現する．提案モデルの中華料理店過程では，対応セグメント集合が店，単語が顧客，トピックが料理に対応し，単語とそのトピックは店の中にあるテーブルにより結び付けられる．具体的には，単語に相当する顧客は順番に店のどこかのテーブルに座り，着席したテーブルに割り当てられた料理がその単語のトピックであると解釈することができる．本表現におけるテーブルに関する統計量は，変数 t で表す．具体的には，t_{igk} は，文書対 i 中の対応セグメント集合 g のトピック k に関するテーブル数を表す．

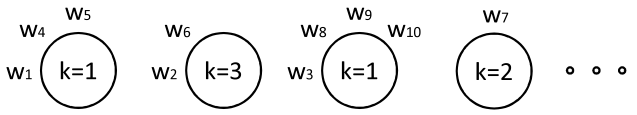


図 4 中華料理店過程の例
Fig. 4 An example of Chinese restaurant process.

中華料理店過程の例を**図 4** に示す．図 4 は，文書対 i 中の対応セグメント集合 g の中にある 10 個の単語 $(w_1, \cdots, w_{10})$ が 4 個のテーブルに着席した状態を示しており，最初のテーブルから 4 番目のテーブルには，それぞれ，料理 1，3，1，2 が割り当てられている．図 4 では，たとえば，単語 w_2 のトピックは $k = 3$ となっている．また，最初と 3 番目のテーブルに，$k = 1$ の料理が割り当てられているので，$t_{ig1} = 2$ であり，同様に，$t_{ig2} = 1$，$t_{ig3} = 1$ である．

この中華料理店過程に従うと，$m + 1$ 番目の単語のトピックの確率分布は，1 から m 番目までの単語が着席した店の状態に基づき，$\frac{b+T^* \times a}{b+m}\boldsymbol{\theta_i} + \sum_{t=1}^{T^*} \frac{n_t^* - a}{b+m}\delta_{k_t^*}(\cdot)$ で規定することができる．ここで，$T^* = \sum_{k=1}^{K} t_{igk}$，$n_t^*$ は t 番目のテーブルに着席している顧客数，k_t^* は t 番目のテーブルに割り当てられている料理，$\delta_{k_t^*}(\cdot)$ はディラック測度である．

3.2 BiSTM における推定

本節では，BiSTM において，モデルパラメータ（$\boldsymbol{\alpha}$，$\boldsymbol{\beta}$）および観測データ（$\boldsymbol{w}$，$\boldsymbol{y}$）を基に，隠れ変数（$\boldsymbol{\theta}$，$\boldsymbol{\nu}$，$\boldsymbol{z}$，$\boldsymbol{\phi}$）を推定する方法を説明する．以降では，言語依存の変数に対して，上付き文字を省略することで e と f の両言語の変数を表すことにする．たとえば，$\boldsymbol{z}$ は $\boldsymbol{z^e}$ と $\boldsymbol{z^f}$ を表す（$\boldsymbol{z} = \{\boldsymbol{z^e}, \boldsymbol{z^f}\}$）．推定では，モデルパラメータと観測データが与えられたときの事後確率 $p(\boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{z}, \boldsymbol{\phi} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{y})$ が最大となる隠れ変数を特定する．しかし，LDA や BiLDA などのその他のトピックモデル同様，BiSTM においても，隠れ変数の事後確率 $p(\boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{z}, \boldsymbol{\phi} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{y})$ を直接計算することはできない．そこで，本節では，Du ら [10] の方法を参考にし，ブロック化ギブスサンプリングにより各隠れ変数を推定する．

3.1.1 項で述べたとおり，BiSTM の階層性は中華料理店過程で表現する．つまり，この過程により，$\boldsymbol{\theta}$，$\boldsymbol{\nu}$ および $\boldsymbol{\phi}$ を積分消去し，代わりに，中華料理店過程のテーブルに関する変数 $\boldsymbol{t}$ を導入する．**表 1** に推定で用いる統計量をまとめる．ここで，$\boldsymbol{W^\ell}$ は言語 ℓ の単語集合を表す．さらに，

表 1 推定で用いる統計量
Table 1 Statistics used in our inference.

t_{igk}	文書対 i 中の対応セグメント集合 g のトピック k に関するテーブル数
$\boldsymbol{t_{ig}}$	k 番目の要素が t_{igk} である K 次元ベクトル
$t_{ig\cdot}$	文書対 i 中の対応セグメント集合 g の総テーブル数（$\sum_k t_{igk}$）
n_{igk}	文書対 i 中の対応セグメント集合 g 中のトピックが k である単語数
$n_{ig\cdot}$	文書対 i 中の対応セグメント集合 g 中の総単語数（$\sum_k n_{igk}$）
M^{ℓ}_{kw}	トピックが k である言語 ℓ の単語 w の数
$\boldsymbol{M^{\ell}_{k}}$	w 番目の要素が M^{ℓ}_{kw} である $\lvert\boldsymbol{W^{\ell}}\rvert$ 次元ベクトル

収束を早めるため，Chen ら [5] に倣い，各単語 w^{ℓ}_{ijm} に対して，w^{ℓ}_{ijm} がテーブルの最初の顧客になる（$\delta^{\ell}_{ijm}=1$）か否（$\delta^{\ell}_{ijm}=0$）かを二値で示す補助変数 δ^{ℓ}_{ijm} を導入し，t_{igk} は $\boldsymbol{\delta}$ に基づき，$t_{igk}=\sum_{s^{\ell}_{ij}\in\boldsymbol{AS_{ig}}}\sum_{m=1}^{N^{\ell}_{ij}}\delta^{\ell}_{ijm}I(z^{\ell}_{ijm}=k)$ のとおり算出する．ここで，$I(x)$ は条件 x を満たすときは 1，満たさないときは 0 を返す関数である．したがって，$\boldsymbol{z}$ と $\boldsymbol{\delta}$ の 2 種類の変数をサンプリングすることで BiSTM の推定を行う．その際，z^{ℓ}_{ijm} と δ^{ℓ}_{ijm} をブロック化し，2 つの変数を同時にサンプリングする．ただし，$\boldsymbol{y}$ が観測データとして与えられていない場合は，$\boldsymbol{z}$ と $\boldsymbol{\delta}$ に加えて $\boldsymbol{y}$ もサンプリングにより推定する．つまり，$(z^{\ell}_{ijm},\ \delta^{\ell}_{ijm})$ と $y_{ijj'}$ の 2 種類のブロックのサンプリングを交互に繰り返し行うことで 3 種類の変数を推定する．以降，各ブロックのサンプリングについて説明する．

3.2.1 （$\boldsymbol{z}$，$\boldsymbol{\delta}$）のサンプリング

$\boldsymbol{z}$，$\boldsymbol{w}$ および $\boldsymbol{\delta}$ の同時事後分布は以下の式 (1) のとおりである．この式 (1) は，Du ら [8] の式 (1) および Du ら [10] の式 (3) の導出と同様に導出できる．

$$
\begin{aligned}
&p(\boldsymbol{z},\boldsymbol{w},\boldsymbol{\delta}|\boldsymbol{\alpha},\boldsymbol{\beta},a,b,\boldsymbol{y})\\
&=\prod_{i=1}^{D}\Bigg(\frac{\mathrm{Beta}_K(\boldsymbol{\alpha}+\sum_{\boldsymbol{A}_i}\boldsymbol{t_{ig}})}{\mathrm{Beta}_K(\boldsymbol{\alpha})}\\
&\prod_{\boldsymbol{A}_i}\Bigg(\frac{(b|a)_{t_{ig\cdot}}}{(b)_{n_{ig\cdot}}}\prod_{k=1}^{K}S\Big(n_{igk},t_{igk},a\Big)\binom{n_{igk}}{t_{igk}}^{-1}\Bigg)\Bigg)\\
&\prod_{k=1}^{K}\Bigg(\frac{\mathrm{Beta}_{W^e}(\boldsymbol{\beta^e}+\boldsymbol{M^e_k})}{\mathrm{Beta}_{W^e}(\boldsymbol{\beta^e})}\frac{\mathrm{Beta}_{W^f}(\boldsymbol{\beta^f}+\boldsymbol{M^f_k})}{\mathrm{Beta}_{W^f}(\boldsymbol{\beta^f})}\Bigg)
\end{aligned}
\tag{1}
$$

ここで，$\mathrm{Beta}_K(\cdot)$ および $\mathrm{Beta}_{W^{\ell}}(\cdot)$ は，それぞれ，K 次元および $|\boldsymbol{W^{\ell}}|$ 次元のベータ関数，$(b|a)_n$ はポッホハマー記号[*3]，$(b)_n$ は $(b|1)_n$ である．また，$S(n,m,a)$ は一般化された第二種スターリング数 [14] であり，$S(n+1,m,a)=S(n,m-1,a)+(n-ma)S(n,m,a)$ のとおり，再帰的に定義できる．ただし，$S(1,1,a)=S(0,0,a)=1$，$S(n,0,a)=S(0,m,a)=0$，$S(n,m,a)=0$（$m>n$ のとき）である．評価実験では，Du ら [8], [10] に倣い，計算コストを削減するため，スターリング数はあらかじめ対数形式 [4] で計算して保存しておき，サンプリング時には保存してある値を使った．ただし，この計算コスト削減処理を行うにあたっては，ディスカウントパラメータ a をあらかじめ定めて固定する必要があることに注意されたい．

[*3] $(b|a)_n=\prod_{t=0}^{n-1}(b+ta)$.

z^{ℓ}_{ijm} と δ^{ℓ}_{ijm} の同時事後分布は，Du ら [10] のように，上記式 (1) の $\boldsymbol{z}$，$\boldsymbol{w}$ および $\boldsymbol{\delta}$ の同時事後分布 $p(\boldsymbol{z},\boldsymbol{w},\boldsymbol{\delta}|\boldsymbol{\alpha},\boldsymbol{\beta},a,b,\boldsymbol{y})$ からベイズの定理に従い，以下の式 (2) および (3) となる．

$$
\begin{aligned}
&p(z^{\ell}_{ijm}=k,\delta^{\ell}_{ijm}=1|\boldsymbol{z}^{-z^{\ell}_{ijm}},\boldsymbol{w},\boldsymbol{\delta}^{-\delta^{\ell}_{ijm}},\boldsymbol{\alpha},\boldsymbol{\beta},a,b,\boldsymbol{y})\\
&=\frac{\beta^{\ell}_{w^{\ell}_{ijm}}+M^{\ell}_{kw^{\ell}_{ijm}}}{\sum_{w\in\boldsymbol{W^{\ell}}}(\beta^{\ell}_{w}+M^{\ell}_{kw})}\frac{\alpha_k+\sum_{\boldsymbol{A}_i}t_{igk}}{\sum_{k=1}^{K}(\alpha_k+\sum_{\boldsymbol{A}_i}t_{igk})}\\
&\frac{b+at_{ig'\cdot}}{b+n_{ig'\cdot}}\frac{S(n_{ig'k}+1,t_{ig'k}+1,a)}{S(n_{ig'k},t_{ig'k},a)}\frac{t_{ig'k}+1}{n_{ig'k}+1}
\end{aligned}
\tag{2}
$$

$$
\begin{aligned}
&p(z^{\ell}_{ijm}=k,\delta^{\ell}_{ijm}=0|\boldsymbol{z}^{-z^{\ell}_{ijm}},\boldsymbol{w},\boldsymbol{\delta}^{-\delta^{\ell}_{ijm}},\boldsymbol{\alpha},\boldsymbol{\beta},a,b,\boldsymbol{y})\\
&=\frac{\beta^{\ell}_{w^{\ell}_{ijm}}+M^{\ell}_{kw^{\ell}_{ijm}}}{\sum_{w\in\boldsymbol{W^{\ell}}}(\beta^{\ell}_{w}+M^{\ell}_{kw})}\frac{1}{b+n_{ig'\cdot}}\\
&\frac{S(n_{ig'k}+1,t_{ig'k},a)}{S(n_{ig'k},t_{ig'k},a)}\frac{n_{ig'k}+1-t_{ig'k}}{n_{ig'k}+1}
\end{aligned}
\tag{3}
$$

ここで，s^{ℓ}_{ij} は $\boldsymbol{A_{ig'}}$ の要素である．

3.2.2 $\boldsymbol{y}$ のサンプリング

中華料理店過程において，対応セグメント集合は中華料理店に相当する．したがって，$y_{ijj'}$ のサンプリングは，中華料理店の分割か統合かを選択することとしてとらえることができる．具体的には，$y_{ijj'}=0$ の場合，1 つの対応セグメント集合 $\boldsymbol{A_m}$ が，2 つの対応セグメント集合 $\boldsymbol{A_l}$ と $\boldsymbol{A_r}$ に分割されるととらえることができる．ここで，$\boldsymbol{A_l}$，$\boldsymbol{A_r}$，$\boldsymbol{A_m}$ は，それぞれ，s^{e}_{ij}，$s^{f}_{ij'}$，s^{e}_{ij} と $s^{f}_{ij'}$ の両方を含むセグメント集合である．また，$y_{ijj'}=1$ の場合，2 つの対応セグメント集合 $\boldsymbol{A_l}$ と $\boldsymbol{A_r}$ が，1 つの対応セグメント集合 $\boldsymbol{A_m}$ に統合されるととらえることができる．

$\boldsymbol{A_l}$ および $\boldsymbol{A_r}$ は，現在の $\boldsymbol{y}$ に基づき，次のとおり具体化する．$A_i(s^{e}_{ij})=A_i(s^{f}_{ij'})$ の場合は，$\boldsymbol{A_l}=\{s^{e}_{ij}\}\cup A^{f}_{i}(s^{e}_{ij})\setminus\{s^{f}_{ij'}\}$，$\boldsymbol{A_r}=\{s^{f}_{ij'}\}\cup A^{e}_{i}(s^{f}_{ij'})\setminus\{s^{e}_{ij}\}$ とし，その他の場合は，$\boldsymbol{A_l}=A_i(s^{e}_{ij})$，$\boldsymbol{A_r}=A_i(s^{f}_{ij'})$ とする．ここで，$A_i(j)$ は，$\boldsymbol{A_i}$ の要素の中でセグメント j を含む対応セグメント集合であり，$A^{\ell}_{i}(j)$ は，$A_i(j)$ に含まれる言語 ℓ のセグメントの集合である．たとえば 3 章の例においては，$A_i(s^{f}_{i1})=\boldsymbol{A_{i1}}=\{s^{e}_{i1},s^{f}_{i1},s^{f}_{i2}\}$，$A^{e}_{i}(s^{f}_{i1})=\{s^{e}_{i1}\}$，$A^{f}_{i}(s^{f}_{i1})=\{s^{f}_{i1},s^{f}_{i2}\}$ である．また，$y_{i11}=0$ の場合，$\boldsymbol{A_m}=\{s^{e}_{i1},s^{f}_{i1},s^{f}_{i2}\}$ が，$\boldsymbol{A_l}=\{s^{e}_{i1}\}\cup A^{f}_{i}(s^{e}_{i1})\setminus\{s^{f}_{i1}\}=\{s^{e}_{i1},s^{f}_{i2}\}$ と $\boldsymbol{A_r}=\{s^{f}_{i1}\}\cup A^{e}_{i}(s^{f}_{i1})\setminus\{s^{e}_{i1}\}=\{s^{f}_{i1}\}$ の 2 つの集合に分割される．$y_{i23}=1$ の場合，$\boldsymbol{A_l}=A_i(s^{e}_{i2})=\{s^{e}_{i2}\}$ と $\boldsymbol{A_r}=A_i(s^{f}_{i3})=\{s^{f}_{i3}\}$ が，$\boldsymbol{A_m}=\{s^{e}_{i2},s^{f}_{i3}\}$ に統合される．

上記のように $\boldsymbol{y}$ のサンプリングを中華料理店の分割か統合かを選択することとしてとらえると，$\boldsymbol{y}$ のサンプリン

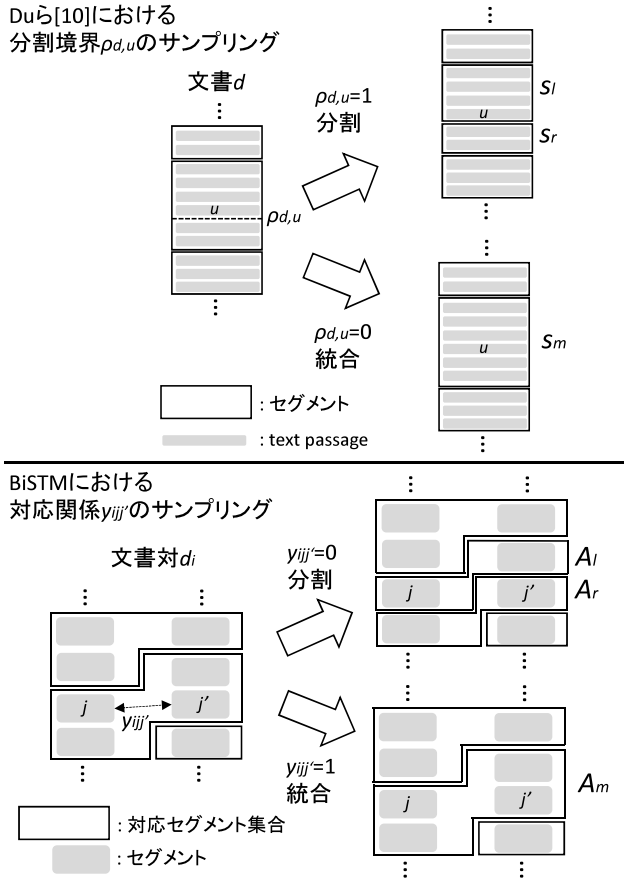


図 5 BiSTM の y のサンプリングと Du ら [10] の ρ のサンプリングの対応関係

Fig. 5 Correspondence between sampling y in BiSTM and sampling ρ in Du et al. [10].

グは，Du ら [10] における分割境界のサンプリング（Du ら [10] の 4.2 節参照）と同様に実現できる．BiSTM における $\boldsymbol{y}$ のサンプリングと Du ら [10] における分割境界 $\boldsymbol{\rho}$ のサンプリングとの対応を**図 5** に示す．Du ら [10] では，「text passage」を最小単位とし，分割境界のサンプリングにより，同じトピックの「text passage」を結合してセグメントを作るのに対して，BiSTM では，セグメントを単位とし，$\boldsymbol{y}$ のサンプリングにより，同じトピックのセグメントを結合して対応セグメント集合を作る，と対応付けることができる．

すると，$y_{ijj'}$ の事後分布は以下の式 (4) および (5) で計算できる．この式 (4)，(5) は，それぞれ，Du ら [10] の式 (6)，(9) に対応する[*4]．

$$p(y_{ijj'}=0|\boldsymbol{y}^{-y_{ijj'}},\boldsymbol{z},\boldsymbol{w},\boldsymbol{\delta},\boldsymbol{\alpha},a,b,\eta_0,\eta_1)$$
$$\propto \frac{\eta_0+c_{i0}}{\eta_0+\eta_1+c_{i0}+c_{i1}}\mathrm{Beta}_K\left(\boldsymbol{\alpha}+\sum_{\boldsymbol{A_i}}\boldsymbol{t_{ig}}\right)$$
$$\prod_{g\in\{\boldsymbol{A_l},\boldsymbol{A_r}\}}\frac{(b|a)_{t_{ig\cdot}}}{(b)_{n_{ig\cdot}}}\prod_{k=1}^{K}S(n_{igk},t_{igk},a) \tag{4}$$
$$p(y_{ijj'}=1|\boldsymbol{y}^{-y_{ijj'}},\boldsymbol{z},\boldsymbol{w},\boldsymbol{t}\setminus\mathbb{T},\boldsymbol{\alpha},a,b,\eta_0,\eta_1)$$
$$\propto\sum_{\mathbb{T}}\left(\frac{\eta_1+c_{i1}}{\eta_0+\eta_1+c_{i0}+c_{i1}}\mathrm{Beta}_K\left(\boldsymbol{\alpha}+\sum_{\boldsymbol{A_i}}\boldsymbol{t_{ig}}\right)\right.$$
$$\left.\frac{(b|a)_{t_{i,\boldsymbol{A_m},\cdot}}}{(b)_{n_{i,\boldsymbol{A_m},\cdot}}}\prod_{k=1}^{K}S(n_{i,\boldsymbol{A_m},k},t_{i,\boldsymbol{A_m},k},a)\right) \tag{5}$$

ここで，「$\boldsymbol{A_l}$ と $\boldsymbol{A_r}$ の両方が $n_{igk}=1$」かつ「$\boldsymbol{A_l}$ と $\boldsymbol{A_r}$ のどちらかあるいはその両方が $t_{igk}=1$」となるトピックと対応セグメント集合の組合せを C とすると，$\mathbb{T}=\{t_{igk}:(k,g)\in C\}$ である．また，c_{i0} は値が 0 である y_i の数，c_{i1} は値が 1 である y_i の数である．

$y_{ijj'}$ のサンプリングでは，$y_{ijj'}$ の値の更新に加えて，選択した行為（分割あるいは統合）を行った後の対応セグメント集合と整合性を取るために，必要に応じて $y_{ijj'}$ 以外の y の値も変更されることに注意されたい．

3.2.3 $\boldsymbol{\theta}$，$\boldsymbol{\nu}$，$\boldsymbol{\phi}$ の推定

提案の推定法では，中華料理店過程により $\boldsymbol{\theta}$，$\boldsymbol{\nu}$ および $\boldsymbol{\phi}$ を積分消去するため，これらの変数は直接推定しない．代わりに，推定された統計量 t_{igk}，n_{igk}，M^{ℓ}_{kw} を使い，以下の式 (6) から (8) の期待値に基づき算出できる．

$$\hat{\theta}_{ik}=\mathbb{E}_{\boldsymbol{z_i},\boldsymbol{t_i}|\boldsymbol{w_i},\boldsymbol{\alpha},\boldsymbol{\beta},a,b,\boldsymbol{y}}\left[\frac{\alpha_k+\sum_{\boldsymbol{A_i}}t_{igk}}{\sum_{k=1}^{K}(\alpha_k+\sum_{\boldsymbol{A_i}}t_{igk})}\right] \tag{6}$$
$$\hat{\nu}_{igk}=\mathbb{E}_{\boldsymbol{z_i},\boldsymbol{t_i}|\boldsymbol{w_i},\boldsymbol{\alpha},\boldsymbol{\beta},a,b,\boldsymbol{y}}\left[\frac{n_{igk}-at_{igk}}{b+n_{ig\cdot}}+\theta_{ik}\frac{at_{ig\cdot}+b}{b+n_{ig\cdot}}\right] \tag{7}$$
$$\hat{\phi}^{\ell}_{kw}=\mathbb{E}_{\boldsymbol{z},\boldsymbol{t}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta},a,b,\boldsymbol{y}}\left[\frac{\beta^{\ell}_{w}+M^{\ell}_{kw}}{\sum_{w'\in\boldsymbol{W^{\ell}}}(\beta^{\ell}_{w'}+M^{\ell}_{kw'})}\right] \tag{8}$$

4. 教師なしトピック分割の BiSTM への導入（BiSTM+TS）

本章では，潜在トピックとともにセグメント境界も推定するため，Du ら [10] により提案されたベイズ的トピック分割手法を BiSTM に導入したモデル（BiSTM+TS）を提案する．BiSTM+TS では，連続した同一トピックの塊，つまり，トピック分割手法で分けられた塊をセグメントとして扱う．本研究では，トピック分割時の最小単位（Du ら [10] における「text passage」）は文とする．具体的には，BiSTM+TS では，各文書 d^{ℓ}_i のセグメントを，各文 u^{ℓ}_{ih} $(h\in\{1,\ldots,U^{\ell}_i\})$ に対して設けられた分割境界を表す二値変数 ρ^{ℓ}_{ih} で定義する．文 u^{ℓ}_{ih} の直後に境界があるときは，ρ^{ℓ}_{ih} は 1 であり，そうでないときは，ρ^{ℓ}_{ih} は 0 である．たとえば，$\boldsymbol{\rho^{\ell}_i}=(0,1,0,0,1)$ は，文書 d^{ℓ}_i が，$\{u^{\ell}_{i1},u^{\ell}_{i2}\}$ と $\{u^{\ell}_{i3},u^{\ell}_{i4},u^{\ell}_{i5}\}$ の 2 つのセグメントで構成されることを意味する．

アルゴリズム 3 にセグメントの生成過程を示す．BiSTM+TS 全体の生成過程は，アルゴリズム 3 をアル

[*4] 式 (4)，(5) の導出は，Du ら [10] の式 (6)，(9) の導出において，ρ に関する統計量を y に関する統計量に置き換え，t と n に関して，セグメントごとではなく対応セグメント集合ごとに集計すればよい．詳細は文献 [10] を参照されたい．

アルゴリズム 3 セグメントの生成過程
1: **for** each document d_i^ℓ ($i \in \{1,\ldots,D\}$) **do**
2: choose $\pi_i^\ell \sim \mathrm{Beta}(\lambda_0,\lambda_1)$
3: **for** each sentence u_{ih}^ℓ ($h \in \{1,\ldots,U_i^\ell\}$) **do**
4: choose $\rho_{ih}^\ell \sim \mathrm{Bernoulli}(\pi_i^\ell)$
5: **end for**
6: $\boldsymbol{s_i^\ell} = \mathrm{concatenate}(\boldsymbol{u_i^\ell}, \boldsymbol{\rho_i^\ell})$
7: **end for**

ゴリズム 2 のステップ 7 と 8 の間に挿入したものである．ここで，文書対 d_i における 2 つの文書 d_i^e と d_i^f は，互いに独立に分割されることに注意されたい．BiSTM+TS では，各文書 d_i^ℓ に対して，確率 π_i^ℓ でトピックが変わると仮定する．この仮定に基づき，まず，各文書 d_i^ℓ に対して，π_i^ℓ が λ_0 と λ_1 をパラメータとするベータ分布から生成される（ステップ 2）．その後，各文 u_{ih}^ℓ ($h \in \{1,\ldots,U_i^\ell\}$) に対して，$\rho_{ih}^\ell$ が π_i^ℓ をパラメータとするベルヌーイ分布から生成される（ステップ 4）．最後に，セグメント $\boldsymbol{s_i^\ell}$ が，$\boldsymbol{\rho_i^\ell}$ にしたがって文を結合させることにより生成される（ステップ 6）．

4.1 BiSTM+TS における推定

BiSTM+TS における推定では，BiSTM におけるサンプリング（($\boldsymbol{z}$, $\boldsymbol{\delta}$) と $\boldsymbol{y}$ のサンプリング）に $\boldsymbol{\rho}$ のサンプリングを加えた，3 種類のブロックのサンプリングを交互に行う．$\boldsymbol{\rho}$ のサンプリングは，Du ら [10] と同様に行う．$\boldsymbol{\rho}$ の事後分布は，以下の式 (9) および (10) に示されるとおり，文 u_{ih}^ℓ の直後に境界を設けることにより，1 つのセグメント s_m が 2 つのセグメント s_r と s_l に分割される（$\rho_{ih}^\ell = 1$）確率と，文 u_{ih}^ℓ の直後の境界を除くことにより，2 つのセグメント s_r と s_l が 1 つのセグメント s_m に統合される（$\rho_{ih}^\ell = 0$）確率で構成される．

$$p(\rho_{ih}^\ell = 1|\boldsymbol{\rho}^{\boldsymbol{\ell}-\rho_{ih}^\ell}, \boldsymbol{z^\ell}, \boldsymbol{w^\ell}, \boldsymbol{\delta^\ell}, \boldsymbol{\alpha}, a, b, \lambda_0, \lambda_1)$$
$$\propto \frac{\lambda_1 + c_{i1}^{\prime\ell}}{\lambda_0 + \lambda_1 + c_{i0}^{\prime\ell} + c_{i1}^{\prime\ell}} \mathrm{Beta}_K\left(\boldsymbol{\alpha} + \sum_{j=1}^{S_i^\ell} \boldsymbol{t'_{ij}}\right)$$
$$\prod_{j\in\{s_l,s_r\}} \frac{(b|a)_{t_{ij\cdot}^{\prime\ell}}}{(b)_{n_{ij\cdot}^{\prime\ell}}} \prod_{k=1}^{K} S(n_{ijk}^{\prime\ell}, t_{ijk}^{\prime\ell}, a) \quad (9)$$

$$p(\rho_{ih}^\ell = 0|\boldsymbol{\rho}^{\boldsymbol{\ell}-\rho_{ih}^\ell}, \boldsymbol{z^\ell}, \boldsymbol{w^\ell}, \boldsymbol{t'^\ell} \setminus \mathbb{T}', \boldsymbol{\alpha}, a, b, \lambda_0, \lambda_1)$$
$$\propto \sum_{\mathbb{T}'} \left(\frac{\lambda_1 + c_{i0}^{\prime\ell}}{\lambda_0 + \lambda_1 + c_{i0}^{\prime\ell} + c_{i1}^{\prime\ell}} \mathrm{Beta}_K\left(\boldsymbol{\alpha} + \sum_{j=1}^{S_i^\ell} \boldsymbol{t_{ij}^{\prime\ell}}\right) \right.$$
$$\left. \frac{(b|a)_{t_{i,s_m,\cdot}^{\prime\ell}}}{(b)_{n_{i,s_m,\cdot}^{\prime\ell}}} \prod_{k=1}^{K} S(n_{i,s_m,k}^{\prime\ell}, t_{i,s_m,k}^{\prime\ell}, a) \right) \quad (10)$$

ここで，$t_{ijk}^{\prime\ell}$ は，文書 d_i^ℓ 中のセグメント j のトピック k に関するテーブル数，$n_{ijk}^{\prime\ell}$ は，文書 d_i^ℓ 中のセグメント j 中のトピックが k である単語数，$c_{i0}^{\prime\ell}$ は，文書 d_i^ℓ 中で値が 0 である ρ の数，$c_{i1}^{\prime\ell}$ は，文書 d_i^ℓ 中で値が 1 である ρ の数である．また，「s_l と s_r の両方が $n_{ijk}^{\prime\ell} = 1$」かつ「s_l と s_r のどちらかあるいはその両方が $t_{ijk}^{\prime\ell} = 1$」となるトピックとセグメントの組合せを C' とすると，$\mathbb{T}' = \{t_{ijk}^{\prime\ell} : (k,j) \in C'\}$ である．式 (9) および (10) は，式 (4) および (5) において，y に関する統計量を ρ に関する統計量に置き換え，t と n に関して，対応セグメント集合ごとではなくセグメントごとに集計したものである．

本研究では，簡単のため，$\boldsymbol{\rho_i}$ のサンプリング時に d_i^e と d_i^f の独立性を仮定していることに注意されたい．つまり，$\boldsymbol{\rho}$ のサンプリングにおいて $\boldsymbol{y}$（他方の言語の対応セグメント）を考慮しない．$\boldsymbol{y}$ を考慮したトピック分割（バイリンガルなトピック分割への拡張）は今後の課題とする．ρ_{ih} のサンプリング後は，選択した行為（分割あるいは統合）に応じて，s_m，s_l および s_r の y を変更する．具体的には，s_m が s_l と s_r に分割される場合，$A(s_l) = A(s_m)$，$A(s_r) = A(s_m)$ となるように y を変更し，s_l と s_r が s_m に統合される場合，$A(s_m) = A(s_l) \cup A(s_r)$ となるように y を変更する．ここで，ρ_{ih} のサンプリングにより分割/統合を行う場合，厳密には，分割/統合後のセグメントに基づいて y を更新する必要があるが，本研究では，簡単のため，上記の近似を行うことに注意されたい．たとえば，$A(s_m) = \{s_m, s_x\}$（s_m と s_x は異なる言語のセグメント）であり，s_m が ρ のサンプリングにより s_l と s_r に分割されるとき，本来は s_l と s_r に基づいて y_{lx} と y_{rx} を更新する必要がある（$y_{lx} = 0$ あるいは $y_{rx} = 0$ となる場合もありうる）が，本研究では，暫定的に y_{lx} と y_{rx} をともに 1 と定める．よりよい近似の模索は今後の課題とする．

5. 評価実験

本章では，提案手法の有効性をパープレキシティと対訳対抽出における性能の観点で評価する．実験データは，Wikipedia 日英京都関連文書対訳コーパス[*5]を基に作成した，3,995 の日本語と英語のウィキペディア記事対からなる日英コンパラブルコーパスと 3,159 のフランス語と英語のウィキペディア記事対からなる仏英コンパラブルコーパスを使った．日英コンパラブルコーパスは，Wikipedia 日英京都関連文書対訳コーパスの日本語記事に対して，対応する英語記事を，英語版ウィキペディアの 2015 年 6 月 2 日のダンプファイル[*6]から言語間リンクに基づき収集することで作成した[*7]．ここで，Wikipedia 日英京都関連文書対訳コーパスは，本来，日本語記事の各文を人手で英語に翻訳した対訳コーパスであるが，この翻訳された英語記事は実験データに含まれていないことを特筆しておく．つまり，実験データは対訳コーパスではなく，文書単位で対応付いているコンパラブルコーパスである．仏英コンパラブルコーパスは，日英コンパラブルコーパスの英語記事に対応する

[*5] https://alaginrc.nict.go.jp/WikiCorpus/
[*6] http://dumps.wikimedia.org/enwiki/
[*7] 対応する英語記事が存在しない日本語記事は除いた．

フランス語記事を，フランス語版ウィキペディアの 2015 年 6 月 2 日のダンプファイル[*8]から Wikipedia の言語間リンクに基づき収集することで作成した[*9]．収集した英語およびフランス語記事からはオープンソースのスクリプト[*10]によりテキストを抽出した．日本語テキストは MeCab[*11]，英語およびフランス語テキストは TreeTagger[*12] [28] により形態素解析した後，機能語は除去し，その他の単語は原形化した．

対訳対抽出実験のために，日英の対訳対の正解セットを，Liu ら [17] に従い自動的に作成した．まず最初に，本来の Wikipedia 日英京都関連文書対訳コーパスに対して，GIZA++ [24] を用いて IBM モデル 4 で $p(w^e|w^f)$ および $p(w^f|w^e)$ を算出し，「$\hat{w}^e = \text{argmax}_{w^e} p(w^e|w^f = \hat{w}^f)$」かつ「$\hat{w}^f = \text{argmax}_{w^f} p(w^f|w^e = \hat{w}^e)$」を満たす単語ペア $(\hat{w}^e, \hat{w}^f)$ を抽出した．その後，日英コンパラブルコーパスの文書対に出現しない単語ペアを除き，残った単語対を正解セットとした．日英対訳対抽出実験では，正解セット中の全日本語単語 7,930 に対して対訳語獲得を行った．

仏英の対訳対の正解セットは，Gouws ら [12] や Coulmance ら [7] に倣い，Google 翻訳サービス[*13]を用いて自動的に作成した．まず最初に，仏英コンパラブルコーパスのフランス語テキストに出現する単語を Google 翻訳で英語に翻訳し，仏英の単語ペアを作成した．その後，仏英コンパラブルコーパスの文書対に出現しない単語ペアを除いた．仏英対訳対抽出実験では，残った単語ペアのフランス語単語の中から，仏英コンパラブルコーパスにおいて頻度が高い上位 1,000 個のフランス語単語に対して対訳語の獲得を行った[*14]．

5.1 実験対象

評価実験では，提案モデル（BiSTM，BiSTM+TS）と従来のバイリンガルトピックモデル BiLDA を比較する．BiSTM では，ウィキペディア記事中の各セクションをセグメントとした．また，実験データにはセクション間の対応関係は付与されていないため，BiSTM および BiSTM+TS では $\boldsymbol{y}$ を推定した．

BiLDA における推定は，BiSTM 同様，ギブスサンプリングにより行った（文献 [18], [22], [33] 参照）．各モデルの推定では，まず，各変数を無作為に初期化した．z^{ℓ}_{ijm} には，1 以上 K 以下の整数を，δ^{ℓ}_{ijm}，$y_{ijj'}$ および ρ^{ℓ}_{ih} には，0 か 1 の値を無作為に割り当てた．その後，一連のサンプリングを 10,000 回繰り返した．ハイパーパラメータ $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}^{\ell}$ は，Vulić ら [31] に倣い，それぞれ，対称なパラメータ $\alpha_k = 50/K$，$\beta^{\ell}_w = 0.01$ を用いた．a，b，λ_0 および λ_1 は，Du ら [8], [10] に倣い，それぞれ，0.2，10，0.1，0.1 に，η_0 と η_1 は，0.2 に設定した[*15]．また，トピック数の影響を調べるため，K は，100，400，2,000 の 3 種類を試した．$K = 100$ および $K = 400$ は，Liu ら [17] に倣い採用した．$K = 2{,}000$ は，Vulić ら [31] において最高性能を示すトピック数であるため採用した．

対訳対抽出実験では，Vulić ら [31] の対訳対抽出手法（*Cue*）と Liu ら [17] の対訳対抽出手法（*Liu*）の 2 種類の手法を試した．両手法とも，まず，バイリンガルトピックモデル（BiLDA，BiSTM あるいは BiSTM+TS）により各単語のトピックを推定し，その後，推定したトピックに基づき確率 $p(w^e|w^f)$ を算出する．そして，$p(w^e|w^f)$ が高い単語対 (w^e, w^f) を抽出することにより対訳対の抽出を行う．*Cue* では，$p(w^e|w^f)$ は $\sum_{k=1}^{K} p(w^e|k)p(k|w^f)$ のとおり算出する．ここで，$p(k|w) \propto \frac{p(w|k)}{\sum_{k=1}^{K} p(w|k)}$ であり，$p(w|k) = \phi_{kw}$ である．*Liu* では，まず，推定したトピックに基づき，コンパラブルコーパスをトピックで対応付けた対訳コーパスに変換する．そして，変換後の対訳コーパスに対して，IBM モデル 1 により $p(w^e|w^f, k)$ を算出し，その後，確率 $p(w^e|w^f)$ を $\sum_{k=1}^{K} p(w^e|w^f, k)p(k|w^f)$ のとおり算出する．以降では，各対訳対抽出手法で用いるバイリンガルトピックモデルを括弧内で示す．たとえば，*Cue*(BiLDA) は，BiLDA を用いる *Cue* である．

5.2 実験結果

各モデルのテストデータに対する予測性能をテストセットパープレキシティで評価した．テストセットパープレキシティは，5 分割交差検定により求めた．このパープレキシティは，値が低いほどモデルの汎化性能が良いことを表す．結果を**表 2** に示す．表 2 より，BiSTM および BiSTM+TS は，BiLDA よりパープレキシティの観点で優れたモデルであることが分かる．

各モデルの対訳対抽出性能を N ベスト正解率（ACC_N）で評価した．N ベスト正解率とは，上位 N 個の対訳対候補に正しい対訳対が含まれる場合に正解とした正解率である．**表 3** に各モデルを用いた場合の ACC_1 と ACC_{10} を示す．表 3 より，*Cue* と *Liu* の両手法において，BiSTM や BiSTM+TS を用いた方が BiLDA を用いた場合よりも正解率が高いことが分かる．この差は，符号検定により有

[*8] http://dumps.wikimedia.org/frwiki/
[*9] 対応するフランス語記事が存在しない英語記事は除いた．
[*10] https://github.com/attardi/wikiextractor/
[*11] http://taku910.github.io/mecab/
[*12] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[*13] http://translate.google.com/
[*14] Google 翻訳サービスを利用した従来の評価 [7], [12] に倣い 1,000 単語対を評価対象にしたが，全単語ペア（12,976 単語対）に対しても，*Liu*($K = 2{,}000$) で，本実験を通じて得られる結論と同じ結論が得られることを確認している．具体的には，*Liu*(BiLDA)，*Liu*(BiSTM)，*Liu*(BiSTM+TS) の ACC_1 は，それぞれ，0.288，0.319，0.305 であった．
[*15] 本稿では，計算量を削減するため，a および b はあらかじめ定めた値に固定した．a，b のサンプリングによる推定は今後の課題とする．

意差水準 1%で有意であった．これより，BiSTM および BiSTM+TS は，より適切なトピックを単語に割り当てることにより，*Cue* と *Liu* の対訳対抽出性能を改善できることが分かる．

パープレキシティおよび対訳対抽出の実験から，セグメント間の対応関係が多言語文書集合のモデル化に役立つことが実験的に確認できる．また，これらの評価実験において，BiSTM+TS は BiSTM に近い性能を達成していることから，あらかじめセグメントに分割されていない多言語文書集合に対しても，セグメントを自動推定することで人手によるセグメントの代替となりうるため，セグメント間の対応関係を活用できることが確認できる．

表 2 と表 3 より，すべてのバイリンガルトピックモデルにおいて，トピック数が大きいほど性能が良いことが分かる．また，用いるバイリンガルトピックモデルにかかわらず，*Liu* の方が *Cue* より性能が良いことが分かる．BiLDA に関するこれらの傾向は，Liu ら [17] で報告されている．今回の評価実験を通じて，提案モデル（BiSTM, BiSTM+TS）においても BiLDA と同様の傾向があることが分かった．

表 2 テストセットパープレキシティ
Table 2 Test set perplexity.

コーパス	モデル	K=100	K=400	K=2,000
日英	BiLDA	693.6	530.7	479.9
	BiSTM	520.1	429.3	394.6
	BiSTM+TS	537.5	445.3	411.8
仏英	BiLDA	616.7	514.2	439.1
	BiSTM	510.8	425.9	379.4
	BiSTM+TS	541.2	456.3	386.6

表 3 対訳対抽出性能（正解率）
Table 3 Performance of translation extraction.

ACC_1				
コーパス	手法	K=100	K=400	K=2,000
日英	*Cue*(BiLDA)	0.024	0.056	0.101
	Cue(BiSTM)	0.055	0.112	0.184
	Cue(BiSTM+TS)	0.052	0.107	0.176
	Liu(BiLDA)	0.206	0.345	0.426
	Liu(BiSTM)	0.287	0.414	0.479
	Liu(BiSTM+TS)	0.283	0.406	0.467
仏英	*Cue*(BiLDA)	0.097	0.169	0.219
	Cue(BiSTM)	0.160	0.242	0.275
	Cue(BiSTM+TS)	0.156	0.223	0.257
	Liu(BiLDA)	0.582	0.662	0.715
	Liu(BiSTM)	0.621	0.701	0.742
	Liu(BiSTM+TS)	0.616	0.695	0.732
ACC_{10}				
コーパス	手法	K=100	K=400	K=2,000
日英	*Cue*(BiLDA)	0.093	0.170	0.281
	Cue(BiSTM)	0.218	0.286	0.410
	Cue(BiSTM+TS)	0.196	0.274	0.398
	Liu(BiLDA)	0.463	0.550	0.603
	Liu(BiSTM)	0.531	0.625	0.671
	Liu(BiSTM+TS)	0.536	0.612	0.667
仏英	*Cue*(BiLDA)	0.379	0.494	0.556
	Cue(BiSTM)	0.462	0.548	0.580
	Cue(BiSTM+TS)	0.456	0.544	0.582
	Liu(BiLDA)	0.716	0.797	0.838
	Liu(BiSTM)	0.748	0.832	0.859
	Liu(BiSTM+TS)	0.753	0.826	0.852

6. 考察

6.1 セグメント間の対応付け

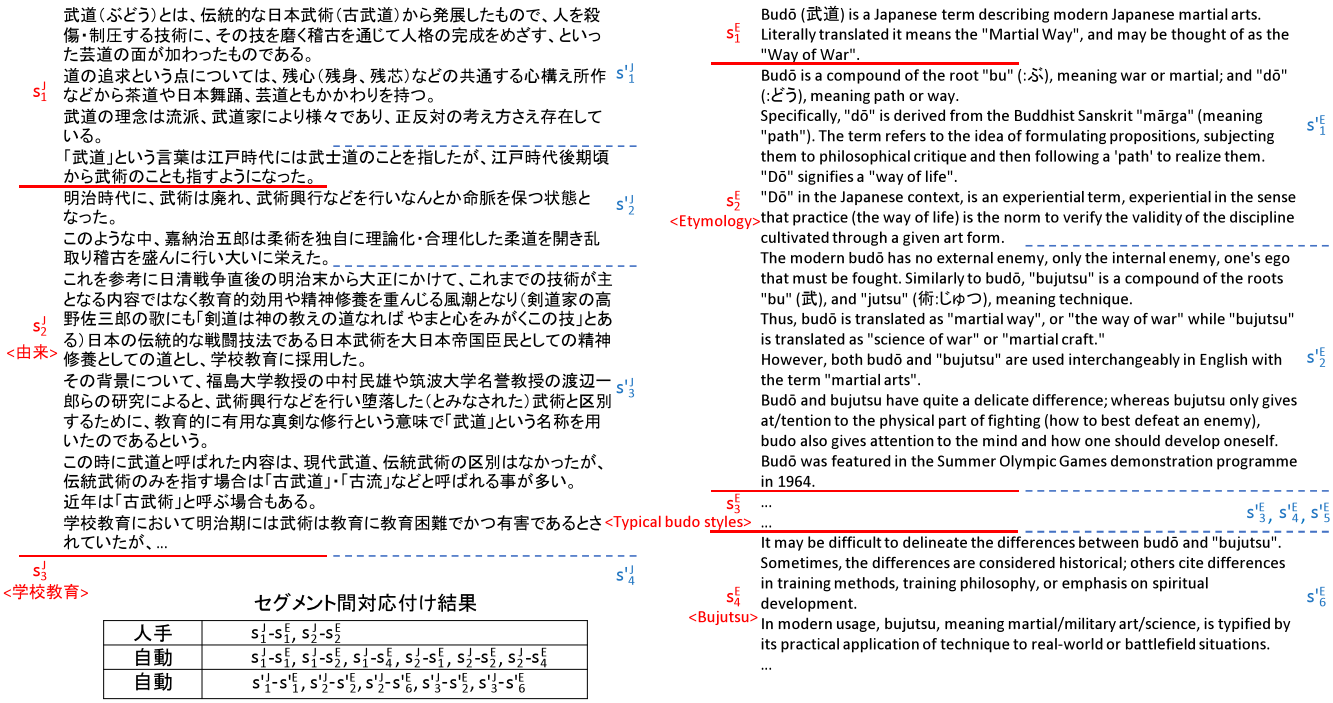
本節では，推定されたセグメント間の対応関係について考察する．セグメント分割性能の影響を除き，セグメント間の対応付け性能を評価するため，人手で設けられたセグメントを用いる BiSTM（K=2,000）により推定された $\boldsymbol{y}$ を評価した．具体的には，日英コンパラブルコーパスから無作為に抽出した 100 文書対に対して，セクション間の対応関係を人手で付与し，人手ラベルと BiSTM（K=2,000）による推定ラベルとを比較した．**表 4** に比較結果を示す．表 4 より，セグメント間の対応付け（$\boldsymbol{y}$）の正解率は，0.859（1,327/1,544）であった．

誤りの中で，人手ラベルが $y=0$ で推定ラベルが $y=1$ である事例の多く（121/174）は，セクション間で完全には対応付かないが，部分的に対応付いている事例であった．具体例を**図 6** に示す．図 6 は，「武道」に関する日英記事対である．図 6 では，人手で設けられたセグメント境界（セクション境界）が赤の実線で示され，セクション名が山括弧内に記載されている．また，提案手法により自動推定されたセグメント境界が青の点線で示されている．日本語および英語の人手で設けられたセクションは，それぞれ，「s_1^J，s_2^J，$\cdots$」，「s_1^E，s_2^E，$\cdots$」で示され，自動推定されたセグメントは，それぞれ，「$s_1'^J$，$s_2'^J$，$\cdots$」，「$s_1'^E$，$s_2'^E$，$\cdots$」で示されている．

図 6 において，英語のセクション「Bujutsu」（s_4^E）の内容は，日本語のセクション「由来」（s_2^J）の一部に書かれていることが分かる．本研究におけるセクション間の対応付けでは，セクション全体としてトピックが同じものを対応付けることを目指しているため，セクション間に部分的な対応関係があったとしても人手の判定は $y=0$ としている．一方，提案手法では，この部分的な対応関係がセグメント

表 4 セグメント間の対応付け性能
Table 4 Performance of segment-level alignment.

		人手ラベル	
		$y=1$	$y=0$
推定ラベル	$y=1$	195	174
	$y=0$	43	1132



セグメント間対応付け結果

人手	s_1^J-s_1^E, s_2^J-s_2^E
自動	s_1^J-s_1^E, s_1^J-s_2^E, s_1^J-s_4^E, s_2^J-s_1^E, s_2^J-s_2^E, s_2^J-s_4^E
自動	s'^J_1-s'^E_1, s'^J_2-s'^E_2, s'^J_2-s'^E_6, s'^J_3-s'^E_2, s'^J_3-s'^E_6

図 6 セグメント分割およびセグメント間対応付けの例
Fig. 6 Examples of segmentation and alignment.

間の対応関係に反映され，s_2^J と s_4^E は対応付く（$y=1$）と判定されていた．この対応関係は，本節で行った厳密な対応付けの評価では誤りと見なされるが，部分的な対応関係も多言語文書集合のモデル化に役立つ可能性があるため，この種の誤りは必ずしも悪い影響を与えるとは限らない．今後は，多言語文書集合のモデル化において部分的な対応関係の活用法を考え，提案モデルの改良につなげたい．

6.2 セグメント分割

本節では，推定されたセグメント境界について考察する．ウィキペディア記事中のセクション境界はトピックの変化点であるので，推定されたセグメント境界の質を調べるため，BiSTM+TS（K=2,000）で推定したセグメント境界のセクション境界に対する再現率[*16]を調べた．その結果，再現率は 0.727 であった．セクション境界は明らかなトピック変化を人手で示した箇所であることを考えると，BiSTM+TS は，極端なトピック変化を高い再現率で検出できたことが分かる．

表 5 に各モデルで用いたセグメント中の平均単語数を示す．BiSTM では，セクションをセグメントに用いたことを確認しておく．表 5 より，BiSTM+TS は，セクションよりも小さいセグメントに分割することが分かる．いい換えると，BiSTM+TS は，文書をセクションよりも細かく分割する．図 6 においても，日本語記事と英語記事とも

[*16] BiSTM+TS で推定したセグメント境界がセクション境界でなかったとしてもトピックの変化点になっていることがあるため，精度は評価しなかった．

表 5 1 セグメント中の平均単語数
Table 5 Average number of words in a segment.

モデル	日英コーパス		仏英コーパス	
	日本語	英語	フランス語	英語
BiSTM	264.7	245.1	276.4	259.3
BiSTM+TS	119.9	173.4	207.3	181.0

に，セクションの数よりも自動分割したセグメントの数の方が多いことが分かる．ウィキペディアでは，1 つのセクションに複数のトピックが記載されることが多々あることから，この分割は理に適っているように思われる．しかしながら，表 2 と表 3 より，自動推定したセグメントがウィキペディアのセクションよりも多言語文書集合のモデル化に有効であるとはいい難い．この理由の 1 つとして考えられるのは，BiSTM+TS の分割単位が細かすぎたことにより，モデル化の際にデータスパースネスの問題が生じてしまった可能性がある．また，表 5 より，BiSTM+TS では BiSTM に比べて，二言語間でセグメントに含まれる単語数の差が大きいことが分かる．これは，セグメントの情報を多言語文書集合のモデル化で使用する際は，言語間で粒度の揃ったセグメントを使うのが好ましい可能性があることを示唆している．言語間で粒度を揃えたセグメント分割への拡張は今後行っていきたい．

7. 関連研究

文書単位で対応付いたコンパラブルコーパスに対する多言語トピックモデルは，BiLDA（2 章参照）以外にも提

案されている．たとえば，Fukumasu ら [11] は，注釈付き画像データなどのマルチモーダルデータをモデル化するために提案されていた SwitchLDA [21] や Correspondence LDA [1] をコンパラブルコーパスのモデル化に適用した．彼らは，コンパラブルコーパスのための多言語トピックモデルとして，Correspondence LDA を対称化したモデルも提案している．また，Platt ら [26] は，PLSA や Principal Component Analysis に基づく単言語のトピックモデルを，対となる文書が近くなるように共通の多言語空間に写像することで，コンパラブルコーパスをモデル化している．Hu ら [15] は，文書間の対応関係に加えて，階層構造を持つ対訳辞書を事前知識として活用するモデルを提案している．ただし，これらのモデルはどれも，セグメント単位の対応関係を考慮しないことを特筆しておく．

文書単位で対応付いたコンパラブルコーパス以外の多言語コーパスに対する多言語トピックモデルも提案されている．たとえば，対訳文集合を解析対象とした，単語アライメントや機械翻訳のためのバイリンガルトピックモデルが提案されている [36], [37]．また，文書間の対応関係がない多言語コーパスに対して，言語間のずれを対訳辞書で埋めることで，言語共通のトピックを解析するモデルも提案されている [16], [20], [35]．対訳辞書に加えて対訳文も使うことで，言語間のずれを埋めるモデルも提案されている [3]．しかしながら，これらのモデルは対訳文や対訳辞書を事前に用意する必要があり，それらの資源をあらゆる言語対，分野で揃えるのは難しいという欠点がある．

単言語文書集合のモデル化においては，単語より大きく文書より小さい単位，つまり，セグメントのトピックを考慮するトピックモデルが提案されている．Du ら [8] は，各文書をセグメントの集合と見なし，各セグメントのトピック分布を，属する文書のトピック分布に基づいた Pitman–Yor 過程により生成するモデルを提案している．また，各文書をセグメントの列と見なすモデルも提案されている．Cheng ら [6] は，各文書に対してトピックの並びに関する分布を考え，セグメントのトピック列を考慮した文書のモデル化を行っている．Wang ら [34] は，セグメントのトピックの並びを，潜在変数の一次マルコフ連鎖でモデル化している．Du ら [9] は，各セグメントのトピック分布を，属する文書のトピック分布と直前のセグメントのトピック分布に基づき生成するモデルを提案している．また，Purver ら [27] は，発話を文書と単語の中間構造の単位とし，複数人による談話（話し言葉）に対するトピック分割とトピック推定のためのトピックモデルを提案している．持橋ら [19], [38] は，文書を意味的変化点により区切られるブロック列と見なし，ブロック単位で文書内部の意味的遷移を推定する文書モデルを提案している．ただし，これらのモデルは，今まで多言語の文書集合に拡張されていないことを特筆しておく．

8. まとめ

本稿では，文書を階層的にモデル化し，セグメント間の対応関係を考慮して多言語文書をモデル化する BiSTM を提案した．具体的には，対応関係のある文書に加えて，対応関係のあるセグメントのトピック分布も共有させることで，セグメント間の対応関係を反映したモデル化を行う．また，本稿では，BiSTM に Du ら [10] の教師なしトピック分割を組み込み，潜在トピックに加えてセグメント境界も推定する BiSTM+TS も提案した．実験を通じて，セグメント単位の対応関係を活用することで，パープレキシティおよび対訳対抽出の性能を改善できることを確認した．また，セグメントが与えられていない場合でも，セグメントを自動推定することで，提案モデルを活用できることを確認した．

今後は，他のデータセットや，言語横断文書分類 [22], [23], [26], [29] や言語横断情報検索 [32] などの多言語処理タスクで提案モデルの有効性を確認したい．また，本稿では BiLDA を拡張したが，セグメント間の対応関係は，7 章で記述した BiLDA 以外の多言語トピックモデルにおいても役立つ情報であると考えられる．BiLDA 以外の多言語トピックモデルの階層化（セグメント層の導入）は今後行いたい．

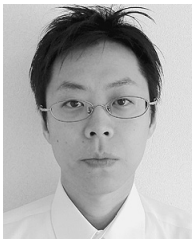
謝辞 本稿は，国際会議 The 54th Annual Meeting of the Association for Computational Linguistics で発表した論文 [30] に基づいて日本語で書き直し，提案手法の説明および評価を追加したものである．

参考文献

[1] Blei, D.M. and Jordan, M.I.: Modeling Annotated Data, *Proc. 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp.127–134 (2003).
[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] Boyd-Graber, J. and Blei, D.M.: Multilingual Topic Models for Unaligned Text, *Proc. 25th Conference on Uncertainty in Artificial Intelligence*, pp.75–82 (2009).
[4] Buntine, W. and Hutter, M.: A Bayesian View of the Poisson-Dirichlet Process (2012), available from ⟨http://arxiv.org/pdf/1007.0296.pdf⟩.
[5] Chen, C., Du, L. and Buntine, W.: Sampling Table Configurations for the Hierarchical Poisson-Dirichlet Process, *Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2011*, pp.296–311 (2011).
[6] Chen, H., Branavan, S., Barzilay, R. and Karger, D.R.: Global Models of Document Structure using Latent Permutations, *Proc. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp.371–379 (2009).
[7] Coulmance, J., Marty, J.-M., Wenzek, G. and

Benhalloum, A.: Trans-gram, Fast Cross-lingual Word-embeddings, *Proc. 2015 Conference on Empirical Methods in Natural Language Processing*, pp.1109–1113 (2015).
[8] Du, L., Buntine, W. and Jin, H.: A Segmented Topic Model Based on the Two-parameter Poisson-Dirichlet Process, *Machine Learning*, Vol.81, No.1, pp.5–19 (2010).
[9] Du, L., Buntine, W. and Jin, H.: Modelling Sequential Text with an Adaptive Topic Model, *Proc. 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp.535–545 (2012).
[10] Du, L., Buntine, W. and Johnson, M.: Topic Segmentation with a Structured Topic Model, *Proc. 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.190–200 (2013).
[11] Fukumasu, K., Eguchi, K. and Xing, E.P.: Symmetric Correspondence Topic Models for Multilingual Text Analysis, *Advances in Neural Information Processing Systems 25*, pp.1286–1294 (2012).
[12] Gouws, S., Bengio, Y. and Corrado, G.: BilBOWA: Fast Bilingual Distributed Representations without Word Alignments, *Proc. 32nd International Conference on Machine Learning*, pp.748–756 (2015).
[13] Hofmann, T.: Probabilistic Latent Semantic Indexing, *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.50–57 (1999).
[14] Hsu, L.C. and Shiue, P.J.-S.: A Unified Approach to Generalized Stirling Numbers, *Advances in Applied Mathematics*, Vol.20, No.3, pp.366–384 (1998).
[15] Hu, Y., Zhai, K., Eidelman, V. and Boyd-Graber, J.: Polylingual Tree-Based Topic Models for Translation Domain Adaptation, *Proc. 52nd Annual Meeting of the Association for Computational Linguistics*, pp.1166–1176 (2014).
[16] Jagarlamudi, J. and Daumé III, H.: Extracting Multilingual Topics from Unaligned Comparable Corpora, *Proc. 32nd European Conference on Advances in Information Retrieval*, pp.444–456 (2010).
[17] Liu, X., Duh, K. and Matsumoto, Y.: Topic Models + Word Alignment = A Flexible Framework for Extracting Bilingual Dictionary from Comparable Corpus, *Proc. 17th Conference on Computational Natural Language Learning*, pp.212–221 (2013).
[18] Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A. and McCallum, A.: Polylingual Topic Models, *Proc. 2009 Conference on Empirical Methods in Natural Language Processing*, pp.880–889 (2009).
[19] Mochihashi, D. and Matsumoto, Y.: Context as Filtering, *Advances in Neural Information Processing Systems 18*, pp.907–914 (2005).
[20] Negi, S.: Mining Bilingual Topic Hierarchies from Unaligned Text, *Proc. 5th International Joint Conference on Natural Language Processing*, pp.992–1000 (2011).
[21] Newman, D., Chemudugunta, C., Smyth, P. and Steyvers, M.: Statistical Entity-topic Models, *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.680–686 (2006).
[22] Ni, X., Sun, J.-T., Hu, J. and Chen, Z.: Mining Multilingual Topics from Wikipedia, *Proc. 18th International World Wide Web Conference*, pp.1155–1156 (2009).
[23] Ni, X., Sun, J.-T., Hu, J. and Chen, Z.: Cross Lingual Text Classification by Mining Multilingual Topics from Wikipedia, *Proc. 4th ACM International Conference on Web Search and Data Mining*, pp.375–384 (2011).
[24] Och, F.J. and Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, Vol.29, pp.19–51 (2003).
[25] Pitman, J. and Yor, M.: The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator, *The Annals of Probability*, Vol.25, No.2, pp.855–900 (1997).
[26] Platt, J., Toutanova, K. and tau Yih, W.: Translingual Document Representations from Discriminative Projections, *Proc. 2010 Conference on Empirical Methods in Natural Language Processing*, pp.251–261 (2010).
[27] Purver, M., Körding, K.P., Griffiths, T.L. and Tenenbaum, J.B.: Unsupervised Topic Modelling for Multi-Party Spoken Discourse, *Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp.17–24 (2006).
[28] Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees, *Proc. International Conference on New Methods in Language Processing*, pp.44–49 (1994).
[29] Smet, W.D., Tang, J. and Moens, M.-F.: Knowledge Transfer Across Multilingual Corpora via Latent Topics, *Proc. 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp.549–560 (2011).
[30] Tamura, A. and Sumita, E.: Bilingual Segmented Topic Model, *Proc. 54th Annual Meeting of the Association for Computational Linguistics*, pp.1266–1276 (2016).
[31] Vulić, I., Smet, W.D. and Moens, M.-F.: Identifying Word Translations from Comparable Corpora Using Latent Topic Models, *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp.479–484 (2011).
[32] Vulić, I., Smet, W.D. and Moens, M.-F.: Cross-Language Information Retrieval Models Based on Latent Topic Models Trained with Document-Aligned Comparable Corpora, *Information Retrieval*, Vol.16, No.3, pp.331–368 (2013).
[33] Vulić, I., Smet, W.D., Tang, J. and Moens, M.-F.: Probabilistic Topic Modeling in Multilingual Settings: An Overview of Its Methodology and Applications, *Information Processing & Management*, Vol.51, No.1, pp.111–147 (2015).
[34] Wang, H., Zhang, D. and Zhai, C.: Structural Topic Model for Latent Topical Structure Analysis, *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp.1526–1535 (2011).
[35] Zhang, D., Mei, Q. and Zhai, C.: Cross-Lingual Latent Topic Extraction, *Proc. 48th Annual Meeting of the Association for Computational Linguistics*, pp.1128–1137 (2010).
[36] Zhao, B. and Xing, E.P.: BiTAM: Bilingual Topic AdMixture Models for Word Alignment, *Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp.969–976 (2006).
[37] Zhao, B. and Xing, E.P.: HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation, *Advances in Neural Information Processing Systems 20*, pp.1689–1696 (2008).
[38] 持橋大地，菊井玄一郎：Gibbs Sampling による確率的テ

キスト分割と複数観測への拡張，言語処理学会第 12 回年次大会発表論文集，pp.212–215 (2006).



田村 晃裕 （正会員）

2005 年東京工業大学工学部情報工学科卒業．2007 年同大学院総合理工学研究科修士課程修了．2013 年東京工業大学大学院総合理工学研究科博士課程修了．日本電気株式会社，情報通信研究機構にて自然言語処理に関する研究に従事した後，2017 年より愛媛大学大学院理工学研究科助教．工学博士．言語処理学会，人工知能学会，ACL 各会員．

隅田 英一郎 （正会員）

1982 年電気通信大学大学院修士課程修了．1999 年京都大学博士（工学）．現在，国立研究開発法人情報通信研究機構フェロー，同 ASTREC 副センター長．日本翻訳連盟理事．アジア太平洋機械翻訳協会理事．機械翻訳，e ラーニングを研究．NLP，ASJ，ACL 各会員．