文書の階層構造を考慮したバイリンガルトピックモデル

田村 晃裕　隅田 英一郎
国立研究開発法人 情報通信研究機構
{akihiro.tamura, eiichiro.sumita}@nict.go.jp

1 はじめに

これまで，文書に隠れた潜在トピックを教師無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 [5] 参照）．

Bilingual Latent Dirichlet Allocation (BiLDA) [3, 4] 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）をモデル化する．具体的には，コンパラブルコーパスの特徴を利用し，対応関係がある文書のトピック分布を共通化することで，文書間の対応関係を反映したモデル化を行う．

一方で，ほとんどの文書は「文書—セグメント（例えば，段落やセクション）—単語」といった階層構造を持ち，文書より小さい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 に対応付く．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ないが，我々は，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べきであると考えた．

そこで，本研究では，BiLDA を拡張し，セグメント間の対応関係を捉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．BiSTM では，各文書をセグメント集合とみなし，「文書—セグメント—単語」の階層構造でモデル化する．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，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．

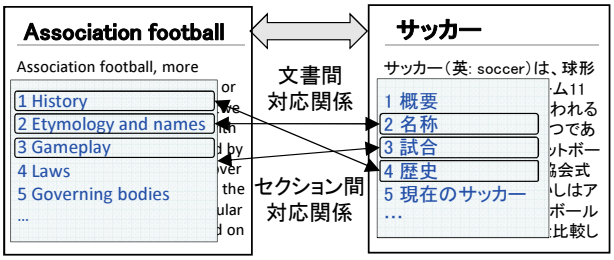

図 1: ウィキペディア記事の例

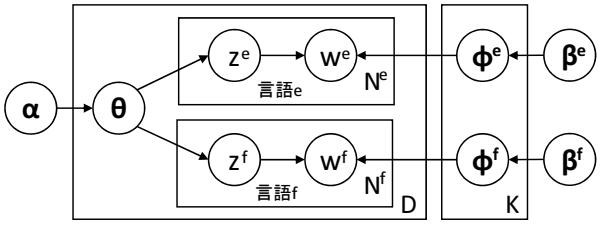

図 2: 従来モデル BiLDA のグラフィカルモデル

英語と日本語のウィキペディア記事から成るコンパラブルコーパスを使った実験により，BiSTM 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事を示し，対訳対抽出の性能も改善できることを示す．

2 従来モデル：BiLDA

BiLDA は，対応関係がある文書のトピック分布を共通化することで，多言語文書に潜む言語共通の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，言語 e と f で記述された D 個の文書対から成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 $(i \in \{1, ..., D\})$ における言語 e の文書を d_i^e，言語

アルゴリズム 1 BiLDA の生成過程

for each topic $k \in \{1, ..., K\}$ **do**
  for each language $l \in \{e, f\}$ **do**
    choose $\boldsymbol{\phi_k^l} \sim$ Dirichlet($\boldsymbol{\beta^l}$)
  end for
end for
for each document pair d_i ($i \in \{1, ..., D\}$) **do**
  choose $\boldsymbol{\theta_i} \sim$ Dirichlet($\boldsymbol{\alpha}$)
  for each language $l \in \{e, f\}$ **do**
    for each word w_{im}^l ($m \in \{1, ..., N_i^l\}$) **do**
      choose $z_{im}^l \sim$ Multinomial($\boldsymbol{\theta_i}$)
      choose $w_{im}^l \sim p(w_{im}^l | z_{im}^l, \boldsymbol{\phi^l})$
    end for
  end for
end for

f の文書を d_i^f と表記する．BiLDA では，各トピック $k \in \{1, ..., K\}$ は，言語 e の単語分布 $\boldsymbol{\phi_k^e}$ と言語 f の単語分布 $\boldsymbol{\phi_k^f}$ を持つ．そして，各単語分布 $\boldsymbol{\phi_k^l}$ ($l \in \{e, f\}$) は，$\boldsymbol{\beta^l}$ 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 $\boldsymbol{\theta_i}$ 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_i^e と d_i^f は共通のトピック分布 $\boldsymbol{\theta_i}$ を持つ．その後，文書 d_i^l の各単語位置 $m \in \{1, ..., N_i^l\}$ に対して，潜在トピック z_{im}^l が，$\boldsymbol{\theta_i}$ 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_{im}^l が，具体化された潜在トピック z_{im}^l と言語 l の単語分布 $\boldsymbol{\phi^l}$ に基づき，確率分布 $p(w_{im}^l | z_{im}^l, \boldsymbol{\phi^l})$ から生成される（ステップ 11）．

3 提案モデル：BiSTM

本節では，セグメント間の対応関係を考慮する 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_i^l は，S_i^l 個のセグメントで構成されているものとする（$d_i^l = \bigcup_{j=1}^{S_i^l} s_{ij}^l$）．BiSTM では，各言語で，セグメントのトピック分布（$\boldsymbol{\nu^e}$，$\boldsymbol{\nu^f}$）が，文書のトピック分布（$\boldsymbol{\theta}$）と単語のトピック（$\boldsymbol{z^e}$，$\boldsymbol{z^f}$）の間に挿入され，文書を階層的に生成する．また，セグメント間に対応関係があるかを示す二値変数 $\boldsymbol{y}$ 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．

まず，BiLDA 同様，各トピックに対して言語固有の

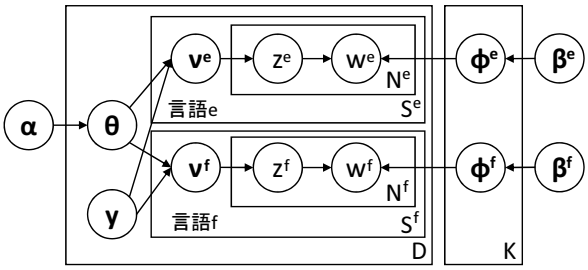

図 3: 提案モデル BiSTM のグラフィカルモデル

アルゴリズム 2 BiSTM の生成過程

for each topic $k \in \{1, ..., K\}$ **do**
  for each language $l \in \{e, f\}$ **do**
    choose $\boldsymbol{\phi_k^l} \sim$ Dirichlet($\boldsymbol{\beta^l}$)
  end for
end for
for each document pair d_i ($i \in \{1, ..., D\}$) **do**
  choose $\boldsymbol{\theta_i} \sim$ Dirichlet($\boldsymbol{\alpha}$)
  generate aligned segment sets $\boldsymbol{AS_i}$ = genAS($\boldsymbol{y_i}$)
  for each set $\boldsymbol{AS_{ig}}$ ($g \in \{1, ..., |\boldsymbol{AS_i}|\}$) **do**
    choose $\boldsymbol{\nu_{ig}} \sim$ PYP($a, b, \boldsymbol{\theta_i}$)
  end for
  for each language $l \in \{e, f\}$ **do**
    for each segment s_{ij}^l ($j \in \{1, ..., S_i^l\}$) **do**
      get index of s_{ij}^l in $\boldsymbol{AS_i}$: g =get_idx($\boldsymbol{AS_i}$,s_{ij}^l)
      for each word w_{ijm}^l ($m \in \{1, ..., N_{ij}^l\}$) **do**
        choose $z_{ijm}^l \sim$ Multinomial($\boldsymbol{\nu_{ig}}$)
        choose $w_{ijm}^l \sim p(w_{ijm}^l | z_{ijm}^l, \boldsymbol{\phi^l})$
      end for
    end for
  end for
end for

単語分布 $\boldsymbol{\phi_k^l}$ 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，$\boldsymbol{y_i}$ に基づき，対応関係のあるセグメント集合の集合 $\boldsymbol{AS_i}$ が生成される (ステップ 8)．ここで，$y_{ijj'} = 1$ は s_{ij}^e と $s_{ij'}^f$ 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．例えば，$d_i^e = \{s_{i1}^e, s_{i2}^e\}$, $d_i^f = \{s_{i1}^f, s_{i2}^f, s_{i3}^f\}$, y_{i11} と y_{i12} が 1，その他の y が 0 の時，ステップ 8 では，$\boldsymbol{AS_i} = \{\boldsymbol{AS_{i1}} = \{s_{i1}^e, s_{i1}^f, s_{i2}^f\}, \boldsymbol{AS_{i2}} = \{s_{i2}^e\}, \boldsymbol{AS_{i3}} = \{s_{i3}^f\}\}$ が生成される．続いて，$\boldsymbol{AS_i}$ 中の各セグメント集合 $\boldsymbol{AS_{ig}}$ ($g \in \{1, ..., |\boldsymbol{AS_i}|\}$) に対して，トピック分布 $\boldsymbol{\nu_{ig}}$ が，基底測

t^l_{ijk}	セグメント s^l_{ij} のトピック k に関するテーブル数
T^l_{ij}	セグメント s^l_{ij} の総テーブル数（$\sum_k t^l_{ijk}$）
n^l_{ijk}	セグメント s^l_{ij} 中のトピック k の単語数
N^l_{ij}	セグメント s^l_{ij} 中の総単語数（$\sum_k n^l_{ijk}$）
M^l_{kw}	トピックが k である言語 l の単語 w の数
$\boldsymbol{M^l_k}$	w 番目の要素が M^l_{kw} である $\lvert\boldsymbol{W^l}\rvert$ 次元ベクトル

表 1: 推定で用いる統計量

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 10）．ステップ 8 から 11 を通じて，$\boldsymbol{y}$ 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．

最後に，セグメント s^l_{ij} の各単語位置 $m \in \{1, ..., N^l_{ij}\}$ に対して，潜在トピック z^l_{ijm} 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 16），単語 w^l_{ijm} が，具体化された z^l_{ijm} と単語分布 $\boldsymbol{\phi^l}$ に基づき生成される（ステップ 17）．ここで，g は，セグメント s^l_{ij} を含むセグメント集合のインデックスであり，ステップ 14 で具体化されている．

3.1 推定

本節では，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．ここでは，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（例えば，$\boldsymbol{z} = \{\boldsymbol{z^e}, \boldsymbol{z^f}\}$）．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，ギブスサンプリングにより各隠れ変数を推定する．

提案の推定手法では，BiSTM の階層性（$\boldsymbol{\nu}$ と $\boldsymbol{z}$ の生成過程）を中華料理店過程で表現する．この過程により，$\boldsymbol{\theta}$，$\boldsymbol{\nu}$，$\boldsymbol{\phi}$ 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．したがって，$\boldsymbol{z}$ と $\boldsymbol{t}$ の 2 種類の変数のサンプリングを交互に繰り返し行うことで BiSTM の推定を行う．ただし，$\boldsymbol{y}$ が観測データとして与えられない場合，$\boldsymbol{y}$ もサンプリングにより推定する．表 1 に推定で用いる統計量をまとめる．$\boldsymbol{W^l}$ は，言語 l の単語集合である．

z^l_{ijm} 及び t^l_{ijk} の事後分布は，[1] と同様の導出により求められる：$p(z^l_{ijm} = k | \boldsymbol{z}^{-z^l_{ijm}}, \boldsymbol{w}, \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b, \boldsymbol{y})$

$$\propto \left(\frac{\alpha_k + \sum^{**} t}{\sum_{k=1}^{K} \left(\alpha_k + \sum^{**} t\right)} \left(b + a \sum^{*} T\right) \right)^{I(\sum^* n = 0)} \left(\frac{S(\sum^* n + 1, \sum^* t, a)}{S(\sum^* n, \sum^* t, a)} \right)^{I(\sum^* n > 0)} \frac{\beta^l_{w^l_{ijm}} + M^l_{kw^l_{ijm}}}{\sum_{w \in \boldsymbol{W^l}} (\beta^l_w + M^l_{kw})},$$

$p(t^l_{ijk} | \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{t}^{-t^l_{ijk}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b, \boldsymbol{y})$

$$\propto \frac{\Gamma(\alpha_k + \sum^{**} t)}{\Gamma\left(\sum_{k=1}^{K} (\alpha_k + \sum^{**} t)\right)} (b|a)_{\sum^* T} \, S(\sum^* n, \sum^* t, a).$$

ここで，$\sum^{**} t$ と $\sum^{*} t / T / n$ は，それぞれ，$\sum_{\boldsymbol{G} \in \boldsymbol{AS_i}} \sum_{j \in \boldsymbol{G}} t^{l_j}_{ijk}$ と $\sum_{j' \in AS_i(j)} t^{l_{j'}}_{ij'k} / T^{l_{j'}}_{ij'} / n^{l_{j'}}_{ij'k}$ を表す．$AS_i(j)$ は，セグメント j と対応関係のあるセグメント集合（例えば，3 節の例では，$AS_i(s^f_{i1}) = \boldsymbol{AS_{i1}}$），$l_j$ は，セグメント j の言語である．また，$(b|a)_n$ はポッホハマー記号，$S(n, m, a)$ は一般化された第二種スターリング数，$\Gamma(\cdot)$ はガンマ関数，$I(x)$ は条件 x を満たす時 1，満たさない時は 0 を返す関数である．

$y_{ijj'}$ は，s^e_{ij} と $s^f_{ij'}$ の類似度をパラメータとするベルヌーイ分布からサンプリングする：$y_{ijj'} \sim \text{Bernoulli}(\text{sim}(s^e_{ij}, s^f_{ij'}))$．$s^e_{ij}$ と $s^f_{ij'}$ の類似度は，s^e_{ij} と $s^f_{ij'}$ の TF-IDF トピックベクトルのコサイン類似度を用いる．ベクトルの各重みは，通常の単語 TF-IDF とは異なり，コーパスを文書集合ではなくセグメント集合と捉え，各セグメントを単語列ではなく潜在トピック列とみなして算出する．

4 実験

本節では，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．3,995 文書対からなる日英コンパラブルコーパスを実験データとして用いた．実験データは，Wikipedia 日英京都関連文書対訳コーパス [1] の日本語記事に対して，対応する英語記事を Wikipedia の言語間リンクに基づき収集することで作成した [2]．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．日本語テキストは MeCab[3]，英語テキストは TreeTagger[4] 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，対訳対の正解セットを，[2] にしたが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，IBM モデル 4 により $p(w^e|w^f)$ 及び $p(w^f|w^e)$ 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，コンパラブルコーパスの文書対

[1] https://alaginrc.nict.go.jp/WikiCorpus/
[2] 対応する英語記事が存在しない日本語記事は除いた．
[3] http://taku910.github.io/mecab/
[4] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

モデル	K=100	K=400	K=2000
BiLDA	693.6	530.7	479.9
BiSTM	522.8	431.1	398.7

表 2: テストセットパープレキシティ

手法	K=100	K=400	K=2000
Liu + BiLDA	0.206	0.345	0.426
Liu + BiSTM	0.282	0.413	0.475

表 3: 対訳対抽出性能（正解率）

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．対訳対抽出実験では，正解セット中の全日本語単語 7,930 に対する対訳語獲得を行った．

4.1 実験対象

実験では，提案モデル BiSTM と従来モデル 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 では $\boldsymbol{y}$ を推定した．

BiLDA の推定は，BiSTM 同様，ギブスサンプリングにより行った [3, 4, 5]．各モデルの推定では，各変数を無作為に初期化した後，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 $\boldsymbol{\beta}^l$ は，それぞれ，対象なパラメータ $\alpha_k = 50/K$，$\beta^l_w = 0.01$ を用い，a と b は，それぞれ，0.2，10 とした．また，トピック数の影響を調べるため，K は，100，400，2000 の 3 種類を試した．

対訳対抽出実験では，Liu らの対訳対抽出手法を用いた [2]．この手法では，まず，多言語トピックモデル（BiLDA あるいは BiSTM）により各単語のトピックを推定する．次に，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し，変換後の対訳コーパスに対して，IBM モデル 1 により，$p(w^e|w^f, k)$ を算出する．そして，確率 $p(w^e|w^f) = \sum_{k=1}^{K} p(w^e|w^f, k)p(k|w^f)$ が高い単語対 (w^e, w^f) を対訳対とする．

4.2 実験結果

表 2 に各モデルのテストセットパープレキシティを示す．このパープレキシティは，5 分割交差検定により求めた．表 2 より，BiSTM 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．

表 3 に各モデルを用いて抽出した対訳対の正解率を示す．表 3 より，BiSTM を用いた方が，BiLDA を用いた場合より正解率が高い．この差は，符号検定により有意差水準 1%で有意であった．これより，BiSTM は，より適切なトピックを単語に割り当てることで，対訳対抽出性能を改善できることが分かる．

表 2 と表 3 より，セグメント間の対応関係を捉えることで，多言語コーパスのモデル化性能を改善できることが実験的に確認できる．また，トピック数が大きいほど，性能が良いことも分かる．

5 おわりに

本研究では，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通化することで，セグメント間の対応関係を反映した．実験により，BiSTM は，パープレキシティ及び対訳対抽出の性能を改善できることを示した．今後は，他のデータセットや多言語処理タスクで提案モデルの有効性を確認したい．

参考文献

[1] Lan Du, Wray Buntine, and Huidong Jin. A Segmented Topic Model Based on the Two-parameter Poisson-Dirichlet Process. *Machine Learning*, Vol. 81, No. 1, pp. 5–19, 2010.

[2] Xiaodong Liu, Kevin Duh, and Yuji Matsumoto. Topic Models + Word Alignment = A Flexible Framework for Extracting Bilingual Dictionary from Comparable Corpus. In *Proc. CoNLL 2013*, pp. 212–221, 2013.

[3] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual Topic Models. In *Proc. EMNLP 2009*, pp. 880–889, 2009.

[4] Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. Mining Multilingual Topics from Wikipedia. In *Proc. WWW 2009*, pp. 1155–1156, 2009.

[5] Ivan Vulić, Wim De Smet, Jie Tang, and Marie-Francine Moens. Probabilistic Topic Modeling in Multilingual Settings: An Overview of Its Methodology and Applications. *Information Processing & Management*, Vol. 51, No. 1, pp. 111–147, 2015.