3H2-OS3-4

文書上の事象に基づいた潜在的トピック推定

Latent Topic Estimation based on Events in a Document

北島　理沙　　小林　一郎
Risa Kitajima　　Ichiro Kobayashi

お茶の水女子大学大学院人間文化創成科学研究科理学専攻
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

Recently, some latent topic model-based methods such as LSI, pLSI, and LDA have been widely used. However, they assign topics to words, therefore, the relationship between words in a document is unconsidered. In our previous study, we have proposed a latent topic extracting method which assigns topics to Events that represent the relationships between words based on dependency relation. Meanwhile, the definition of an Event was decided heuristically. In this paper, we reconsider how to define an Event grasping latent topics in a document, and compare proposed event types each other with a common document retrieval task. Moreover, we examine which constraint to define an Event works effectively in the proposed method.

1. はじめに

近年，文書上の潜在的トピックを扱う機会が増え，LSI（Latent Semantic Indexing）[Deerwester 90]，pLSI（probabilistic LSI）[Hofmann 99]，LDA（Latent Dirichlet Allocation）[Blei 03] などの潜在的意味解析手法が利用されるようになってきた．しかしこれらの手法において，トピックが割り当てられるのは単語であり，単語間の依存関係は考慮されていない．この問題に対して筆者らは，文書上の各事象をイベントとして定義し[*1]，文書をイベントの集合として扱うモデルを提案した [北島 10]．潜在的意味解析手法としては潜在的ディリクレ配分法（LDA）を用い，トピックの割り当て対象を単語からイベントに変更することで，文書検索，要約文生成課題において従来の手法よりも精度が高くなることを示した．先行研究 [北島 10] では，イベントの定義方法として文節の係り受け関係を利用したが，この定義は経験的に定めたものであり，イベントの定義に関しては精査する必要があると考える．従って本稿では，新たに 4 タイプのイベントの定義方法を提案し，共通の文書検索課題を通じて比較を行う．そしてその結果から，イベントという単位を構成する際にどのような条件が有効であるのかについて考察する．

2. 関連研究

従来の単語から他の対象に潜在的トピックの割り当て対象を変更して処理を行っている研究としては，鈴木らによる研究 [鈴木 10] がある．彼らは，潜在的ディリクレ配分法においてトピックの割り当て対象を単語列に変更したことによって，より柔軟なトピック割り当てが出来ることを報告している．単語の依存関係を利用した研究としては，藤村らによる研究 [藤村 05] や，松本らによる研究 [松本 04] がある．前者は，文節の n-gram による素性を用いることによって，評判分類における再現率が向上することを報告しており，後者は，単語の部分木パターンや系列パターンを素性として扱うことによって，文書分類の精度が向上することを報告している．これらの研究から，潜在的トピックの割り当て対象を単語以外のものにしても文書の持つ意味を捉えることができ，また，単語の依存関係を考慮することで文書分類の精度が向上することが示されている．文書上の単語に対してトピックを割り当てる場合，単語の出現頻度が等しい 2 つの文書は，その語の依存関係にかかわらず，同じトピック分布を持つと推定されてしまう．しかし，単語の出現頻度よりもむしろ語と語の関係性が文書を表わす特徴量として重要となる場合がある．例えば，評価分類をする場合では，何に対してどのような意見を持っているか，という情報が重要になると考えられる．以上のような理由に基づき，本研究では文書上のイベントを単位としたトピック割り当てを提案する．

連絡先: 北島理沙，お茶の水女子大学大学院人間文化創成科学研究科理学専攻情報科学コース小林研究室，〒 112-8610　東京都文京区大塚 2-1-1，03-5978-5708，kitajima.risa@is.ocha.ac.jp

*1　イベントの定義については，4 章で詳述する．

3. 潜在的ディリクレ配分法

本研究では，潜在的意味解析手法として，潜在的ディリクレ配分法を用いる．潜在的ディリクレ配分法とは，一つの文書に対して複数のトピックが存在すると想定した確率的トピックモデルであり，それぞれのトピックがある確率を持って文書上に生起するという考えの下，そのトピックの確率分布を導き出す手法である．

図 1 に，潜在的ディリクレ配分法のグラフィカルモデルを示す．各文書は，トピック分布 θ を持ち，文書上の各単語の位置について，θ に従ってまずトピック z が選ばれ，そのトピック z に対応する単語分布 ϕ に従って，その位置の単語 w が生成される．K はトピック数，D は文書数，N_d は文書 d 上の単語の出現回数を表わしており，トピック分布 θ は各文書ごとに生成され，単語分布 ϕ は各トピックごとに生成され，単語 w とその単語のトピックを表わす z は各単語の出現する位置ごとに生成される．また，α と β はハイパーパラメータであり，それぞれ，パラメータ θ が従うディリクレ分布のパラメータ，パラメータ ϕ が従うディリクレ分布のパラメータを示す．これらの変数の中で，実際に観測される変数は文書上に現れている単語 w であり，実用的には，この観測変数を用いて潜在変数の推定を行っている．潜在的ディリクレ配分法における文書の生成過程は，以下のような手順である．

1. 各トピック $k = 1,\cdots,K$ について：

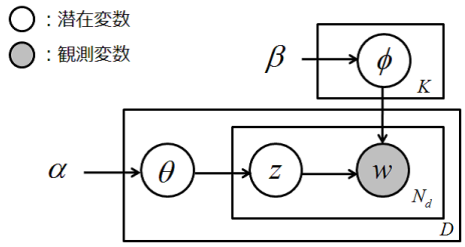

図 1: LDA のグラフィカルモデル

(a) ディリクレ分布に従って単語分布 ϕ_k を生成
$\phi_k \sim Dir(\beta)$

2. 各文書 $d = 1,\cdots,D$ について：

(a) ディリクレ分布に従ってトピック分布 θ_d を生成
$\theta_d \sim Dir(\alpha)$

(b) 文書 d における各単語 $n = 1,\cdots,N_d$ について：

i. 多項分布に従ってトピックを生成
$z_{dn} \sim Multi(\theta_d)$

ii. 多項分布に従って単語を生成
$w_{dn} \sim Multi(\phi_{z_{dn}})$

なお，ϕ_k はトピック k の単語分布，θ_d は文書 d のトピック分布，z_{dn} は文書 d の n 番目の単語の潜在的トピック，w_{dn} は文書 d の n 番目の単語を表わし，$Dir(\cdot)$ はディリクレ分布，$Multi(\cdot)$ は多項分布を表わす．トピック集合 Z と文書集合 W の完全尤度は，式 (1) で示される．ここで，$P(W|Z,\beta)$ と $P(Z|\alpha)$ は独立に扱うことができ，式 (2) と式 (3) によってそれぞれ表わされる．なお，V は語彙数，$\Gamma(\cdot)$ はガンマ関数を表わしている．

$$P(Z,W|\alpha,\beta) = P(W|Z,\beta)P(Z|\alpha) \quad (1)$$

$$P(W|Z,\beta) = \left(\frac{\Gamma(\beta V)}{\Gamma(\beta)^V}\right)^K \prod_{k=1}^{K} \frac{\prod_{w=1}^{V}\Gamma(N_{kw}+\beta)}{\Gamma(N_k+\beta V)} \quad (2)$$

$$P(Z|\alpha) = \left(\frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K}\right)^D \prod_{d=1}^{D} \frac{\prod_{k=1}^{K}\Gamma(N_{kd}+\alpha)}{\Gamma(N_d+\alpha K)} \quad (3)$$

トピック集合 Z の推定手法としては，変分ベイズ法 [Blei 03]，Collapsed 変分ベイズ法 [Teh 06]，ギブスサンプリング [Grifiths 04] などが提案されているが，ギブスサンプリングは十分な反復回数が得られるならば変分ベイズ法よりも高い精度でモデル推定を行えることが分かっており [Teh 06]，本研究でもギブスサンプリングによる推定を行うこととする．

ギブスサンプリングによって得られたサンプルから，各文書のトピック分布 θ と各トピックの単語分布 ϕ の予測分布を計算する．文書 d においてトピック k が生成される確率の推定量 $\hat{\theta}_d^k$，トピック k が選択されたときに単語 w が生成される確率の推定量 $\hat{\phi}_k^w$ は，それぞれ式 (4)，式 (5) によって求められる．

$$\hat{\theta}_d^k = \frac{N_{dk}+\alpha}{N_d+\alpha K} \quad (4)$$

$$\hat{\phi}_k^w = \frac{N_{kw}+\beta}{N_k+\beta V} \quad (5)$$

4. イベントの定義

イベントとは，文書上に存在している事象のことを指しており，2 つの単語の組として表現する．先行研究 [北島 10] では，文書上の係り受け関係から抽出される語の関係について経験的にルールを定め，イベントという単位を以下のように設定した．まず，文書に対して構文解析器 CaboCha[*2] を用いて文節の係り受け関係を取り出す．そして，係り受け関係にある 2 つの文節からそれぞれ単語を抽出し，(主語，述語)，(述語 1，述語 2) の条件を満たす組をイベントと定義する．主語には名詞，未知語が，述語には動詞，形容詞，形容動詞がそれぞれ該当する．(述語 1，述語 2) をイベントとして選んだ理由は，予備実験にて実際に抽出されたイベントと文書を見比べることによりその必要性を確認したこと，および，主語が省略されている文に対しては前者の条件を満たすイベントが抽出できないことによる．本稿では，このようなイベントの定義をイベントタイプと呼ぶ．また，上で示した先行研究にて用いたイベントタイプを event0 と定義する．

先行研究では，このように経験的に定めた 1 種類のイベントタイプを素性として取り扱い，文書検索課題や要約文生成課題を用いることによって，対象が文書であっても文であっても，単語を素性として扱う場合よりも高い精度で潜在トピックを推定することができることを示した．しかし，どのような単語の組み合わせをイベントという 1 つの単位として扱うかによって，潜在トピックの推定精度は左右されることが予想され，本研究においてこのイベントの定義について精査することは重要であると考えられる．したがって，本稿ではイベントの定義方法について検討を行い，具体的には，以下で説明するような文書上のイベントを表現するのに有効と思われる 4 つのタイプを設定して，実験により比較を行う．なお，以下の説明で挙げられている「自立語」は，動詞-自立，名詞，助動詞の「ない」，形容詞，連体詞，副詞のことを指すとし，未知語は名詞として扱うことにする．また，名詞の中でも，名詞–数，名詞–接尾，名詞–非自立に関しては，今回は名詞から除外することとした．

1. 文内の自立語の共起
同文内で共起する 2 つの語は，同文書内で共起する 2 つの語よりも関連性が高いと考え，1 文中で共起する 2 つの語を組とする．対象とするのは自立語であり，その総当たりを 1 文に対する素性とする．このイベントタイプを event1 と定義する．

2. 事象内の自立語の共起
event1 では，同文内で共起する語の組み合わせを全通り抽出しているが，接続詞や接続助詞によって複数の事象が 1 文に含まれることがある．例えば，「部屋はきれいだが，お風呂は汚い」という文について考えてみると，この文の中には「部屋はきれい」という事象と「お風呂は汚い」という事象が含まれていることが分かる．この文に対して，event1 によるイベント抽出を行うと，(部屋，汚い) や (お風呂，きれい) のような，元の文の持つ意味と異なる意味を持ったイベントが抽出されてしまう．この問題を回避するために，共起関係をとる段階を文よりもより意味を捉えることのできるような細かい単位にしたいと考え，文を接続詞と接続助詞で区切ってから，その区切られた範囲の中で共起関係をとる．event1 と同様に，対象とするのは自立語であり，その総当たりを 1 文

*2 http://chasen.org/ taku/software/cabocha/

に対する素性とする．このイベントタイプを event2 と定義する．

3. 係り受け関係にある自立語の共起
　共起関係を持つ 2 つの語の組み合わせの中には，直接的な関連性のないものも含まれることがあり，ときにそのような組み合わせはノイズとなることがあると考えられる．したがって，共起関係よりもさらに親密な関連性を持った組み合わせをとる必要があると考え，係り受け関係にある 2 つの自立語の共起を組とする．例えば,「朝食のパンはとても美味しかったです」という文に対しては,（パン，朝食）（パン，美味しい）（とても，美味しい）という 3 つのイベントが抽出される．このイベントタイプを event3 と定義する．

4. 経験的に定めた規則を満たす共起
　event3 では，係り受け関係にある全ての自立語の共起をイベントとして抽出した．しかし，この中には文の内容を捉えるにあたって重要でない組み合わせも存在すると考えられる．したがって，文の内容を捉えるために必要と考えられる品詞の組み合わせを経験的に定めることとし，具体的には,（名詞，名詞）,（名詞，形容詞）,（動詞，名詞）,（動詞，副詞）のいずれかを満たす係り受け関係にある語の組み合わせをイベントとして抽出する．また，イベント抽出の前処理として，サ変接続名詞と動詞「する」は接続させて 1 つの動詞として扱うという規則，および，動詞と助動詞「ない」は接続させて 1 つの動詞として扱うという規則を設けた．これは，単語の組み合わせ条件に対して自立語よりも詳細な品詞情報を考慮するにあたって，これらの規則を設けることが必要であると考えたためである．例えば，項目 3 で例に挙げた文に対しては,（パン，朝食）（パン，美味しい）という 2 つのイベントが抽出される．このイベントタイプを event4 と定義する．

5. イベントに基づいたトピック推定

文書検索において，各文書は文書を構成する単語とその重要度の積からなる文書ベクトルとして表現され，その重要度は索引となる単語の出現頻度を用いることが多い．しかし本研究では，イベントという単位で文書を扱うとするため，各文書に対してイベントを抽出し，文書群全体について索引となるイベントを決め，そのイベントの出現頻度を要素としたイベント－文書行列を作成する．そして，それに基づいてトピック推定を行う．

5.1 イベント－文書行列の作成

文書を単語集合として扱う場合，各文書について単語を抽出した後，その中から不要な単語を除去して単語文書行列を作成するための索引となる単語を決定する．このとき，ストップワードと呼ばれるどのような文書においても一般的に頻出する単語と，文書群において極端に出現頻度の少ない語は除去されることが多い．提案手法では，先行研究において前者のような除去すべき頻出イベントは見受けられなかった．これは，イベントを構成する各単語は不必要である機能語として捉えるべきであっても，イベントという単語の組にすることで機能語にも意味が付与され，結果的にどのイベントも文書を特徴づける素性として扱う必要性が出てくるためであると考えられる．一方，後者のような出現頻度の少ないイベントは非常に多く見受けられた．このことは，単語の組を一つの単位として扱うというイベントの性質から明らかであり，素性の持つ意味が単語の場合と異なるため，同様の処理では対応できない場合が存在する．具体的には，文書群において出現頻度が 1 であるイベントを全て除去してしまうと，文書内容の再現性の低い文書ベクトルが生成されてしまうことがある，ということが予備実験により確認されている．そこで，このことを踏まえ，それを除去してしまうと文書ベクトルの要素が消えてしまうようなイベントは，たとえ出現頻度が 1 であっても残し，文書としての再現性を保つことにした．本研究においても，先行研究と同様の手順を全てのイベントタイプに対して用い，イベント－文書行列を作成する．

5.2 トピック分布の推定

イベント－文書行列の作成後，潜在的ディリクレ配分法によってトピック推定を行う．本研究では，トピックの割り当て対象はイベントとなるため，各トピックはイベントの多項分布として表現される．また，クエリのトピック分布については，クエリに含まれる各イベントの持つトピック分布の総和とする．

6. 文書検索精度に基づくイベント評価実験

共通の文書検索課題を通じて，各イベントタイプにおける潜在トピック推定の性能を比較および評価する．具体的には，クエリの持つトピック分布と類似するトピック分布を持った文書を検索結果とし，検索結果の精度を調べることで，推定されたトピック分布が各文書の持つ意味を捉えられているかを確かめる．トピック分布の類似度判定指標としては，Jensen-Shannon 距離を適用する．

6.1 実験仕様

対象データとしては，人の意見や評価などの裏に隠れた潜在トピックを扱いたいと考え，楽天トラベル[*3]のホテル・施設に関するレビュー・評価データを用いた．レビューには,「部屋」や「立地」などの各対象につき 1（悪い）〜5（良い）の 5 段階評価があり対象と評価の関係性が保持されているため，本実験に適していると考える．レビューの長さに関しては，より多くのトピックを扱っている文書を対象にすべきであると考え，様々な対象に対して意見が述べられていると考えられる長さである，100 字以上のレビューを利用することにする．文書検索課題として使用するクエリは「部屋が良かった」とし，対象文書群は「部屋」の評価が 1 のレビューから無作為に選んだ 1000 件，5 のレビューから無作為に選んだ 1000 件の合計 2000 件とする．正解文書は，評価が 5 のレビュー 1000 件である．多くのレビューで「部屋」に関するコメントがされており，また，評価を 1 や 5 としているユーザは特に「部屋」についての意見を述べている可能性が高いと考え，上記のクエリ，対象文書群にて実験を行うとした．評価指標には，11 点平均適合率を使用する．

本実験では，4 章にて設定した各イベントタイプを用いた提案手法による文書検索精度の比較を行う．トピック数 k は，先行研究 [北島 10] において最も高い精度を示した値を用い，$k = 5$ とする．試行回数は 20 回とし，その平均をとる．LDA において潜在変数の推定を行うために用いているギブスサンプリングの反復回数は，先行研究の結果から 200 回とする．先行研究にて用いたイベントタイプである event0 についても同様の実験を行い，その結果を提案する 4 種類のイベントタイプによる実験結果と比較する．

*3　http://travel.rakuten.co.jp/

表 1: イベントタイプの比較

イベントタイプ	次元数	11 点平均適合率
event0	5198	0.6256
event1	84635	0.6536
event2	36916	0.8175
event3	12199	0.7901
event4	8408	0.7641

6.2 実験結果

表 1 に，各イベントタイプを提案手法に用いたときの次元数と 11 点平均適合率の結果を示す．提案した 4 つのイベントタイプは，先行研究よりも高い精度を示していることが分かる．一方で，その次元数はより大きくなり，特に，係り受け関係を用いた event3 や event4 よりも，共起関係を用いた event1 や event2 において高次元となった．11 点平均適合率の値が最も高いのは event2 であり，本研究で提案したイベントタイプの中で最も精度が低くなったのは，event1 であった．

6.3 考察

実験結果から，イベントを構成する際にどんな条件が精度の向上に役立つのかを調査する．具体的には，どの単位において共起関係をとると良いのか，係り受け関係を考慮した方が良いのかどうか，そして，経験則の効果について考察を行う．

まず，event1 と event2 を比較することで，共起関係をとる単位について考察する．event2 は event1 と比べて次元数が半分以下に抑えられながらも，高い精度を示している．このことから，単語の共起関係を利用するときに接続詞や接続助詞で文を区切ってから組み合わせを取ることによって高い精度を見込めることが確認できた．また，次元数は半分以下となっていることから，それだけ役に立たないノイズとなる素性が event1 では存在していることが分かる．従って，共起関係をとる単位としては，接続詞などで区切った少し狭い範囲を対象とすることが，より精度の向上につながるのではないかと考える．

次に，event2 と event3 を比較することで，係り受け関係を考慮した方が良いのかどうかについて考察する．ここで event2 を比較対象として用いているのは，共起関係を用いる際には文内の共起よりも事象内の共起を考慮した方が良いということが前の考察にて分かったからである．event3 は，event2 に次いで 2 番目に高い結果を出しており，その次元数は 3 分の 1 程度に抑えられている．その精度の差は僅差であることから，係り受け関係を素性に利用することは大いに意味があると考えられる．

最後に，event0，event3，event4 を比較することで，経験則の効果について考察する．event0 で用いた経験則を経験則 1，event4 で用いた経験則を経験則 2 と呼ぶことにする．また，event3 は全係り受け関係を抽出したものであり，経験則を導入していない．event0，event4 は，重要度が高そうな品詞の組み合わせを経験則として設定したものの，event3 と比べてみると精度は低いという結果になった．一方で，その次元数は少ないという利点もある．また，event0 と event4 を比較すると，event0 は次元数が小さい一方で精度は低く，event4 は精度が高い一方で次元数が大きい，という精度と次元数のトレードオフの関係が見られた．従って，今回提案した経験則においてどの経験則が効いてくるのか，という判断は難しく，また，経験則を導入した方が良いのかどうかという点に関しては，経験則の使用，不使用の両者の間にも見られる同様のトレードオフの関係，および，経験則の設定コスト次第と考える．

7. おわりに

本研究では，先行研究におけるイベントの定義についての再検討として，4 章にて 4 つのイベントタイプを提案した．そして，6 章において文書検索を用いた性能評価実験を行い，イベントを構成するにあたって，どんな条件が提案する eventLDA において効いてくるのか，ということを調査した．

今後の課題としては，潜在トピック推定対象が文の場合についても提案したイベントタイプを用いて実験を行い，さらなる知見を得たいと考えている．それにより，文書のもつ潜在トピックと文のもつ潜在トピックの性質の違いや，イベントタイプの違いによる影響について深く知ることができるのではないかと考える．また，様々なデータセットやクエリを用いて実験を行い，考察を行っていきたいと考えている．

謝辞

本研究では，楽天株式会社の許諾を頂き “楽天トラベル” のデータを利用させて頂きました．ここに深く感謝の意を表します．

参考文献

[Blei 03] D. M. Blei, A. Y. Ng, and M. I. Jordan: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003).

[Deerwester 90] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman: Indexing by Latent Semantic Analysis, *Journal of the American Society of Information Science*, Vol. 41, No. 6, pp. 391–407 (1990).

[Grifiths 04] T. Grifiths and M. Steyvers: Finding scientific topics, *Proc. of the National Academy of Sciences*, Vol.101, pp. 5228–5235 (2004).

[Hofmann 99] T. Hofmann: Probabilistic Latent Semantic Indexing, *Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57 (1999).

[Teh 06] Y. W. Teh, D. Newman, and M. Welling: A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation, *Advances in Neural Information Processing Systems Conference*, Vol.19, pp. 1353–1360, (2006).

[北島 10] 北島理沙，小林一郎：文書内の事象を対象にした潜在的トピック抽出手法の提案とその応用，言語処理学会年次大会（2010）.

[桜井 04] 桜井俊彦，内海彰：情報検索のためのクエリに基づく文書自動要約，言語処理学会年次大会発表論文集，Vol. 10, pp. 265–268 (2004).

[鈴木 10] 鈴木康広，上村卓史，喜田拓也，有村博紀：潜在的ディリクレ配分法の単語列への拡張，第 2 回データ工学と情報マネジメントに関するフォーラム，I-6, (2010).

[藤村 05] 藤村滋，豊田正史，喜連川優：文の構造を考慮した評判抽出手法，電子情報通信学会第 16 回データ工学ワークショップ，6C-i8, (2005).

[松本 04] 松本翔太郎，高村大也，奥村学：単語の系列及び依存木を用いた評価文書の自動分類，情報科学技術フォーラム一般講演論文集，Vol. 3, No. 2, pp. 213–214, (2004).