# 文書内の事象を対象にした潜在的ディリクレ配分法による要約

北島　理沙†　　小林　一郎††

† お茶の水女子大学理学部情報科学科　〒 112–8610 東京都文京区大塚 2-1-1
†† 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　〒 112–8610 東京都文京区大塚 2-1-1
E-mail: †{g0720520,koba}@is.ocha.ac.jp

**あらまし**　近年，文書内のトピックを推定する手法として，LSI，pLSI，LDA といった潜在的意味解析手法が利用されているが，トピックは単語に割り当てられ，語の関係について考慮されないという問題がある．そこで本稿では，係り受け関係に基づいた語の関係をイベントという単位で扱い，イベントにトピックを割り当てる潜在的トピック抽出手法を提案する．また，その応用として，本手法により潜在的な意味に基づいた要約が生成できることを示す．
**キーワード**　イベント抽出，潜在的ディリクレ配分法，文書分類，複数文書要約

# Summarization using Latent Dirichlet Allocation based on Events in a Document

Risa KITAJIMA† and Ichiro KOBAYASHI††

† Department of Information Science, Faculty of Sciences, Ochanomizu University
2-1-1 Ohtsuka, Bunkyo-ku, Tokyo, 112–8610 Japan
†† Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University
2-1-1 Ohtsuka, Bunkyo-ku, Tokyo, 112–8610 Japan
E-mail: †{g0720520,koba}@is.ocha.ac.jp

**Abstract**　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 this paper, we propose a latent topic extracting method which assigns topics to events that represent the relationships between words based on dependency relation. We also show that our proposed method can generate a document summary based on a latent topic.
**Key words**　Event Extraction, Latent Dirichlet Allocation, Document Classification, Multi-Document Summarization

## 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）[1]，pLSI（probabilistic LSI）[2]，LDA（Latent Dirichlet Allocation）[3] 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．そこで本研究では，文書上の各事象をイベントとして定義し(注1)，文書をイベントの集合として扱うモデルを提案する．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．提案手法の性能を検証するための実験として，まず，実際に抽出されたトピックに対応するイベントの分布から，提案手法によって潜在的トピックがどのように抽出されるかを調べる．次に，共通の文書検索課題を通じて，従来の単語にトピックを割り当てる手法と比較をすることで，提案手法が文書に対して潜在的トピックを推定でき，文書検索にも有用であることを示す．提案手法の性能検証後，トピック推定の対象を文書から文に置き換え，近年盛んになっているクエリに特化した要約 [4] を対象とし，提案手法を応用した要約文生成を示す．

以下，本稿では，2 章では関連研究，3 章では潜在的ディリクレ配分法，4 章ではイベントにトピックを割り当てる提案手法について説明する．5 章ではトピック抽出実験について，6 章では文書検索を用いた実験について，7 章ではテキスト要約を用いた実験についてまとめる．最後に，8 章で本研究のまと

(注1)：イベントの定義については，3 章で詳述する．

めと今後の課題について述べる．

## 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 [5] 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 [6] や，松本らによる研究 [7] 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．

また，テキスト要約に関する研究としては，従来の基本的な重要文抽出法以外に潜在的意味解析手法を用いた手法が提案されている [8] [9]．これらにおいては，対象が文書である場合と同様にして文のトピック分布が推定され，それに基づいた要約文が生成される．本研究でも，提案手法をテキスト要約に用いることで，提案手法が対象を文としたときの潜在的トピック推定にも有効であることを示す．

## 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 $\theta$ を持ち，文書上の各単語の位置について，$\theta$ 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 $\phi$ 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 $\theta$ は各文書ごとに，単語分布 $\phi$ は各トピックごとに，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，$\alpha$ と $\beta$ はハイパーパラメータであり，それぞれ，パラメータ $\theta$ が従うディリクレ分布のパラメータ，パラメータ $\phi$ が従うディリクレ分布のパラメータを示す．実際に観測される変数は単語 $w$ であり，実用的には，この観測変数から潜在変数の推定を行う．

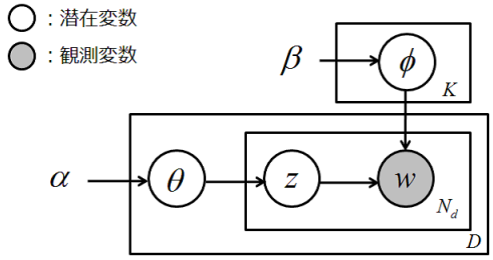


図 1　LDA のグラフィカルモデル

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：
  (a) ディリクレ分布に従って単語分布 $\phi_k$ 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 $\theta_d$ 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
      $w_{dn} \sim Multi(\phi_{z_{dn}})$

$\phi_k$ ： トピック $k$ の単語分布
$\theta_d$ ： 文書 $d$ のトピック分布
$z_{dn}$ ： 文書 $d$ の $n$ 番目の単語の潜在的トピック
$w_{dn}$ ： 文書 $d$ の $n$ 番目の単語
$Dir(\cdot)$ ： ディリクレ分布
$Multi(\cdot)$ ： 多項分布

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．

$$P(Z, W|\alpha, \beta) = P(W|Z, \beta)P(Z|\alpha) \tag{1}$$

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

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

トピック集合 $Z$ の推定手法としては，変分ベイズ法 [3]，Collapsed 変分ベイズ法 [10]，ギブスサンプリング [11] 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 [10]，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．式 (4) に，潜在的ディリクレ配分法におけるギブスサンプリングの更新式を示す．

$$P(z_i|z_{\backslash i}, W) \propto \frac{p(w|z)p(z)}{p(w_{\backslash i}|z_{\backslash i})p(z_{\backslash i})} = \frac{(n^v_{\backslash i,j}+\beta)(n^d_{\backslash i,j}+\alpha)}{(n^{\cdot}_{\backslash i,j}+W\beta)(n^d_{\backslash i,\cdot}+T\alpha)} \quad (4)$$

ここで，$z_{\backslash i}$ は，トピック集合 $Z$ からトピック $z_i$ を除いたものを表わしている．また，$n^v_{\backslash i,j}$，$n^d_{\backslash i,j}$，$n^{\cdot}_{\backslash i,j}$，$n^d_{\backslash i,\cdot}$ はそれぞれ位置 $i$ の情報を除外した場合の，トピック $j$ から単語 $v$ が生成された頻度，文書 $d$ においてトピック $j$ が割り当てられた頻度，コーパス全体においてトピック $j$ が割り当てられた頻度，文書 $d$ において単語が生成された頻度を表わしている．

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 $\theta$ と各トピックの単語分布 $\phi$ 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 $\hat{\theta^k_d}$，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 $\hat{\phi^w_k}$ は，それぞれ式 (5)，式 (6) によって求められる．

$$\hat{\theta^k_d} = \frac{N_{dk}+\alpha}{N_d+\alpha K} \quad (5)$$

$$\hat{\phi^w_k} = \frac{N_{kw}+\beta}{N_k+\beta V} \quad (6)$$

## 4. 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．

### 4.1 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，何が起こったか，誰がどのように感じたか，などの出来事を表わすような単語の組として表現する．その抽出方法について述べる．

まず，文書に対して構文解析器 CaboCha(注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タイプのイベントが抽出できないことによる．

### 4.2 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，予備実験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．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する．

### 4.3 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．

## 5. トピック抽出実験

ここで，対象文書群に存在する潜在的トピックを抽出する実験を行い，実際に抽出されるトピックに対応するイベントの分布から，提案手法によって抽出されるトピックの特徴を調べる．

### 5.1 実 験 仕 様

対象データとしては，語と語の関係性がより重要な役割を果たすような，意見や評価を含んだ文書が良いと考え，楽天トラベル(注3)のホテル・施設に関する評価・レビューを用いた．レビューには様々な長さのものがあるが，文書中により多くのトピックが扱われている文書を対象にした方が本実験に適していると考え，様々なジャンルに対して評価が行われていると考えられる 100 字以上のレビューを利用することにする．対象文書数としては，その中から無作為に選んだ 2000 件とする．与えるトピック数 $k$ は $k = 50$ とし，推定手法として用いるギブスサンプリングの反復回数は 500 回とした．ここでは，対数尤度の変化にも着目し，ギブスサンプリングによる結果がどの程度の反復回数をもって収束するかということについても，調査を行う．

(注2)：http://chasen.org/ taku/software/cabocha/

(注3)：http://travel.rakuten.co.jp/

### 5.2 実験結果

表 1 に，抽出されたトピックを表わす代表的なイベントを示す．ここでは，トピック $k$，イベント $e$ に対して $P(e|k)$ の大きい順に並べ，上位 5 位までを表示している．

表 1　抽出されるトピックの例

| topic | イベント |
|---|---|
| 0 | （食事，美味しい）（心，こもる）（温泉，良い）（感じる，不便）（サービス，良い） |
| 1 | （ホテル，利用する）（部屋，ない）（ホテル，ない）（出張，利用する）（窓，開ける） |
| 2 | （思う，宿泊する）（ホテル，言う）（価格，安い）（旅館，古い）（おいしい，食べる） |
| 3 | （対応，丁寧）（対応，親切）（嬉しい，思う）（欲，言う）（ホテル，利用する） |
| 4 | （部屋，狭い）（風呂，入る）（すごい，良い）（残念，無い）（仕方ない，思う） |
| 5 | （思う，良い）（できる，言う）（雰囲気，良い）（お願いする，思う）（頂ける，おいしい） |
| 6 | （温泉，できる）（無料，できる）（風呂，利用する）（思う，利用する）（部屋，戻る） |
| 7 | （部屋，ない）（場所，便利）（思う，良い）（窓，見える）（見る，思う） |
| 8 | （思う，利用する）（ホテル，利用する）（部屋，広い）（できる，利用する）（音，聞こえる） |
| 9 | （いい，思う）（お部屋，きれい）（値段，考える）（部屋，普通）（プラン，できる） |
| 10 | （風呂，できる）（部屋，広い）（思う，利用する）（部屋，できる）（景色，楽しむ） |
| 11 | （建物，古い）（風呂，入る）（ありがたい，大変）（ホテル，到着する）（おいしい，大変） |
| 12 | （音，聞こえる）（感じ，よい）（部屋，感じる）（いう，思う）（バス，広い） |
| 13 | （雰囲気，良い）（チェックイン，遅い）（朝食，美味しい）（風呂，広い）（お湯，良い） |
| 14 | （料金，安い）（立地，よい）（いい，思う）（値段，安い）（宿，泊まる） |
| 15 | （いい，思う）（プラン，ない）（人，多い）（言う，思う）（良い，綺麗） |
| 16 | （感じる，思う）（感じる，不便）（思う，利用する）（ホテル，良い）（言う，出る） |
| 17 | （思う，良い）（ホテル，泊まる）（ホテル，良い）（大変，良い）（利用する，良い） |
| 18 | （温泉，入る）（ホテル，感じる）（浴槽，行く）（言う，良い）（施設，思う） |
| 19 | （感じ，良い）（お願いする，思う）（ホテル，いい）（できる，良い）（思う，利用する） |
| 20 | （ホテル，利用する）（声，かける）（しれる，思う）（よい，わかる）（ホテル，いう） |
| 21 | （ホテル，できる）（できる，良い）（部屋，きれい）（ホテル，泊まる）（ホテル，思う） |
| 22 | （いい，思う）（思う，泊まる）（荷物，持つ）（広い，良い）（感じ，よい） |
| 23 | （過ごせる，快適）（近い，便利）（駅，近い）（安い，泊まれる）（部屋，過ごせる） |
| 24 | （ホテル，紹介する）（思う，利用する）（ホテル，探す）（朝食，とる）（ホテル，見つける） |
| 25 | （駅，近い）（ビジネス，普通）（気持ち，いい）（写真，撮る）（できる，よい） |
| 26 | （くだ，さる）（思う，利用する）（掃除，行き届く）（種類，豊富）（お部屋，広い） |
| 27 | （部屋，入る）（ほしい，思う）（印象，良い）（タバコ，臭い）（気，遣う） |
| 28 | （残念，思う）（風呂，広い）（お湯，出る）（部屋，広い）（対応，良い） |
| 29 | （ホテル，思う）（思う，利用する）（ホテル，泊まる）（ホテル，思える）（部屋，持つ） |
| 30 | （頂く，美味しい）（子供，いる）（大変，美味しい）（お部屋，広い）（子供，小さい） |
| 31 | （宿，思う）（風呂，入る）（行く，思う）（家族，宿泊する）（口コミ，見る） |
| 32 | （部屋，用意する）（いう，行く）（チェックイン，遅い）（過ごす，快適）（立地，良い） |
| 33 | （部屋，入る）（よい，思う）（人，いる）（ホテル，出る）（客，多い） |
| 34 | （対応，良い）（感じ，良い）（残念，良い）（広い，綺麗）（部屋，綺麗） |
| 35 | （お湯，熱い）（部屋，広い）（サービス，ない）（残念，非常）（外，出る） |
| 36 | （思う，利用する）（思う，便利）（広い，清潔）（広い，十分）（機会，利用する） |
| 37 | （フロント，言う）（仕方，ない）（苦情，言う）（部屋，使える）（ホテル，思う） |
| 38 | （立地，良い）（気持ち，良い）（思う，良い）（思う，欲しい）（駅，近い） |
| 39 | （荷物，預かる）（アウト，預かる）（チェック，預かる）（行く，利用する）（気，つける） |
| 40 | （思う，良い）（ホテル，宿泊する）（ホテル，思う）（駅，近い）（寝心地，良い） |
| 41 | （声，かける）（いただく，美味しい）（ホテル，選ぶ）（大変，満足する）（アウト，行く） |
| 42 | （部屋，見える）（印象，良い）（思う，良い）（感じ，良い）（地図，見る） |
| 43 | （よい，思う）（布団，敷く）（よい，大変）（申し分，ない）（お湯，よい） |
| 44 | （説明，ない）（場所，分かる）（部屋，広い）（足，悪い）（場所，にくい） |
| 45 | （言う，行く）（チェックイン，行く）（感じる，清潔）（言う，無い）（よい，思う） |
| 46 | （部屋，狭い）（頂く，美味しい）（できる，大変）（朝食，頂く）（言葉，ない） |
| 47 | （部屋，良い）（女性，若い）（景色，見える）（お願いする，よろしい）（量，多い） |
| 48 | （機会，利用する）（過ごす，良い）（できる，過ごす）（機会，行く）（親切，対応する） |
| 49 | （思う，良い）（風呂，良い）（サービス，良い）（値段，思う）（風呂，入れる） |

また，図 2 に，ギブスサンプリングの反復回数に伴う提案手法における対数尤度の変化をグラフで示す．ここでいう対数尤度とは，式 (1) の対数をとったものである．反復回数が 200 回を超えると，対数尤度の値がほぼ収束していることが分かる．

### 5.3 考　　察

抽出されたトピックを見てみると，意見や評価を表わすトピックと，出来事を表わすトピックの大きく分けて 2 種類に分類できることが分かる．例として，topic 0，topic 13，topic 30 では，肯定的な意見が示されており，一方で，topic 4，topic 35，topic 44 では，否定的な意見が散見される．これらは，意見や評価を表わすトピックと考えられる．また，topic 1 では，ホテルを利用したという出来事が，topic 24 では，ホテルを探して

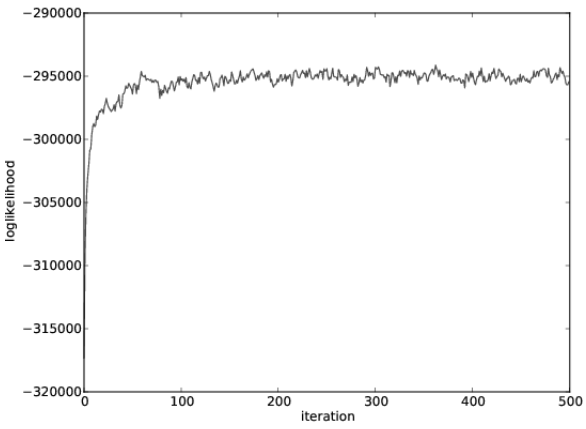


図 2　対数尤度の収束の様子

見つけたという出来事が示されている．これらは出来事を表わすトピックと考えられる．また，意見や評価を表わすトピックについてさらに細かく分析すると，部屋に関するトピック，立地条件に関するトピック，という分類ではなく，部屋に対して肯定的な評価を表わすトピック，立地条件に対して否定的な評価を表わすトピック，というように，対象と評価の関係性が考慮された分類となっており，このことは提案手法により抽出されたトピックの特性と考えられる．

また，ギブスサンプリングの反復回数としては，およそ 200 回の反復により結果が収束していると分かり，続く実験においてもギブスサンプリングの反復回数は 200 回と設定することにする．

## 6. 文書検索精度の比較による性能評価実験

共通の文書検索課題を通じて，従来手法と提案手法の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意味を捉えられているかを確かめる．以後，従来手法を “wordLDA”，提案手法を “eventLDA” と呼ぶ．

### 6.1 トピック分布類似度判定指標

トピック分布の類似度判定指標としては，Kullback-Leibler 距離，Symmetric Kullback-Leibler 距離，Jensen-Shannon 距離，cosine 類似度を用いて比較を行う．wordLDA においては，Jensen-Shannon 距離を用いたときが最も精度が高いと報告されており [9]，提案手法でも同様にして比較を行うことにする．Kullback-Leibler 距離を $D_{KL}$ で表わすとき，Symmetric Kullback-Leibler 距離，Jensen-Shannon 距離は，それぞれ式 (7)，式 (8) で定義される．

$$D_{symKL}(S,Q) = D_{KL}(S \parallel Q) + D_{KL}(Q \parallel S) \tag{7}$$

$$D_{JS}(S,Q) = \frac{1}{2}D_{KL}(S \parallel M) + \frac{1}{2}D_{KL}(Q \parallel M) \tag{8}$$

$$M = \frac{1}{2}(S+Q)$$

### 6.2 実験仕様

対象データには，前章での実験と同様に，楽天トラベルのホテル・施設に関する評価・レビューを用いた．レビューには，「部屋」や「立地」などの各対象につき 1〜5 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，提案手法の性能評価に適していると考える．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」についてのコメントを残している確率が高いと考え，上記のクエリ内容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，適切なトピック数と有効な類似度判定指標の調査の 2 つの観点から両手法の比較を行う．まず，第 1 段階として，最適なトピック数について調べる．トピック数 $k$ を $k = 5, 10, 20, 50, 100, 200$ と変化させ，類似度指標は Jensen-Shannon 距離に固定する．Jensen-Shannon 距離に固定する理由は，単語を単位としたトピック分布の類似度測定において用いる指標を比較した際，最も精度が高いと報告されていることによる [9]．次に，第 2 段階として，類似度判定に用いる指標の変化による結果の違いを比較する．第 1 段階において適するトピック数が決定するため，トピック数としてはその値を用い，次の段階として類似度に用いる指標を変化させる．いずれの条件についても試行回数は 20 回とし，その平均をとった．wordLDA についても同様の実験を行い，その結果を提案手法と比較した．

### 6.3 実験結果

表 2 に，トピック数 $k$ を変化させたときの wordLDA と eventLDA の 11 点平均適合率の結果を示す．eventLDA では $k = 5$ のとき，wordLDA では $k = 50$ のときに精度が最も高くなっている．このことから，対象文書群を複数のトピックで表わす際，単語を素性とする場合には 50 トピック，イベントを素性とする場合には 5 トピックと設定した場合に，文書において持っていた意味を保ったまま柔軟にトピック分類ができている，ということが分かる．また，全体的にも eventLDA は wordLDA に勝る精度を保っている．

表 2　トピック数による比較

| トピック数 | wordLDA | eventLDA |
|---|---|---|
| 5 | 0.5152 | 0.6256 |
| 10 | 0.5473 | 0.5744 |
| 20 | 0.5649 | 0.5874 |
| 50 | 0.5767 | 0.5740 |
| 100 | 0.5474 | 0.5783 |
| 200 | 0.5392 | 0.5870 |

次に，表 3 に，類似度判定指標を変化させたときの eventLDA と wordLDA の 11 点平均適合率の結果を示す．どの指標を用いた場合でも，eventLDA は wordLDA に勝る精度を保っていることが分かる．指標の違いによる精度比較としては，wordLDA では Jensen-Shannon 距離を用いたときに精度が最も高いのに対し，eventLDA では cosine 類似度を用いたときに最も精度が高くなっている．逆に精度が低くなるのは両手法とも Kullback-Leibler 距離を用いた場合で共通であった．

表 3　類似度判定指標による比較

| 類似度判定指標 | wordLDA | eventLDA |
|---|---|---|
| Kullback-Leibler 距離 | 0.5009 | 0.5056 |
| Symmetric Kullback-Leibler 距離 | 0.5695 | 0.6762 |
| Jensen-Shannon 距離 | 0.5753 | 0.6754 |
| cosine 類似度 | 0.5684 | 0.6859 |

### 6.4 考　　察

実験結果より，提案手法は従来手法に比べて高い性能を示しており，文書の内容をより細かく捉えたトピック推定が行えていることが分かった．また，提案手法の特性として，イベント単位にトピック割り当てを行うことにより少ないトピック数で分類が行えていることが分かった．その理由として，単語からイベントとしたことによって，各素性の持つトピックがある程度小さい範囲に絞られ，結果として，意味をなさない誤差としてのトピックが生成されないのではないか，と考える．

一方で，提案手法における最適な類似度判定指標は cosine 類似度となり，確率分布の類似度判定指標として用いられている指標の方が精度が低くなるという，予想に反した結果となった．このことから，トピック分布の確率分布としての性質についても調査が必要であると考える．

## 7. テキスト要約による性能評価実験

次に，対象を文とした場合の提案手法の性能を検証するために，テキスト要約による実験を行う．要約手法の種類としては，文書から重要箇所を抽出することによって文書の全体を要約するものの他に，与えられたクエリに関する要約文を生成する研究が近年盛んになってきている [12] [13] [4]．提案手法においては，クエリのトピック分布と類似のトピック分布を持つ文書の検索性能が高いことが前実験により検証されていることから，本実験もクエリに特化した要約文を生成することを目指す．要約対象は複数テキストとし，与えられたテキストデータにおける，あるクエリに関する要約文を生成する．

### 7.1 MMR-MD に基づく重要文抽出判定

複数文書の要約において，クエリとの類似度が高い順に文を抽出していくと，抽出された文の内容が重複し冗長性のある要約文が生成される可能性があり，その問題を解決するための，MMR-MD (Maximal Marginal Relevance Multi-Document) という指標が提案されている [14]．これは，クエリとの類似度だけでなく既に抽出された文との類似度をペナルティとして与えることで，内容の重なる文の抽出を妨げる指標であり，式 (9) で定義される [15]．本実験でも，複数文書を対象とした冗長性のない要約文生成を目標とし，この指標を利用する．潜在的トピックに基づいてクエリとの類似度が高い文を選びつつ，表層的には冗長性を削減することを目指し，クエリとの類似度判定

$Sim_1$ には 5 章の実験で用いた 4 種類の指標によるトピック分布間の類似度を用い，既に抽出された文との類似度判定 $Sim_2$ には素性を単位とした cosine 類似度を用いる．

$$MMR\text{-}MD \equiv argmax_{C_i \in R \setminus S}[\lambda Sim_1(C_i, Q) - (1-\lambda) max_{C_j \in S} Sim_2(C_i, C_j)] \quad (9)$$

$C_i$ ： 文書集合中の文
$Q$ ： クエリ
$R$ ： 文書集合からクエリ $Q$ によって検索された文集合
$S$ ： $R$ の内，既に重要文として抽出されている文集合
$\lambda$ ： 重み調整パラメータ

$\lambda$ は，クエリとの類似度と既に抽出された文との類似度の重みを調整する，0 から 1 の値をとるパラメータであり，0 に近いほど冗長性の削減を重視し，1 に近いほどクエリとの類似度を重視した要約文が生成される．この値に関しては，対象となるテキストデータの性質や，目標とする要約文の性質によって適切な値が異なると考えられ，経験的に $\lambda = 0.5$ などと定められることが多い [9]．本研究では，まず最初に $\lambda = 0.5$ と固定した場合において実験を行う．これは，提案手法と従来手法との公平な比較をすることを目標とし，MMR-MD が結果に大きく影響することを防ぐためである．次に，$\lambda$ の値を変化させた場合において実験を行い，$\lambda$ が提案手法においてもたらす影響を，従来手法の場合と比較することにより，提案手法の特性について調べ，適切な $\lambda$ の値を求める．

### 7.2 $\lambda = 0.5$ の場合

#### 7.2.1 実 験 仕 様

本実験では，評価型ワークショップである NTCIR4 TSC3(注4) で用いられたテストセットを利用する．毎日新聞と読売新聞が混在した約 10 記事から成る文書セットが 30 トピック分用意されており，総文数は 3587 文である．各文書セットには，生成した要約文を評価するために，文書集合中の主要な情報に関する質問集合が用意されており，正解として与えられている要約文は，この質問集合の回答群を含んでいる．今回は，文書群全体の要約文ではなくあるクエリに特化した要約文の生成を目指しているため，この質問集合をまとめて 1 つのクエリとし，用意された正解要約文をクエリに特化した要約と見なすことで，これらのデータを利用することにした．図 3 に，クエリの例を示す．

ガルヒ猿人の化石はどの国で発見されたか？ガルヒ猿人の化石が発見された地層はいつごろのものか？現代人の直接の祖先とされる約２００万年前の原人の名前は？エチオピア北部の約２５０万年前の地層で発見された新種の猿人は何と名づけられたか？ガルヒ猿人の化石とともに、石器を使用した最古の証拠として何が見つかったか？

図 3 クエリの例

評価方法としては，TSC3 において用いられた Precision と Coverage を用いる．Precision はシステムが出力した文の内，正解要約文集合に含まれる文の割合であり，Coverage はシステムが出力した文集合中の冗長度合いを考慮しつつ，それが正解要約文集合の内容にどれだけ近いかを測る指標である [16]．

本実験では，用意された質問集合を 1 つのクエリと見なし，クエリとの適合度が高い文を抽出し要約文を生成する．30 文書セット中，5 セットについて同様の実験を行い，平均を求める．抽出する文数は，TSC3 で定められた文数である．5 章の実験と同様に，トピック数，類似度による比較を行い，各文書セットにつき試行回数を 20 回としてその平均をとる．提案手法の比較対象として，MMR-MD を重要文抽出の際の評価指標として wordLDA を用いた場合の実験も行い，その結果と比較する．

#### 7.2.2 実 験 結 果

類似度判定指標による差は現れず，どの指標を用いた場合も同一の結果となった．表 4 に wordLDA と eventLDA による Precision と Coverage の比較を示す．最も精度の高いトピック数 $k$ については，wordLDA では $k = 5$，eventLDA では $k = 10$ となっている．

表 4 トピック数による比較

| トピック数 | wordLDA | | eventLDA | |
|---|---|---|---|---|
| | Precision | Coverage | Precision | Coverage |
| 5 | 0.314 | 0.249 | 0.404 | 0.323 |
| 10 | 0.264 | 0.211 | 0.418 | 0.340 |
| 20 | 0.261 | 0.183 | 0.413 | 0.325 |
| 50 | 0.253 | 0.171 | 0.392 | 0.319 |

さらに，潜在的意味解析を利用しない，基本的なテキスト要約手法との比較を表 5 に示す．手法としては，文書を時系列順に並びかえ各文書の先頭から順に 1 文ずつ重要文として抽出する手法である Lead 手法 [17]，文のスコアを単語重要度の和で定義しスコアの高い文から順に重要文として抽出する手法である，TF-IDF に基づいた重要文抽出手法を比較対象とし，それらの精度に関しては，先行研究 [16] で示されている実験結果の値を用いた．

表 5 要約手法間の比較

| 手法 | Precision | Coverage |
|---|---|---|
| Lead | 0.426 | 0.212 |
| TF-IDF | 0.454 | 0.305 |
| wordLDA (k=5) | 0.314 | 0.249 |
| eventLDA (k=10) | 0.418 | 0.340 |

図 4 に，eventLDA において生成された要約文の例を示す．与えたトピック数は $k = 10$，クエリとの類似度判定に使用した指標は Jensen-Shannon 距離，用いたクエリは図 3 である．このとき，Precision は 0.5，Coverage は 0.571 である．

#### 7.2.3 考 察

実験結果より，全ての条件において，トピックを単語に割り当てるよりもイベントに割り当てる方が精度が高く，提案手法は文に対してもトピック推定を行えていることが分かった．

それぞれの類似度について比較を行うと，トピック分布の類

(注4)：http://research.nii.ac.jp/ntcir/index-en.html

エチオピア北部の約２５０万年前の地層で米国・エチオピア・日本の研究チームが発見した人骨化石が、猿人から原人へ進化する過程にある新種の猿人であることが分かった。しかし今回、頭骨や歯が見つかったことで、同時代の猿人化石の系統の整理は進むと期待される。２３日付の米科学誌「サイエンス」に発表される。二足歩行に適するように脚が長くなっていたとみられる。新種の新人はどう進化するだろうか。人類は約５００万年前にチンパンジーなどの類人猿と分岐し、猿人から原人を経て現代人（新人）に至ったと考えられる。

図 4　生成された要約文の例（$\lambda = 0.5$）

似度の指標の違いによる影響がなく，どの指標を用いても同じ精度を保てている．その理由としては，文に対して推定されたトピック分布が，あるトピックに偏った分布となっており，類似度の指標による影響が現れなかったのではないかと考える．

適切なトピック数は，提案手法の方が大きくなっており，対象を文書とした場合とは異なる性質が見られた．これについては，対象としたデータがあるテーマについて書かれた新聞記事群であり，一つの単語に対するトピックがある程度決まっていたことから，単語にトピックを割り当てた場合のトピック数は少なくても分類が行えたと考えられる．対象文書に対する適切なトピック数の推定には，様々なタイプの文書データを用いて今後実験を行っていく必要がある．

潜在的な意味を考慮しない手法との比較においては，提案手法は近い精度を示しており，表層的な情報を直接扱った場合と同じ程度の性能を持てていることが分かった．特に，Coverage においては高い精度を示しており，潜在的トピックを扱ったことでより網羅的な要約文生成が行えたと考える．

### 7.3　$\lambda$ を変化させた場合

#### 7.3.1　実 験 仕 様

対象データとしては，7.2 の実験と同じ 5 セットの文書群を用いる．提案手法である eventLDA と従来手法である wordLDA によりトピック分布を推定し，推定されたトピック分布に対して MMR-MD における重み調整パラメータ $\lambda$ の値を $\lambda = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ と変化させる．与えるトピック数は，7.2 節の実験で得られた結果を用いて，eventLDA では $k = 10$，wordLDA では $k = 5$ とする．また，クエリとの類似度判定に用いる指標は，Jensen-Shannon 距離とする．評価方法に関しても，7.2 節の実験と同様に Precision と Coverage を用いる．各文書セットにつきトピック推定の試行回数は 20 回とし，その平均をとる．

#### 7.3.2　実 験 結 果

図 5 に，$\lambda$ の値の変化に伴った，wordLDA と eventLDA による Precision と Coverage の変化を示す．

Precision と Coverage のどちらについても，$\lambda$ の値に関わらず，eventLDA は wordLDA よりも高い精度を保っていることが分かる．また，Precision に着目すると，eventLDA の方が wordLDA よりも $\lambda$ の値による影響が大きく，wordLDA では $\lambda = 0.6$ のとき最大値をとっているのに対し，eventLDA では $\lambda$ の値が大きくなるのに伴って増加している．一方，Coverage

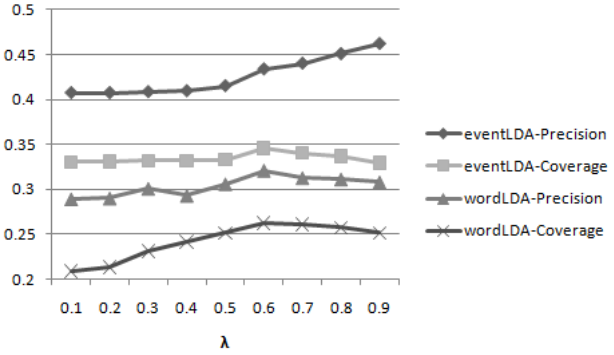


図 5　$\lambda$ に基づく Precision と Coverage の変化

に着目すると，eventLDA と wordLDA のどちらにおいても，$\lambda = 0.6$ のときに最大値をとっている．

Precision と Coverage の両方を考慮した評価を行うために，それらの調和平均を算出して比較を行う．図 6 に，$\lambda$ の値の変化に伴った，wordLDA と eventLDA による要約生成における Precision と Coverage の調和平均の値の変化を示す．

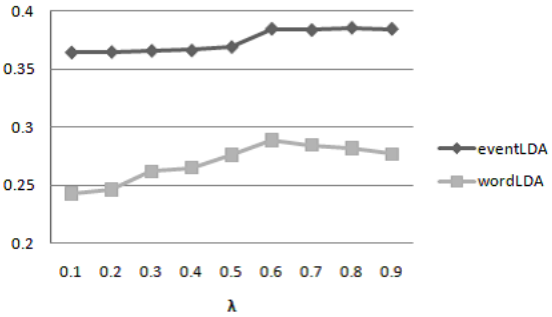


図 6　$\lambda$ に基づく調和平均の変化

Precision と Coverage の両方を考慮すると，wordLDA では $\lambda = 0.6$ の場合，eventLDA では $\lambda = 0.8$ の場合に最も精度が高いという結果になった．また，$\lambda$ の値の変化に伴う調和平均の変化の様子について見てみると，eventLDA は wordLDA と近い変化の仕方を示していることが分かる．

図 7 に，eventLDA において生成された要約文の例を示す．用いたクエリは図 3 であり，このとき，Precision は 0.667，Coverage は 0.571 である．

#### 7.3.3　考　　察

実験結果より，$\lambda$ の値を $\lambda = 0.5$ 以外に変えた場合においても，提案手法による要約文生成が高い精度で行えており，対象となるテキストデータや目標とする要約文の性質に応じて $\lambda$ の値が変わった場合においても，提案手法による要約生成が有効であることが分かった．また，Precision に着目したときに eventLDA の方が wordLDA よりも $\lambda$ の値による影響が大きかったことについては，潜在的トピックに基づいてクエリとの類似度を重視することが Precision に反映されやすく，イベントにトピックを割り当てたことによって，潜在的トピックをより正確に推定できたのではないかと考える．eventLDA による要約文生成と wordLDA による要約文生成の共通点として，

エチオピア北部の約２５０万年前の地層で米国・エチオピア・日本の研究チームが発見した人骨化石が、猿人から原人へ進化する過程にある新種の猿人であることが分かった。研究チームのメンバーの諏訪助教授は「猿人から原人へという空白を埋める数十年ぶりの発見だ」と話している。見つかった化石から体格が類推できることや、当時の食生活が分かるため、人類の進化の解明に役立つと期待されている。チームを率いる米カリフォルニア大バークリー校のティム・ホワイト教授らは「栄養価の高い肉や骨髄を食べられるようになり、その後の人類の発展につながったようだ」と説明している。人類は約５００万年前にチンパンジーなどの類人猿と分岐し、猿人から原人を経て現代人（新人）に至ったと考えられる。空白の時代を埋めるだけでなく、頭骨や手足の骨、石器使用の跡など重要な資料もあり、人類の系統を整理するうえで価値も高い。

図 7　生成された要約文の例（$\lambda = 0.6$）

Coverage がどちらにおいても $\lambda = 0.6$ で最大となっていることについては，クエリとの単語の一致ではなく，その潜在的トピックを扱っていることで冗長性が既に取り除かれており，冗長性を削減するためのペナルティをあまり与えなくても，冗長性が抑えられるからではないかと考える．また，Precision と Coverage の調和平均による比較においては，eventLDA は wordLDA に近い変化の仕方を示しており，提案手法が潜在的トピックに基づいて要約文生成を行っていることの正当性を確認できた．

## 8. おわりに

本研究では，係り受け関係に基づいた単語の対をイベントと定義し，イベントにトピックを割り当てることで，文書内の事象を捉えた潜在的トピック抽出手法を提案した．そして，5 章において提案手法によって抽出される潜在的トピックの例を示し，6 章において文書検索に提案手法を適用することで，提案手法の性能検証を行い，その応用として，7 章では提案手法による要約文生成を示した．

対象が文書であっても文であっても，提案手法である eventLDA は，wordLDA よりも高い性能を持っていることを示すことができ，トピックをイベントという単位に割り当てた場合でも潜在的なトピックを推定できていることが分かった．そして，テキスト要約に提案手法を適用することにより，潜在的トピックを考慮した，より網羅的な要約文生成が行えたことが分かった．また，本研究によって，素性をイベントのような情報量の大きいものにした場合でも潜在的なトピックを推定できることが分かり，単語以外の素性の有用性も示すことができた．

今後は，様々なタイプのデータ，クエリを用いて実験を行い，提案手法の特性についてさらに考察を行うつもりである．また，対象を文とした場合においては，抽出イベントの少なさの影響が大きくなることが考えられ，イベント抽出方法やイベント－文書行列の作成方法についても，より深く考察を行っていくつもりである．

## 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．また，国立情報学研究所の許諾を頂き NTCIR4 のデータセットを使用させて頂きました．ここに深く感謝の意を表します．

## 文　献

[1] 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.

[2] 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.

[3] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, vol.3, pp.993–1022, 2003.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Q. Bing, L. Ting, Z. Yu, and L. Sheng, “Research on Multi-Document Summarization Based on Latent Semantic Indexing,” Journal of Harbin Institute of Technology, vol.12, no.1, pp.91–94, 2005.

[9] L. Henning, “Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis,” International Conference RANLP 2009-Borovers, pp.144–149, Bulgaria, 2009.

[10] 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.

[11] T. Grififths and M. Steyvers, “Finding scientific topics,” Proc. of the National Academy of Sciences, vol.101, pp.5228–5235, 2004.

[12] A. Tombros, M. Sanderson, “Advantages of query biased summaries in information retrieval,” Proc. of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp.2–10, 1998.

[13] 森辰則，野澤正憲，浅田義昭，“質問応答エンジンを利用した複数文書要約手法，” 言語処理学会年次大会発表論文集, vol.10, pp.189–192，2002.

[14] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz, “Multi-document summarization by sentence extraction,” Proc. of ANLP/NAACL Workshop on Automatic Summarization, pp.40–48, 2000.

[15] 奥村学，難波英嗣，知の科学　テキスト自動要約，人工知能学会（編），オーム社，東京，2005.

[16] 平尾努，奥村学，福島孝博，難波英嗣，“TSC3 コーパスの構築と評価，” 言語処理学会年次大会発表論文集, vol.10, pp. A10B5-02, 2004.

[17] R. Brandow, K. Mitze, and L.F. Rau, “Automatic condensation of electronic publications by sentence selection,” Information Processing and Management, vol.31, no.5, pp.675–685, 1995.