Vol.2011-NL-201 No.3
Vol.2011-SLP-86 No.3
2011/5/16

文書上の潜在トピックを捉える事象の検討とその応用

北島理沙[†1]　小林一郎[†1]

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。しかし，これらの手法において，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ていないという問題がある．そこで著者らは，2 つの単語の組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した．一方で，そのイベントの定義は係り受け関係に基づいて経験的に定めたものであるため，検討が必要であると考えられる．本稿では，文書上の潜在トピックを捉える事象の取り方について検討し，文書検索実験を用いて比較を行う．また，その応用として，提案手法を用いた潜在的な意味に基づく要約生成を示す．

An Examination for Proper Events grasping Latent Topics in a Document and its Application

Risa Kitajima[†1] and Ichiro Kobayashi[†1]

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 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 determined 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. As an application of our proposed method, additionally, we also show a multi-document summarization based on latent topics.

†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
Faculty of Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

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[⋆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．

先行研究[5] では，イベントの定義の仕方として文節の係り受け関係を利用した．しかし，この定義の仕方は経験的であり，イベントの定義に関しては検討する必要があると考える．

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 章ではイベントの定義について述べる．5 章ではイベント単位に潜在トピックを割り当てる提案手法について説明し，6 章では文書検索を用いた性能評価実験について，7 章ではテキスト要約への応用についてまとめる．最後に，8 章で本研究のまと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[6] 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[7] や，松本らによる研究[8] 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

⋆1 イベントの定義については，4 章で詳述する．

© 2011 Information Processing Society of Japan

Vol.2011-NL-201 No.3
Vol.2011-SLP-86 No.3
2011/5/16

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[9)10)]．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 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．

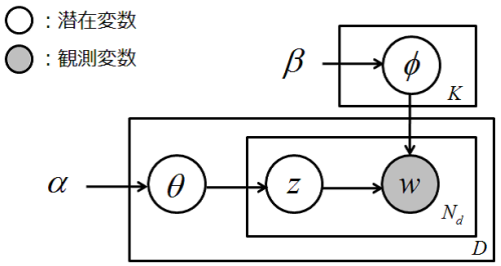


図 1 LDA のグラフィカルモデル
Fig. 1 Graphical of LDA.

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 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．

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

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

© 2011 Information Processing Society of Japan

Vol.2011-NL-201 No.3
Vol.2011-SLP-86 No.3
2011/5/16

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

トピック集合 Z の推定手法としては，変分ベイズ法[3]，Collapsed 変分ベイズ法[11]，ギブスサンプリング[12] 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[11]，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．

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

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 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 $\hat{\theta^k_d}$，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[5] 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[⋆1] 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，本研究に置いてこのイベントの定義について熟考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 4 つのイベントタイプを設定して，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．

4.1 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．

4.2 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

⋆1 http://chasen.org/ taku/software/cabocha/

© 2011 Information Processing Society of Japan

Vol.2011-NL-201 No.3
Vol.2011-SLP-86 No.3
2011/5/16

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．

4.3 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．

4.4 経験的に定めたルールを満たす共起

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ルール，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ルールを設けた．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，これらのルールを設けることが必要であると考えたためである．例えば，4.3 節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. 文書検索精度の比較による性能評価実験

6.1 トピック分布の類似度判定

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．Kullback-Leibler 距離を D_{KL} で表わすとき，Jensen-Shannon 距離は，式 (7) で定義される．

© 2011 Information Processing Society of Japan

Vol.2011-NL-201 No.3
Vol.2011-SLP-86 No.3
2011/5/16

$$D_{JS}(S,Q) = \frac{1}{2}D_{KL}(S \parallel M) + \frac{1}{2}D_{KL}(Q \parallel M) \quad (7)$$
$$M = \frac{1}{2}(S+Q)$$

6.2 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[⋆1]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．

6.3 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．

⋆1 http://travel.rakuten.co.jp/

表 1 イベントタイプの比較
Table 1 Comparison of Event types.

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

6.4 考　　察

実験結果より，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が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おり，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．event3 は，event2 に次いで 2 番目に高い結果を出しており，その次元数が 3 分の 1 程度で済んでいることを考慮すると，その精度の差は僅差であり，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．event4 は，経験的に重要度が高そうな品詞の組み合わせをルールとして設定したものの，全ての係り受け関係にある自立語の組み合わせを利用した 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 event4 では event3 に比べ，次元数が少ないという利点もある．event4 を利用するのは精度と次元数のトレードオフの考慮と，品詞のペアによって構成する経験的なルールの設定コスト次第とも言える．

7. テキスト要約への応用

次に，提案手法の応用例として，6 章の実験において高い精度を示した 2 つのイベントタイプ，event2 と event3 を用いて，複数文書を対象としたテキスト要約を行う．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[13)14)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実験により検証されていることから，先行研究においてもクエリに特化した要約文生成を行っている．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を対象とした，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

© 2011 Information Processing Society of Japan

Vol.2011-NL-201 No.3
Vol.2011-SLP-86 No.3
2011/5/16

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[15]．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 (8) で定義される[16]．なお，C_i は文書集合中の文，Q はクエリ，R は文書集合からクエリ Q によって検索された文集合，S は R の中で既に重要文として抽出されている文集合を表わし，λ は Sim_1 と Sim_2 の重みを調整するパラメータである．

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 6.1 節にて説明した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．

$$MMR\text{-}MD \equiv argmax_{C_i \in R \setminus S}[\lambda Sim_1(C_i, Q) - (1-\lambda) max_{C_j \in S} Sim_2(C_i, C_j)] \quad (8)$$

λ 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[10]．本実験では，λ 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．そして，λ の値の変化に伴う精度の変化を観察することにより，各イベントタイプの特性について調べる．

7.2 実験仕様

本実験では，評価型ワークショップである NTCIR4 TSC3[⋆1] 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 2 に，クエリの例を示す．複数文書からクエリとの適合度が高い文を MMR-MD を指標とすることで抽出し，要約文を生成する．

⋆1 http://research.nii.ac.jp/ntcir/index-en.html

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 2 クエリの例
Fig. 2 An example of a query.

評価方法としては，TSC3 において用いられた Precision（精度）と Coverage（被覆率）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[17]．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．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り，さらに 5 文書セットの平均値をとる．与えるトピック数 k は，先行研究[5] で得られた結果を用いて $k = 10$ 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．なお，精度の比較のために，先行研究で得られた実験結果である，単語を素性として扱った結果を word として示し，また，event0 による実験結果も同様にして示す．

7.3 実験結果

図 3 に，λ の値の変化に伴った，各イベントタイプに基づく Precision の変化を示す．

結果を見ると，先行研究で提案した event0 が最も良い結果となっており，4 章にて提案したイベントタイプである event2 や event3 を素性として扱った場合の実験結果は精度が低くなっていることが分かる．また，今回提案している event2 や event3 は λ の値の増加に伴って精度が高くなっており，先行研究にて提案したイベントタイプと同様の特徴が見られる．

図 4 に，λ の値の変化に伴った，各イベントタイプにおける Coverage の変化を示す．Coverage においても，先行研究にて提案したイベントタイプである event0 が最も高い精度を示しており，今回提案したイベントタイプを用いた場合は単語を素性としたときと同程度の精度となっていることが分かる．λ の値の変化に伴う Coverage の変化の様子を見てみると，提案した event2 は先行研究で出た結果と同様に $\lambda = 0.6$ のときに最大となっており，

© 2011 Information Processing Society of Japan

Vol.2011-NL-201 No.3
Vol.2011-SLP-86 No.3
2011/5/16

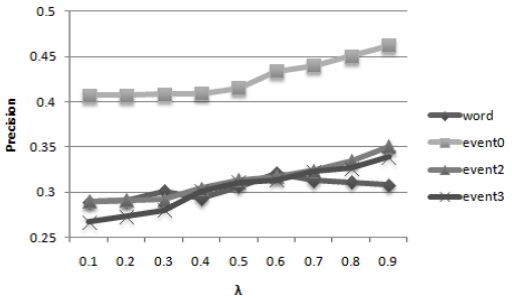


図 3 イベントタイプに基づく Precision の変化
Fig. 3 Precision of each Event type based on λ.

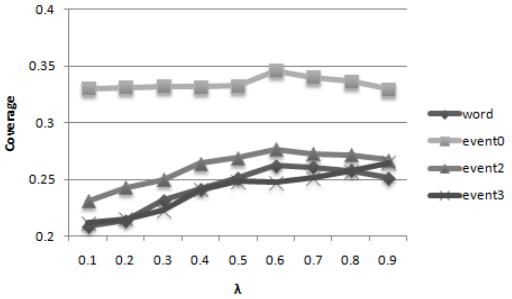


図 4 イベントタイプに基づく Coverage の変化
Fig. 4 Coverage of each Event type based on λ.

これは単語を素性とした場合，また，先行研究にて提案した event0 と同様である．一方で，event3 は λ の値の増加に伴って Coverage の値も増加しており，他の 3 つのイベントタイプとは異なる特徴がみられた．

最後に，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 5 に，λ の値の変化に伴った，各イベントタイプを用いた要約生成における Precision と Coverage の調和平均の値の変化の様子を示す． Precision と Coverage の両方を考慮すると，event2 は event3 よりも精度が高く，また，どちらも λ の値が増加するのに伴って値が増加していることが分かる．この変化の様子は word，event1 とは異なっており，event2，event3 におけるイベントの構成による精度への影響として捉える必要があると考える．

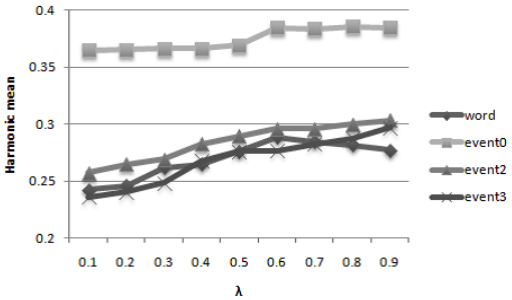


図 5 λ に基づく調和平均の変化
Fig. 5 Harmonic mean of each Event type based on λ.

7.4 考　　察

実験結果より，潜在トピックを捉える対象を文にした場合には，今回提案したイベントタイプよりも先行研究にて提案したイベントタイプの方が高い精度を示すことが分かった．しかしその一方で，λ の値の変化に伴う Precision，Coverage の値の変化の様子に着目してみると，提案したイベントタイプを素性として用いた場合の実験結果は，先行研究で提案した素性を用いた場合と同様の変化を示しており，その有効性を確認できたといえる．さらに，Precision に着目してみると，提案したイベントタイプを用いた場合には λ の値による影響が大きいという結果が見られた．潜在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 の値の増加に反映されることから，このことは，クエリとの類似度を考慮することが文のトピック分布を扱うにあたって重要となっていることを意味しており，提案するイベントタイプがより文の内容を捉えた素性であると考えられる．また，word，event0，event2 の共通点として，Coverage が $\lambda = 0.6$ 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，表層的な表現の一致による冗長性が既に取り除かれており，λ の値を小さくとること，つまり冗長性削減を重視することは，かえって Coverage の値を低下させる原因となるのではないかと考える．この特徴は event3 においては見られず，λ の値の増加に伴い Coverage も増加するという結果になった．これは，event3 におけるイベントの構成が，表層的情報よりも潜在トピックに反応しやすいものになっており，冗長性削減の部分よりもクエリとの類似度を重視する部分の影響の方が大きくなったためではないかと考える．また，Precision と Coverage の調和平

© 2011 Information Processing Society of Japan

Vol.2011-NL-201 No.3
Vol.2011-SLP-86 No.3
2011/5/16

均による比較においては，event3 では word と同程度の性能となったものの，event2 ではそれ以上の精度を示すという結果になり，提案したイベントタイプによる要約生成の有効性を確認できた．しかし，その精度は先行研究で示したものの方が高いという結果となった．今回は，レビューを対象にしたデータセット 1 つに対する実験結果であるため，この結果が汎用的なものであるか今後様々な文書を対象に実験を行い調査するつもりである．

8. 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，その応用として，7 章では提案手法を用いたテキスト要約生成について示した．結果として，潜在トピックを扱う対象が文書である場合には，提案したイベントタイプの優位性を示すことができた．その一方で，潜在トピックを扱う対象が文である場合には，先行研究にて用いた素性の方が高い精度を示すことが分かった．今回は，対象としたテキストデータの種類が前者と後者では異なっており，対象を文書とした実験においては人の意見や評価が書かれたレビューを使い，また，対象を文とした実験においては事実が書かれた新聞記事を使っているため，そのことが実験結果に影響をもたらしたとも考えられる．また，文書の内容を表わすのに適したイベントタイプと，文の内容を表わすのに適したイベントタイプは異なることも考えられる．

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全イベントパターンで実験を行い，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．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．

参考文献

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) 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）.
6) 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).
7) 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).
8) 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).
9) 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).
10) L.Henning: Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis, *International Conference RANLP 2009-Borovers*, pp.144–149, Bulgaria (2009).
11) 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).
12) T. Grififths and M. Steyvers: Finding scientific topics, *Proc. of the National Academy of Sciences*, Vol.101, pp.5228–5235 (2004).
13) 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).
14) 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).
15) 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).
16) 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).
17) 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).

© 2011 Information Processing Society of Japan