係り受け関係の階層化に基づいた構文木モデルを利用したトピック推定手法の提案

大野　一樹†　　波多野賢治††

† 同志社大学大学院文化情報学研究科　〒 610-0394 京都府京田辺市多々羅都谷 1-3 夢告館 502
†† 同志社大学文化情報学部　〒 610-0394 京都府京田辺市多々羅都谷 1-3 夢告館 502
E-mail: †ohno@ilab.doshisha.ac.jp, ††khatano@mail.doshisha.ac.jp

あらまし　文書集合内における各文書のトピックを推定する際に用いられるモデリングの分野では，常に文書に対するトピックの推定精度向上を目指した研究が行われている．従来は語の出現頻度である n-gram に基づいてトピックが各語に割り当てられ，それらの語を含んだ文書に各語が属するトピックを対応させることで文書のトピック分類が行われてきた．最近では， n-gram 以外に語の順序やフレーズ間の依存構造がモデリングの際の素性として用いられており，これらの併用によってトピック推定精度の向上が確認されている．そこで，この素性に着目し，階層化を行った係り受け関係をモデリングの素性として利用することで，日本語におけるトピック推定において精度の向上を目指した．
キーワード　トピックモデリング，文書分類，構文解析

1. はじめに

World Wide Web (WWW) 上には膨大な量の Web 文書が存在している．Web 文書はそれぞれ，一定のトピックに基づいて記述されているため， Web 文書がどのようなトピックに基づいて記述されているかを高精度に分類することができれば，トピックに基づいた文書検索を行うことができ，これにより効率的に情報検索を行うことができる [1].

文書に関してトピック分類を行う手法は Latent Dirichlet Allocation (LDA) [2] を中心としてその発展形が多岐にわたりいくつも提案されている．しかし， LDA の根本的な概念は単語の出現分布である n-gram 分布に対して潜在トピックを割り当て，それらの単語を含む文書に対応するトピックを推定することでトピック分類を行うというものである．このような単語の n-gram 分布に基づいたトピックモデリング手法では，単語の出現頻度が同じであれば，単語の出現順が異なっても，異なる意味を持つ文書に対して同じトピックが付与されるといった問題があった．

最近になって，この問題を解決するために語の結合規則に則った階層トピックや文法構造を表す構文木の構造に基づいたトピックのモデリング手法が提案されている [3], [4]．これらは自然言語の文法を構成する規則の一つである統語構造に基づくことにより，単語の n-gram 分布だけでなく，文の生成過程の一部である語の結び付きの関係性をトピックの推定に利用している．これにより，高精度なトピック推定が行えることが確認されている．

一方で日本語においても，文書中に含まれる文の構造である係り受け関係をトピックの推定に取り入れたモデリング手法が北原，小林 [5] によって提案されており，単語の出現分布に基づいた手法と比較してトピック推定精度の向上が確認されている．しかし，この北原，小林の手法はトピックを構成する素性として文節間の単一の係り受け関係が考慮されているのみであった．そのため，二つの文節の間の係り受け関係のみがモデリングの際に利用され，その前後の文節の係り受け関係が示す文の文脈性や複雑で文節間の距離が遠い係り受け関係がもたらす修飾関係が排除されてしまいがちである．よって，北原，小林の手法では構文構造を網羅できておらず，トピックのモデリングの際にすべての統語構造をカバーできているとは言い難い．

そこで，本稿では日本語を対象に，以前に我々が提案した係り受け関係の階層化に基づいた構文木モデル [6] を用いて構文解析を行うことにより，文節間の任意の長さの係り受け関係を素性として考慮したトピックモデリング手法の提案を行う．生成したトピックモデルに基づくことで，文全体を構成する統語構造を取り入れ，トピック分類の明文化を行ったトピック推定手法を目指す．

以降の節では基本的事項として，日本語の構文構造とその解析手法について述べた後，関連研究として係り受け関係を持つ文節間の組を素性としたトピックモデリング手法とその問題点について述べる．その後，関連研究の問題点について解決を図るため，係り受けの階層化に基づく構文木モデルを利用したトピックモデリング手法を提案し，評価実験をもとに本稿についてまとめる．

2. 日本語構文解析

日本語の構文構造を考慮したトピックモデリングの基本的事項として，日本語構文解析の特徴と，現在に至るまで提案されてきた構文解析手法に関して説明を行う．また，これらについて踏まえた上で，我々が以前に提案した係り受け関係の階層モデルに基づいた構文解析手法に関してその特徴を記述する．

2.1 日本語構文解析の特徴

日本語は Subject (主語), Object (目的語), Verb (動詞) で表される SOV の順序構造に基づいた主要部終端型言語 (Head-

final Language) と呼ばれる言語体系の一つである．そのため，単語の出現順序が自由であるという特徴がある．一方で，英語は Subject (主語), Verb (動詞), Object (目的語) で表される SVO の構造の構造に基づいた主要部先導型言語 (Head-initial Language) と呼ばれる体系に分類される．英語の場合，単語の並びによって文法構造は表現されるが，日本語では語と語の統語的役割によって文法構造が表現される．したがって，日本語構文解析の目的は文節間の修飾関係や意味役割を表す依存構造である係り受け関係の発見であり，この係り受け関係こそ，語と語の統語的役割を果たすものである．そのため，語の並びから構成される文節間の係り受け関係の発見が日本語の構文解析の目的となる．

文節は語の並びによって構成されるが，この語の中でも統語的役割を果たす語として重要な役割を持っているのが格を表す役割を持つ助詞である．助詞は主に名詞に付属することにより，その文節の役割を表す．一般的に文節間の関係はこの格によって決定付けられる．したがって，文節間の並びによって構文が表されるのではなく，文節の格によって構文が表現されるため，文節の並びに制約はないものである．

例えば，次の二つの文「トムは女性に本を渡した。」と「トムは本を女性に渡した。」という文は同じ文法構造あるいは係り受けの構造を持つ．これらの文法構造を文節から文節に対する係り受け関係の存在を文節から文節への矢印で表現した係り受け木と呼ばれる構造で表すと図 1 のように表される．

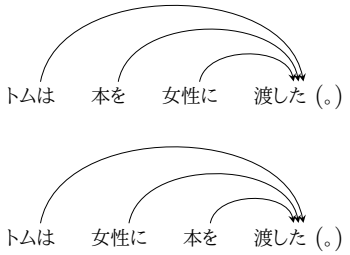

図 1　同じ構文構造を持つ文の例

ここで，どちらの係り受け木においても文節「トムは」は文節「渡した」に対して係り受け関係を持っている．同様に文節「女性に」は文節「渡した」に対して係り受け関係を持っており，文節「本を」も文節「渡した」に対して係り受け関係を持っている．よって，これらの文はすべての文節間において同じ係り受け構造を持つ．このように，日本語は語順が異なったとしても，格が統語的役割を果たすために同じ文法構造，つまり同じ意味を持つ文が存在する．

したがって，日本語の係り受け関係の規則は次の制約を持つものである．

（1） 係り受け関係は左の文節から右の文節に対して生じるものである．後方の文節が前方の文節に対して修飾関係や意味役割を持つことはない．格によって，文法構造である係り受け関係が表現され，格は常に後方の文節に対してのみ修飾関係や意味関係を示す役割を果たす．

（2） 係り受け関係は交差しない．係り受け木においてある文節から出た矢印が他の文節から出た矢印と交わることはない．

（3） 文末の文節を除いたすべての文節は唯一つの係り受け関係を持つ．格が果たす修飾関係や意味役割は唯一つである．

以上の係り受け関係の規則に従うことで，日本語の地の文に対して構文解析を行うことができる．以降の節では，この係り受け関係の制約を考慮して現在までに提案されてきた構文解析手法について説明する．

2.2　文節間の二値分類に基づく構文解析手法

文節間に対して係り受け関係が存在するかどうかの二値分類に基づき， SVM を利用した格文法における構文木モデルの生成手法が Kudo and Matsumoto によって提案されている. [7]. 構文木モデルの生成に SVM を用いているため，この構文木を利用した構文解析手法は文節間の多様な特徴量を考慮しつつ，高速かつ高精度に構文解析を行うことができる．また， Kudo and Matsumoto の手法では文中の文節について直後の文節との間に係り受け関係が生じるかどうかを判定することで，分類器だけでなくアルゴリズムの観点からも高速な構文解析手法を行うことができる. [8] この構文解析アルゴリズムは以下の操作として定義されている．

（1） すべての文節において，右方の直後にある文節に対して係り受け関係を持つかどうか，判定する．

（2） ステップ 1 の処理において係り受け関係を持つ文節があれば，この係り受け関係を解析結果とする．また，この係り受け関係を持つ係り元の文節を 1 のステップにおける処理の対象から除外する．

（3） 1 と 2 のステップを末尾の文節以外が係り受け関係を持つようになるまで繰り返すことで，解析結果を得る．

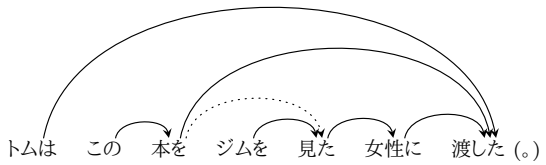

図 2　日本語構文解析手法における誤解析例

一方で， Kudo and Matsumoto の手法は対となっている文節間における係り受け関係の存在の有無は，それぞれの文節を構成する語の形態素に依存する．そのため，類似した形態素を持つ文節同士が並んでいる場合，図 2 に破線で表されるように係り受け関係が発生し得ない文節に対して係り受け関係が発生し，誤解析結果が出力される場合がある．このような誤解析結果は動作の主体とその動作を表す文節の対が一文に複数存在することで，増加する傾向が見られる．

2.3　係り受け関係の階層化に基づく構文解析

2.2 節で述べたように Kudo and Matsumoto の構文解析手法では複雑な構造を持つ文の解析に失敗するいった問題があった．従来手法が係り受け関係を持つ文節間の表層，形態素といった要素を素性として構文木モデルを構成していたのに対し，さらにその前後の係り受け関係を持つ文節を文法要素を構成する素性として，これを取り入れた構文木モデルの生成を行っている [6]．そうして生成した構文木モデルを利用して構文解析を

行うことにより，従来手法の問題点が解決されることを確認している．

以下に構文木モデルの構築過程とこれを利用した構文解析アルゴリズムを示す．

2.3.1 構文木モデルの構築

係り受けの生起するコンテキストとして係り受け関係を持つ文節間の前後の文節を考慮するため，n-gram を元に係り受け関係の生起確率に基づいた構文木モデルを構築する．係り受け関係の生起を n-gram のマルコフ過程として考えると，係り先の文節をコンテキストとしたとき，係り受け関係を持つ係り元の文節が生起する確率を計算することができる．そのため，文末の文節を始点とした n-gram によって，文節間の係り受け関係の生起確率を表現し，構文木モデルを構成する．

しかし，n-gram モデルでは n を小さく設定すると学習データのパープレキシティが大きくなり，言語モデルの精度が下がる問題がある．逆に n が大きいと状態数が爆発的に増加し，モデルのサイズが大きくなってしまう．そこで，本手法では階層 Pitman-Yor 過程 [9] の拡張であり，n を任意の変数として扱うことで可変長の n-gram の共起を考慮した Nested 階層 Pitman-Yor 過程 [10] を利用している．これにより，文末の文節を初期のコンテキスト，つまり，根としてこれに対する係り受け関係を持つ文節の生起確率を計算するとき，任意の係り先の文節を考慮した構文木モデルを生成することができる．なお，構文木モデル生成時の素性には日本語構文解析を最大エントロピー問題として扱い，素性を事前条件として定めて構文木の分布を選択して，高い解析精度を達成している Uchimoto ら [11] の研究を参考にした．Nested 階層 Pitman-Yor 過程を利用して構文木モデルを生成する際，係り元の文節について付属語の表層と形態素を，係り先の文節については先頭の形態素を利用している．

2.3.2 構文解析アルゴリズム

2.3.1 節に示した過程によって構築された構文木モデルは文末の文節を根として階層化されている．そのため，構文解析の際には入力文に対して文末の文節から前方の文節へと動的アルゴリズムである CKY アルゴリズム [12] を用いて最適な係り受け関係を持つ文節を探索する．CKY アルゴリズムを用いた探索の際に，それまでの探索結果をメモリに保持しておく必要があるが，このとき保持する探索の数は Beam Search Algorithm [13] によって限定し，ボトムアップに解析を行うことで，構文解析を行う．このとき，文末の文節はその直前の文節から係り受け関係を得るため，文末の文節とその直前の文節をコンテキストとして，文末の文節の直前の文節に対して係り受け関係を持つ文節の探索を行う．

このプロセスを文頭の文節に到達するまで再帰的に繰り返すことにより，入力文に対して係り受け関係を持つ任意の長さの文節をコンテキストとして考慮した構文解析を行うことが可能となっている．これにより， 従来の日本語構文解析手法の複雑な構文構造を持つ文を解析対象とした際における問題点の解決を示している．

3. 関 連 研 究

本節では従来手法として，基本的な LDA の概念と言語の文法構造である統語規則や依存構造といった文法要素を素性としたトピックモデリング手法を関連研究として述べる．

3.1 単語 ***n*-gram** 分布に基づいたトピック推定

LDA あるいはその派生を利用した一般的なトピック推定手法では各文書ごとの単語 n-gram 分布を観測し，観測された n-gram 分布にトピックを対応させることで各文書に対して潜在的トピックを割り当てていた．文書ごとの単語の出現分布を基に文書に対するトピックを推定することで，推定されるトピックは単語の分布のみに依存してしまう．原因としては，文書において語の出現順序や依存構造といった文法的要素が未考慮ということがある．その結果，本来，同一のトピックに所属しないと考えられる文書に対して共通のトピックが割り当てられるという問題があった．

3.2 イベントに基づいたトピック推定

3.1 節で述べた語の n-gram 分布に基づいたトピック推定手法の問題点として，生成されたトピックモデルは文書内の語の関係である統語構造や依存構造といった文法要素から影響を受けないことが挙げられる．トピックモデルの生成時には単語の出現分布のみが考慮されるため，トピックを単語の出現分布のみで表すことができることに対して疑問があった．そのため，最近では文法規則をトピックモデリングの素性として取り入れた統語構造の階層化に基づいたモデリング手法 [3] や構文木に基づいたモデリング手法 [4] が提案されている．

同様に日本語においても，トピックモデリングの際の素性の候補として文法規則に着目したトピックモデリング手法が提案されている．日本語における単語 n-gram 分布である形態素の出現分布以外の素性をイベントとして文節間の係り受け関係に基づいたトピックモデリング手法が北島，小林によって提案されている [5]．北島，小林の手法では，単語単位ではなくイベントという単位をトピックを形成する概念として定義することで，文書を扱い，各文書に対してイベントを抽出する．そして，抽出したイベントを各文書の索引とし，その出現分布に基づいてトピックの推定を行うことができるとされている．

3.2.1 イベントの定義

北島，小林の手法はトピックモデリングの際に形態素の出現分布以外の素性の利用として，文書ごとにイベントを利用している．ここでイベントとは，文書上に存在している事象を指しており，何が起こったか，誰がどのように感じたかと出来事を表す文書中の要素であると定義されている．

以上の定義より，北島，小林らの手法のアプローチはイベントである文書中の出来事を係り受け関係を持つ文節の組から抽出することを中心に構成されている．係り受け関係を持つ文節の組において，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して定義する．(主語，述語) の条件を満たす組は主語の動作や様子をイベントとして表し， (述語 1 ，述語 2) の条件を満たす組は述語 2 を述語 1 が修飾する関係にあるため， 述語 2 の特徴や様子をイベント

として表している．そのため，これらの条件を見たす組に着目することにより，文書中の出来事を係り受け関係から抽出できるとされている．

出来事を表す際の要素である主語と述語について，実際の文書中に表れる出来事に基づくと，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ものとされる．ゆえに，組を構成する要素について，文節に対して係り受け関係を獲得した後に抽出し，これらを組としてその分布を蘇生として利用したトピックモデルを作成する．そして，それに基づいて文書ごとのトピック推定を行うことで，文法的な要素を取り入れたトピック推定手法となっている．

3.2.2 係り受け関係からのイベントの獲得

北島，小林らの手法では実際に文書から文節間の依存構造である係り受け関係を獲得するために，構文解析器 CaboCha (注1) [8] を用いて係り受け解析を行っている．形態素の出現分布に基づいて文書のトピック推定を行う場合，各文書から抽出された形態素については不要語処理を行う．しかし，係り受け関係を文中から抽出した際には一般的に不要語として削除されるような助詞や助動詞のような語でも格を表現したり，時制を表現したりする文法を構成する役割が存在する場合が存在する．そのため，係り受け関係から文書の出来事を獲得する場合，これらの不要語とされる語は出来事を表す重要な役割を果たしているといえる．

文の主語に注目することで，文の主題に関する出来事の抽出を行うことができている一方で，(主語，述語), (述語 1，述語 2) といった係り受け関係を持つ文節の対を出来事としているために主題に対して並列に修飾語がいくつも付与されるような複雑な修飾関係についてはイベントとして抽出することができない．例えば,「オーシャンビューで景色の美しいホテル」というような名詞句が出現した場合，それぞれ (オーシャンビューで，景色の),(景色の，美しい),(美しい, ホテル) というように修飾関係を表す文節が複数存在するが，これらはまとめると結果としてすべて「ホテル」を修飾する文節になっているが，係り受け関係を持つ文節の対としてイベントに注目すると「ホテル」を修飾しているのは「美しい」という文節だけである．よって，係り受け関係の対に基づいてイベントを推定し，そのイベントに基づいてトピックのモデリングを行うと修飾関係が考慮されない文節が存在することがあり，文書に対して推定精度の高いトピックモデルを作成することは困難である．

4. 提案手法

従来手法である北島，小林の手法ではトピックモデリングの際の素性として主題と述語，述語と述語とが文書中のイベントを表現しているとし，これによって各文書のトピックを形成するとされていた．しかし，このイベントに基づいたトピックモデリング手法では文節間の述語項構造といった主語と述語の関係は考慮されていても，文節間の係り受け関係修飾関係が考慮されていない．その結果，生成されたトピックモデルは各文書の潜在的トピックを網羅できているとはいえないものであった．また，修飾関係を持つ文節をトピックモデルの素性として考慮したとき，ある名詞を含む文節に対して修飾関係として係り受け関係を持つ文節が存在し，さらにその文節に対して修飾関係を持つ文節が存在する場合について考える．この場合，係り受け関係を持つ文節の組によって，名詞に対する修飾関係を持つ文節を獲得するのは文節長が長くなりこれをトピックの素性とすると各文書にこの文節が含まれなくなってしまう．そのため，考えられる文節の組合せによってトピックを表現することは困難である．

そこで，係り受け関係の階層化した構文木モデルに基づいた構文解析手法を利用することにより，任意の長さの係り受け関係によって構成される文節の並びを素性としたトピックモデリング手法の提案を行う．これにより，名詞に対して修飾関係を持つ文節をトピックモデルに対して可変長な文節の並びとして割り当てることで，文書に対するトピック推定の精度を向上できると考える．

4.1 係り受け関係の階層化に基づいた構文木

係り受け関係の階層化に基づいた構文木モデルは文末の文節を根として，その文節に対して係り元となっている文節を辿ることで，階層化を行い構文木モデルが生成される [6] このとき，係り受け関係を持つ最適な長さの文節を Nested 階層 Pitman-Yor 過程 [10] に基づいて決定することで，学習データから最適な任意の長さの係り受け関係を持つ文節を構文木モデルとして生成する．構文木モデルを生成するにあたり，使用したデータは従来手法が CaboCha を使用して係り受け解析を行っているため，これと条件を揃えるため CaboCha において構文木モデルの生成に利用されている京都大学テキストコーパス [14] に含まれる，人手によって係り受けのアノテーションが付与された約 1 万文を用いる．そうして，生成した構文木モデルを利用することにより，任意の長さの係り受け関係に基づいた文節を単位とした係り受け関係を持つ文節に解析を終えた文を分割することができる．

分割を行う構文木の例を図 3 に示す．この図の左部では「私達が泊まったのはオーシャンビューで景色のキレイなホテルで、室内も落ち着いた空間でした。」という文の構文解析例を示している．ここでは文末の文節である「空間でした」から構文木に基づいて任意の長さの係り受け関係を持つ文節に分割することができる．しかし，ここでの目的はこのような文全体の構文構造を示す構文木を求めることではない．なぜならば，目的は構文解析ではなく，構文木モデルを利用した最適な係り受け関係を持つ文節によって表現される構文木の獲得であるためである．

よって，係り受け関係の階層化に基づいた構文木モデルを利用した係り受け解析手法を利用して図 3 の右図のように文節「空間でした」を根として，最適な長さを持つ構文木を選択していく必要がある．この文を係り受けの階層化に基づく構文木モデルを利用して，最適な長さの構文木を抽出したときの構文木の集合を図 4 に示す．

このように文法モデルに基づいて最適な単位に分割された構文木の分布に潜在的トピックを割り当てることで，トピックに

(注1)：http://chasen.org/taku/software/cabocha/ 2014/01/10 確認.

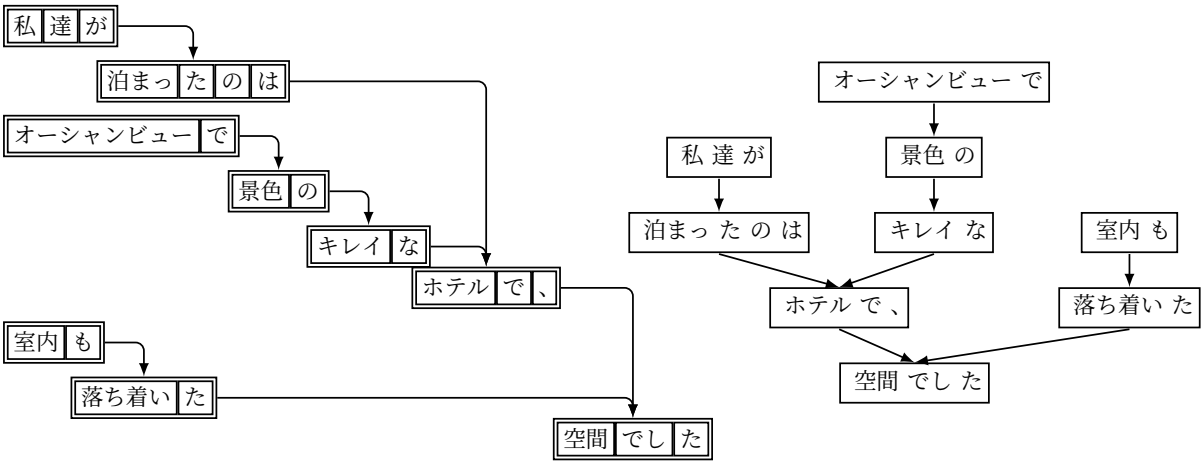

図 3　Web 文書に含まれる構文木の例

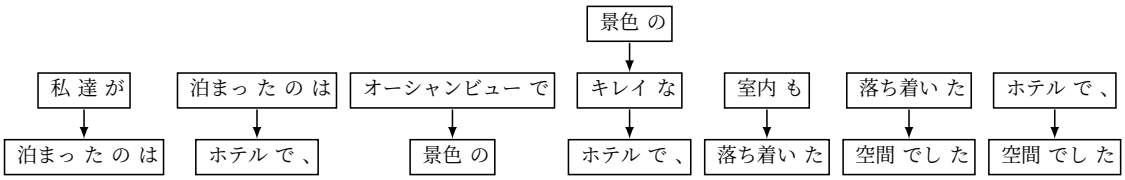

図 4　係り受けの階層化に基づいた構文木モデルによる構文木の分割例

ついてモデリングを行う．なお，このとき従来手法と同様に文節に含まれる形態素の表層すべてを素性として利用はしない．そして，従来手法と同様に，文節の格を表す助詞を除き，文節の主部である名詞，動詞，形容詞などの自立語の基本形を文節を構成する要素として，トピックモデリングの際の素性に利用することで，トピックの主題を明確にすると同時に共通のトピックを文書が包含しやすくする．

4.2 LDA のグラフィカルモデル

本稿における提案手法では，トピック推定手法として LDA を用いる．以下の図 5 に LDA のグラフィカルモデルを示す．

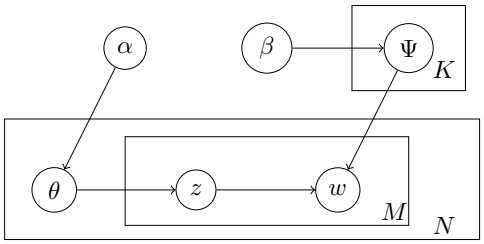

図 5　LDA のグラフィカルモデル

文書群に含まれる各文書にはディリクレ過程によって生成されるトピック分布 θ が割り当てられている．4.1 節で導出した任意の長さの係り受けで構成される文節の出現分布 w は，θ に従ってまずトピック z が選ばれ，そのトピック z に対応する任意の係り受け関係を持つ文節の分布 ϕ に従って，生成される．K, D, N_d はそれぞれトピック数，文書数，文書 d 上の係り受け関係を持つ文節の出現度数を表している．また，α と β は各ディリクレ分布 θ, ϕ のパラメータを表しており，これらはマルコフ連鎖モンテカルロ法によって推定される．

5. トピックモデリングの精度に関する評価実験

本節では，文書に存在する潜在的トピックを推定する実験を行う．実験は，単語 n-gram 分布を素性とした一般的なトピックモデリング手法と比較しても，文書ごとのトピックを高精度に推定が可能である北島，小林のイベントに基づいたトピック推定手法を従来手法として，比較実験として実験を行う．提案手法である任意の長さの係り受け関係に基づくトピックモデリング手法と従来手法の比較は，作成したトピックモデルの精度と，そのモデルを利用した文書のトピック推定の精度に関して評価を行う．本節では，作成したトピックモデルの精度に関する評価について述べる．

5.1 使用データ

使用データとして，従来手法を参考に Web 文書に分類される楽天トラベル(注2)のホテル・施設に関する評価およびレビューデータを用いる．これらを用いる理由として，レビューには各宿泊施設の部屋や立地など各対象に対して五段階評価がなされており，対象のレビューとその評価の間に相関関係が存在するため，提案手法の性能評価に適していると考えられる．

レビューには様々な文書長のものが存在するため，文書中により多くのトピックが含まれている文書はトピック推定の対象として評価に最適であると考えた．そのため，文書長が長いものほどトピックを多く含み，宿泊施設の様々な対象に対して評価が行われていると考えられる．よって，関連研究と同様に 100 字以上の文書長から構成されるレビューを対象とすることにする．実験の条件として，対象文書数はその中から無作為に選んだ 2000 件，トピック数 k は 50 件，ハイパーパラメータ

(注2)：http://travel.rakuten.co.jp/ 2014/01/10 確認.

表 1　従来手法によるトピックモデリング

トピック数	係り受け
0	窓口-予約 - ツイン-部屋 ある-思う 良い-良い
1	*-* コンビニ-ある 部屋-きれい ツイン-シングル 私-泊まる
2	ある-便利 近く-ある 歩く-ある する-思う きれい-広い
3	*-* *-宿泊 ある-思う *-泊まる この-ホテル
4	対応-良い 音-なる 対応-よい ホテル-ある 音-聞こえる
5	気-なる 広い-清潔 この-料金 すぐ-近く 大-ある
6	部屋-広い 泊まる-思う 値段-安い 目-前 部屋-なる
7	目-前 フロント-対応 この-ホテル ホテル-思う そば-ある
8	*-ある シングル-予約 ある-利用 この-ホテル ある-便利
9	ところ-ある 近く-ある この-値段 利用-思う 他-方
10	気-する *-* とても-便利 ある-ある フロント-方
11	こと-できる ある-便利 この-値段 すぐ-近く フロント-人
12	駅-近い ある-便利 ある-いい ホテル-宿泊 朝食-バイキング
13	この-値段 *-宿泊 *-ある 隣-部屋 2-宿泊
14	ある-思う 思う-思う （-* 予約-する こと-できる
15	ある-ある 「-窓口 *-お コストパフォーマンス-高い 1-1
16	機会-ある 部屋-バス *-なる 利用-利用 きれい-ホテル
17	近い-便利 この-ホテル 近い-ある できる-思う ホテル-利用
18	部屋-きれい ある-ある よく-ある 方-いい 夜-遅い
19	交通-便 フロント-対応 泊まる-こと 言う-こと 冷蔵庫-ある
20	とても-良い ホテル-ある ある-ある ダブル-部屋 言う-こと
21	この-ホテル とても-便利 こと-ない 思う-こと ところ-ある
22	利用-思う 気-なる 行く-とき こと-ない 利用-ホテル
23	近く-ある 気-なる *-ある *-利用 ある-ホテル
24	便利-思う *-* とても-きれい とても-親切 気持ち-良い
25	気-なる 部屋-泊まる 好感-持てる 思う-ある 部屋-狭い
26	部屋-広い 値段-安い （-ある 問題-ある こと-ある
27	ホテル-窓口 「-窓口 気-する なる-思う 感じ-する
28	近く-ある *-ところ ちょっと-遠い *-する 窓-見える
29	部屋-広い こと-できる よい-思う 広い-きれい ホテル-思う
30	近く-ある ある-利用 また-泊まる *-安い ホテル-ある
31	泊まる-思う 泊まる-部屋 *-* 音-聞こえる 値段-安い
32	コンビニ-ある 便-良い 前-ある 利用-思う ある-よい
33	いい-思う （-* 良い-思う お-思う ある-泊まる
34	*-ない この-ホテル ホテル-思う ホテル-* 近く-便利
35	こと-ない *-宿泊 「-窓口 ない-思う 思う-思う
36	*-ある 部屋-広い この-値段 この-価格 すぐ-ある
37	感じ-する 泊まる-こと 距離-ある シングル-宿泊 ツイン-部屋
38	ある-良い 機会-ある 部屋-ある ホテル-利用 ツイン-泊まる
39	この-ホテル する-思う 少し-狭い 近く-ある 広い-ある
40	*-ある 良い-ホテル ホテル-ある ある-思う 思う-思う
41	フロント-対応 部屋-きれい ホテル-思う 部屋-ある 駅-近い
42	ある-ある ホテル-思う この-ホテル ある-利用 なる-思う
43	いい-ホテル 良い-ホテル 前-ある *-* シングル-予約
44	部屋-狭い 利用-思う 思う-思う また-泊まる 気-なる
45	コンビニ-ある こと-ある ある-便利 部屋-広い 清潔-ある
46	部屋-狭い また-ある 少し-離れる *-* ある-ホテル
47	気-なる この-ホテル 利用-思う *-思う ホテル-ある
48	駅-近い 大-ある 交通-便 ちょっと-高い 気-なる
49	気-なる ある-ある 「-窓口 良い-ホテル 思う-*

表 2　提案手法によるトピックモデリング

トピック数	係り受け
0	部屋-便利 予約-* *-する-立つ いい-思う 1-位置-思う
1	こと-思う *-する-する ビジネス-思う *-送る-する すぐ-思う
2	思う-思う 方-思う 利用-思う （-思う ホテル-する
3	ホテル-思う-思う この-思う ホテル-ある-思う この-思う-思う ホテル-利用-思う
4	方-思う 広い-思う 泊まる-ホテル コンビニ-思う *-思う-思う
5	ある-* する-思う ある-ホテル 良い-思う ホテル-ある
6	（-思う 良い-* *-はず 部屋-便利 *-是非
7	部屋-思う コンビニ-ある-便利 大変-思う 部屋-なる とても-思う
8	値段-思う 料金-思う お-思う *-探す-到着 *-感じ-庶民
9	*-* 部屋-思う 利用-* *-便利 もの-思う
10	事-事-事 *-思う-同じ *-部屋-便利 （-思う 場所-思う
11	ある-なる 朝食-思う-思う ない-事-事 ある-お ある-ある
12	*-ホテル *-*-ある ある-ある-ある する-思う 和歌山-思う
13	朝食-思う 部屋-* *-ご ホテル-お コンビニ-思う
14	「-思う 私-思う *-何える-はず *-なる ここ-思う
15	*-思う とき-思う お-思う-思う *-付ける-（ *-寂しい-わかる
16	ホテル-ホテル 方-思う 利用-思う-思う こと-思う ある-思う
17	ホテル-思う 部屋-思う とても-思う 1-思う *-ある-思う
18	部屋-思う-思う 部屋-ホテル 宿泊- ある-する *-できる
19	*-思う *-ある-ある この-思う *-満足 利用-ホテル
20	ホテル-思う 利用-思う 思う-ある *-感じ *-*-する
21	朝食-思う ホテル-ある 駅-思う *-おすすめ 予約-思う
22	部屋-ホテル 部屋-思う *-思う-ある ホテル-ある-できる *-する
23	部屋-思う 部屋-なる-思う 時-思う 対応-思う また-ホテル
24	*-事-事 とても-思う-思う 2-思う ある-思う 「-思う
25	フロント-思う ホテル-ホテル ある-ある コンビニ-ある-便利 1-ある
26	「-* 泊まる-思う （-思う *-ある ある-利用
27	*-思う *-思う-思う *-宿泊-思う この-思う *-思う-なる
28	*-思う-思う *-利用 宿泊-思う 2-思う 良い-思う-思う
29	（-* ある-ある-* *-便利-伺える ない-思う 広い-ホテル
30	思う-思う *-いける-つく （-ホテル *-*-いく 部屋-ホテル
31	お-思う こと-思う *-感じる 部屋-ある *-いい-是非
32	きれい-思う *-ない-良い *-行く *-ある-感じ 私-思う
33	1-ホテル *-思う-嬉しい ホテル-行く 部屋-思う また-思う
34	ホテル-思う とても-お *-いく 方-利用 ある-いい-ある
35	*-*-* *-する 予約-思う *-* *-ソ
36	*-良い 安い-思う 朝食-良い ある-する-する 利用-思う
37	1-思う 近く-思う *-*-わかる ところ-思う 使う-思う
38	部屋-思う *-思う *-泊まる ホテル-* ある-ある-なる
39	ある-思う-思う 良い-思う 部屋-事-事 ホテル-思う ある-ある-思う
40	ある-思う 部屋-お する-思う ホテル-ある-便利 コンビニ-便利
41	*-うれしい ある-思う 狭い-思う また-思う *-到着-満足
42	*-内訳-寝る *-残念 *-名称-同じ 思う-思う *-する-思う
43	ツイン-思う 気-思う *-つぶれる できる-思う とても-なる-思う
44	*-お よい-思う こと-* 泊まる-思う 泊まる-思う-思う
45	*-思う-* 行く-思う *-ある-* また-思う *-コイン-利用
46	ちょっと-思う ある-便利 広い-思う ホテル-ある-ある *-魅力
47	*-ホテル ホテル-思う ホテル-* 部屋-思う *-ある
48	部屋-* 歩く-思う 思う-思う-思う 人-思う *-寝る-*
49	フロント-思う 部屋-* なる-思う とても-思う この-思う

θ, ϕ の決定にはギブスサンプラの一種として知られている崩壊型ギブスサンプリングを用いて，従来手法と提案手法においてレビューのトピック分類を行った．

5.2　実 験 結 果

北島，小林のトピックモデリング手法を従来手法とし，提案手法によるトピックモデリングの結果を比較する．使用データに対して，トピックモデリングを行った結果を表 1，表 2 に示す．

5.3　各トピックモデルの評価

5.1 節の使用データで示したようにレビューデータとその評価の間には相関関係があり，一般的に肯定的な意見を含むレビューは評価が高く，否定的な意見を含むレビューは評価が低いと考えられる．ゆえに，肯定的な意見を含むレビューと否定的な意見を含むレビューデータはそれぞれ各トピックに分類される可能性が高い．したがって，こうした評価に応じてトピックの分類がなされているか考察することで評価を行う．

これら二つのトピックモデルは語と語との依存関係あるいは修飾関係である係り受け関係の分布に基づいているため，どの対象に対して，どのような形容詞が使用されているかといったことに着目する．

6.　お わ り に

本稿では，文書集合に含まれる各文書に存在するトピックを推定する際のトピックのモデリング手法について日本語を対象とし，新たなトピックモデリング手法の提案を行った．

日本語においては語の間に存在する依存構造が文節間の係り受け関係によって表現されているが，これを利用したトピックモデリング手法がイベントに基づいたトピック推定手法として北島，小林によって提案されていた．しかし，北島，小林のトピックモデリング手法では，係り受け関係の対を単純にトピックを構成する要素として考慮しているのみであり，係り受け関係が生起するときの文節のその他の関係である，前後の係り受け関係や文節の並びといった要素は考慮されていない．本稿ではこの従来の依存構造の対に基づいたトピックモデリング手法に対し，我々が以前に提案した係り受けの階層化に基づいた構文木モデルを利用することにより，より高精度なトピックモデリング手法を提案することができると考え，これを提案した．

評価実験では，トピックの分類精度として主観による評価以

外に検索課題に基づいた評価を行うことで，従来手法と比較して提案手法においてより潜在的なトピックを捉えることができることを確認した．

なお，今後の課題として提案したトピックモデルの客観的評価として，文書検索課題における評価が必要であり，今後はテストデータの作成およびに検索課題の設定し，その精度の評価を行う予定である．

謝　　辞

本研究では，楽天株式会社及び国立情報学研究所の許諾を頂き，楽天トラベルのデータを利用させて頂きました．ここに感謝の意を表します．

文　　献

[1] Xing Wei and W. Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pp. 178–185. ACM, 2006.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[3] Jordan Boyd-graber and David Blei. Syntactic Topic Models. In *Neural Information Processing Systems*, pp. 185–192, 2008.

[4] Yuening Hu and Jordan Boyd-graber. Efficient Tree-Based Topic Modeling. In *Association for Computational Linguistics*, No. July, pp. 275–279, 2012.

[5] 北島理沙, 小林一郎. 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. In *DEIM Forum 2011*, No. F4-2, pp. 1–8, 2011.

[6] 大野一樹, 波多野賢治. 係り受け関係の階層化とその共起に基づいた構文木モデルを利用した構文解析手法の提案. 情報処理学会研究報告, 第 2013-NL-214 巻, pp. 1–6, 2013.

[7] Taku Kudo and Yuji Matsumoto. Japaese Dependency Structure Analysis based on Support Vector Machines. In *Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 18–25, 2000.

[8] Taku Kudo and Yuji Matsumoto. Japanese Dependency Analysis using Cascaded Chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pp. 63–69, 2002.

[9] Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Association for Computational Linguistics*, pp. 985–992. Association for Computational Linguistics, 2006.

[10] Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *In Proc. of ACL*, 2009.

[11] Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. Japanese Dependency Structure Analysis Based on Maximum Entropy Models. In *Conference of the European Chapter of the Association for Computational Linguistics*, pp. 196–203, 1999.

[12] Tadao Kasami. An efficient recognition and syntax-analysis algorithm for context-free languages. Technical Report AFCRL-65-758, 1965.

[13] Satoshi Sekine, Kiyotaka Uchimoto, and Hitoshi Isahara. Backward beam search algorithm for dependency analysis of Japanese. In *Association for Computational Linguistics*, pp. 754–760, 2000.

[14] Sadao Kurohashi and Makoto Nagao. Building a Japanese Parsed Corpus while Improving the Parsing System. In *Language Resources and Evaluation Conference*, pp. 719–724, 1998.