共起語を利用した事象系列に基づくトピック推定

Identifying Topics Based on Event Sequence Using Co-occurrence Words

若林 啓♥ 三浦 孝夫♠
Kei WAKABAYASHI Takao MIURA

本稿では，文書のトピック推定の手法を提案する．文書分類では一般的に，文書を単語の出現頻度ベクトルなどでモデル化することが多いが，本研究では，文書として一連の事件を報じた新聞記事の系列を用いることで，系列構造によるモデル化を行う．また，事件は事象の系列であると考え，文書に出現する動詞およびそれを特徴づける共起語の集合を事象に対応させ，確率過程の手法に基づいて事象系列の尤度を与える方法を提案する．この方法によるトピック推定のアルゴリズムを示し，実験によりその有用性を確認する．

In this paper, we propose a sophisticated technique for topic identification of documents based on event sequences using co-occurrence words. There have been many investigations for document classification based on vector space modeling, but here we consider each document as an event sequence each event as a verb and words correlated with the verb. We propose a new method for topic classification of event sequences by using Markov stochastic process modeling. We show some experimental results to examine the method.

1. はじめに

近年，計算機上で利用可能な文書データの増加に伴い，より高度な文書処理技術が必要とされている．この現状を背景にして，文書分類技術に関する研究が盛んに行われている．計算機による文書分類は，一般的に文書データを出現単語ベクトルでモデル化し，文書の類似度をコサイン尺度などで定義することによって自動的な分類を実現する．

しかしベクトルモデルでは単語の順序の情報を直接扱わないため，文書が表現するトピックを特定できないことがある．例えば，特定の事件や出来事を時系列に沿って客観的に記述した文書の場合，内容の理解には順序を考慮する必要がある．

文章の順序が時間的な順序に対応している文書を内容に基づいて分類するためには，系列構造による文書のモデル化を行うことが望ましい．本研究では，特定の事件に関する新聞記事の系列を，ひとつの「事件」を記述した文書と考え，そのような文書を対象にしたトピック分類を目的とする．

トピックを扱う代表的なアプローチの一つに，Topic Detection and Tracking（TDT）がある[1]．TDT では，トピックは事象（event）によって特徴付けられる．事象とは，位置的，時間的に特定の，個々の発生した事実を意味する．TDT の Event tracking タスクでは，ある事象に関して述べている文書を逐次的に分類する[11]．Makkonen らは，出来事を分岐する事象の系列と考え，日付のある文書集合から出来事を発見する手法を提案している[6]．これらの研究では，文書の類似度に基づいてトピックを発見する．

本研究では系列情報を反映したトピック分類の手法を提案する．事象系列を考慮した分類手法は過去にあまり積極的な提案はなされていない．これは，決定木や SVM，自己組織化マップ（SOM），単純ベイズ[7] といった従来の分類手法では，ベクトルの分類に問題を帰着させるため，系列情報を反映させることが容易ではないことによる．本稿では，トピックを事象系列のクラスと考え，確率過程に基づく文書分類を提案する．

我々はこれまでに，新聞記事の系列のトピック推定を行う手法を提案している[12]．ここでは，文末の動詞が文書の事象としての要約になっていることを利用して，隠れマルコフモデルの出力シンボルに動詞を対応させ，事象系列のトピック推定を行う．実験では，隠れマルコフモデルが文書の分類器としてうまく働くことを示している．

しかし，表現する事象によっては，動詞単独では事象の記述として十分でないことがある．例えば，「捜査を始めた」と「事情聴取を始めた」など，同じ動詞でも共起する格によって意味が大きく異なることがある．本稿では，動詞の共起語を事象の特徴量として利用し，より文書の内容を反映させたトピック分類の手法を提案する．

関連研究として，確率過程を用いて文書をモデル化する手法に，Mulbregt らがある[9]．ここでは文書を話題の系列と考え，系列の隠れ状態を推定することで文書のトピックセグメンテーションを行っている．彼らはセグメンテーションを目的としているため，系列そのものには意味を対応させず，話題の遷移確率は話題の境界に対するペナルティーとしてのみ作用している．

文書の構造を系列でモデル化して推定する研究に，Barzilay らがある[3]．ここでは，地震などを報じる文書には報じる内容の順序に特徴があることを利用して，確率過程モデルを用いて文書の構造を推定する．彼らは．文章の bigram 分布を内容に対応させ，あらかじめクラスタリングした文章を学習データとして分布のパラメタを推定する．この手法は単語分布に依存するため，文書の表現のパターンを扱うのに適しているが，この方法をそのままトピックのパターンに適用することはできない．

確率過程による文書のモデル化を談話構造の解析に応用する研究に，柴田らがある[10]．ここでは日本語の料理番組のナレーションを対象として，用言の格フレームをシンボルとみなした隠れマルコフモデルを用いている．実験では，動詞に着目することで，事象の遷移をうまく捉えられることを示している．

2 章では事象系列に基づくトピック分類について述べる．3 章では本研究で用いる確率過程モデルについて説明する．4 章で本稿で提案するトピック推定のアルゴリズムの説明を行う．5 章で実験結果を示し，6 章で結びとする．

2. 事象系列によるトピック分類

本稿では，「事象」は位置的，時間的に特定される個々の事柄，発生した事実を意味する．「トピック」は，一連の事

♥ 学生会員 法政大学大学院工学研究科修士課程 kei.wakabayashi.bq@gs-eng.hosei.ac.jp
♠ 正会員 法政大学工学部情報電気電子工学科 miurat@k.hosei.ac.jp

件やテーマに関する事象系列のクラスを意味する．

２つの事件が似ているかどうかは，類似性の定義によって異なる．本研究では事象の系列が似ている事件を「類似する」とする．例えば，東京で起きた強盗事件と広島で起きた強盗事件は，場所も犯人も盗品も違う．しかし，どちらも「盗まれた」，「指名手配された」，「犯人が逮捕された」のように事象の系列が類似しているため，ここで両事件は「類似する」とする．

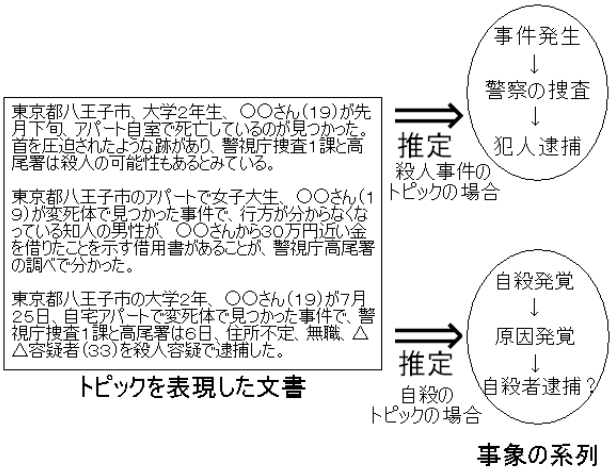


図１　文書のトピック推定

Fig.1　Estimating Topic in a Document

図１は，トピック推定の例である．「殺人事件」のトピックを表現している文書集合に対し，新聞記事の第一段落を日付順に連結して時系列に並べる．この例は，ある殺人事件に関する新聞記事を連結したものである．

図の右側には，この文書でたどっている「事象の系列」を示す．個々の事象は，この事象系列が何のトピックであるかによって解釈が異なる．記事内容は，ある人物の死亡の発見，警察による犯人の手がかりの発覚，容疑者が逮捕されるという事象を表し，これらの事象が，「殺人事件」という特性を述べていると考えることができる．他方，この文書がある人物の自殺に関するトピックならば，推定される事象の種類として，自殺の発覚のあとに自殺の原因の特定が続くであろう．殺人事件の最後では，「逮捕」に関する事象が通常生じる．自殺事件の場合でも，逮捕された容疑者が獄中で自殺を図ることが考えられるが，すでに自殺発覚の事象があるため，発生順序が不自然である．即ち，殺人事件トピックは自殺トピックとは通常両立しない系列を有している．

このように，事象の発生順序にはトピックごとに特徴があると考えられるため，本稿では，事象の遷移に着目することで事件のトピック分類が可能であることを論じる．

3. 確率過程モデル

ここでは，トピック推定に用いる確率過程モデルを定義する．本研究では，確率過程モデルとして確率過程オートマトンを用いる．確率過程オートマトンは，確率的な状態遷移と各状態からの確率的な出力をもつオートマトンである．ここでは状態は，単純マルコフモデルに基づいて遷移する．モデルの枠組みは隠れマルコフモデル[5] と同じであるが，隠れマルコフモデルでは状態は観測できず，主に状態の推定問題に適用される．これに対し，本稿で用いるモデルでは状態は観測可能であり，系列の確率を求めることのみを目的とする．

3.1　モデルの定義

与えられた状態の集合 Q と，状態から出力されるシンボルの集合 V について，確率過程オートマトンを定義する．図 2 は，本稿で提案する確率過程モデルの例である．図中の楕円は状態を表し，それぞれの状態は１つ以上の共起語を出力する．本稿では，状態には動詞が対応し，シンボルには動詞の共起語集合が対応する．例えば，「警察が捜査を始めた」という文章からは,状態として「始める」,シンボルとして{「警察」,「捜査」}を得る．

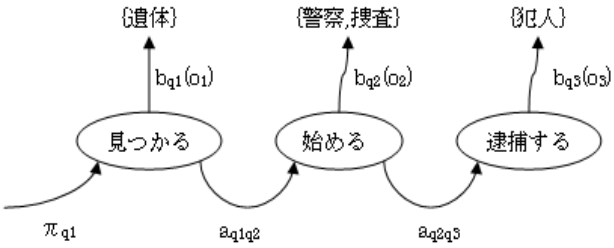


図２　確率過程モデル

Fig.2　Markov Stochastic Model

本稿で用いる確率過程オートマトンは，以下に示す 5 つのパラメタ(Q,Σ,A,B,π) で定義する．

(1) $Q = \{q_1,\ldots,q_N\}$：状態の有限集合．

(2) $\Sigma = \{o_1,\ldots,o_M\}$：出力の有限集合．Σはシンボルの集合 V の冪集合である．すなわち，$\Sigma \subseteq 2^V$であり，o_t は V の部分集合である．

(3) $A = \{a_{ij}, i,j = 1,\ldots,N\}$：状態遷移確率分布．$a_{ij}$ は状態 q_i から状態 q_j への遷移確率であり，$a_{i1}+\ldots+a_{iN} = 1.0$ である．動詞 i が観測された後，次に観測される動詞の確率分布に対応する．

(4) $B = \{b_i(o_t), i = 1,\ldots,N, t = 1,\ldots,M\}$：シンボル出力確率分布．$b_i(o_t)$ は状態 q_i が o_t を出力する確率であり，$b_i(o_1)+\ldots+b_i(o_M) = 1.0$ である．動詞 i が観測されたとき，同時に観測される共起語の確率分布に対応する．ひとつの動詞が複数の共起語を伴って出現しているとき，$b_i(o_t)$ はそれぞれの共起語 $o_{t,k}$ の出力確率の積として求める．すなわち，シンボル集合 $o_t = \{o_{t,1}, o_{t,2}, \ldots, o_{t,|t|}\}$ が状態 i から出力される確率は，$b_i(\{o_{t,1}, o_{t,2}, \ldots, o_{t,|t|}\}) = \Pi_k b_i(o_{t,k})$とする．

(5) $\pi = \{\pi_i, i = 1, \ldots, N\}$：初期状態確率分布．$\pi_i$は状態 q_i が初期状態である確率である．文書で最初に観測される動詞が i である確率に対応する．

本稿では，確率行列 A は単純マルコフ過程に基づく状態遷移確率に対応する．これは，次の状態の確率は現在の状態にのみ依存することを意味する．例えば，「見つかる，始める，逮捕する」という状態列において「逮捕する」が観測される確率は，前の状態「始める」にのみ依存して決定される．同様に，シンボル集合の出力確率は現在の状態にのみ依存する．例えば，{「警察」,「捜査」}というシンボル集合が出力される確率は，現在の状態「始める」にのみ依存して決定される．本研究では，状態とシンボル集合は語として文書から観測できる．

3.2　モデルの算出

本稿の確率過程モデルは 5 つのパラメタ(Q,Σ,A,B,π) から成るが，Q および Σは事前に与える．ここではモデルの算出として，状態遷移確率分布 A，シンボル出力確率分布 B，および初期状態確率分布πを学習によって計算する方法に

ついて述べる．本研究では，モデルの算出に教師あり学習を用いる[8]．各トピックについて，事前に学習データを用意する．これらは人手により正しく分類されているものとする．これらの学習データを用いることで，トピックにおける事象系列パターンをモデルのパラメタに反映させる．

まず，学習データ D から状態とシンボルの系列を抽出し，状態遷移回数，シンボルの出力回数をカウントする．それぞれの頻度から，相対頻度を確率値として用いる．すなわち，D_i を最初の動詞が i である文書の集合，V_{ij} を動詞 i から j への遷移回数，W_{ik} を動詞 i が語 o_k と共起して出現した回数とすると，それぞれのパラメタは以下の推定式で求める．

$$\pi_i = \frac{|Di|}{|D|},\quad a_{ij} = \frac{Vij}{\sum_j Vij},\quad b_i(k) = \frac{Wik}{\sum_k Wik}$$

モデルのパラメタが決定しているとき，系列の確率を求めることができる．次節に示すように，系列は動詞列に対応する状態列 $q = \{q_1,q_2,\ldots,q_T\}$ と，共起語集合列に対応するシンボル集合列 $o = \{o_1,o_2,\ldots,o_T\}$ の組から成る．モデル M が<q,o>を出力する確率 $P(q,o|M)$ は以下のように求める．

$$P(q,o|M) = \pi_{q1} b_{q1}(o_1) \prod_{t=1}^{T-1} a_{qtqt+1} b_{qt+1}(o_{t+1})$$

4. トピック推定

ここでは本稿で提案する，文書のトピック推定アルゴリズムについて述べる．本研究では，尤度原理に基づいてトピック推定を行う．それぞれのトピックに対応する確率過程モデル M_1，…，M_L においてテストデータの尤度を求め，尤度を最大にするトピック m を文書のトピックと推定する．図３は，本手法によるトピック推定の例である．本章では，推定アルゴリズムの詳細を示す．

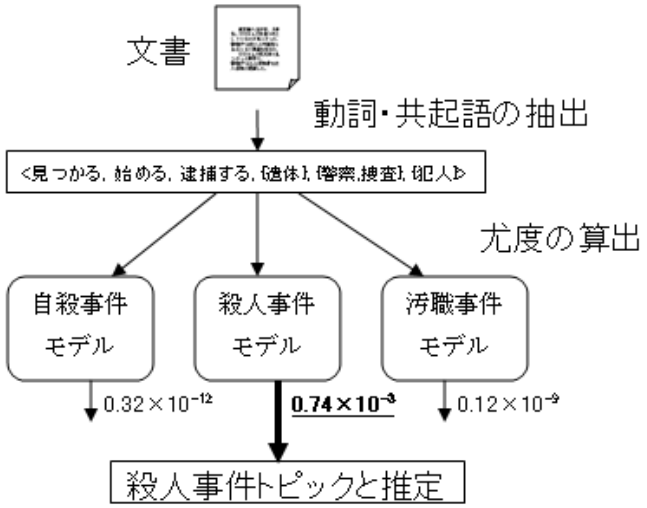


図３　確率過程モデルによるトピック推定

Fig.3　Estimating Topic using Markov Stochastic Model

4.1　確率過程モデルの適用

本研究では，各トピックに確率過程モデルを対応させ，尤度原理を適用することでトピック推定を行う．本稿では，文書を一つの事件に関する記事系列とする．それぞれの新聞記事において第一段落には最も重要な内容が含まれていると考え，各記事の第一段落を時系列順に連結したものを文書として扱う．

事件は，トピックの内容に依存した事象の種類から成る．例えば，殺人事件トピックでは「容疑者の逮捕」，「被害者の発見」といった事象が考えられるが，自殺事件トピックでは「遺書の発見」など異なる事象が生じる．しかし，入力文書に出現する全ての語をモデルに反映させると特徴的な事象系列パターンの推定が困難になる．このため本研究では，図４に示すように，その時点で起きた事象を表現する部分として文末の動詞とそれに関連する共起語のみを用いる．これは，日本語の文章において，状況の変化を表現するために各文章末の動詞を用いることが多いという特徴による．

しかし日本語の場合，動詞に関連する共起語は主格や目的格に関わらず，動詞よりも前の，どの位置にも出現する可能性がある．これは，助詞によって格を表現するという日本語の特徴による．

このため本研究では，共起語の抽出に EDR 電子化辞書の日本語共起辞書を用いる[13]．この辞書は，コーパスの解析により抽出された語句同士の係り関係を列挙したものであり，各語句に対して品詞や意味の情報が付与されている．本稿では，この共起辞書から名詞句から動詞句への係り関係のみを取り出したものを動詞の共起語の抽出に用いる．共起辞書の一部を図４右に示す．

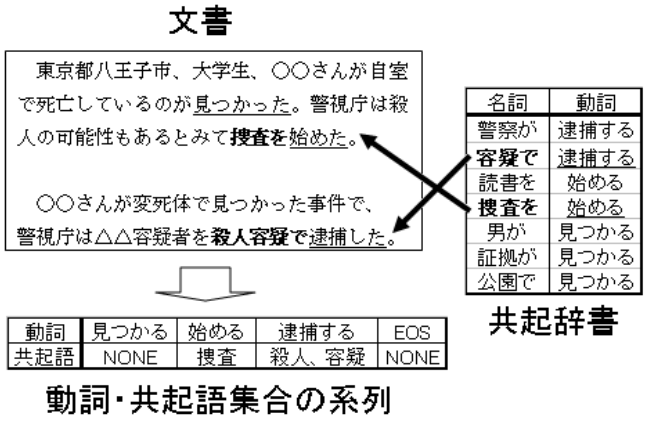


図４　動詞と共起語の抽出

Fig.4　Extracting Verbs and Co-occurrence Words

本研究では，各文書は動詞と共起語集合の系列であると考える．また，動詞を状態に，共起語集合をシンボルに対応させた確率過程モデルを用いて，各トピックに固有の事象系列パターンを形式化する．

4.2　状態およびシンボルの抽出

ある文書 D が与えられたとき，対応する状態およびシンボルの系列<$(q_1,q_2,\ldots,q_n),(o_1,o_2,\ldots,o_n)$> を与える関数 g を考える．ここで，$q_i$ は i 番目の状態であり，o_i は i 番目のシンボル集合である．

文書は読点（。）で区切られた文章列とする．まず，それぞれの文章に対して形態素解析を行い，最後の形態素が過去を表す助動詞「た」でない文章を取り除く．これは，「死因の特定を急ぐ」，「可能性もあるとみている」，など状況の変化を伴わない記述を除去するためである．最後の形態素が「た」である場合は，その直前に動詞があれば，その動詞を取り出し，v_t とする．

次に，v_t と共起関係にある名詞句を抽出する．v_t と同じ文章に含まれる v_t より前の全ての名詞句について，一番後ろに現れている名詞以降の形態素を文字列として取り出す．例えば，「太郎を殺人容疑で逮捕した」という文章からは，動詞

として「逮捕」，名詞句として「太郎を」，「殺人容疑で」を抽出する．「殺人容疑で」という名詞句は，「殺人」「容疑」「で」の 3 形態素から成るため，一番後ろに現れている名詞である「容疑」以降の文字列「容疑で」を取り出す．この文字列について共起辞書を参照し，v_t との共起レコードが存在するか調べる．レコードが存在する場合，その名詞句に含まれる名詞を全て共起語として抽出し，その集合を n_t とする．この例では，「容疑で，逮捕」のレコードが存在する場合，{「殺人」，「容疑」}を共起語として抽出する．共起語がひとつも抽出されず，n_t が空集合である場合は，共起語が無いことを意味するシンボル「NONE」を n_t に加える．

この操作を文書 D の全ての文章に対して行い，得られた v_t の系列を状態列 q，n_t の系列をシンボル列 o とする．さらに，q の末尾に終端を意味する状態「EOS」を加え，それに対応するシンボルとして「NONE」を o の末尾に加える．q_t および o_t の順序は，文書中の出現順序に一致させる．文書 D について関数 g は次のようになる．

$g(D) = <(q_1,q_2,\ldots,q_T,EOS),(o_1,o_2,\ldots,o_T,NONE)>$

4.3 モデルの学習およびトピックの推定

ここでは，学習によってモデルのパラメタを算出し，トピック推定を行う手法を示す．

トピック c に対応するモデルを M_c とし，M_c の学習用としてトピック c の文書集合 D_c が与えられているとする．D_c に含まれる全ての文書について g(D)を求め，状態遷移回数，シンボル出力回数をカウントする．この頻度から，3．2 節で示した推定式を用いてモデル M_c のパラメタを決定する．

本研究では尤度原理に基づいてテスト文書 D のトピックを推定する．g(D) = <q,o>について，モデル M_c が<q,o>を生成する確率 $P(q,o|M_c)$を求める．文書 d のトピックは，$P(q,o|M_c)$の値を最大にするようなトピックと推定する．すなわち，文書 d のトピック c_d は，

$$c_d = \mathrm{argmax}_c \, P(q,o|M_c)$$

と推定する．

5. 実験

5.1 実験方法

本稿では 3 つのトピックに分類した 256 件の文書を，毎日新聞 2001 年，2002 年の 2 年分から人手で用意する．その内訳を表 1 に示す．ここで扱うトピックは次の 3 つである．

単独犯事件：犯人が一人あるいは少数による殺人や強盗事件のトピック

組織犯事件：組織的な殺人や強盗事件のトピック

汚職事件：企業や政府などの要人による汚職事件のトピック

それぞれのトピックで，用意した文書の一部を学習用，残りをテスト用とする．提案アルゴリズムに従い，学習文書を用いてモデルの学習を行い，テスト文書それぞれに対してトピックの推定を行う．トピック推定の結果と人手による分類との一致率で提案アルゴリズムの評価を行う．

表 1 実験データ
Table 1 Test Corpus

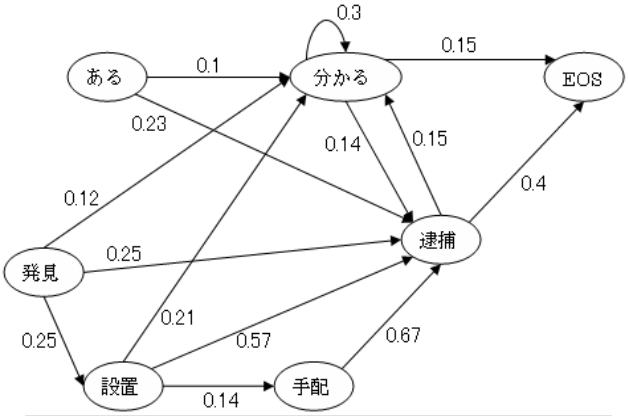
トピック	学習文書数	テスト文書数
単独犯事件	91	45
組織犯事件	35	17
汚職事件	46	22

形態素解析には日本語形態素解析ツール「Chasen」[2] を用いる．共起辞書には，EDR 電子化辞書の「日本語共起辞書」[13]を用いる．

5.2 実験結果

まず，学習アルゴリズムで得られた確率過程モデルの構造を示す．ここで，モデルの構造とは，モデルの状態間の関係および状態が出力するシンボルの確率分布を意味する．この構造を見ることで，トピックの事象系列の特徴をモデルが表現していることを確かめる．

図 5 に，単独犯事件のモデルの構造の一部を示す．円は状態を表し，矢印は遷移確率の大きい状態遷移を確率値と共に示している．また図の下には，それぞれの状態から出力されるシンボルのうち，確率の大きいものを列挙してある．



動詞	共起語
ある	通報、110番
発見	遺体、一室、路上
設置	捜査、本部、県警、署
手配	容疑、殺人、強盗、傷害
分かる	こと、調べ、供述、可能性、疑い
逮捕	容疑、疑い、殺人、死体、遺棄、県警、強盗、傷害

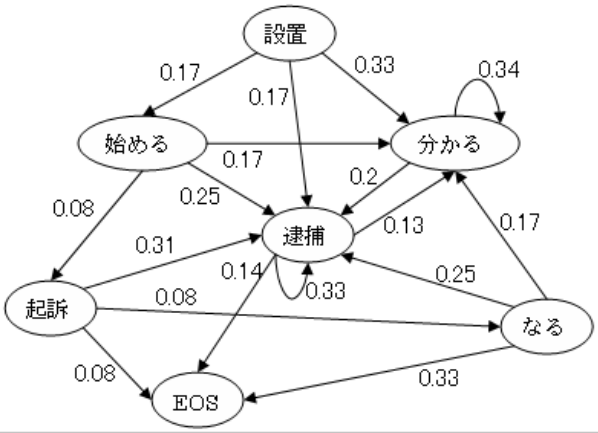
図 5 単独犯事件の構造

Fig.5 One-man Crime

この構造から，単独犯事件モデルの特徴を考察する．まず，出力される確率の高い共起語から，「ある」は「通報があった」，「発見」は「遺体が発見された」という文脈で出現することが多い．これらの動詞は初期状態である確率が高く，他の状態からの遷移確率は低い．これは単独犯事件における事件発生の特徴を表している．また，「逮捕」から「EOS」への遷移確率が高いことは，単独犯事件の特性から，一人の犯人が逮捕されることで事件が解決する特徴を表していると言える．

図 6 は，組織犯事件のモデルの構造である．このモデルでは，「逮捕」から再び「逮捕」に遷移する確率，「起訴」から「逮捕」に遷移する確率が高い．これは，組織犯事件トピックでは複数の犯人が存在し，複数の逮捕が起こるという特徴を表していると考える．

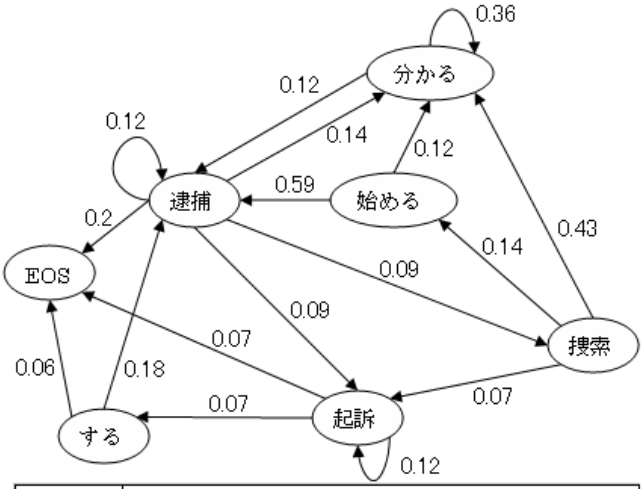
図 7 は，汚職事件のモデルの構造である．このモデルは他のトピックに比べ，共起語に特徴がある．例えば，「逮捕」には「贈賄」「収賄」「詐欺」，「始める」には「家宅捜索」，「する」には「懲戒免職」といった単語が共起しやすい．これらの単語は汚職事件の特徴を表している．



動詞	共起語
設置	捜査、本部、県警、特別、対策
起訴	容疑者、地検、罪、事件、殺人
始める	捜査、調査、容疑、死体、遺棄、殺人、本部
なる	浮き彫り、事件、捜査、可能性、明確、段階
分かる	こと、調べ、供述、実行、疑い、可能性、目撃
逮捕	容疑、疑い、殺人、死体、遺棄、保険金、違反

図 6　組織犯事件の構造

Fig.6　Organizational Crime



動詞	共起語
始める	取調べ、家宅、捜索、容疑、聴取、疑い
捜索	自宅、特捜、地検
起訴	事件、地検、容疑者、地裁、罪
する	処分、こと、懲戒、免職、事件、証言、電話
分かる	こと、調べ、話す、供述、疑い
逮捕	容疑、疑い、贈賄、収賄、背任、違反、詐欺、妨害

図 7　汚職事件の構造

Fig.7　Corruption Scandal

表 2　分類結果
Table 2 Classification Ratio

トピック	テスト文書数	正解数	正解率
単独犯事件	45	32	71.1(%)
組織犯事件	17	8	47.1
汚職事件	22	14	63.6
合計	84	53	64.3

テスト文書のトピック分類を行った結果を表２に示す．表にはトピック別に正解率を示している．全トピックの合計で64.3%の正解率を得た．単独犯事件の分類率が 71.1% と最も高く，組織犯事件の分類率は 47.1% と最も低い．

5.3　考察

実験結果の考察と，提案アルゴリズムの評価を行う．

第一に，本実験の正解率が示すように，単独犯事件に関する結果がよい．文書を詳しく見ると，単独犯事件の特徴として犯人が一人であるため，逮捕に関する事象で事件が終わることが多い．これは単独犯事件モデルの特徴と一致しており，分類率が高い一因と考えられる．

先行研究[12]との比較では，汚職事件の分類率が先行研究で 45.5%であったのに対し，本実験では 63.6%と大きく改善されている．これは文書の特徴量として共起語を新たに用いたことによる．実際，学習したモデルの構造を見ると，汚職事件で頻出する共起語は他のトピックと比べ特徴的である．

一方，組織犯事件は分類率が最も低い．これは，出現する動詞および共起語が単独犯事件と類似しており，さらに単独犯事件に比べ，組織犯事件には多様な事象系列パターンが存在するためであると考えられる．しかしながら，文書に現れる語が類似していても 47.1% とある程度の分類が可能であることから，提案アルゴリズムが出現する語に依存しておらず，従来の文書分類とは本質的に異なることが言える．

学習したモデルの共起語を見ると，動詞と共起語の関係が不明な場合がある．例えば，「設置」の共起語として「本部」が高頻度で現れているが，「本部が設置した」，「本部を設置した」など，複数の関係で共起することが考えられる．しかし，日本語文章では同一の意味でも異なる格助詞を伴うことや，同一の格助詞でも異なる関係を意味することがあるなど，表層の情報から動詞との関係を与えることは容易ではない．

また，本手法では固有名詞を共起語としてモデルに反映することが難しい．これは，発生頻度が少ないため共起辞書に含まれていない可能性が高いことと，シンボルとしての発生頻度が少ないためシンボル出力確率が大きくならないことによる．

6. 結論

本研究では，事象系列の分類という目的に基づいて，確率過程を用いたモデル化によるトピック分類の手法を提案した．また，動詞の共起語を用いることにより，動詞だけでは表現できない事象の記述をモデルに反映させる手法を示した．従来の文書分類とは異なり，順序を考慮した分類を行えることを実験によって確かめた．

本研究では，完結した事象系列をトピック分類の対象とした．しかし，未完結系列について本手法を適用することにより，事件途中での予測が可能である．この推測から「今後の展開」に沿って捜査情報・手法の提示や対象の絞込みが行えるものとなろう．本研究では，教師有り学習としてモデルを予め構築し，いずれかのモデルに分類する方法をとったが，ニューストリームからの逐次学習など，モデルの構築とモデル推定を同時に行わせることにより，過去に例を見ない事件への適用も可能となる．

[文献]

[1] Allan, J., Carbonell,J., Doddington, G., Yamron,J. and Yang,Y.: “Topic Detection and Tracking Pilot Study:

Final Report ”, proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.
[2] Asahara, M. and Matsumoto, Y.: Extended Models and Tools for High Performance Part-of-Speech Tagger, COLING, 2000
[3] Regina Barzilay and Lillian Lee: “Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization”, In Proceedings of the NAACL/HLT, pp. 113-120, 2004.
[4] D. M. Blei and P. J. Moreno.:“Topic segmentation with an aspect hidden Markov model”, In Int. Conf. Research and Dev. Inf. Retrieval, pp. 343-348, New York, 2001.
[5] 北研二: “確率的言語モデル”, 東京大学出版会, 1999.
[6] Makkonen, J.: Investigations on Event Evolution in TDT, In Proceedings of HLTNAACL 2003 Student Workshop, May 2003, Edmonton, Canada, pp. 43-48.
[7] Manning, C.D. and Schutze, H.: Foundations of Statistical Natural Language Processing, MIT Press, 1999
[8] Mitchell, T.: Machine Learning, McGrawHill Companies, 1997
[9] Mulbregt, P.van; Carp, I.; Gillick, L.; Lowe, S.; and Yamron, J. 1998. Text Segmentation and Topic Tracking on Broadcast News Via a Hidden Markov Model Approach. In Proceedings of the ICSLP’98, volume 6. 2519-2522.
[10] 柴田知秀, 黒橋禎夫: “隠れマルコフモデルによるトピックの 遷移を捉えた談話構造解析”, 言語処理学会第 11 回年次大会, 2005.
[11] Yiming Yang, Tom Ault, Thomas Pierce, Charles W. Lattimer: “Improving Text Categorization Methods for Event Tracking”, In Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval, 2000.
[12] 若林啓, 三浦孝夫: HMM を用いた文書における事象系列の推定, 日本データベース学会論文誌（DBSJ Letters）,Vol.6,No.3,2006, pp.17-20, 平成 19 年(2007) 12 月
[13] 情報通信研究機構： EDR 電子化辞書 http://www2.nict.go.jp/r/r312/EDR/index.html

若林 啓 Kei WAKABAYASHI
法政大学大学院工学研究科修士課程在学中.

三浦 孝夫 Takao MIURA
京都大学理学部, 工学博士(東京大学). 現在, 法政大学工学部情報電気電子工学科教授. データモデル, 知識表現, 演繹データベース, 複合オブジェクトなどの分野の研究に従事. 電子情報通信学会, ACM 各会員. 著書に"データモデルとデータベース"(全 2 巻, サイエンス社)