HMMを用いた文書における状況系列の推定

若林　啓†　三浦　孝夫†

† 法政大学 工学部 情報電気電子工学科　〒184–8584 東京都小金井市梶野町 3-7-2
E-mail: †kei.wakabayashi.8f@eng.hosei.ac.jp, ††miurat@k.hosei.ac.jp

あらまし　本稿では, 文書で表現されたトピックを分類する手法を提案する. これまでに文書をモデル化する手法については多く論じられてきたが, 直接文書の内容を扱った研究は少ない. 一連のトピックは状況の系列によって表現できる. 本研究では, 一連の新聞記事からの状況の系列の推定を HMM によるタグ付け問題として扱い, 実験により手法の妥当性を示す.

キーワード　隠れマルコフモデル

On Event Sequences for Documents using Hidden Markov Model

Kei WAKABAYASHI† and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
E-mail: †kei.wakabayashi.8f@eng.hosei.ac.jp, ††miurat@k.hosei.ac.jp

Abstract　In this paper, we propose a technique to classify topics appeared in documents. There have been many investigation proposed so far, but few investigation which capture contents directly. Here we consider a topics as a *sequence* of events and a classification problem as a tagging problem based on Hidden Markov Model (HMM). We show some experimental results to see the validity of the method.

Key words　Hidden Markov Model

1. 前　書　き

近年, 計算機上で利用可能な文書データの増加に伴い, より高度な知識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. 文書分類技術は, 一般的に文書データを出現単語ベクトルにモデル化することで分類を行う. しかしこのベクトルモデルではその文書が述べているトピックを扱うことは難しい. 本研究では文書そのものではなく, その文書が表現しているトピックを対象にモデル化を考える. 本稿の目的は, 文書分類ではなくトピックの分類である.

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 [7].

本研究ではトピックを事象の系列と考えることで隠れマルコフモデル (Hidden Markov Model, HMM) を適用し, トピックを形式化することを考える. 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. というのは, 決定木や SVM といった従来の分類手法では, ベクトル分類に帰着させることが多く, 系列情報を反映させることは容易ではない.

本研究では事象系列を分類するための新しい手法として, 確率過程に基づいた文書の表現モデルを提案する. 確率過程は事象の系列をモデル化したものであり, 事象間の遷移を扱うことができる. このため, 文書を確率過程と考えることで, トピック分類を行うことができる.

文書をサブトピックの系列と考えることで確率過程としてモデル化する手法は, 今までにいくつか提案されている. HMM を用いてトピックセグメンテーションを行う研究には, Mulbregt らがある [5]. ここでは HMM の状態にトピックを対応させ, 文書の内部トピック系列を推定することで, 同じトピックが続く部分を一つのセグメントとしている.

Blei らは,Mulbregt らの手法に Aspect モデルを応用した Aspect HMM を用いてトピックセグメンテーションを行っている [3]. ここでは, 状態からの単語の出力確率分布に着目することで状態のラベル付けを行っている. 例えば,peace, israeli, palestinian といった単語が高い確率で出力される状態は, イスラエル・パレスチナ紛争のトピックを表現していると解釈できる.

これらの研究では, 複数のトピックをランダムな順序で含んだ文書を対象にしてトピック系列の特定を行う. 一方本研究では, 文書は全体が一貫した内容について論じており, 内容の順序

が意味を持っているものと仮定する．この仮定の下では, 系列そのものが特徴を持つ．この特徴を表現するため, 我々はトピックの種類ごとに HMM を用意する．ここでトピックの種類とは, 系列が似ているトピックの集合である．複数の HMM を用いて複数の事象系列の推定を行い, その中で最適な系列を選ぶことによって事象系列を特定する．

一貫した内容の文書を確率過程とみなしてモデル化する研究には, Barzilay らがある [2]．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,HMM によるモデル化を行う．この研究は, トピックの種類を一つに限定して内容系列を特定することを目的としている．

確率過程による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 [6]．ここでは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出力シンボルとみなした HMM を用いている．用言の格フレームとは, 動詞とその格によって分類される文章の大づかみの意味のことである．柴田らは動詞に着目することで, トピックの遷移をうまく捉えられることを実験により示している．このため本研究では, 文書の特徴量として動詞を用いる．

2 章では本研究が扱うトピックの分類について述べる．3 章では隠れマルコフモデルについて説明する．4 章で具体的な事象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の分類

ここでは, 本稿で提案する手法のアイデアを例を用いて述べる．本論文中では,「状況」(「事象」とも言う）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．「出来事」(「トピック」とも言う）は, 一連の事件やテーマに関する状況の系列を意味する．単に状況の系列（事象の系列）と言う場合, 特に一連の事件やテーマとは無関係な状況の系列を意味する．

２つのトピックが似ているかどうかは,「似ている」の解釈によって判断が分かれる．このためトピックの分類方法は一般に一意ではない．本研究ではトピックを事象の系列と考えるため, 事象の系列が似ているトピックを「似ている」ものとする．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, と事象の系列が似ているならば, 両トピックは似ているとする．

図 1 は, 本研究で考えるトピックの推定の例である．いま, 図中の左側に示されているような,「トピックを表現している文書」が与えられている．本稿では, 新聞記事の第一段落を日付順に連結し時系列に並べた文書を考える．この例では, ある殺人事件に関する新聞記事を連結させたものである．この文書は「殺人事件」というトピックを表現している文書となっている．

図中の右側に示しているのは, このトピックでたどっている「事象の系列」である．文書の解読から, ある人物の死亡の発見という事象から始まる．次に, 警察の調べによって不審な人物の手がかりが明らかになる事象が続く．最後に, 容疑者が逮捕されるという事象が得られる．これらがこのトピックの構成である．無論このトピックは,「殺人事件」という特性に依存する．ある人物の自殺に関するトピックの場合, 推定される事象として，最初に自殺の発覚, 次は自殺の原因が明らかになる等の系列になるであろう．

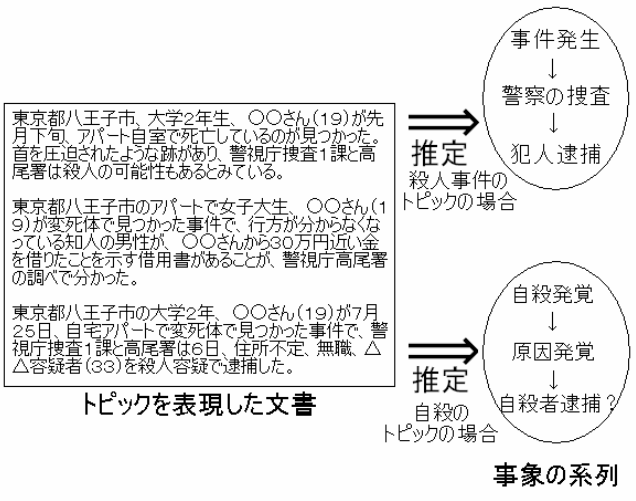

図 1　文書からのトピックの推定

殺人事件トピックの最後では逮捕という事象が見られるが, 自殺トピックとの関連からは, 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．しかし, ここではすでに自殺発覚の事象があるため, 事象の順序としては不自然である．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．

事象系列が, トピックの種類ごとに存在すると考えられることから，本研究では,（事象系列に基づく）トピックを推定することによって当該トピックの分類を行うことができることを論じる．

3. 隠れマルコフモデル

隠れマルコフモデルは, 確率的に遷移する内部状態をもつオートマトンである．内部状態は単純マルコフ過程に従って遷移する．通常, 内部状態は直接観測できない．その代わりそれぞれの内部状態は, 一つの観測可能なシンボルを確率的に出力する．

3.1 モデルの定義

隠れマルコフモデルは次の５つのパラメータによって定義される [4].

(1) $Q = \{q_1, \cdots, q_N\}$：状態の有限集合

(2) $\Sigma = \{o_1, \cdots, o_M\}$：出力シンボルの有限集合

(3) $A = \{a_{ij}\}$：状態遷移確率分布

a_{ij} は状態 q_i から状態 q_j への遷移確率である．

(4) $B = \{b_i(o_t)\}$：シンボル出力確率分布

$b_i(o_t)$ は状態 q_i でシンボル o_t を出力する確率である．

(5) $\pi = \{\pi_i\}$：初期状態確率分布

π_i は状態 q_i が初期状態である確率である．

本研究では, 状態は事件発生, 容疑者の逮捕, 自殺発覚などの事象の種類に対応する．状態集合 Q はトピックの種類によって

異なる集合をもつ．出力シンボルは観測可能な情報であり，文書に該当するが，次章で述べる特徴的な単語の抽出によって得る単語のみをシンボルとする．

状態遷移確率分布 A はある状態から次の状態へ遷移する確率であるため，ある事象が起きた後，次に起こる事象の確率分布である．シンボル出力確率分布 B は，ある事象が起きたとき，文書中にどのような単語が出現するかを表す確率分布である．初期状態確率分布 π は，最初に起きる事象の確率分布である．

3.2 状態列の推定

隠れマルコフモデルは，観測したシンボル列から，隠れた内部状態列を推定する目的で用いることが多い．モデルのパラメータに基づいて，与えられたシンボル列に対して最適な内部状態列を求める問題を，隠れマルコフモデルの復号化問題と呼ぶ．Viterbi アルゴリズムは，復号化問題を効率的に解くアルゴリズムである．

最適な状態列とは，最もシンボル列の生成確率が高くなるような状態列のことである．あるモデル上において状態列とシンボル列が決定すれば，モデルがその状態列とシンボル列を生成する確率（尤度）は一意に求まる．具体的には，シンボル列 $o_1 o_2 \cdots o_T$，状態列 $q_1 q_2 \cdots q_T$ が与えられたときの尤度は

$$\pi_{q_1} b_{q_1}(o_1) \times a_{q_1 q_2} b_{q_2}(o_2) \times \cdots \times a_{q_{T-1} q_T} b_{q_T}(o_T)$$

と求まる．Viterbi アルゴリズムは，ある時刻 t でそれぞれの状態に到達する状態列のうち，最も尤度の高い状態列のみを記憶していくことで最適な状態列を得る．Viterbi アルゴリズムは以下のように再帰的に尤度の最大値をとる計算を行う．

$$\delta_{t+1}(j) = \max_i (\delta_t(i) a_{ij}) b_j(o_{t+1})$$

この計算と同時に最大値を与える状態を記憶していけば，最終的に最適な状態列を得ることができる．

3.3 モデルの算出

隠れマルコフモデルは 5 つのパラメータから成るが，このうち状態集合 Q とシンボル集合 Σ は事前に与えるパラメータである．一方で状態遷移確率分布 A，シンボル出力確率分布 B，および初期状態確率分布 π は一般的に自明ではない．モデルの算出とは，これらの確率値を学習によって計算することである．モデルの算出を行うには，シンボル列に内部状態をなんらかの方法（通常人手）によって与えたサンプルデータが必要になる．しかし，そのようなデータを利用できない場合でも，シンボル列のみの学習データによってパラメータを学習する Baum-Welch アルゴリズムによる学習が可能である．

Baum-Welch アルゴリズムは EM アルゴリズムの一種である．Baum-Welch アルゴリズムは，モデルが学習データとして与えられたシンボル列を生成する尤度が大きくなるようにパラメータの更新を繰り返すことで学習を行う．各繰り返しにおいて，現在のパラメータによって各時刻における状態遷移の確率を求め，その期待値を最大化するようにパラメータを更新する．

$\overline{\pi}_i =$ 初期状態が状態 i の回数の期待値

$$\overline{a}_{ij} = \frac{\text{状態 } i \text{ から状態 } j \text{ へ遷移する回数の期待値}}{\text{状態 } i \text{ から遷移する回数の期待値}}$$

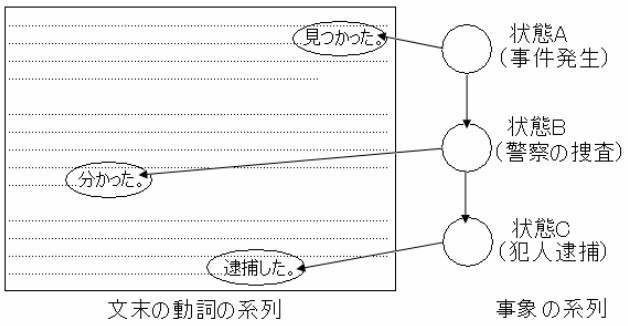

図 2　トピックの推定モデル

$$\overline{b}_i(k) = \frac{\text{状態 } i \text{ に滞在し記号 } k \text{ を出力する回数の期待値}}{\text{状態 } i \text{ に滞在する回数の期待値}}$$

この再推定式をパラメータが収束するまで繰り返し計算する．一般的に尤度は最大ではなく極大になるため，初期パラメータの分布に依存して収束するパラメータは異なる場合がある．

4. トピック推定

ここでは本稿で提案する，文書から事象系列を推定するアルゴリズムについて述べる．

4.1 HMM 手法の適用

我々は，トピックに隠れマルコフモデル（HMM）を適用してモデル化する．図 2 は，事象を内部状態，文章を出力シンボルに当てはめた HMM によるトピックの推定モデルである．前章で述べたように，HMM でモデル化することによって事象系列を Viterbi アルゴリズムに基づいて求めることができる．

しかし出力シンボルとして全ての単語を与えると，効果的なトピック推定が困難になる．例えば，東京都やアパートや事件といった単語は，トピックの状況の変化に対して意味を持っていない．そこで図中の左側に示すように，文書中において特に状況の変化を表現する部分だけをモデルに反映させる．日本語の文章の場合，特に状況の変化を表現する部分は各文章末の動詞である．このため図に例示されているように，文書から各文章末の動詞部分のみを抽出し HMM のシンボルとして与える．

またトピックは，トピックの種類によって異なる事象で構成される．例えば，殺人事件では事件発生や犯人逮捕といった事象があるが，自殺事件では自殺発覚や理由発覚などといった異なる事象がある．つまり，HMM のパラメータの一つである状態の有限集合 Q がトピックの種類によって異なる．このため，トピックの種類ごとに違う HMM を用意する必要がある．

ここで，本節以降で使用する用語を定義する．

- 文書

文書は，トピックを表現した文書という意味で用いる．本稿では新聞記事の第一段落を日付順に連結し時系列に並べた文書のみを扱うが，ここでは特にその意味に限定するものではない．

- カテゴリ

カテゴリは，トピックの種類という意味で用いる．例えば，殺人事件，自殺事件，汚職事件などがカテゴリとなる．本研究では，カテゴリに依存して異なる HMM を用意する．

4.2 シンボル列の抽出

ある文書 d が与えられたとき, それに対応するシンボル列 $o_1o_2\cdots o_n$ を与える関数を考える. すなわち,

$$Symbol(d) = o_1o_2\cdots o_n$$

となるような関数 $Symbol$ を定義する.

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, 単語列にする. 次に, 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シンボル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, その動詞をシンボル列に加える. この操作を文書 d の全ての文章に対して行う. これによって得られたシンボル列の末尾に, 終端を意味するシンボル「EOS」を加えたシンボル列を $Symbol(d)$ の値とする. このシンボル列のシンボルの順序は, 文書中の出現順序と一致していることを必ず保証する.

4.3 HMM モデルの学習

あるカテゴリ c に対応する HMM を M_c とする. M_c の学習用としてカテゴリ c の文書集合 $D_c = \{d_{c1}, d_{c2}, \cdots, d_{c|D_c|}\}$ が与えられたとき, M_c のパラメータを学習によって決定する方法を考える. M_c の状態集合 Q は任意の状態数 N_{M_c} 個の要素をもち, シンボル集合 Σ は全てのカテゴリの HMM で共通の集合とする.

まず,M_c の状態遷移確率分布 A, シンボル出力確率分布 B, 初期状態確率分布 π を乱数で初期化する. この M_c について,D_c のそれぞれの要素をシンボル列に変換して得られるシンボル列集合 L_c

$$L_c = \{Symbol(d_{c1}), \cdots, Symbol(d_{c|D_c|})\}$$

を学習データとして Baum-Welch アルゴリズムを実行し, パラメータを決定する.

なお, 最初に A,B,π を乱数で初期化するため, それぞれの状態がどのような事象の種類を意味しているかを事前に知ることができない. このため, この学習の後,HMM の確率分布を直接見ることで状態の解釈を後から加える.

4.4 トピックの推定

カテゴリの不明な文書 d が与えられたとき, d のトピックを推定する方法を考える.

まず,d によって与えられるシンボル列 $Symbol(d) = o_1o_2\cdots o_n$ に対して, 全てのカテゴリの HMM で状態列を推定する. いま, カテゴリ c の HMM,M_c が推定する状態列が $s_{c1}s_{c2}\cdots s_{cn}$ であるとする. このとき得た状態列とシンボル列の組を Mc が生成する確率 $P(o_1o_2\cdots o_n, s_{c1}s_{c2}\cdots s_{cn}|M_c)$ を最大にするような c が, 文書 d の所属するカテゴリである. すなわち,d の所属するカテゴリ c_d は

$$c_d = argmax_c P(o_1o_2\cdots o_n, s_{c1}s_{c2}\cdots s_{cn}|M_c)$$

であり, このカテゴリの HMM,M_{c_d} が推定した状態列が d のトピックとなる.

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

表 1　実験データ

5. 実　　験

ここでは, 提案アルゴリズムの評価実験について述べる. まず実験方法について述べ, 次に実験結果を示し, 最後に考察および提案アルゴリズムの評価を行う.

5.1 実 験 方 法

我々は 3 つのカテゴリに分類した 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した. その内訳を表 1 に示す. それぞれのカテゴリで, 用意した文書のうち $\frac{2}{3}$ を学習用に, $\frac{1}{3}$ をテスト用に割り振る. カテゴリは次の 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

前節のアルゴリズムに従い, 学習文書を用いて HMM 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.

また今回, 全てのカテゴリで HMM の状態数 N_{M_c} を 5 で行った実験の結果を示す. 状態には事前にどのような事象の種類であるかという解釈を与えないため, 状態数は任意の値で学習を行わざるを得ない. このため, 予備実験によりカテゴリの分類率が最も高かった状態数 5 を用いる.

なお, 単語の切り分けおよび品詞の同定には日本語形態素解析ツール Chasen を用いる.

5.2 実 験 結 果

ここでは実験結果を示す. まず, 学習によって得られた HMM の構造を示し, 状態の解釈の結果を示す. 次に, カテゴリの不明な文書のトピックの推定を行った結果の一例を示す. 最後に, テスト文書をカテゴリ分類した分類率を示す.

5.2.1 モデルの構造

それぞれのカテゴリについて,HMM のパラメータを学習した結果得られたモデルの構造を示す. ここで, モデルの構造（トポロジー）とは,HMM のパラメータが表現している状態間の関係を意味する. また, ここでは特に, 状態が出力するシンボルの分布も含める. モデルの構造を見ることは,HMM のパラメータの特徴的な部分に着目することである. このためモデルの構造は HMM の学習アルゴリズムによって得られたパラメータそのものであり, 事前に与えたものではない. この構造を見ることで, 状態を事象として解釈する.

図 3 は, 単独犯事件の HMM の構造である. 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, 円の近くにはその状態が出力するシンボルのうち, 出力確率の大きいものを列挙してある.

これらの確率分布から, それぞれの状態が意味している事象

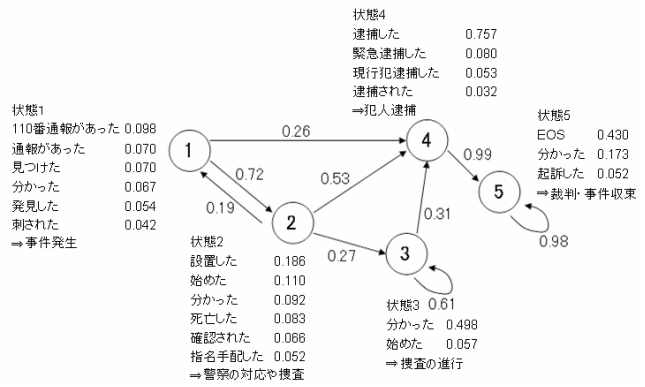

図 3　単独犯事件の構造

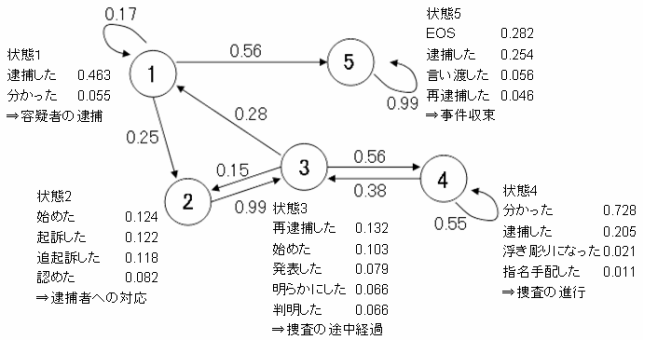

図 4　組織犯事件の構造

を解釈する．例えば, 状態 1 は 110 番通報があった, 見つけた, 刺された, などのシンボルが出力される確率が高いことから, 事件発生の事象を意味している．また, 状態 4 は逮捕した, 緊急逮捕した, 現行犯逮捕した, などのシンボルが出力されることから, 犯人の逮捕の事象を意味している．このように状態を解釈した結果を, それぞれの状態について図中の出力シンボルの下に示してある．解釈は主観的な判断に基づいて行うが, このカテゴリのモデルの構造は比較的解釈が容易である．

次に, 状態の遷移に着目して遷移確率の高い状態をたどっていく．事件発生, 警察の対応, 犯人逮捕, 事件収束と事象の変化として自然な状態遷移が起こりやすくなっている．また, 自分自身に遷移する確率の高い状態 3 を通るパスをもつ事件は, 捜査の長引く事件を表現している．状態遷移の点からも, このカテゴリのモデルの構造は解釈が容易になっている．

図 4 は, 組織犯事件の HMM の構造である．このモデルでは単独犯に比べ, 逮捕したのシンボルが複数の状態から出力される点で特徴的である．例えば, 捜査の進行と解釈した状態 4 からも高い確率で逮捕したのシンボルが出力されている．これは, 組織犯事件と分類したトピックには逮捕された犯人が複数いるような事件が多いため, 逮捕もまた捜査の進行の一部であると解釈する．

図 5 は, 汚職事件の HMM の構造である．このモデルは, 他のカテゴリに比べても複雑な構造である．例えば, 図中には捜査の進行と解釈した状態が二つある．これらの状態は出力シンボルの分布が似ており, 解釈を分けることができない．また, 状態遷移も複雑であり, 特徴的なパターンの発見が困難である．

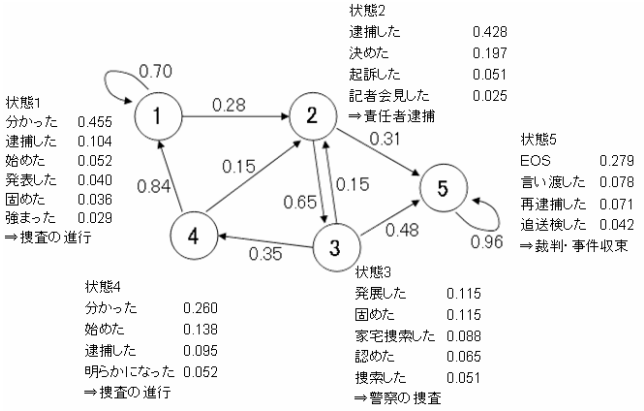

図 5　汚職事件の構造

２３日午前２時３５分ごろ,... 署員が女性の焼死体を 発見した 。... 殺人事件と断定, 三島署に捜査本部を 設置した 。... ... , ○○さん（当時１９歳）が焼殺された事件で,... 疑いを強め, 事情聴取を 始めた 。... , ○○さん（当時１９歳）が殺害された事件で,... , △△容疑者（３０）を逮捕・監禁, 強盗などの疑いで 逮捕した 。... ... , △△容疑者（３０）が３０日,... ○○さん殺害を認める供述を 始めた 。... ... , 三島署捜査本部は１３日,... △△容疑者（３０）を殺人容疑で 再逮捕した 。 △△容疑者は容疑を認めている。

表 2　トピックを推定する文書

シンボル列	単独犯モデル	組織犯モデル	汚職モデル
発見した	事件発生	容疑者の逮捕	責任者逮捕
設置した	警察の対応や経過	逮捕者への対応	捜査の進行
始めた	捜査の進行	捜査の途中経過	捜査の進行
逮捕した	犯人逮捕	捜査の進行	捜査の進行
始めた	事件収束	捜査の途中経過	捜査の進行
再逮捕した	事件収束	容疑者の逮捕	事件収束
EOS	事件収束	事件収束	事件収束
尤度	6.25×10^{-9}	2.79×10^{-10}	5.87×10^{-18}

表 3　推定された事象

5.2.2　トピックの推定

表 2 に示すカテゴリの不明な文書に対して, トピックの推定を行った結果を一例として示す．

この文書から, シンボル列を抽出するアルゴリズムによって得られたシンボル列を四角で囲んで示している．このシンボル列に対して推定された事象系列を表 3 に示す．なお, 表中で表記されている事象は, 図 3〜図 5 の図中に示した状態の解釈に対応する．

単独犯モデルが推定した事象は, 文書の解読から, 納得のいく結果になっている．特に, 逮捕したのシンボルが出現した後の事象が事件収束と推定されている．これは単独犯モデルが, トピックを犯人が一人の事件に限定しているためである．一方組織犯モデルでは, 逮捕したのシンボルが捜査の進行と推定されている．これは組織犯モデルが, トピックを犯人が複数いる事件に限定しているためである．

トピック	正解数	テスト文書数	正解率 (%)
単独犯事件	34	45	75.6
組織犯事件	9	17	52.9
汚職事件	10	22	45.5
合計	53	84	63.1

表 4　分 類 結 果

表の最後に示した尤度は, それぞれのモデルがこの状態列とシンボル列の組を生成する確率 $P(o_1 o_2 \cdots o_n, s_{c1} s_{c2} \cdots s_{cn} | M_c)$ である. この $argmax_c$ をこの文書のカテゴリと推定するため, この文書は単独犯事件のカテゴリに分類する.

5.3　カテゴリ分類

テスト文書のトピックを分類した結果を表 4 に示す. 表にはカテゴリ別に正解率を示している.

3 つのクラスへの分類であるため,63.1% の正解率は評価できる数値である. 特に単独犯事件の分類率が 75.6% と高い. 一方, 汚職事件の分類率は 45.5% と最も低い結果である.

5.4　考察・評価

実験結果の考察と, 提案アルゴリズムの評価を行う.

まず実験結果から言えることは, 単独犯事件に関する結果がよいことである. モデルの構造の解釈が容易であり, 分類精度が特に高い. 逆に, 汚職事件はモデルの構造の解釈が難しく, 分類率が最も低い. この原因として, 汚職事件のトピックの事象系列が一定のパターンに従うことが少ないことがある. HMM は Baum-Welch アルゴリズムによって, 学習シンボル列の尤度を大きくするようにパラメータを収束させる. これによって得られたモデルは, 当然, 学習データと同じパターンのシンボル列の尤度を大きくする. もし他の汚職事件のシンボル列の尤度が小さいならば, それは学習したシンボル列とは異なったパターンのシンボル列である. モデルの構造の解釈が難しい原因も同様に, 学習シンボル列集合に多くのパターンが混在しているためである. ただし,HMM で近似したためにパターンが発見できなかった可能性はある. しかしながらこのモデルで, 単独犯事件のパターンはよく表現できている. これより, 少なくとも,HMM でトピックを表現できるケースが存在すると言える.

また, どのカテゴリにおいても出現するシンボルが類似していることにも着目する. 今回用意した文書は, どれも事件に関するものであり, 逮捕した, 起訴した, 分かったなど類似したシンボルが出現する. このことから, 提案アルゴリズムは出現したシンボルの種類に依存してトピックを分類しているのではない. 提案アルゴリズムが考慮しているのはシンボルの出現順序であり, 従来の文書分類とは大きく異なっている.

6.　結　　論

本研究では, トピックが事象の系列であるというアイデアに基づいて, トピックを確率過程としてモデル化する手法を提案した. また, 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カテゴリ分類の対象とした. 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 HMM 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] 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.

[3] 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.

[4] 北 研二: “確率的言語モデル”, 東京大学出版会, 1999.

[5] 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.

[6] 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.

[7] 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.