HMM を用いた文書における事象系列の推定

Identifying Event Sequences using Hidden Markov Model

若林 啓♥ 三浦 孝夫♠

Kei WAKABAYASHI Takao MIURA

本稿では，文書を正しくトピックに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，事象系列の分類を扱った研究は少ない．本研究では，トピックを事象系列のクラスと考え，一連の新聞記事からの事象の系列の推定を 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．

In this paper, we propose a sophisticated technique for classification of documents into topics. There have been many investigations proposed so far, but few investigation which capture event sequences directly. Here we consider a topic as a class of *event sequences* and a classification as segmentation (or tagging) problem based on *Hidden Markov Model* (HMM). We show some experimental results to see the validity of 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，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し，コサイン尺度に基づいて分類を行う．しかしこの方法は単語の分布しか扱わず，文脈や発生順序を考慮しないため，その文書が述べているトピックを扱うことは難しい．本稿の目的は，文書を事象の系列とみなしてトピックに分類することである．

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]．Makkonen らは，事件を分岐する事象の系列と考え，日付のある文書集合から一連の事件を発見する手法を提案している[5]．

本研究ではトピック分類の手法を提案する．ここでは，トピックを事象系列のクラスと考え，隠れマルコフモデル(Hidden Markov Model，HMM)を適用して，事象の発生順序を基に系列を形式化する．事象系列を考慮した分類手法に関してはあまり積極的な提案はない．これは，決定木や SVM，自己組織化マップ(SOM)，単純ベイズ[3]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．

本研究では，事象系列を正しくトピックに分類するため，確率過程に基づいた文書の表現モデルを提案する．確率過程を用いて文書を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，HMM を用いて文書の構造を推定する．彼らは問題領域固有の知識を得るため，2-gram 出現分布を用いている．本研究ではどのトピックに適合するかを推定するため，単語分布に依存するこの手法を直接用いることはできない．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[6]．ここでは日本語による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．実験では，動詞に着目し，事象の遷移をうまく捉えられることを示している．本研究でも，日本語文書の特徴量として動詞を用いる．

2. 事象系列のトピック分類

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．また，「トピック」は，一連の事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とは，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」等のように，事象の系列が類似しており，ここでは両事件は「類似する」とする．

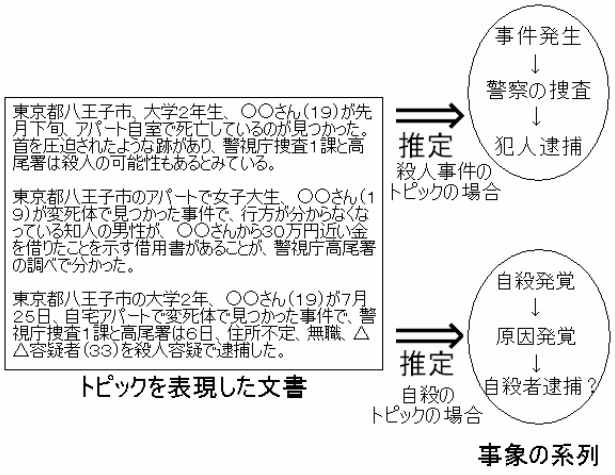

図１ 文書のトピック推定

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」というトピックを述べ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１右には，この文書でたどっている「事象の系列」を示す．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」というトピック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事件を表すならば，推定される事象は，自殺の発覚，自殺の原因の特定と続くであろう．殺人事件の最後では，「逮捕」の事象が通常生じる．自殺事件の場合でも，逮捕された

♥ 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

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，事象系列がトピックごとに存在すると考えられるため，本研究では，事象系列パターンを推定することによって，文書のトピック分類が可能であることを論じる．

3. 隠れマルコフモデル(HMM)

HMM 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が，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．また，モデルのパラメータが与えられているとき，観測されたシンボル列に対して最も確からしい内部状態列を計算するアルゴリズムが存在する[4].

HMM は５つのパラメータで定義される．

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

(2) $\Sigma = \{o_1, \ldots, o_M\}$：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 への遷移確率を表す．

(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 を出力する確率を示す．

(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を示す．

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毎に存在する．状態集合 Q はトピック毎に異なる集合をもつ．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，本稿では次章で述べる特徴的な単語の抽出によって得る単語を対応させる．状態遷移確率分布 A は次に生じる事象の確率分布である．シンボル出力確率分布 B は，一つの事象が起きたとき，そこで生じる単語の出現確率分布であり，その状態にのみ依存すると仮定する．

4. トピック推定

ここでは文書から事象系列を抽出し、文書のトピック推定を行うアルゴリズムについて述べる．

4.1 HMM 手法によるモデル化

本稿では，トピック推定を隠れマルコフモデル(HMM)に基づいてモデル化する．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推定モデルである．前章で述べたように，このモデルによって，観測したシンボル列に対して最適な事象の遷移系列を求めることができる．

入力文書に生じる全ての語を出力シンボルとして用いると，特徴的な事象の推定が困難になる．実際，「東京都」や「アパート」，「事件」といった単語は，単独では殺人事件の状況の記述とは言い難い．そこで本研究では，文書中において状況の変化を表現する部分だけをモデルに反映させる．日本語の文章では，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．このため例のように，形態素解析により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ものとする．本稿では，文書を一つの事件に関する記事の系列とし，更にそれぞれの記事の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文書を扱う．

また各トピックに対応して個々に HMM を与える．例えば，殺人事件では事件発生や犯人逮捕といった事象の集合，また自殺事件では自殺発覚や理由発覚などのような異なる事象の集合が対応する．このため，トピックごとに違う HMM を用意する必要がある．

HMM は各事象系列が表す内容を確率的に分類する「分類器」として働く。即ちトピック t に分類される確率 p_t を最大にするトピックを推定結果とする。

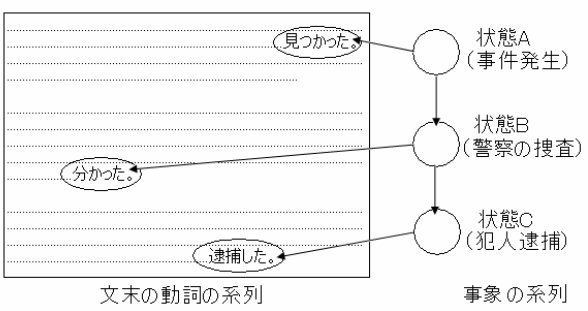

図 2 トピック推定のモデル

Fig.2 Estimating Topics

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 $o_1o_2\ldots o_n$ を与える関数を考える．すなわち，

$Symbol(d) = o_1o_2\ldots o_n$

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．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，$Symbol(d)$ の値とする．シンボルの順序は，文書中の出現順序と一致させるものとする．

4.3 HMM モデルの学習

トピック c に対応する HMM を M_c とする．M_c の学習用としてトピック c の文書集合 $D_c = \{d_{c1}, d_{c2}, \ldots, d_{c|Dc|}\}$ 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．M_c の状態集合 Q は任意の状態数 N_{Mc} 個の要素をもち，シンボル集合Σは全てのトピックの 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

$L_c = \{Symbol(d_{c1}), \ldots, Symbol(d_{c|Dc|})\}$

を学習データとして Baum-Welch アルゴリズムを実行する．Baum-Welch アルゴリズムは EM アルゴリズムの一種であり，学習データの尤度を大きくするようにモデルのパラメータを更新していく学習手法である．これによって収束したパラメータを M_c の学習結果とする．

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 の確率分布を直接分析し，状態の解釈を与える．

4.4 文書のトピック推定

文書 d が与えられたとき，尤度原理によって d のトピックを推定する．

まず，d によって与えられるシンボル列 $Symbol(d)=o_1o_2\ldots o_n$ に対して，全てのトピックの HMM で，最も確からしい状態列を動的プログラミング手法により推定する[4]. いま, トピック c の HMM, M_c が推定する状態列が $s_{c1}s_{c2}\ldots s_{cn}$ であるとする．このとき得た状態列とシンボル列の組を M_c が生成する確率 $P(o_1o_2\ldots o_n, s_{c1}s_{c2}\ldots s_{cn}/M_c)$ を最大にするような c が，文書 d のトピックであると推定する．

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，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)の 3 種類である．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．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推定を行う．

また，全てのトピックで HMM の状態数 N_{Mc} を 5 で行った実験の結果を示す．事前に事象の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，予備実験により文書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．

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

5.2 実験結果

ここでは，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，テスト文書のトピック推定の結果を例と共に示す．

5.2.1 モデル構造の解釈

それぞれのトピックについて, 学習アルゴリズムで得られた HMM の構造を示す．ここで，モデルの構造(トポロジ)とは，HMM の状態間の関係および各状態が出力するシンボルの分布を意味する．この構造を見ることで，状態を事象として解釈することができる．

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．

これらの確率分布を見ることで, それぞれの状態を事象として解釈できる. 例えば, 状態 1 は「110 番通報があった」，「見つけた」，「刺された」などの出力確率が高いことから，事件発生の事象と解釈できる. また, 状態 4 は「逮捕した」，「緊急逮捕した」，などのシンボルが出力されることから，犯人の逮捕の事象と解釈できる．このようにそれぞれの状態を解釈した結果を，図中の出力シンボルの下に示す．解釈は主観的であるが，このモデルの構造は比較的解釈が容易であると言ってよい．

状態の遷移に着目して構造を解釈することもできる. 遷移確率の高い状態は,「事件発生」,「警察の対応」,「犯人逮捕」,「事件収束」であり，事象の変化として自然である．また, 自己遷移する確率の高い状態 3 を通る事件は, 捜査の長引く事件を表している．

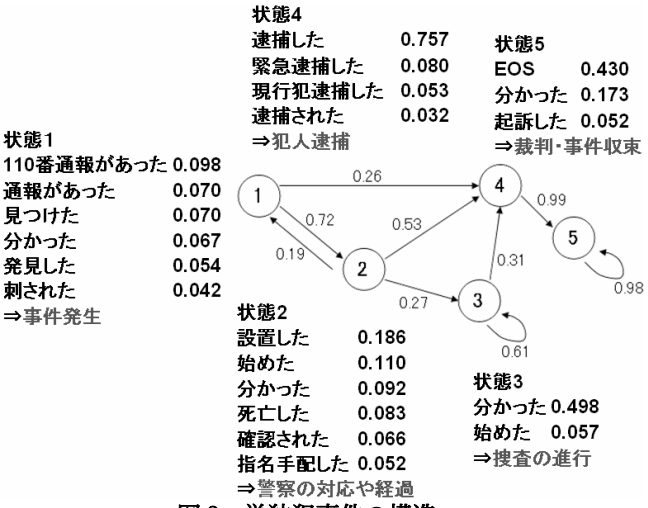

図 3 単独犯事件の構造

Fig.3 One-man Crime

図 4 は，組織犯事件の HMM 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できる．

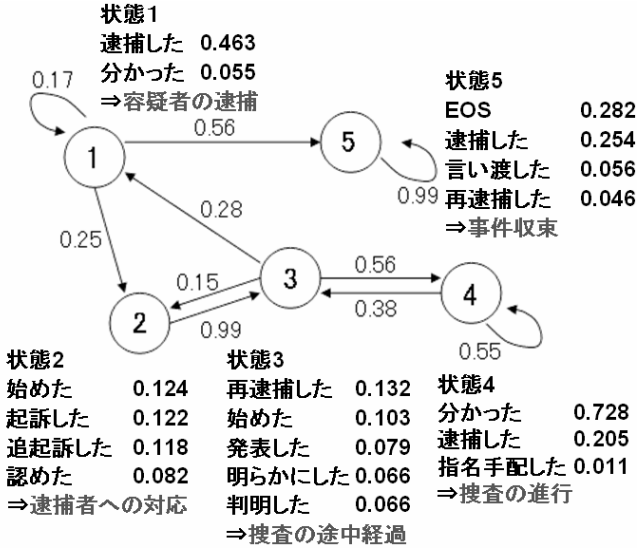

図 4 組織犯事件の構造

Fig.4 Organizational Crime

汚職事件の HMM の構造は，他のモデルに比べても複雑な構造である．例えば，ある二つ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．

5.2.2 トピック推定

表 2 に示すテスト文書に対して, トピック推定を行った結

果を示す.

表2 トピック推定を行う文書
Table 2 Document for Estimating Topics

２３日午前２時ごろ、... 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**。...
... 、○○さんが焼殺された事件で、... 事情聴取を**始めた**。
... 、○○さんが殺害された事件で、... 、△△容疑者を逮捕・監禁、強盗などの疑いで**逮捕した**。...
... 、△△容疑者が３０日、... 殺害を認める供述を**始めた**。...
... 、捜査本部は１３日、... △△容疑者を殺人容疑で**再逮捕した**。△△容疑者は容疑を認めている。

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す. 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, 前節で行った状態の解釈の結果を反映している.

表3 推定された事象
Table 3 Estimated Events

シンボル	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	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}

単独犯モデルが推定した事象は, 文書の解読から, 自然な結果となっている. 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, 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, 犯人が複数いる事件に限定しているためである.

各トピックの尤度は, 状態列とシンボル列の組を生成する確率 $P(o_1o_2\ldots o_n, s_{c1}s_{c2}\ldots s_{cn}|M_c)$である. 尤度原理によって, この文書は単独犯事件のトピックに分類する.

テスト文書のトピックを推定した結果を表 4 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 63.1% の正解率を得た. 単独犯事件の分類率が 75.6% と最も高く, 汚職事件の分類率は 45.5% と最も低い.

表4 分類結果
Table 4 Classification Ratio

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

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, 本方式の正解率が示すように, 単独犯事件に関する結果がよい. これは, モデルの構造の解釈が容易であり, 分類精度が特に高いことによる. 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. これは, 汚職事件には様々なパターンが存在するためであろう. 実際, 汚職事件トピックの文書を詳しく見ると, テストデータには学習データに含まれていないパターンをもつシンボル列が含まれている. このことが分類率を低くしている.

また, どのトピックにおいても共通して出現するシンボルが多い. 実際, 3つのトピック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. このことは, 提案アルゴリズムがシンボルの種類に依存しておらず, 従来の文書分類とは本質的に異なることを意味する.

6. 結論

本研究では, トピックは事象系列のクラスであるというアイデアに基づいて, 確率過程によりトピック推定を行う手法を提案し, 実験によりこの有用性を確かめた.

本研究では, 完結した事象系列をトピック推定の対象とした. 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.

[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] Barzilay, R. and Lee, L: “ Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization ”, In Proceedings of the NAACL/HLT, pp. 113-120, 2004.

[3] Manning, C.D. and Schutze, H.: Foundations of Statistical Natural Language Processing, MIT Press, 1999

[4] 北研二: “ 確率的言語モデル ”, 東京大学出版会, 1999.

[5] Makkonen, J.: Investigations on Event Evolution in TDT, In Proceedings of HLTNAACL 2003 Student Workshop, May 2003, Edmonton, Canada, pp. 43–48.

[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] Yang, Y., Ault, T., Pierce, T., Lattimer, C. W.: “ Improving Text Categorization Methods for Event Tracking ”, In Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval, 2000.

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
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.