キーストロークデータに基づくライティングプロセス推定のための執筆者依存型隠れマルコフモデル

Author-dependent hidden Markov model to analyze writing processes using keystroke log data

宇都雅輝 *1，宮澤芳光 *2，加藤嘉浩 *3，中島功滋 *3，桑田一 *3
Masaki Uto*1, Yoshimitsu Miyazawa*2，Yoshihiro Kato*3，Koji Nakajima*3，Hajime Kuwata*3
*1 電気通信大学，*2 東京学芸大学，*3 ベネッセ教育総合研究所
*1University of Electro-Communications, *2Tokyo Gakugei University,
*3Benesse Education Research and Development Institute
Email: uto@ai.lab.uec.ac.jp

あらまし： 近年，ライティングスキルの育成手法の一つとしてライティングプロセスに基づく指導が注目されている．ライティングプロセスは「計画」,「文章化」,「推敲」などの作業の系列とみなすことができ，プロセスに基づく指導は一般にこれらの出現パタンに基づいて行われる．しかし，執筆者数が増加すると，指導者が全ての執筆者の執筆パタンを把握することは困難になる．そこで本研究では，キーストローク・ログデータを用いて，各執筆者の文章執筆パタンを推定できる機械学習手法を開発する．提案モデルは，隠れマルコフモデルにおける潜在状態を作業内容とみなし，作業内容の遷移確率に執筆者依存性と時間依存性を導入したモデルとして定式化する．また，本発表では，提案モデルの実データへの適用例と有効性評価について報告する．

キーワード： ライティングプロセス，キーストローク・ログ，隠れマルコフモデル，MCMC

1 はじめに

近年，高等教育においてライティングスキル育成の重要性が広く指摘されている．ライティングの指導法としては，完成した文章に指導者がフィードバックを与える形式が一般的であったが，近年では，ライティングのプロセスに着目した指導も注目されるようになってきた [1]．ライティングプロセスは「計画」,「文章化」,「推敲」などのサブタスク（以降では「作業内容」と呼ぶ）の系列とみなすことができ [2, 3]，ライティングプロセスに基づく指導は，一般に各執筆者の作業内容の出現パタンに基づいて行われる．作業内容の出現パタンの分析には，執筆者に思考内容を発話させる口頭内省法やビデオを用いた再生刺激法が古くから用いられてきた．しかし，これらの方法は時間的・経済的なコストが大きいため，執筆者数が増加すると実用できない問題がある．この問題を解決するアプローチの一つとして，コンピュータ上で執筆者にライティングを行わせ，その際のキーストローク・ログデータを用いてライティングプロセスを分析する手法が注目されている [1]．しかし，先行研究ではライティングプロセスの類型化を行うに留まっており，作業内容の出現パタンを執筆者ごとに推定できる手法は見当たらない．

そこで，本研究では，キーストローク・ログデータから執筆者ごとの作業内容の出現パタンを推定する手法を提案する．提案手法は，隠れマルコフモデル（HMM）における潜在状態を作業内容とみなし，作業内容の遷移確率が執筆開始からの経過時間と執筆者に依存すると仮定したモデルとして定式化する．提案モデルの利点は次のとおりである．1）状態出現の執筆者依存性と時間依存性の導入により，一般の HMM に比べてデータ適合が向上する．2）各執筆者の作業内容の出現パタンを確率分布として推定でき，各執筆者のライティングプロセスを定量的に把握できる．3）作業内容の出現分布の差異を執筆者間で比較することで，一般的なクラスタリング手法を用いたライティングプロセスの類型化も可能となる．

2 キーストローク・ログデータと特徴量抽出

本研究では，キーストローク・ログデータを収集するためにキーロガーシステムを開発した．開発したシステムは，執筆者に提示したテキストエリア上でキー入力操作が行われるたびに，文字列とカーソル位置の情報をタイムスタンプとともに記録する．本研究では，このシステムにより得られたキーストローク・ログデータに機械学習手法を適用するために，各執筆者のログデータを特徴量の時系列データとして定義する．具体的には，時系列データから移動窓方式（微小な時間幅を持つ分析フレーム単位で特徴量を抽出する方法）により，時点ごとの特徴量を抽出する．特徴量には，関連研究で共通して利用されている次の 7 種類の特徴量（以降ではライティング特徴量と呼ぶ）を用いる．1）平均文字数，2）バースト（1 秒以内での連続入力）回数，3）5 秒以上停止回数，4）文字追加操作の回数，5）文字削除操作の回数，6）相対カーソル位置の平均値，7）カーソル移動の回数．

以上による，各執筆者のデータは 7 次元のライティング特徴量の系列データとして定義される．具体的には，執筆者 $i \in \mathcal{I} = \{1, \cdots, I\}$ の $j \in \mathcal{J} = \{1, \cdots, J\}$ 番目の分析フレームに対する $f \in \mathcal{F} = \{1, \cdots, F = 7\}$ 次元目の特徴量を変数 $X_{ijf} \in \mathbb{R}$ とし，$\boldsymbol{X}_{ij} = \{X_{ij1}, \cdots, X_{ijF}\}$ で表すと，執筆者 i のライティング特徴量の系列データは $\boldsymbol{X}_i = \{\boldsymbol{X}_{i1}, \cdots, \boldsymbol{X}_{iJ}\}$ で定義できる．本研究では，データ $\boldsymbol{X}_i$ の集合 $\boldsymbol{X} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_I\}$ から，執筆者ごとの作業内容の出現パタンを推定する機械学習手法を開発する．具体的には，ガウシアン隠れマルコフモデルにおける状態遷移確率に執筆者依存性と時間依存性を導入したモデルを提案する．

3 執筆者依存型時間依存隠れマルコフモデル

提案モデルでは，執筆者 i の j 番目の分析フレームにおける特徴量 $\boldsymbol{X}_{ij}$ に対して「状態」と呼ばれる潜在変数 $S_{ij} \in \mathcal{S} = \{1, \cdots, S\}$ を想定する．状態 S_{ij} は直前の状態 $S_{i,j-1}$ に依存する遷移確率に従って得られると仮定する．また，提案モデルでは，各データ $\boldsymbol{X}_i$ を一定の時間幅を持つ T 個の時間区分 $\mathcal{T} = \{1, \cdots, T\}$ に分割したとき，遷移確率が時間区分と執筆者にも依存すると仮定する．具体的には，状態 s から状態 s' に遷移する確率を $A_{ss'}$ で表し，時間区分 t において執筆者 i が状態 s を取る確率を ϕ_{its}，データ $\boldsymbol{X}_{ij}$ が属する時間区分を $t_{ij} \in \mathcal{T}$ とすると，潜在状態 S_{ij} は一つ前の分析フレームの状態 $S_{i,j-1} = s$ に従う遷移確率と，執筆者と時間区分に依存する状態出力確率の積として，$P(S_{ij}|S_{i,j-1} = s, \boldsymbol{A}, \boldsymbol{\phi}_i) = A_{s,S_{ij}} \cdot \phi_{i,t_{ij},S_{ij}}$ で得られると仮定する．また，初期状態 S_{i1} の確率は，S 次元の多項分布で表される初期確率 $\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_S\}$ と状態出力確率 $\phi_{i,1,S_{i1}}$ の積として，$P(S_{i1}|\boldsymbol{\pi}, \boldsymbol{\phi}_i) = \pi_{S_{i1}} \cdot \phi_{i,1,S_{i1}}$ で生起すると考える．他方，各時点のデータは状態 S_{ij} に依存した正規分布に従うと仮定する．具体的には，$f \in \mathcal{F}$ 次元目の特徴量 X_{ijf} は，状態 s を所与として，平均値 μ_{sf}，分散 σ^2_{sf} を持つ正規分布

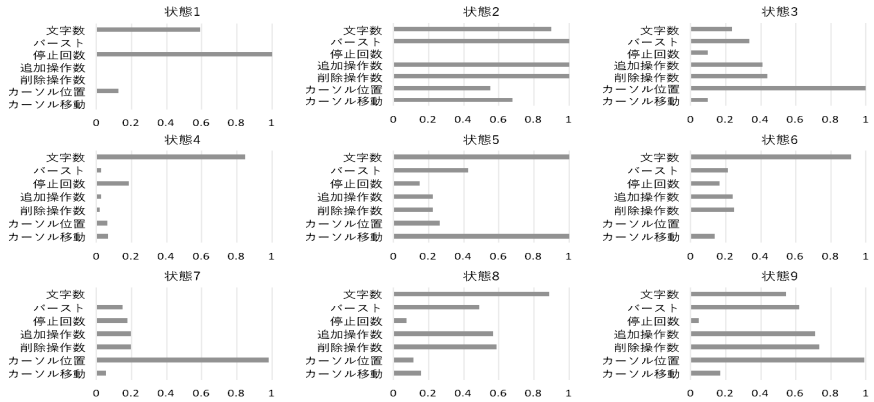

図 1　状態ごとの特徴量出力分布における平均値パラメータの正規化値

表 1　各状態の解釈

大区分	小区分	対応する状態
待機	-	状態 1
文章化	速筆型	状態 2，9
	遅筆型	状態 3，7
推敲	修正量多	状態 8
	修正量少	状態 4
	全体修正	状態 5
	個別箇所修正	状態 6

表 2　クラスタごとの評価得点

	レトリカル評価	メカニカル評価
クラスタ 1	2.467 (.146)	3.065 (.431)
クラスタ 2	2.673 (.114)	3.615 (.198)
p 値*	.027	<.001

*ウィルコクソンの順位和検定

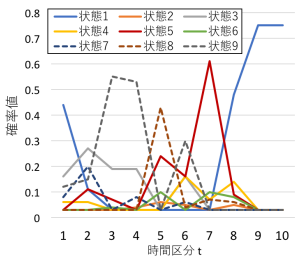

図 2　執筆者 1 の状態出現分布

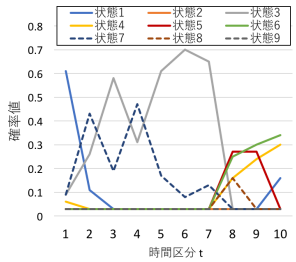

図 3　執筆者 2 の状態出現分布

$p(X_{ijf}|\mu_{sf}, \sigma^2_{sf}) = N(\mu_{sf}, \sigma^2_{sf})$ に従うと仮定する．

4　評価実験

本章では，実際のキーストローク・ログデータを用いた提案モデルの適用例と性能評価について述べる．本実験では，あるテーマについて論述させる課題を 72 名の被験者に行わせた際のデータを用いた．回答時間は 45 分とし，特徴量はフレーム幅 30 秒，ステップ幅 10 秒の移動窓方式で抽出した．

このデータを用いて，まずは，提案モデルの最適な状態数 S と時間区分数 T を求めた．対数周辺尤度が最大となる値を求めたところ，$S = 9$，$T = 10$ が最適値として選択された．このとき，提案モデルのパラメータ推定手法には周辺化ギブスサンプリングを用いた．次に，$S = 9$，$T = 10$ の提案モデルに基づいて状態の解釈を行った．ここで，各状態からのデータ出力分布（正規分布）の平均値を正規化した値を図 1 に示す．図 1 から，各状態の特徴を表 1 のように解釈した．

提案モデルでは，状態の解釈と状態出現分布 ϕ_i に基づいて各執筆者の作業内容の出現パタンを分析できる．例として，図 2，図 3 に代表的な執筆者の状態出現分布 ϕ_i を示した．図の横軸は時間区分を表し，縦軸が各状態の生起確率を表す．図より各執筆者のライティングプロセスについて次のように解釈できる．執筆者 1（図 2）は，初期の時間区分では待機状態の割合が大きく，時間進行とともに速筆型文章化，推敲作業の順に出現割合が高くなり，後半の時間区分では待機状態に戻っている．このことから，この執筆者は，計画・文章化・推敲作業を順に行なった上で，余裕を持って書き上げている理想的な書き方の執筆者であると解釈できる．これに対し，執筆者 2（図 3）は，遅筆型の文章化作業の出現比率が時間区分全体を通して高く，推敲状態の出現頻度が著しく少ないことが読み取れる．提案モデルでは，このように執筆者ごとの作業内容の出現パタンを時間区分ごとに定量的に把握できる．

次に，提案モデルで推定された作業パタンの妥当性の評価について述べる．ライティングのプロセスとスキルには関係があることが知られている [4]．そこで，提案モデルに基づきライティングプロセスのクラスタリングを行い，得られたクラスタとライティングスキルの関係を分析し，先行研究と類似した知見が導けるかを評価する．具体的には，状態出現分布の Jensen-Shannon(JS) 擬距離を執筆者間で求めて階層的クラスタリングを行い，各クラスタに属する執筆者の成果物の質を評価した．このとき，クラスタ数は擬似 F 値により決定された 2 クラスタとした．各クラスタに属する執筆者の状態出現分布を比較したところ，各クラスタの特徴は次のように解釈できた．クラスタ 1 は，一定の待機ののち，速筆型の文章化段階が続き，後半では全体的な推敲，軽微な修正を一定割合で含みながら待機状態に移行している．計画，文章化，推敲がバランスよく生起しており，余裕を持って執筆が完了する理想的なプロセスと解釈できる．クラスタ 2 は，相対的に書き出しが遅く，執筆開始後には遅筆型の文章化作業が長く続き，終了時間直前で軽微な修正や特定箇所の修正などが生起する傾向がある．文章化が遅く，推敲に十分に時間を割くことができないクラスタと解釈できる．

一般に，執筆が早く，推敲に時間をかける執筆者はスキルが高いことが知られている [4]．このことを確認するために，クラスタごとに成果物の質を評価した．具体的には，各執筆者の成果物を「構成の読みやすさ（レトリカル）」と「表層的な読みやすさ（メカニカル）」の 2 つの観点で 2 名の専門家に採点させた．評価カテゴリには，レトリカル評価では 5 段階，メカニカル評価では 4 段階で採点を行わせ，2 名の専門家の平均点を最終的な得点とした．結果を表 2 に示す．実験結果より，クラスタ 1 の執筆者群の得点が両方の観点においてクラスタ 2 より優位に得点が高いことが確認でき，先行研究の知見と一致したことがわかる．この結果から，提案モデルを用いて推定した執筆者の状態出現分布が実態を正しく反映しており，その分布に基づくライティングプロセスの分析は適切な知見を導ける妥当なものといえることが示された．

5　まとめ

本研究では，キーストローク・ログデータから各執筆者の作業内容の出現パタンを推定する執筆者依存型時間依存隠れマルコフモデルを提案し，実データ適用を通して提案モデルを用いた分析例とその妥当性を示した．紙面の都合で割愛した提案モデルと推定法の詳細などは宇都ら [5] を参照されたい．

参考文献

[1] Mo Zhang, Jiangang Hao, Chen Li, and Paul Deane. *Classification of Writing Patterns Using Keystroke Logs*, pp. 299–314. Springer International Publishing, Cham, 2016.
[2] S. Flower and R. Hayes. A cognitive process theory of writing. *College Composition and Communication*, Vol. 32, pp. 365–387, 1981.
[3] Nihat Bayat. The effect of the process writing approach on writing success and anxiety. *Educational Sciences: Theory & Practice*, Vol. 14, No. 3, pp. 1133–1141, 2014.
[4] Miyuki Sasaki. *Building An Empirically-Based Model of Efl Learners' Writing Processes*, pp. 49–80. Springer Netherlands, Dordrecht, 2002.
[5] 宇都雅輝, 宮澤芳光, 加藤嘉浩, 中島功滋, 桑田一．キーストロークデータからライティングプロセスを推定する執筆者依存型隠れマルコフモデルの提案．先進的学習科学と工学研究会（人工知能学会），2018.