LM は日本の時系列構造をどうエンコードするか

佐々木睦史 [1]　鴨田豪 [1]　高橋良允 [1]　Benjamin Heinzerling[2,1]　坂口慶祐 [1,2]
[1] 東北大学　[2] 理化学研究所
{sasaki.mutsumi.q5, go.kamoda, takahashi.ryosuke}@dc.tohoku.ac.jp
benjamin.heinzerling@riken.jp　keisuke.sakaguchi@tohoku.ac.jp

概要

LM の西洋の人物の内部表象に時系列的な方向が存在することが観察されている．では，和暦という独自の暦法体系を持つ日本の人物の内部表象はどのような構造を持つだろうか．本研究では江戸から平成までの LM 内部での時代表現を別々に取り出し，時代間の方向と位置を比較することで日本の時系列構造を調べた．実験の結果，LM 内部では，日本の時代間の方向はバラバラに表現されるが，時代間の位置は江戸から平成まで単調な順番に配置されることが示された．また，本研究で提案する時系列構造を同一平面で可視化する方法を用いて，日本の時系列構造を簡単な図に示すことができた．

1　はじめに

世界知識に関する質問に高精度で答えることができる言語モデルについて，知識がどのように構造化されているかの研究が行われている [1–3]．例えば「アインシュタインはいつ生まれましたか？」と問い合わせると，言語モデルの中間表現に時系列情報を格納した方向が存在することが観察される [4,5]．

本研究では，時系列情報のうち，日本特有の和暦や時代の区分が明確である時系列が LM 内部でどのように表現されるかを調査する．LM 内部における各時代の時系列情報には，図 1 に示されるように，1) 時代間で位置と方向が揃う場合や，2) 全く揃わない場合など，さまざまな可能性が考えられる．

本研究では，江戸，明治，大正，昭和，平成の各時代に生まれた人物の生年のデータセット § 2.1 を用いて，各々の時代に属する人物により構成される LM の時代表現を別々に取得する．そして，時代間の表現を比較する実験を行う．具体的には江戸，明治，大正，昭和，平成の時代の表現を同一平面に可視化する実験，時代間の方向を比較する実験，時代間の位置を比較する実験を行った．その結果，可視化実験では，LM 内部において各時代の表現は江戸から平成の順につながるように配置されるが，時代間の方向には相関がないという時系列の全体像が図示された．時代間の方向，位置を比較する実験では，時代間の方向はバラバラであること，時代間の位置は江戸から平成の順にそろうことが示された．

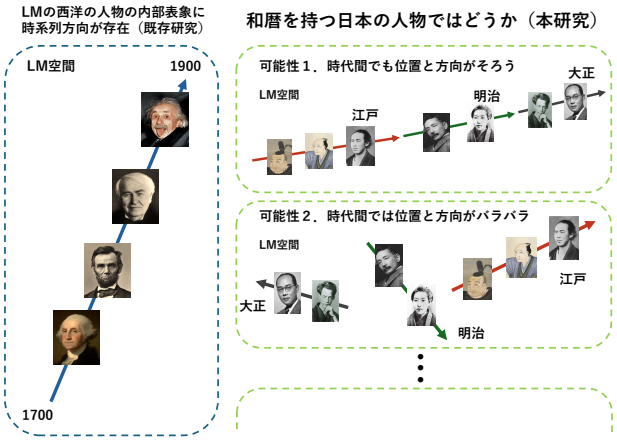

図 1　LM の内部表象において，西洋の人物の生年に関する時系列的な方向性が存在することが先行研究により示されている．一方，和暦という独自の暦法体系を有する日本の人物について，その生年の時系列的な表現が LM の内部でどのように構築されているかは明らかになっていない．

本研究で行った手法は日本以外の特有の暦法体系を持つ国の時系列を調べる際も適用できるため，多種多様な国の時系列を調べることで LM の時系列構造の深い理解に貢献する．

2　時代方向の検出

2.1　データセット

まず，LM による各時代の表現を別々に取得するために，日本人の名前と生年からなるデータセットを時代ごとに作成する．具体的には，Wikidata を用いて江戸 (1603-1868)，明治 (1868-1911)，大正 (1911-1926)，昭和 (1926-1988)，平成 (1988–2019) の

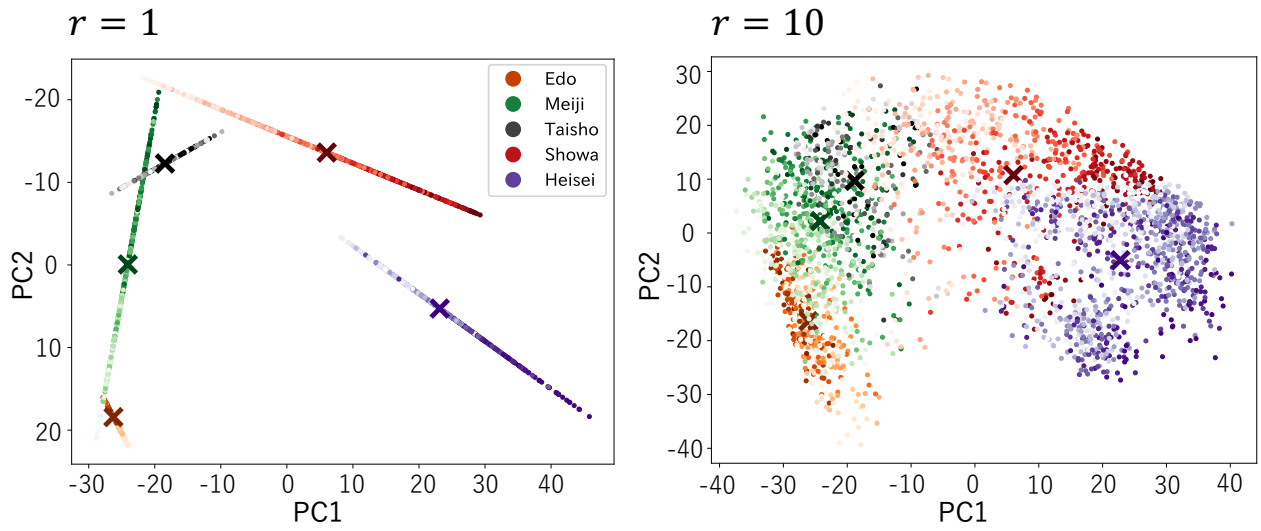

図 2　Swallow-13b における日本の時系列構造の PCA による可視化結果．左は 1 個の潜在変数 ($r = 1$) で PLS の逆変換を使った場合の可視化結果であり，右は 10 個の潜在変数 ($r = 10$) で PLS の逆変換を使った場合の可視化結果である．

各時代に生まれた日本人の名前とその生年 (西暦) のデータを収集した．ここで，人名を介して各時代の表現を取得する本実験では，LM が知識を形成できる程度に有名な人物をデータに用いる必要がある．そのため，本実験で用いるモデルである Swallow-13b [6, 7] が「何年に(人名)は生まれましたかという質問の答えは、」という入力に対して正しい生年（西暦）を回答できるデータに絞る操作を時代ごとに行った．その後のデータの件数は江戸，明治，大正，昭和，平成でそれぞれ 611 件，1297 件，555 件，1426 件，4744 件である．

2.2　PLS による時代方向の検出

PLS(部分的最小二乗法) [8] は，もとの説明変数をより目的変数を考慮しながら低次元な潜在変数に圧縮するアルゴリズムである．PLS で N 個の d 次元ベクトルそれぞれから r 個の潜在変数を抽出する際，説明変数 $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ と目的変数 $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$ に対して，次の圧縮による誤差 $\boldsymbol{E} \in \mathbb{R}^{N\times d}$ と $\boldsymbol{f} \in \mathbb{R}^{N\times 1}$ を最小化する $\boldsymbol{T} \in \mathbb{R}^{N\times r}, \boldsymbol{P} \in \mathbb{R}^{d\times r}, \boldsymbol{c} \in \mathbb{R}^{r\times 1}$ を学習する．

$$\frac{\boldsymbol{X} - \bar{\boldsymbol{x}}}{\boldsymbol{s_x}} = \boldsymbol{T}\boldsymbol{P}^\top + \boldsymbol{E} \tag{1}$$

$$\frac{\boldsymbol{y} - \bar{y}}{s_y} = \boldsymbol{T}\boldsymbol{c} + \boldsymbol{f} \tag{2}$$

ここで，$\bar{\boldsymbol{x}} \in \mathbb{R}^{1\times d}$ は $\boldsymbol{X}$ の平均ベクトル，$\bar{y}$ は $\boldsymbol{y}$ の平均，$\boldsymbol{s}_x \in \mathbb{R}^{1\times d}$ は $\boldsymbol{X}$ の次元ごとの分散，s_y は $\boldsymbol{y}$ の分散である．

PLS 学習後，圧縮した説明変数である T から d 次元空間でのベクトル群 $\hat{\boldsymbol{X}}$ を再構成する操作は以下の式に従う．これを PLS の逆変換と呼ぶことにする．

$$\hat{\boldsymbol{X}} = \boldsymbol{T}\boldsymbol{P}^\top \odot (\mathbb{1}^\top \boldsymbol{s_x}) + \bar{\boldsymbol{x}} \tag{3}$$

ここで，$\odot$ は要素積を表す．

学習後の PLS モデルは，d 次元ベクトル $\boldsymbol{x}$ に対して，以下の計算で回帰に用いることができる．

$$y = (\boldsymbol{x} - \bar{\boldsymbol{x}})\boldsymbol{w}^\top + \bar{y} \tag{4}$$

なお，$\boldsymbol{w} \in \mathbb{R}^{1\times d}$ は以下の式で与えられる．

$$\boldsymbol{w} = \frac{s_y \left(P^{\top\dagger} c\right)^\top}{\boldsymbol{s}_x} \tag{5}$$

$\dagger$ は擬似逆行列を表し，割り算は要素ごとに行う．この w が d 次元空間における X の y に相関する方向を表している．

後述の通り，本実験では X は各時代を表現する LM の隠れ状態，y は生年であるため，PLS により w を取得することは LM の隠れ状態の時代方向を検出することである．

3　時系列の全体像

モデルの隠れ状態が時代や和暦の情報をエンコードしているか，またもしエンコードしている場合，どのような構造になっているのかを調査する．

まず，§ 2.1 に基づき作成した時代ごとのデータセットを用いて，「何年に(人名)は生まれましたか？」を LM に入力し，隠れ状態を集める．

ここで，隠れ状態は本実験で用いる Swalllow-13b, LLM-jp-3-13b [9] のような 40 層のモデルでは第 24

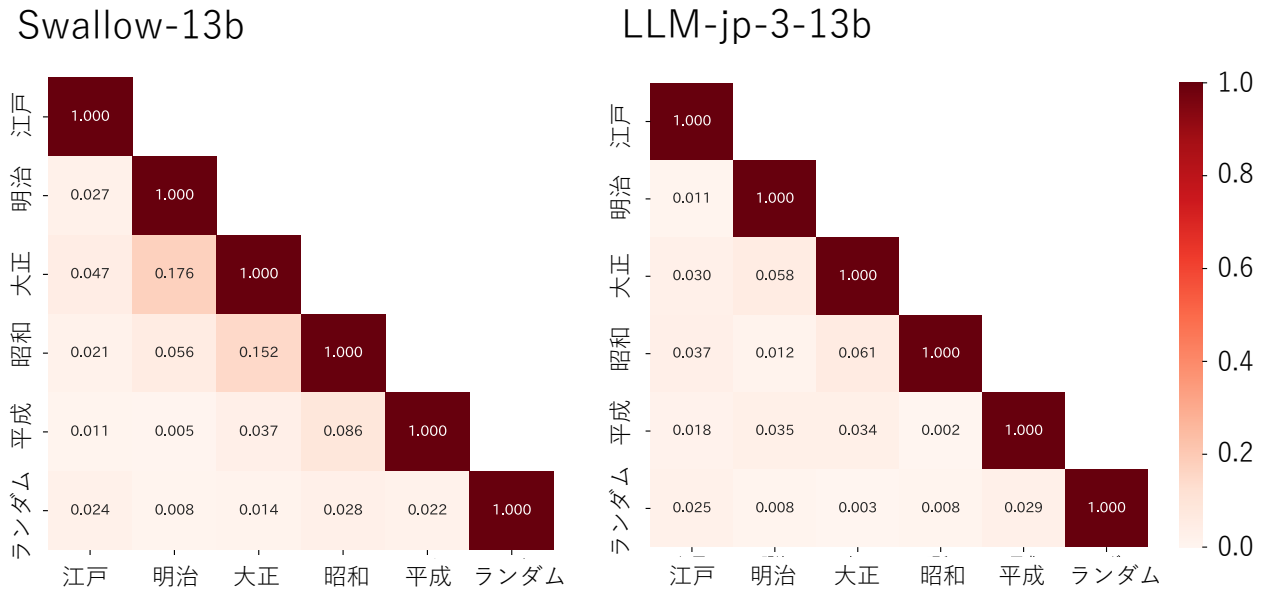

図 3　方向を表すベクトル w の時代間におけるコサイン類似度の絶対値の比較．左は Swallow-13b，右は LLM-jp-3-13b を用いた場合の結果であり，いずれのモデルでも時代間の方向に相関がないことが分かる．

層，トークン位置は人名に続く助詞の「が」の位置で収集した．隠れ状態の収集位置に関する詳細は付録 A を参照されたい．

収集された隠れ状態 $\boldsymbol{X}$ は時系列以外の情報を含むため，時系列情報の抽出が望ましい．そこでまず，収集された隠れ状態 X と，データセットから得られる生年の正解ラベル y を用いて § 2.2 の PLS を行い，低 (r) 次元の潜在変数 T を得る．$\boldsymbol{T}$ は $\boldsymbol{X}$ から生年を回帰する上でもっとも重要な情報を表現したものであるが，次元圧縮後の r 次元空間は時代ごとに異なる空間になってしまうため，時代間の比較ができない．そこで，式 3 に定めた逆変換を行うことで $\hat{\boldsymbol{X}}$ を得る．$\hat{\boldsymbol{X}}$ は生年の回帰に最も重要な情報のみを保持しつつ，元の LM (d 次元) の空間に戻したものであるため，異なる時代のデータで学習した PLS の $\hat{\boldsymbol{X}}$ でも同じ空間で扱うことができる．

それによって形成された $\hat{X}_{\text{江戸}}$ から $\hat{X}_{\text{平成}}$ を PCA により同一平面に可視化した．1 個，および 10 個の潜在変数を用いて PLS を学習し，逆変換を行った場合の可視化結果を図 2 に示す．ここで，$\bar{\boldsymbol{x}}_{\text{江戸}}$ から $\bar{\boldsymbol{x}}_{\text{平成}}$ が PCA 空間上でどこに対応するかを印で示した．また，図 2 では，時代ごとに人物の生年の値によって色の濃さを変えている．

可視化結果から，PLS 空間では 1) 時代間の方向にあまり相関が見られず，時代間で方向はそろっていないこと、2) $\bar{\boldsymbol{x}}$ により表現される各時代の位置が江戸，明治，大正，昭和，平成の順に配置されること，3) 全体の時系列の方向にある程度沿って各時代の表現が並ぶことが観察され，LM 内部空間においても各時代の表現は同様の特徴を持つ可能性が示唆される．

4　時代間の方向の比較

LM が表現する各時代の方向を時代間で比較した時，時代間の方向がそろっているか，バラバラであるかを調べることで LM が表現する日本の時系列構造を調査する．

まず，§ 2.1 に基づき作成したデータセットを用いて，§ 3 と同様の方法で時代ごとに LM の適切な位置での隠れ状態 X を収集する．

LM の各時代を表現する方向は，収集した各時代の隠れ状態 X に対して PLS 回帰を用いて生年 y に強く相関する方向を抽出することで取得できる．ここで，§ 2.2 で説明したように式 4 におけるベクトル w が d 次元空間におけるもとの説明変数 X の目的変数 y に相関する方向であるから，ここでは w が各時代の隠れ状態 X の生年 y に相関する方向である．

よって，時代間の方向は各時代の w を時代間のコサイン類似度を比較することで調べることができる．モデルに Swallow-13b と LLM-jp-3-13b を用いた場合の比較結果を図 3 に示す．ここで，図 3 のランダムとは w と同じ形状を持つ，ランダムなベクトルを表す．

図 3 を見ると，例えば大正と明治，昭和と大正は

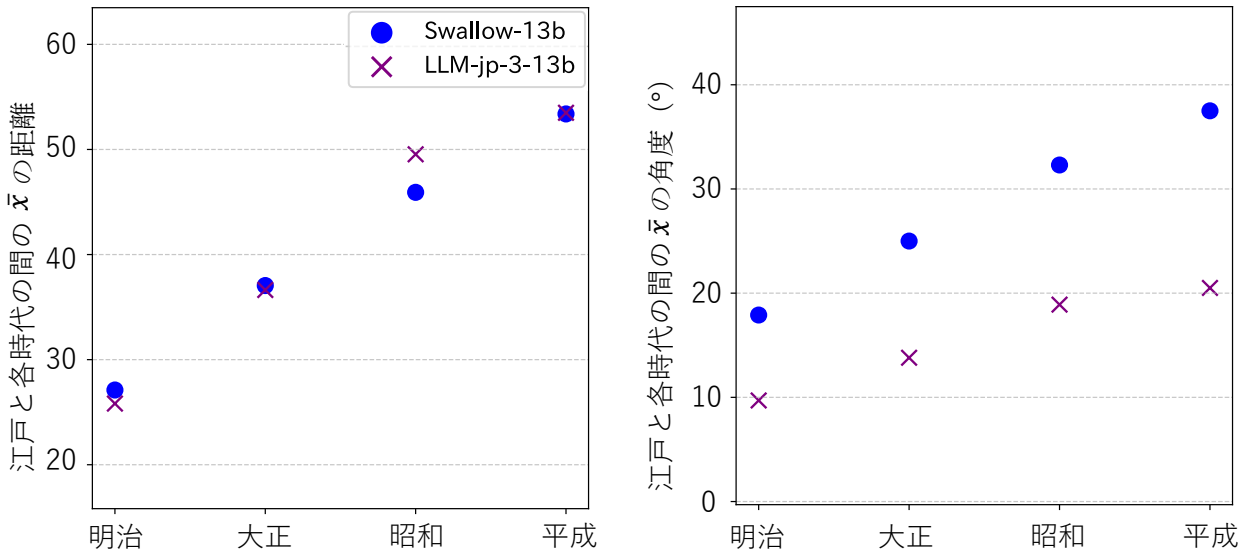

図 4　江戸と各時代（明治，大正，昭和，平成）の間の $\bar{x}$ の距離（左）と角度（右）の比較．時代間の $\bar{x}$ 同士の距離，時代間の $\bar{x}$ の間の角度はいずれも時代が離れるほど大きくなる．

ランダムベクトルとの比較結果に比べると，有意な相関が確認できる．これは大正時代は他と比べて短い分，時代間の共通な概念を反映している可能性が考えられる．一方，それ以外の時代間の w については，ランダムベクトルと比較した場合とあまりコサイン類似度の絶対値が変わらないため，時代間の方向に相関がないと考えられる．以上の結果は LM が表現する日本の時系列では，一部の隣接する時代間の方向に若干の関係があるものの，時代間の方向はバラバラであることを示す．

なお，この結果は図 2 の PCA による可視化結果において時代間の方向にあまり相関が見られなかったことが LM の内部空間で実際に示されたことを意味する．

5　時代間の位置の比較

LM が表現する各時代の方向を時代間で比較した時，どのような関係があるかを調べることで LM が表現する日本の時系列構造を調査する．

図 2 では，PCA による圧縮後の空間で $\bar{x}$ を時代間で比較すると，江戸，明治，大正，昭和，平成の順に配置されていることが観察された．したがって，§ 2.1 に基づき作成したデータセットを用いて，§ 3 と同様の方法で時代ごとに収集した隠れ状態 X の $\bar{x}$ に各時代の位置が表現されていると考えられる．

そこで，江戸と各時代（明治，大正，昭和，平成）の間の $\bar{x}$ 同士の距離と時代間の $\bar{x}$ の角度を Swallow-13b, LLM-jp-3-13b を用いて調べた．その結果が図 4 である．

まず，時代間の $\bar{x}$ 同士の距離は，江戸に対して時代を遡るほど大きくなることがどちらのモデルでも確認できる．次に，時代間の $\bar{x}$ の角度も，江戸に対して時代を遡るほど大きくなることがどちらのモデルでも確認できる．以上の結果は LM が表現する日本の時系列では，時代間の位置は江戸から平成の順番にそろっていることを示す．なお，この結果も図 2 による PCA の可視化結果における時代間の $\bar{x}$ の位置関係と対応がついている．

6　終わりに

本研究は，和暦という特殊な暦法体系を持つ日本の時系列は LM 内部にどうエンコードされるかをテーマに設定した．PLS という出力変数を考慮した次元圧縮手法を使うことで LM が各時代に属する人物の生年を表現する際の隠れ状態から，その時代表現を抽出でき，様々な形で取得することができる．

これを利用して，最初に，圧縮により抽出した各時代表現を LM 内部に再構成する形で取得することで時代表現の同一平面での可視化を行い，LM が表現する日本の時系列の全体像を調べた結果，江戸から平成までの時代表現が単調に並ぶ構造を持ちつつ，各時代の方向には相関がないことが示唆された．その後，各時代表現を生年方向へ回帰させることで各時代の方向を取得することで時代間の方向を実際に比較し，これらに相関がないことを確認した．

謝辞

本研究は JSPS 科研費 JP22H00524, JP24K03236 の助成を受けたものです.

参考文献

[1] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In **The Eleventh International Conference on Learning Representations**, 2023.

[2] Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. In **International Conference on Learning Representations**, 2022.

[3] Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Sø gaard. Can language models encode perceptual structure without grounding? a case study in color. In **Proceedings of the 25th Conference on Computational Natural Language Learning**, pp. 109–132, November 2021.

[4] Wes Gurnee and Max Tegmark. Language models represent space and time. In **The Twelfth International Conference on Learning Representations**, 2024.

[5] Benjamin Heinzerling and Kentaro Inui. Monotonic representation of numeric properties in language models. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, 2024.

[6] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In **Proceedings of the First Conference on Language Modeling (COLM)**, 2024.

[7] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a Large Japanese Web Corpus for Large Language Models. In **Proceedings of the First Conference on Language Modeling (COLM)**, 2024.

[8] Svante Wold, Michael Sjöström, and Lennart Eriksson. Pls-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, Vol. 58, No. 2, pp. 109–130, 2001.

[9] LLMjp: Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, et al. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs, 2024.

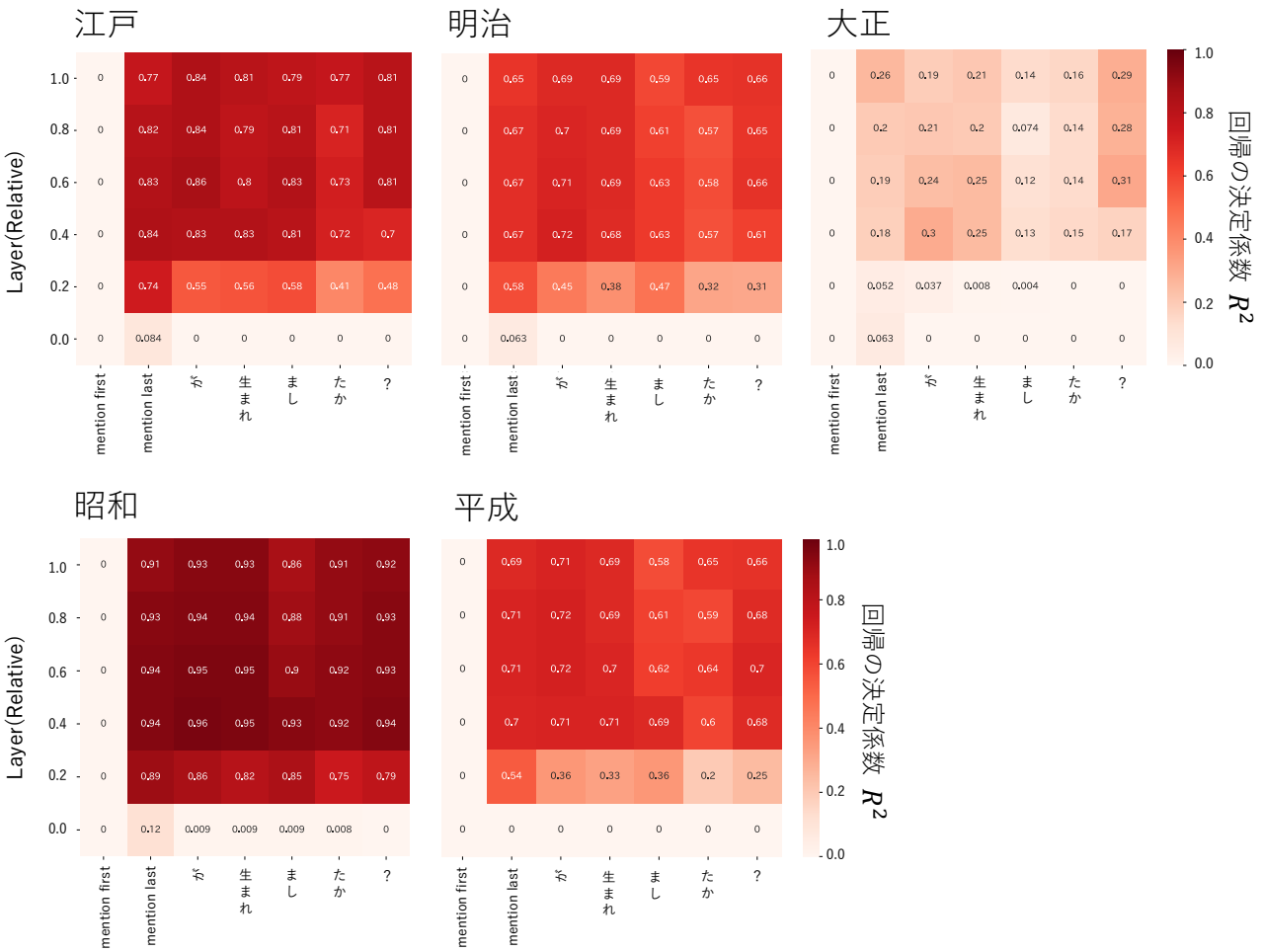

図 5　時代表現がエンコードされる隠れ状態の選択．すべての時代に共通して回帰の決定係数 R^2 が大きくなる「が」の部分の相対レイヤー位置 0.6 の隠れ状態を時代表現が最も検出される隠れ状態として実験に用いた．

A　時代表現がエンコードされる隠れ状態の選択

本実験では § 2.1 に基づき作成したデータセットを用いて LM の隠れ状態 X を収集し，PLS により生年 y を考慮した潜在変数 T への次元圧縮を行うことで，その X の時代表現を検出する．この時，PLS により時代表現を検出できるような隠れ状態を選択する必要がある．そこで，「何年に(人名)は生まれましたか？」を LM に入力した隠れ状態について時代ごとに網羅的に PLS 回帰を行い，全体的に回帰の決定係数 R^2 が高くなるのはどの位置の隠れ状態かを探索した．その結果を図 5 に示す．これを見るとすべての時代で全体的に，人名の接尾辞に相当する「が」の部分の相対レイヤー位置 0.4 や 0.6 において R^2 が大きくなる傾向が確認できる．ゆえに，「が」の部分の相対レイヤー位置 0.6 の隠れ状態を時代表現が最も検出される隠れ状態として実験に用いた．